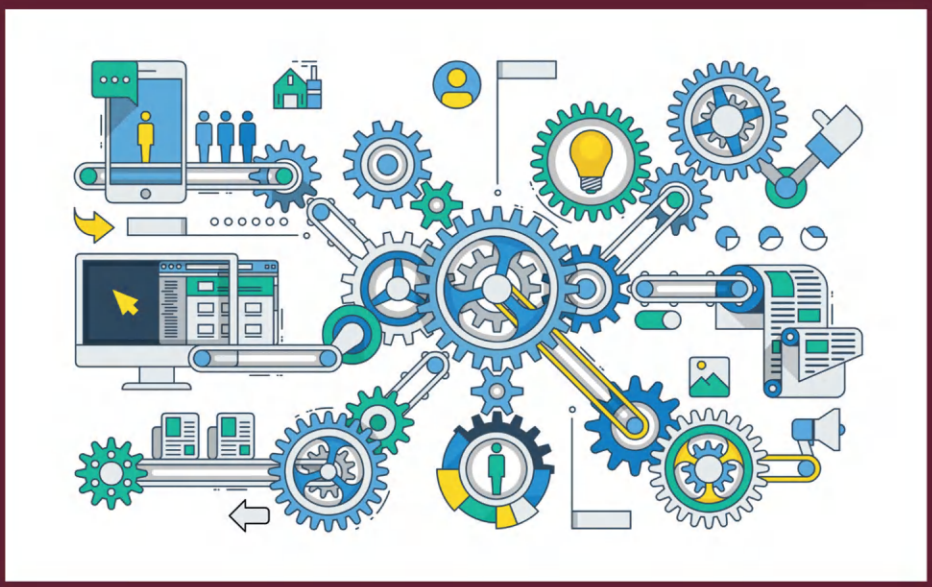
SERIES ON STATISTICS IN BUSINESS AND ECONOMICS

CAUSE AND EFFECT BUSINESS ANALYTICS AND DATA SCIENCE



Dominique Haughton, Jonathan Haughton, and Victor S. Y. Lo

A **Chapman & Hall** Book

CRC Press
Taylor & Francis Group

Cause and Effect Business Analytics and Data Science

Among the most important questions that businesses ask are some very simple ones: If I decide to do something, will it work? And if so, how large are the effects? To answer these predictive questions, and later base decisions on them, we need to establish causal relationships.

Establishing and measuring causality can be difficult. This book explains the most useful techniques for discerning causality and illustrates the principles with numerous examples from business. It discusses randomized experiments (aka A/B testing) and techniques such as propensity score matching, synthetic controls, double differences, and instrumental variables. There is a chapter on the powerful AI approach of Directed Acyclic Graphs (aka Bayesian Networks), another on structural equation models, and one on time-series techniques, including Granger causality.

At the heart of the book are four chapters on uplift modeling, where the goal is to help firms determine how best to deploy their resources for marketing or other interventions. We start by modeling uplift, discuss the test-and-learn process, and provide an overview of the prescriptive analytics of uplift.

The book is written in an accessible style and will be of interest to data analysts and strategists in business, to students and instructors of business and analytics who have a solid foundation in statistics, and to data scientists who recognize the need to take seriously the need for causality as an essential input into effective decision-making.

CHAPMAN & HALL/CRC Series on Statistics in Business and Economics

The Chapman & Hall/CRC Series on Statistics in Business and Economics is a comprehensive collection of cutting-edge books dedicated to advancing the understanding and application of statistical methodologies in the realms of business and economics.

Empirical Research in Accounting
Tools and Methods
Ian D. Gow and Tongqing Ding

Game Theory for Applied Econometricians
Data Analytics with R
Chritopher P. Adams

Bayesian Econometric Modelling for Big Data
Hang Qian

Cause and Effect Business Analytics and Data Science
Dominque Haughton, Jonathan Haughton, and Victor S. Y. Lo

Risk and Predictive Analytics in Business with R
Ozgur M. Araz and David L. Olson

For more information about this series, please visit: https://www.routledge.com/Chapman-and-HallCRC-Series-on-Statistics-in-Business-and-Economics/book-series/CHSBE

Cause and Effect Business Analytics and Data Science

Dominique Haughton,
Jonathan Haughton,
and Victor S.Y. Lo

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Front cover image: filip robert/Shutterstock

First edition published 2026
by CRC Press
2385 NW Executive Center Drive, Suite 320, Boca Raton FL 33431
and by CRC Press

4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-1-482-21647-9 (hbk)
ISBN: 978-1-041-07411-3 (pbk)
ISBN: 978-0-429-17258-8 (ebk)

DOI: 10.1201/9780429172588

Typeset in Palatino
by KnowledgeWorks Global Ltd.

Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

About the Authors

Dominique Haughton (PhD MIT 1983) is a Professor Emerita of Mathematical Sciences and Global Studies at Bentley University near Boston, and an Affiliated Researcher at Université Paris 1 (Pantheon-Sorbonne, SAMM) and at Université Toulouse 1 (TSE-R). Her widely published work concentrates on how to best leverage modern analytics techniques to address questions of business or societal interest. She is an alumna of the Ecole Normale Supérieure and a Fellow of the American Statistical Association.

Jonathan Haughton earned his PhD in economics from Harvard University in 1983. He has published widely in the areas of economic development, taxation, the environment, and the analysis and measurement of poverty. Now emeritus, he chaired the economics department at Suffolk University, Boston, and he has taught or worked as a consultant in over 20 countries on five continents.

Victor S.Y. Lo is an executive with over three decades of consulting and corporate experience employing data-driven solutions in a wide variety of business areas, including Marketing, Risk Management, Financial Econometrics, Insurance, Product Development, Transportation, Healthcare, Operations Management, and Human Resources, and is a pioneer of uplift modeling. He is currently SVP, Data Science and AI at Fidelity Investments, and has led data science and analytics teams in various organizations. Victor earned a master's degree in Operational Research and a PhD in Statistics and was a Postdoctoral Fellow in Management Science.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Acknowledgments

We would like to thank David Grubbs and his editorial team from CRC Press/Taylor & Francis for initiating the process for the book and for endless support, patience, and encouragement over the past decade. Thanks are also due to Mayank Sharma and the team at KnowledgeWorks Global for their expert editing.

We received very helpful reviews of drafts of our chapters from Alison Kelly, Jongbyung Jun, and Le (Sarah) Tang, all of from Suffolk University; and from Mingfei Li of Bentley University.

Dominique would like to thank Bentley University for its support over the years. And Jonathan is very grateful for support, including a sabbatical, from Suffolk University, which helped move the work forward.

First and foremost, Victor would like to thank the Lord Almighty for providing the opportunity, inspiration, and protection over the past many years during the writing of this book. He also would like to express his gratitude to his family for their gracious support over the countless number of weekends and their encouragement throughout this project. Last but not least, he would like to thank Jane Zheng and Karl Rexer for their valuable suggestions.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

Introduction to Cause-and-Effect Business Analytics

"In God we trust, all others bring data"

W. Edwards Deming

This book discusses two major related topics: Methods of causal analysis applied to business problems and uplift analytics. It is written to be accessible to anyone with a basic foundation in analytics, economics, statistics, marketing, and similar fields. It is likely to be useful for students in master's programs in related fields and for practitioners of analytics who want to expand their knowledge and skills.

The emphasis is on the concepts and on the essentials of important techniques. This should get the reader started, but mastery of any given technique will require further reading, digging, practice, and experimentation. We aim to provide the right path and the initial steps, but we have not written a programming cookbook, as we believe that readers should have little trouble finding the appropriate routines in Stata, SAS, Statistical Package for the Social Sciences (SPSS), R, Python, or in specialized software, once they know broadly what to look for.

The rest of this chapter discusses the importance of causality in thinking about business and suggests how readers may want to pick their way through the chapters in order to profit most fully from the book.

1.1 The Trend toward Business Analytics and Data Science

In this age of Big Data, the ability to draw valuable insights and optimize decisions based on massive data is seen as a competitive advantage in business. As a result, there is tremendous demand for employing Artificial Intelligence (AI), Machine Learning, Data Science, or Advanced Analytics, resulting in several documented success stories in multiple industries ranging from technology to financial services to insurance to healthcare to retail.

DOI: 10.1201/9780429172588-1

Taylor (2024) reports that the quantity of data created, captured, copied, and consumed worldwide rose from 2 zettabytes in 2000 to an estimated 149 zettabytes in 2024. This explosion of data has created a situation where the opportunities for analytics are only limited by our imagination and capacity. In today's terminology, there are multiple terms that are closely related, including Data Analytics, Business Analytics, Data Science, Data Mining, Machine Learning, and AI. In particular, AI, a field that was born in the 1950s and has strong renewed popularity today, loosely means mimicking human behavior with computers; see Jordan (2019). As a subset of AI, Machine Learning is a class of techniques that learns from data, unlike the older generation form of AI, where human expertise is required to hard-code specific rules into the system. While AI and Machine Learning speak to the techniques and tools, Data Science is a broader term and a multidisciplinary field (see Meng 2019) that integrates Computer Programming or Computer Science (including extracting data from any forms, data processing, and transformations), Mathematics and Statistics, subject matter expertise (e.g., marketing knowledge is required for analyzing marketing data, and knowledge of risk is needed for risk analytics), and soft skills (e.g., consulting, communication, and presentation). In Figure 1.1, data scientists or analysts who possess all the skills in the overlapping portion of the Venn Diagram are sometimes known as unicorns that are not easy to find (Lo 2019a).

In this book, we choose the term "Business Analytics" to represent our content, but the other terms or fields mentioned above are sometimes

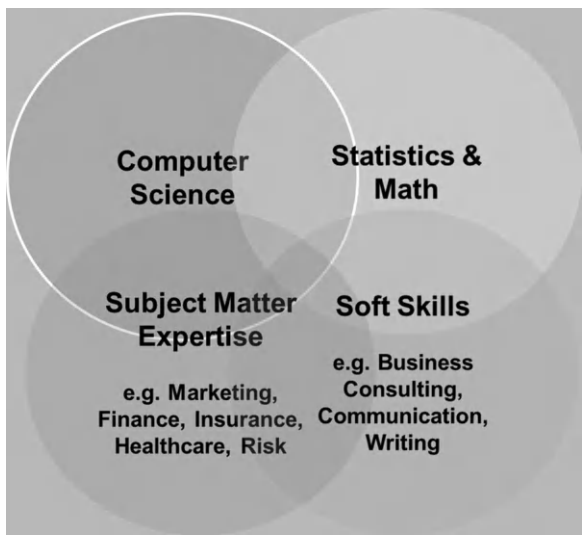


FIGURE 1.1
Venn Diagram of data science. (Lo 2019a.)

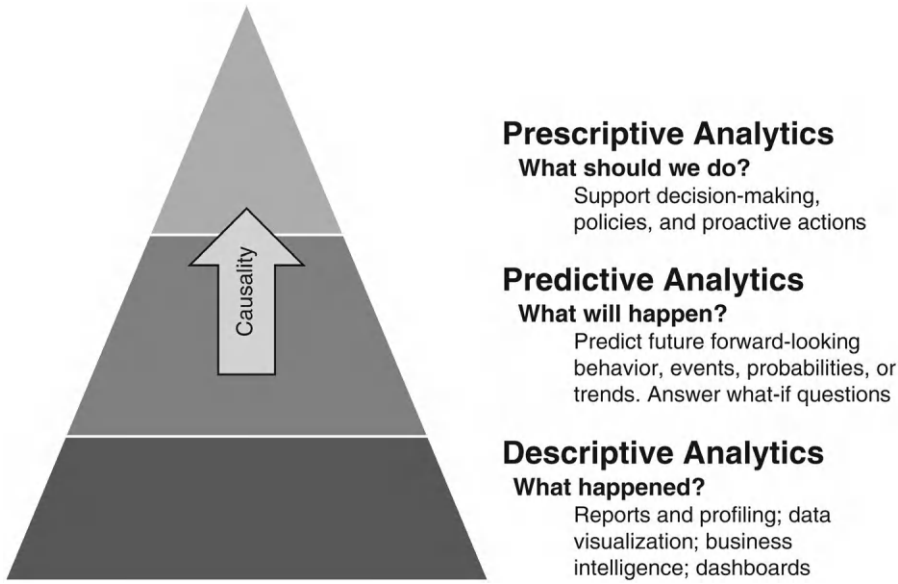


FIGURE 1.2
Three types of analytics.

synonymous in the industry. By definition, "Business" Analytics is an application of "Analytics" for business purposes. The field of Analytics (also known as Data Analytics) can be classified into the following three types (Figure 1.2):

1. **Descriptive Analytics:** It describes what happens with data analysis or reports, often including summary statistics and data visualization (e.g., scatterplots, pie charts, histograms, box plots, and line charts), and is highly associated with Business Intelligence. Often, data are presented in the form of dashboards, which may be updated frequently. Multidimensional tables and graphs are sometimes included to provide in-depth analysis of data. For example, reporting the weather condition yesterday is a form of descriptive analytics, as are efforts to measure worldwide income or consumption. Good descriptive analytics is not straightforward and requires clear protocols, attention to detail, and imagination in how best to present potentially enormous quantities of data in ways that an audience can understand usefully.
2. **Predictive Analytics:** It predicts what will happen, such as economic growth over the coming year, future customer behavior, or probabilities of certain events happening in the future. For example, weather forecasting itself is a form of predictive analytics.

Predictive Analytics (or Predictive Modeling) is typically based on some form of Statistical Analysis (or Statistical Modeling) and/or Machine Learning (in particular, Supervised Learning). Predictive analytics is also at the heart of policy analysis, where one is trying to answer questions such as what would happen if the tax on gasoline were doubled or if the minimum wages were raised. Causal reasoning is central to answer policy questions such as these.

3. **Prescriptive Analytics:** It gives some knowledge about the future, considers potential alternative decisions, and then determines the best decision. For example, if you know there is a good chance of heavy snow tomorrow (through predictive analytics), you may evaluate the potential risk of studying or working from home versus going to school or work and then determine the right decision to achieve your goal while considering the risk. This field is traditionally a large subset of Operations Research, Management Science, or Industrial Engineering, and is also linked to a subset of modern Machine Learning methods (e.g., Reinforcement Learning). Some make a distinction between positive economics, which seeks to answer the "what if" questions, and normative economics, which tackles the question of "what should be done."

The three types of analytics are closely related instead of being independent from each other. Learning about what happened in the past (descriptive analytics) is often a prerequisite to predicting the future (predictive analytics). Knowing something about the future (predictive analytics), even with a high degree of uncertainty, enables us to evaluate alternative choices and select the right decision quantitatively (Lo 2019b).

These three types of analytics have been around for several decades. What drives the rise of analytics are:

1. **Big Data:** Data increases not only in volume but also in many different forms including structured and unstructured data (e.g., text, voice, and image).
2. **Increased Machine Power:** Computational power has increased and continues to increase tremendously over time, with the latest advances in graphics processing unit (GPU) machines and cloud computing.
3. **Better Algorithms:** Increased sophistication and flexibility of algorithms have been observed over the past decade for processing and analyzing data, including massively parallel processing (MPP), deep learning, reinforcement learning, causal inference, and uplift modeling. The latter two are the focus of this book and will be introduced in Sections 1.2 and 1.3.

1.2 Introduction to Causality in Business

In the previous section, we described the three types of analytics. In order to perform prescriptive analytics, some form of predictive analytics is usually required. In fact, "causality" is the link between predictive analytics and prescriptive analytics (see Figure 1.2), in that we need to establish a causal relationship if we are to be able to make successful changes; this is equivalent to being in a position to apply Pearl's "do-operator," which we discuss further in Chapter 3. Let us illustrate the link between predictive and prescriptive analytics with a classical business problem – how to set the price of your product.

Suppose you are selling premium bubble tea in a small local tea shop. Assume you have some idea about the relationship between the quantity of daily sales and unit price, which is approximately: $D = 500 - 50\,P$, where D = daily sales and P = unit price (bounded between 0 and 10). This is a classical demand curve, as shown in Figure 1.3; by convention, economists put the price on the vertical axis and the quantity demanded on the horizontal axis (left panel), while marketers tend to put the quantity on the vertical axis (right panel). This demand curve shows, for instance, that for every dollar increase in price, you expect to sell 50 fewer cups daily. Since you aim at maximizing daily profit, you would like to determine the optimal P such that the profit function is maximized. You know the cost of making bubble tea is \$2 per cup. Therefore, your objective function is maximizing the profit function $f(P) = D(P-2) = (500-50P)(P-2)$ or $= -50P^2 + 600P - 1000$. We can solve this through standard differential calculus. By differentiating the function $f(P)$ with respect to P and setting it to zero, we have $P^* = \$6$. Since the second derivative of the profit function is negative, that is, $f''(P) < 0$, $P^* = \$6$ is the optimal price to achieve the maximum profit, which will be $f(6) = 200 * 4 = \$800$. Any other prices will result in a lower daily profit. In

FIGURE 1.3
Sales as a function of unit price.

this exercise, establishing the sales function of price is a form of predictive analytics, while determining the optimal price belongs to the third type of analytics: Prescriptive analytics. While the prescriptive analytics (optimization) process is straightforward, the key is how to come up with the quantity of sales as a function of price: $D = 500 - 50\ P$ In particular, the coefficient 50 reflects the sensitivity of the quantity demanded to price, and it is the key piece of information needed to determine the optimal price.[1] How can we know it is 50? This is a large portion of what this book will cover.

Let us take a visit to various data-gathering methods that may help determine the price sensitivity.

1. **Survey Data:** In this approach, also known as market research or primary data collection, researchers or analysts directly go to the field and gather self-reporting data from survey respondents. For causality, we can ask survey respondents why something causes something, for example, what caused them to do what they did. In our example, we could use a contingent valuation survey – asking prospective consumers how much they would want to buy at different prices – which is one of the most widely used tools in cost-benefit analysis, particularly when trying to put a value on non-marketed goods and services (such as the value of a public park or of clean air). Market researchers have also come up with techniques such as conjoint analysis and best-worst scaling to estimate such relationships. While conjoint analysis and related techniques are quite established (e.g., Green and Srinivasan 1990), they are based on self-reporting data, and it does take some cost and effort to conduct a well-designed survey. The question arises of whether we could simply observe it with "real" data instead of doing a survey.
2. **Experimental Data:** The gold standard of answering causality is through randomized experiments, frequently known as randomized controlled trials (RCTs), or as A/B testing in the marketing, analytics, and data science literatures. To do that, for our example, one would systematically test various price points in a randomized setting and observe the sales level at each price. It takes some effort to conduct such an experiment. This is usually the best method if it is feasible, although in reality it typically falls well short of the ideal, for the reasons discussed in Chapter 3.
3. **Observational Data:** If a randomized experiment is not practically feasible or we just want to obtain an answer faster, could we simply utilize available historical data? It is possible if historical data has price variability. In fact, a significant portion of this book (Chapters 3–5 and 9–12) discusses how to infer causality from observational data. In our example, let us assume that you are only able to gather a few data points from some past time periods (days), and the data are plotted in Figures 1.4a and 1.4b with a linear regression line fit

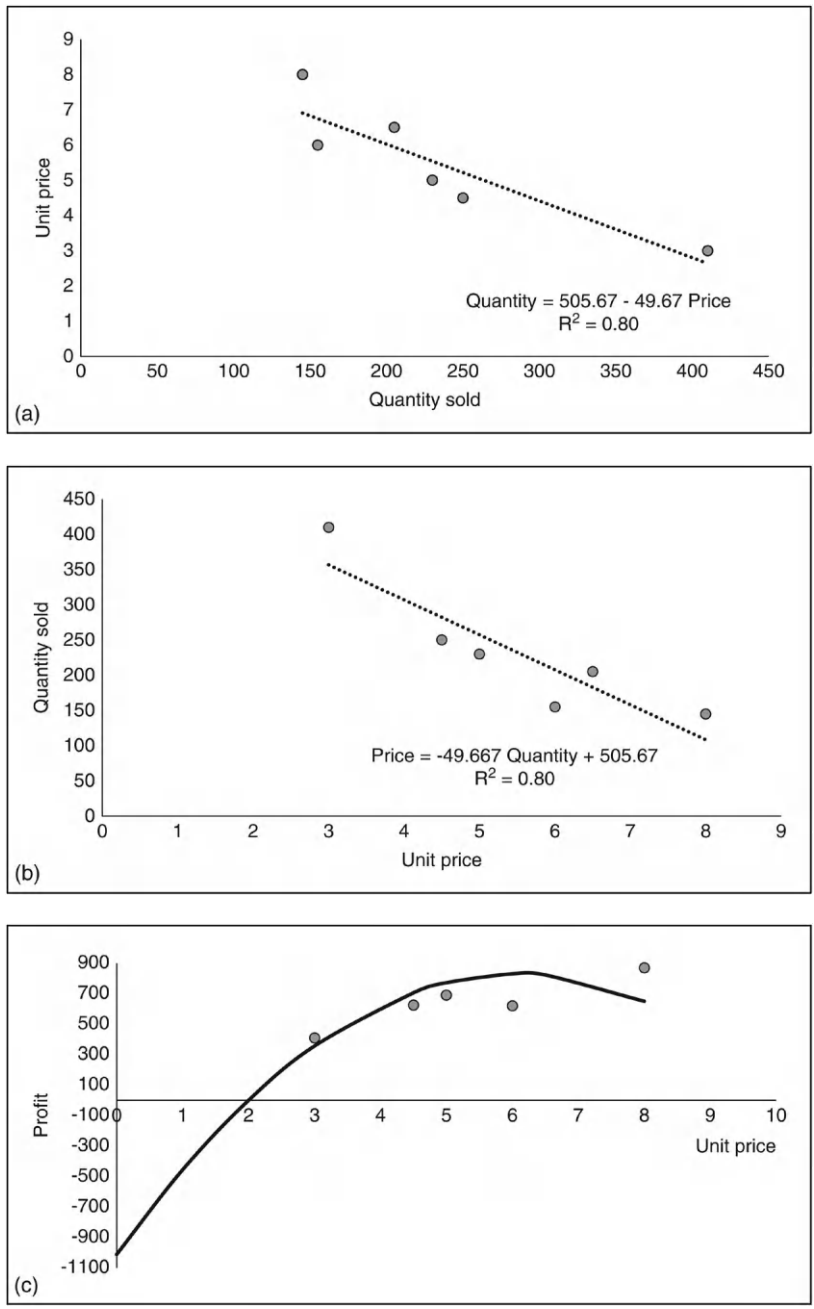


FIGURE 1.4
Observational data and its fitted regression line: (a) Economists and (b) marketers. (c) Profit as a function of unit price. Each dot is actual profit = actual sales * (unit price – 2), and the curve is based on the fitted sales, that is, profit function = fitted sales * (unit price – 2).

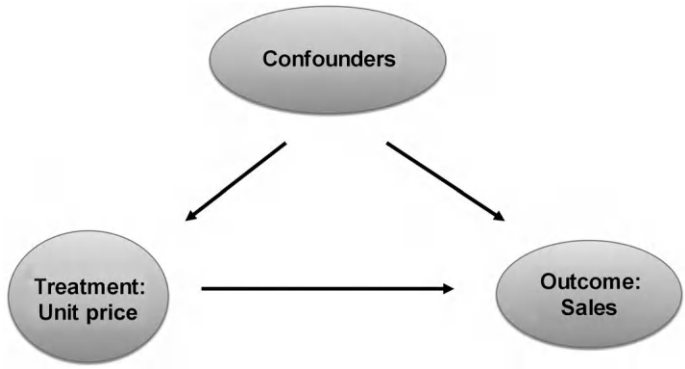


FIGURE 1.5
Back-door path: Potential confounding in determining price sensitivity.

to the data. If we take the regression line (predictive analytics) as an input to optimization (prescriptive analytics), we can plot the profit as a function of price in Figure 1.4c. Applying standard differential calculus, the optimal price based on this data would be $6.09 with an estimated daily profit of $831, based on the fitted sales function from Figure 1.4a (or Figure 1.4b). If the data were from a randomized experiment (i.e., the price points were randomly assigned to different time periods), this should be a reasonable analysis. However, since the data is observational rather than experimental, it is possible that there are other "confounders" that affect both the treatment (price) and outcome (sales), (see Figure 1.5).

A confounder is defined as a variable that predicts or drives both the treatment and outcome. In this case, it would be driving both the unit price and sales. Since we need to assess the causal effect of unit price change on sales, our interest is in quantifying the direct (horizontal) link between price and sales. However, the presence of confounders implies that a "back-door path" exists between price and sales (via the confounders-to-price and confounders-to-sales links), leading to a biased estimation of the causal treatment effect. We treat this issue in more detail in Chapter 4.

In the bubble tea sales example, suppose your manager tells you that some of the data points collected were not a good representation of the usual situation. For example, in Figure 1.4b, the data point ($3, 410) was actually from the opening day of the shop with plenty of promotion running on and before that day (e.g., banner and email marketing), and because it was the first day of business, we also started with a very low price of $3 as part of the promotion. The opening day special promotion may have both lowered the price and increased the sales. As a result, the opening day can be a strong confounder that drives both the treatment (price) and outcome (sales). Likewise,

you have just discovered that the data point ($8, 145) was from a major holiday – with so many shops closed, yours was one of a few shops that were in business, and as a result, you raised the price, and the sales level was still relatively high (when compared to the fitted regression line). In this case, the holiday is a confounder that had a positive effect on both the unit price and sales. A simple solution is to remove both data points and refit the regression line, resulting in the sales function in Figure 1.6a (or Figure 1.6b) with a much smaller price sensitivity, followed by the profit function in Figure 1.6c.

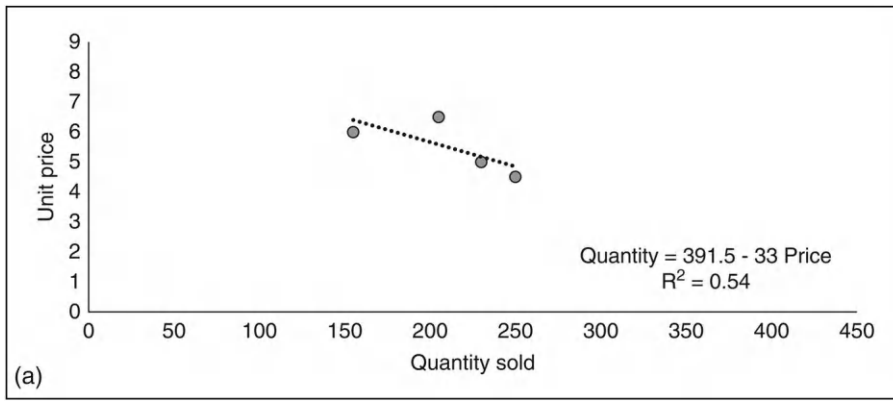


FIGURE 1.6a
Observational data and its fitted regression line (excluding two special days): Economists.

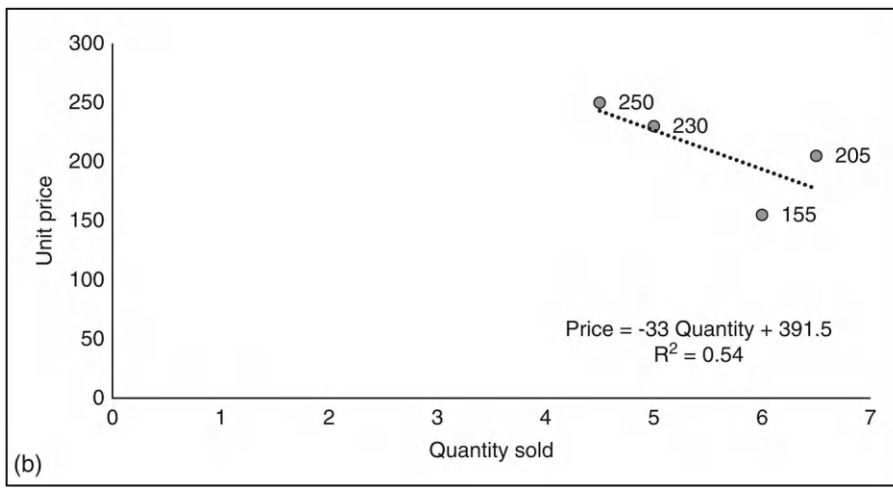


FIGURE 1.6b
Observational data and its fitted regression line (excluding two special days): Marketers.

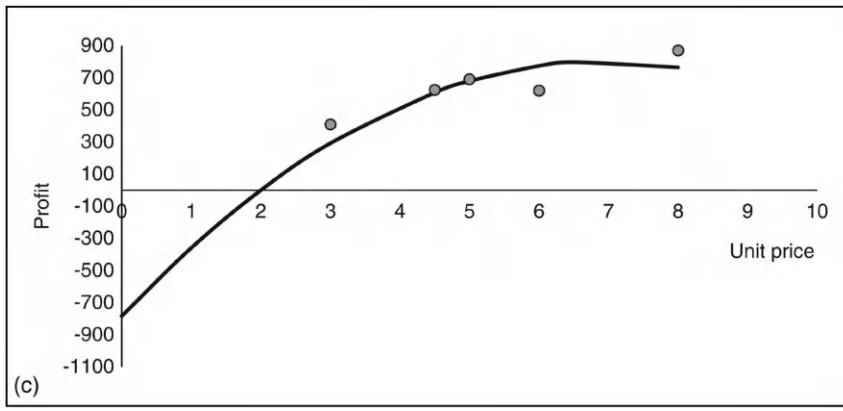


FIGURE 1.6c
Profit as a function of unit price, using the fitted sales function from Figure 1.6a (excluding two special days). Each dot is actual profit = actual sales * (unit price – 2), and the curve is based on the fitted sales, that is, profit function = fitted sales * (unit price – 2).

By applying differential calculus again, we obtain the optimal price which is \$6.93. Substituting this price to the profit function, we have the daily optimal profit = $f(6.93) = (391.5 – 33 * 6.93) * (6.93 – 2) = \$802.65$.[2]

While the above is an illustration of specific confounding situations, there are many other possible confounders, such as:

1. **Competitor Pricing:** If there is a major coffee shop nearby, coffee may serve as a substitute for bubble tea for some customers. As a result, any pricing change of this coffee shop may drive your sales and your price of bubble tea.
2. **Product Innovation:** On certain days, you may have tried different flavors of bubble tea, which may have impacted both the demand and price.
3. **Coupon Strategy:** If some selected customers receive a coupon as part of a promotion, it would simultaneously drive price and sales.
4. **Underlying Market Trend:** With time series data, it is possible that there is an underlying trend (inertia) that pushes both the unit price and sales over time, confounding the actual causal relationship between them.

The above illustrates the importance of handling causality, using price as an example. We can also consider other "treatments" or interventions, such as promotion, product, and place.

1. Promotion is a key marketing function, and for large corporations, there can be many types of promotion happening at the same time – for example, TV advertising, online promotion, email marketing, out-of-home (billboard), radio, and beyond. It is important to

understand the impact of each of these (through predictive analytics and causal inference) and then optimally allocate budget based on their return on investment (via prescriptive analytics). Without proper causal business analytics, managers running each of these promotional channels could claim the same revenue gained, and the sum of their claims would be far higher than the actual overall revenue. We develop an example along these lines in Chapter 3.

2. Since each product may have a variety of product features, knowing which ones would drive the desirable customer outcome and revenue enables us to optimally assign features for each product. For instance, to design a laptop computer, there are multiple features the business and consumers will consider, for example, size, color, sound quality, keyboard quality, CPU, and RAM. The business may optimize the features that are most important to the consumers, which would require some form of causal business analytics to assess the impact of various features on consumer purchases.
3. Place involves where to open a store and how to decorate it, which all involve causal questions requiring causal business analytics to estimate the impact (of opening a store in a particular location or decorating a store with a specific theme or color).

Methodologies for measuring causal effects are not from one academic field but several academic fields, ranging from Economics and Econometrics (Lee 2005, 2016, Angrist and Pischke 2009, 2014, and Gertler et al. 2016), Statistics (Rosenbaum 2002, 2010, Rubin 2006, Weisberg 2010, and Imbens and Rubin 2015), Computer Science and Artificial Intelligence (Pearl 2000, Scutari and Denis 2015, Pearl et al. 2016, Peters et al. 2017, and Pearl and Mackenzie 2018), Epidemiology (Vanderweele 2015 and Hernan and Robins 2020), Sociology (Morgan and Winship 2015), Psychology (Glymour 2001 and Sloman 2005), Political Science (Dunning 2016), to Philosophy (Spirtes et al. 2000, Cartwright 2007, and Cartwright and Hardie 2012). While the books from these fields cover a variety of related methodologies from different academic disciplines, they are not specialized in business applications, which is the focus of our book. Chapters 3–5 and 10–12 provide a wide variety of methodologies to help answer these causality questions in business.

1.3 From Population Causality to Individual Causality: Uplift Modeling

The previous section introduced causal measurement at the population or group level, which enables the selection of the best treatment for part or all of the overall population. Researchers and practitioners have not stopped at estimating the overall treatment effect for the population or a group (or for

the "average" person). With the advances of Machine Learning and predictive analytics, along with randomized experiments and causal inference, researchers have taken a further step to estimate treatment effects at the individual level. This increasingly popular subfield known as uplift modeling was separately created in the late 1990s and early 2000s by researchers and practitioners such as Radcliffe and Surrey (1999) and Lo (2002) with the objective of finding individuals who are truly positively influenced by a treatment or intervention through Machine Learning and predictive modeling by uncovering heterogeneous treatment effects in available data. This technique enables us to potentially identify the "persuadables" and thus optimize target selection in order to maximize treatment impact; see Siegel (2011, 2013b). Motivated by the trend of personalization in many industries, this subfield has gained tremendous attention in recent years with seemingly unlimited applications such as personalized marketing (original usage of uplift), personalized medicine, political elections, personalized insurance, and healthcare programs, with growing numbers of publications and presentations from both industry practitioners and academics across the globe. Independently, researchers from Economics, Epidemiology, and Statistics have considered similar or related methodologies applied to Social Sciences and Medical Sciences, for example, Athey and Imbens (2015), Yong (2015), and Zink et al. (2015).

The most prominent application of uplift modeling so far is perhaps in political elections. In the 2012 Obama re-election campaign, the campaign team had a limited budget to spend on targeting voters. Their analytics team designed an intelligent approach – instead of communicating with those who would almost surely vote for Obama or his opponent (as it would be ineffective to market to these two groups), they applied uplift modeling to uncover the swing voters who are most "persuadable" by marketing communication, enabling their campaign to maximize overall impact, as documented in Stedman (2013) and Siegel (2013a). This success has led to wider applications of uplift modeling in political elections. As in marketing and sales campaigns, where traditional predictive modeling focuses on predicting the outcome, uplift modeling estimates the effectiveness of the treatment or intervention at the individual level, allowing one to focus resources on the subjects that are likely to be positively impacted by the treatment.

Figure 1.7 provides a classification framework for uplift modeling and causal inference. The vertical axis extends from association only (what may be correlated) to causal measurement (which are causing which), while the horizontal axis moves from population (or group or subpopulation) level summary statistics to personalized level modeling. We will describe the four quadrants of Figure 1.7 in the following:

a. **Reporting/Summary Statistics** – This is a subset of Descriptive Analytics in Figure 1.2, where the past data (sales reports, financial reports, and summaries of other historical data) are summarized

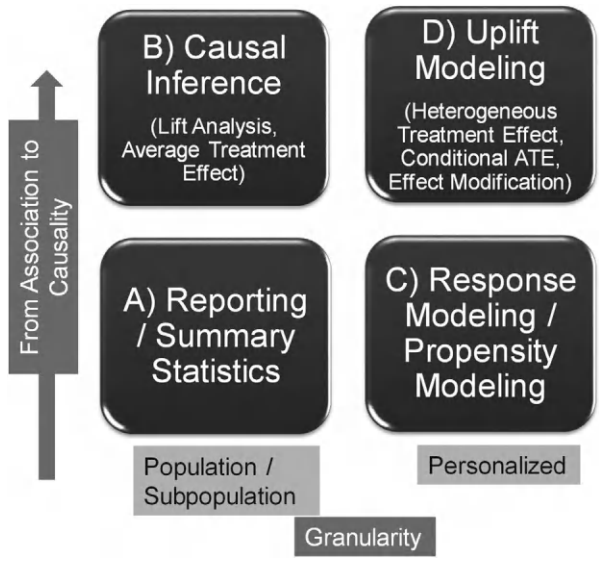


FIGURE 1.7
Framework for causal and association analysis.

visually or statistically to describe what happened in the population (or subpopulation) as a whole.

b. **Causal Inference** – As described in Section 1.2, this type of analytics extends beyond association or correlation-based summary statistics and provides insight into the causal link between a treatment or intervention and an outcome for the overall population (or a group). The estimate is known as the Average Treatment Effect (ATE) in academia or simply "lift" in business.

c. **Response Modeling or Propensity Modeling** – Using historical customer data at the individual level, the traditional usage of analytics is to apply predictive modeling, also known as supervised learning, to target customers who are likely to take a desirable action regardless of whether or not they receive an intervention or treatment, as documented in marketing analytics textbooks such as Jackson and Wang (1996) and Roberts and Berger (1999). Such a model is known as a Response Model or Propensity Model, which almost guarantees, by design, that the model targets are better than random targets in terms of response rate by design. It improves efficiency by increasing the proportion of responders within the treatment group. It extends Reporting/Summary Statistics from the population or group level to the individual level, leading to targeting at a personalized level.

d. **Uplift Modeling** – While response or propensity modeling aims at predicting the outcome, uplift modeling is about predicting the lift

between treatment (contacted) and control (uncontacted)[3] at the individual level. It requires an application of predictive modeling in a nontraditional way, using both the treatment and control data from historical campaigns. It is also an extension of causal inference from the population level to the individual (personalized) level. To illustrate with a common marketing example, in a customer cross-sell campaign where the goal is to sell additional products to existing customers, a response or propensity model is developed to differentiate between those who responded (purchased) and those who did not respond (no purchase) to a historical campaign, and then applying such a model to a future campaign. Measuring the effectiveness of the treatment requires an A/B test or a randomized experiment. The experiment would have two separate target groups: (1) Model targets (say, using the top 30% of those who are likely to purchase as determined by the traditional model, also known as the top three model deciles) and (2) random targets (for comparison and potential model refinement). In each of the two target groups, customers would be randomly split into treatment (receiving the marketing campaign) and control (not receiving the campaign) groups, so any difference in measurement results can be attributable to the treatment (campaign). Since the traditional predictive model is designed to focus on customers who are likely to buy the product, success over random targets is expected. However, those customers who are likely to buy (as determined by the top three model deciles here) may respond *naturally*, regardless of whether they receive the campaign, resulting in no lift over control. Table 1.1 shows an illustrative example. While there is a difference between the model group and the random group, that is, the top three model deciles have a higher purchase rate than the random targets, there is no difference in purchase rate between the treatment and control groups within each target group; as a result, the campaign generates no lift. This happens because the objective of the traditional predictive model is to predict the likelihood of purchase rather than estimating and optimizing lift over control; that is, *what is modeled does not match what is measured*. Uplift modeling is thus needed to estimate lift over control, enabling the

TABLE 1.1

Example of Possible Campaign Result from Traditional Predictive Modeling

	Model Targets (Top Three Deciles) (%)	Random Targets (%)
Treatment	2.5	1.0
Control	2.5	1.0
Lift	0.0	0.0

business to select those who are likely to generate higher lift values, as in the 2012 Obama re-election. Chapters 6–9 described uplift modeling in detail. In the example above, a randomized experiment was available so we could properly estimate the lift and develop an uplift model. In the scenario where randomized experiments are not feasible, an integration of causal inference on observational data with uplift modeling will be required, and this approach is introduced in Chapter 9.

Chapter 2 of this book describes the traditional predictive modeling approach before we dive into causal inference and uplift modeling in subsequent chapters.

1.4 Organization of This Book

This book discusses two major topics, Causal Business Analytics and Uplift Analytics, with a range of methodologies illustrated with practical business problems and data. The first topic is about applying proper causal analytics methodologies, known as causal inference in academia, in the business setting. As mentioned in Section 1.2, there are many textbooks on causal inference from various academic fields, but there has not been one seen so far that touches on practical business applications with a wide range of methodologies. The second topic is the emerging field of uplift analytics – given its relatively emerging state, there are very few books on this topic. Our coverage on uplift is broad: From experimental design for uplift to model measurement, predictive modeling for uplift, and prescriptive uplift analytics (treatment optimization).

A suggested flow of reading the book chapters is outlined in Figure 1.8. Chapter 2 serves as a review of common statistical, econometric, and Machine Learning techniques, on which subsequent chapters on causal business analytics and uplift modeling are built. Readers who are familiar with these common techniques can skip to other chapters. Chapter 3 introduces causal business analytics (causal inference in business) with more standard methodologies such as randomized experiments (also known as A/B testing), potential outcomes approach, and propensity score matching. Chapter 4 describes additional causal inference techniques that are widely used in economics and sociology, including synthetic controls, double differences, and instrumental variables, but they can also be utilized in business. Chapter 5 discusses a powerful AI-based approach known as Directed Acyclic Graphs or Bayesian Networks that can be employed to discover and quantify causal relationships between variables, with or without prior knowledge of the likely causal pathways.

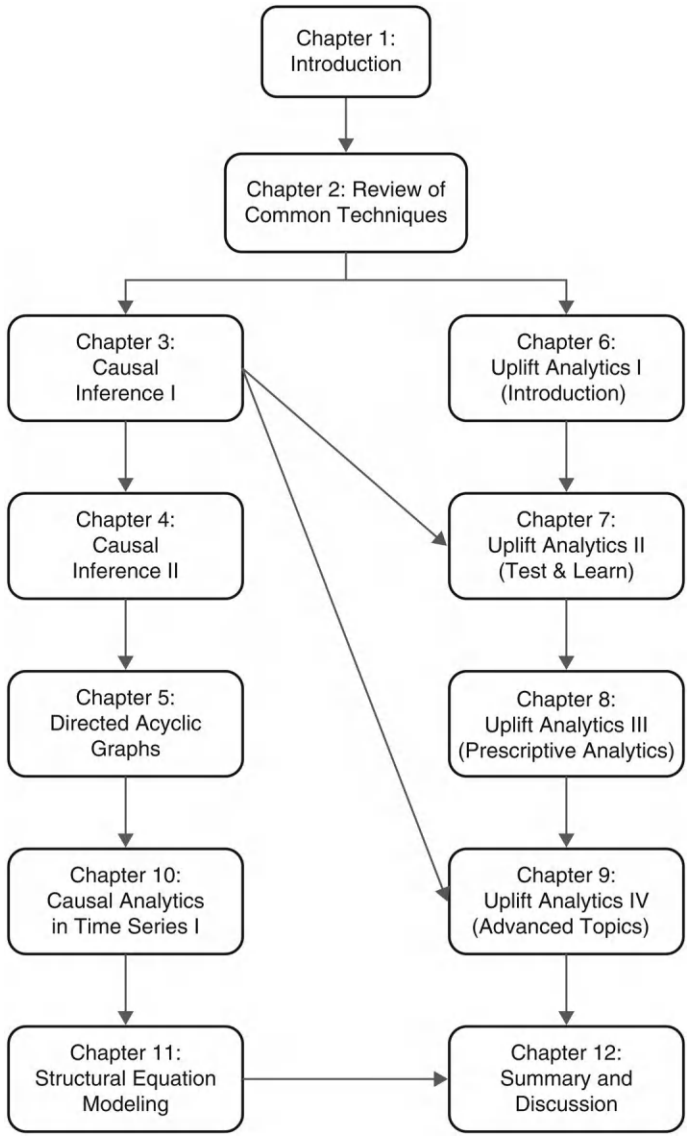


FIGURE 1.8
Suggested flow of book chapters.

Uplift analytics is discussed in Chapters 6–9. Chapter 6 introduces the concept of uplift analytics along with a simple modeling approach. Chapter 7 discusses the Test and Learn process for uplift, including sample size determination, experimental design, and measurement metrics. Chapter 8 provides an overview of prescriptive analytics for uplift, employing mathematical

techniques to optimally match individuals to treatments. The discussion includes techniques for handling the uncertainty of model estimates, based on methods from Operations Research, Finance, and Risk Management. Chapter 9 discusses selected advanced topics in uplift analytics, including some of the latest uplift modeling techniques, developing models on observational data (without randomized experiments) by combining causal inference with uplift modeling, and integrating direct response data with experimental data for uplift modeling.

Chapter 10 takes us back to causal business analytics, where data are in the form of time series, with methodologies from Marketing Science, Econometrics, and Time Series Analysis. Chapter 10 discusses various advanced topics for analyzing time series and related data. Chapter 11 introduces Structural Equation Modeling (SEM), originally from Psychology and Psychometrics, which can be employed to handle business problems. We conclude the book with a discussion and summary in Chapter 12. Practical business problems and data are used for illustration throughout the book.

Notes

1. Economists typically measure sensitivity using the own-price elasticity of demand, which is defined as $\frac{\partial lnS}{\partial lnP}$ and has the virtue of being unit-free.
2. Using the previous fitted sales function from Figure 1.5a, where confounders are included, the price would be \$6.09 and the daily profit would be $f(6.09)$ = (391.5 – 33*6.09) * (6.09 – 2) = \$779.27, resulting in a reduction of \$23.38 in daily profit.
3. For simplicity, we assume that the control group receives no treatment, that is, remains uncontacted. In practice, the control group may receive an older treatment (e.g., a legacy intervention method) in marketing or a "placebo" in clinical trials.

References

Angrist, Joshua, and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Angrist, Joshua, and Jorn-Steffen Pischke. 2014. *Mastering Metrics: The Path from Cause to Effect*. Princeton, NJ: Princeton University Press.

Athey, S., and G. W. Imbens. 2015. "Machine learning methods for estimating heterogeneous causal effects". Working Paper, Graduate School of Business, Stanford University.

Cartwright, Nancy. 2007. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. New York, NY: Cambridge University Press.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York, NY: Oxford University Press.

Dunning, Thad. 2016. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M.J. Vermeersch. 2016. *Impact Evaluation in Practice*, 2nd edition. Washington, DC: World Bank Group.

Glymour, Clark. 2001. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: MIT Press.

Green, Paul E., and V. Srinivasan. 1990. "Conjoint Analysis in Marketing Research: New Developments and Directions". *Journal of Marketing*, 54(4): 3–19.

Hernan, Miguel A., and James M. Robins. 2020. *Causal Inference*. Boca Raton, FL: Taylor & Francis.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Jackson, Robert, and Paul Wang. 1996. *Strategic Database Marketing*. Chicago, IL: NTC Publishing.

Jordan, Michael I. 2019. "Artificial Intelligence – The Revolution Has Not Happened Yet". *Harvard Data Science Review*, issue 1.1. https://hdsr.mitpress.mit.edu/pub/wot7mkc1

Lee, Myoung-Jae. 2005. *Micro-Econometrics for Policy, Program, and Treatment Effects*. New York, NY: Oxford University Press.

Lee, Myoung-Jae. 2016. *Matching, Regression Discontinuity, Difference in Differences, and Beyond*. New York, NY: Oxford University Press.

Lo, Victor S. Y. 2019a. "Searching for the Perfect Unicorn". *Informs: Analytics*, 1(1). https://doi.org/10.1162/99608f92.ba20f892

Lo, Victor S. Y. 2019b. "Three Types of Analytics Used in Practice and their Links". *CIO Review*.

Meng, Xiao-Li. 2019. "Data Science: An Artificial Ecosystem". *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.ba20f892

Morgan, Stephen, and Christopher Winship. 2015. *Counterfactuals and Causal Inference Methods and Principles for Social Research*, 2nd edition. New York, NY: Cambridge University Press.

Rosenbaum, P. R. 2002. *Observational Studies*. New York, NY. Springer.

Rosenbaum, P. R. 2010. *Design of Observational Studies*. New York, NY. Springer.

Rosenbaum, Paul R., and Donald B. Rubin 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika*, 70(1): 41–55.

Rubin, D. B. 2006. *Matched Sampling for Causal Effects*. New York, NY: Cambridge University Press.

Rubin, D. B., and R.P. Waterman. 2006. "Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology". *Statistical Science*, 21: 206–222.

Samuelson, Douglas A. 2013. "Analytics: Key to Obama's Victory". *OR/MS Today*, Feb, 20–24.

Scherer Michael. 2012. How Obama's Data Crunchers Helped Him Win". CNN News. http://www.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team/index.html?hpt=hp_bn5

Scutari, Marco, and Jean-Baptiste Denis. 2015. *Bayesian Networks: With Examples in R*. Boca Raton, FL: CRC Press.

Siegel, Eric. 2011. "Upilft Modeling: Predictive Analytics Can't Optimize Marketing Decisions Without It". Prediction Impact white paper sponsored by Pitney Bowes Business Insight.

Siegel, Eric. 2013a. "The Real Story Behind Obama's Election Victory". *The Fiscal Times* 01/21/2013. http://www.thefiscaltimes.com/Articles/2013/01/21/The-Real-Story-Behind-Obamas-Election-Victory.aspx#page1

Siegel, Eric. 2013b. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. New Jersey: Wiley.

Sloman, Steven. 2005. *Causal Models: How People Think about the World and Its Alternatives*. New York: Oxford University Press.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*, 2nd edition. Cambridge, MA: MIT Press.

Taylor, Petroc. 2024. "Amount of Data Created, Consumed, and Stored 2010-2023, with Forecasts to 2028". Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

Vanderweele, Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.

Weisberg, Herbert I. 2010. *Bias and Causation: Models and Judgment for Valid Comparisons*. Hoboken, NJ: Wiley.

Wolpert, David H. 1995. The Relationship Between PAC, the Statistical Physics Framework, the Bayesian Framework, and the VC Framework. In D. H. Wolpert (ed.), *The Mathematics of Generalization*. Reading, MA: Addison-Wesley, 117–214.

Wolpert, David H. 2002. The Supervised Learning No-Free-Lunch Theorems. In R. Roy, M. Köppen, S. Ovaska, T. Furuhashi, and F. Hoffmann (eds.), *Soft Computing and Industry*. London: Springer. https://doi.org/10.1007/978-1-4471-0123-9_3

Yong, Florence H. 2015. "Quantitative Methods for Stratified Medicine". *PhD Dissertation*, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University.

Zink, Richard C., Lei Shen, Russell D. Wolfinger, and H.D. Hollins Showalter. 2015. Assessment of Methods to Identify Patient Subgroups with Enhanced Treatment Response in Randomized Clinical Trials. In Zhen Chen, Aiyi Liu, Yongming Qu, Larry Tang, Naitee Ting, and Yi Tsong (eds.), *Applied Statistics in Biomedicine and Clinical Trials Design*. Switzerland: Springer International Publishing

2

Review of Common Data Mining Techniques

2.1 Introduction

One might think that data mining or data reduction techniques should have no place in a book that emphasizes causality. After all, data mining mainly seems to ask large amounts of data to simply come up with patterns – recognizing faces, seeking anomalies in business accounts, or identifying an individual's spending patterns.

Sometimes this is indeed enough, especially in applications of unsupervised learning. Recognizing common elements in photos so they may be compiled into albums is useful but does not need to address cause and effect.

However, in many cases, we want to harness the tools of data mining in the interest of identifying or not rejecting potential causal pathways. We supervise learning for this purpose. For instance, we may want to know why some people switch vendors for phone service. The interesting question here is what the company could do to keep its customers from moving, thereby reducing "churn." It is not enough simply to find the correlates of churn; we need to identify which variables (under our control) can reasonably be thought to affect churn and measure the direction, strength, and form of these influences.

In this chapter, we examine several of the main techniques that help us better address causality. They do not show or prove causality – that may not be truly possible – but they can often test whether our prior assumptions of causality are plausible; and where causal effects seem to be working, they can often help quantify them.

We start with the most widely used and familiar technique, linear regression, because this is where almost all analysis begins; readers familiar with the subject may want to skim this section. We then turn to the problem of classifying observations – for instance, identifying the determinants of whether someone will buy a car. Often this is done with logistic regression, but several other techniques have become popular, including classification and regression trees (CART), random forests, gradient-boosted trees, neural nets, and support vector machines (SVMs). These techniques may also be used to create lift tables, which are discussed in greater detail in Chapters 6–9.

DOI: 10.1201/9780429172588-2

None of these techniques can prove causality; instead, they reflect back to us the causal assumptions we make, whether implicitly or explicitly. Almost all causal analysis requires us to begin with some serious thought, often formalized in the form of analytical modeling, and later chapters address this more fully. Nonetheless, it is sometimes possible to rule out causal effects, and this can be useful.

To illustrate the techniques and ideas in this chapter, we draw heavily on models of employee earnings and employee attrition. In the regression context, we will examine what variables plausibly "explain" earnings. And then we will look at classification techniques, where the key question is: What could the company do to reduce employee attrition, to keep employees on board?

In both cases, we need to have a model in mind. For example, we may believe that employees who are better paid, get stock options, or do more overtime are more willing to stay. The data also allow us to look at the proximate drivers of earnings at the firm, which will prove to be useful in our discussions below.

Where do these ideas – theories or models – come from? They arise from our own personal experience with employment, from talking to experts, and from reviewing the academic literature. But we rarely, if ever, start with a blank canvas, which is fortunate because the model helps us organize the data and guides our choice of techniques.

The data for the employee earnings and attrition example simulate the results of a survey of 1,470 individuals at a U.S. company. The dataset was created for training purposes by data scientists at IBM, and we have modified it somewhat.[1] While the results look plausible, they should not be used to infer anything new about the "real" world. Table 2.1 summarizes the information. These are the data we have to work with, and they constrain our ability to answer all the questions of interest, but this is realistic, as we almost never have information on everything that we would like. From Table 2.1, we see that the employees are, on average, 37 years old, earn $6,503 per month, and have worked for 11 years.

2.2 Linear Regression

An important question for any Human Resources Department is whether there is gender discrimination in the workplace.

A natural place to start is to ask whether men earn more than women, after controlling for other relevant factors. In our dataset, the mean monthly income for men was $6,687, and it was $6,381 for women. Starting with the null hypothesis that both men and women are paid the same, a t-test shows that there is a fairly low probability (11%) of obtaining these results – the p-value is 0.11 – if the null hypothesis is true. There is thus a strong suggestion that there really is a difference between the incomes of men and women.

TABLE 2.1

Summary Statistics for Employee Data

Variable	Total or %
Monthly income (USD)	6,503
Age	36.9
Gender (% female)	60.0
Total years worked	11.3
Years worked at current company	7.0
Educational level (%)	
No college	11.6
Some college	19.2
Bachelor's degree	38.9
Master's degree	27.1
Doctorate	3.3
Job classification (%)	
Level 1 (low)	36.9
Level 2	36.3
Level 3	14.8
Level 4	7.2
Level 5 (high)	4.7
Educational background	
Human resources	1.8
Life sciences	41.2
Marketing	10.8
Medical	31.6
Technical degree	5.6
Other	9.0

Note: Based on modified data from the Kaggle attrition dataset, used here for illustrative purposes.

The issue is, what? For instance, older and more-experienced employees typically earn more, and if they are disproportionately male, then men will (on average) be earning more than women, but not necessarily because of gender. Again, if male employees are overrepresented in roles such as research scientist that require high educational skills, we risk confounding high pay (for being a research scientist) with high pay (for being male).

Regression, including linear regression, can help us disentangle these effects. Let Y_i be a measure of the variable of interest for individual i (the left-hand-side variable, or "target" or "dependent" or "outcome" variable), and $X_{1i}, \ldots, X_{ki}$ be a set of "explanatory" variables (also known as control variables, independent variables, or regressors). Then we may hypothesize that:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \tag{2.1}$$

This says that Y_i results from a linear combination of X_i variables. The weights are the coefficients ($\beta_0 \ldots \beta_k$), and these may be estimated (giving $\hat{\beta}_0 \ldots \hat{\beta}_k$) if we have the appropriate dataset. The ε_i represents an error term

since we do not expect the linear relation in Eqn. 2.1 to be exact. Every statistical package allows one to estimate equations like this easily and efficiently.

To illustrate, suppose Y_i represents monthly income, X_1 is a binary ("dummy") variable that represents gender, set to 1 for females and 0 for males; and the other X variables control for age, education, and experience. Based on our dataset (discussed above), the estimated values of the coefficients of these, and some other variables, are shown in column (1) of Table 2.2. For gender, the value is 16.1, which might be taken to indicate that women earn $16 per month more than men, holding other things constant. However, this estimate is not statistically significantly different from zero; we know this from the p-value, given in column (2), which shows a very high probability (0.81) that the estimated coefficient is consistent with the null hypothesis of no effect (i.e., of a coefficient of zero). Based on this model, we do not find evidence of a gender effect on wages.

In this case, the direction of potential causality is clear, since gender maps to income and not income to gender, although even here, if high incomes disproportionately attract high-performing women, the observed relationship could go in the other direction.

The measurement of any effect is contingent on the rest of the model being appropriate: For instance, have we included all the relevant controls, are they measured correctly, is the linear form appropriate, and is the sample unbiased? The adjusted R^2 here is 0.926, so this model is "explaining" most of the variation in monthly incomes. Indeed, there may be too many variables in this model: As working people age, they also get more experience, so it is almost certainly inappropriate to include variables for both age and experience as regressors.

In his classic model, Jacob Mincer (1974) argues that the appropriate form of an earnings function would use the log of monthly earnings as the dependent variable, with measures of education and experience on the right-hand side. It should also allow for some curvature related to experience. The idea is that as workers gain experience, their earnings first rise quickly and then more slowly, perhaps even peaking and declining before they retire. One way to model this is to include both years of experience and the square of experience as right-hand variables.

Estimates of a Mincerian model of wages are shown in column (2) of Table 2.2. The gender variable now appears to be negative but again is not statistically significant. Holding other variables constant, experience is associated with higher earnings for the first 12.5 years, after which the effect of further experience flattens and declines.[2]

A popular, if more mechanical, approach to developing an estimating model is to use stepwise regression, either forward (where one adds variables one by one if they meet certain thresholds of statistical significance) or backward (where one removes variables one by one if they do not meet certain thresholds). We apply a backward stepwise procedure to the Micerian model, removing variables if they are not statistically significant at the 20% level or better, and show the results in column (3) of Table 2.2. It is a more parsimonious model than the one in the middle columns and fits equally well.

TABLE 2.2

Regression Estimates for Income Model

	(1) Linear Model		(2) Mincer Model		(3) Mincer, Stepwise	
Dependent Variable	**Monthly Income**		**Ln (Monthly Income)**		**Ln (Monthly Income)**	
Variable	**Coeff.**	**p-value**	**Coeff**	**p-value**	**Coeff**	**p-value**
Gender (F = 1, M = 0)	16.1	0.814	–0.001	0.992	–0.0004	0.972
Age	43.1	0.172				
Age squared	–612.0	0.124				
Total years worked	81.4	<0.001	0.025	<0.001	0.024	<0.001
Years worked, squared	–1.7	0.003	–0.001	<0.001	–0.0006	0.000
Years worked at current co.	–12.9	0.076	–0.0002	0.862		
Educational level						
No college (ref.)						
Some college	–267.2	0.036	–0.035	0.133	–0.040	0.018
Bachelor's degree	–228.2	0.046	–0.029	0.172	–0.033	0.015
Master's degree	-46.3	0.706	–0.008	0.730		
Doctorate	–187.8	0.380	–0.002	0.968		
Job classification						
Level 1 (low) (ref.)						
Level 2	2500.7	0.000	0.616	0.000	0.619	<0.001
Level 3	6732.2	0.000	1.186	0.000	1.188	<0.001
Level 4	12387.6	0.000	1.656	0.000	1.657	<0.001
Level 5 (high)	16066.6	0.000	1.877	0.000	1.879	<0.001
Educational background						
Human resources (ref.)						
Life sciences	–128.0	0.613	0.0004	0.993		
Marketing	–119.1	0.658	0.020	0.682		
Medical	–76.3	0.764	0.007	0.876		
Technical degree	–132.2	0.627	–0.004	0.943		
Other	–182.4	0.523	0.002	0.976		
Intercept	1990.5	0.001	7.797	0.000	7.806	<0.001
Adjusted R-squared	0.926		0.873		0.874	

Note: Based on modified data from the Kaggle attrition dataset, used here for illustrative purposes.

Increasingly, the preferred method of mechanical variable selection, also referred to as feature selection, is to use a Lasso (least absolute shrinkage and selection operator), which is a method for regularizing the regression coefficients and dropping unhelpful variables, thereby making the model easier to interpret as well as more accurate at prediction (Tibshirani 1996). There are several generalizations of the basic Lasso. A particularly useful one is the elastic net, which is applicable in cases where the number of variables exceeds the number of observations so that the number of variables has to be seriously trimmed (Zou and Hastie 2005).

While it is often the case that most of the control variables are continuous (such as age), this need not be the case. For instance, in our example, jobs are classified into five levels, from lowest-skilled (1) to highest-skilled (5). To include such information in the regression, first pick one job level as the reference point, and then include separate dummy variables for each of the other categories. The model whose results are shown in Column (1) of Table 2.2 shows that those whose jobs are classified at level 5 can expect to make $16,067 more per month than someone at job level 1, or about 6.5 times as much, according to the log-linear models.[3]

2.2.1 Judging the Model

Regression is the traditional workhorse of modeling. But how do we know whether the model that we have estimated is reasonable? This is as much an art as a science, but there are a number of things to check. We first need to ask whether the underlying logic is solid, a topic to which we turn in Chapter 3.

Some argue that if a model forecasts well, then it is a good model (Friedman 1953). This is unsatisfactory at both a philosophical and a pragmatic level. To see why, consider the following model: Proposition 1 is that all dogs are cats. Proposition 2 is that all cats bark. And so we conclude logically that all dogs bark. While the prediction is quite good (except for our husky, who only howled), the theory is decidedly odd. Note too that we cannot infer anything about the veracity of the assumptions based merely on the outcome.

Pragmatically, a model built on sand will, in due course, fail to perform. It will lack external validity, which means that it will be difficult to apply it to situations that are similar, but not identical, to the ones for which it was designed.

If a model is a priori reasonable, we typically look at the goodness of fit, measured by the coefficient of determination, R^2, which measures the proportion of the variation (around the mean) of the dependent variable that is "explained" by the regression. The value of R^2 varies from 0 (no relationship) to 1 (perfect fit), but there is no absolute standard for what constitutes a good fit. When more variables are added to the right-hand side of a regression model, R^2 will necessarily rise, but this may be at the expense of greater complexity, even confusion. One fix is to report the adjusted R^2, which takes into account the number of variables in the model and will only rise if a newly added variable is at least moderately statistically significant (with a p-value

of about 0.34 or lower); or to use an information criterion such as the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC). An over-exuberant search for a good fit – hunting for R, or chasing R – can lead to models that may have shaky foundations and poor external validity.

Given our prior beliefs, we then need to see whether the estimated coefficients are of the "right sign" and statistically significant. The estimated coefficients ($\hat{\beta}$s) are random variables: If we had randomly taken a different sample of data, we would have come up with slightly different values for the $\hat{\beta}$s. So the $\hat{\beta}$s have a distribution, which will be normal (i.e., Gaussian) if the sample size is sufficiently large, and has a standard error (i.e., an estimated standard deviation). It also raises the possibility that we have an estimate of $\hat{\beta}$ that is only nonzero by chance. If the associated p-value is small, then there is a strong indication that the coefficient is indeed not zero – or put another way, legitimately belongs in the equation or model. Many analysts use a p-value cutoff of 0.1 or 0.05 when assessing statistical significance, but these thresholds are, of course, somewhat arbitrary.

Even if an estimated coefficient appears to be statistically significant, it may have the "wrong" sign. For instance, we would be surprised if employees with more education regularly had lower incomes than those with less education. Sometimes these surprises stand up to scrutiny, and there is a plausible explanation, but often they are a symptom of a problem with the model, or perhaps the data.

2.2.2 Regression Diagnostics

When regression is used for inference – to test hypotheses – then there are a number of diagnostic checks that are called for. There is a huge literature on the subject, so we only review the issues briefly as a form of refresher; a more complete treatment may be found in any econometrics textbook (e.g., Wooldridge 2020, Cameron and Trivedi 2022) or in some more specialized monographs (Haughton and Haughton 2011).

2.2.2.1 Multicollinearity

When two (or more) right-hand-side variables are correlated, we have multicollinearity. If, for instance, X_1 and X_2 more together too closely, then it is difficult to separate the effects of X_1 and X_2 on the outcome (Y) variable, and the coefficient estimates will be imprecise. For instance, in the Mincer model estimated above, educational levels and job classifications are likely closely related, making it hard to disentangle the effects of one from the other. The problem becomes more serious when there are scores of independent variables to choose from, as it may not be clear which variables to pick.

One solution is to use more data, in the hope of seeing more movement in X_1 that is independent of X_2, but this is rarely realistic. A simple matrix of correlation coefficients of the X_i variables allows one to see whether

multicollinearity risks being a problem. If it does not affect the variables on which we are focusing, or if we are simply interested in making predictions, then the multicollinearity need not be fatal. If it does, one may choose the X variable that is more correlated with the outcome variable to enter into the model and leave out the other one.

2.2.2.2 Heteroskedasticity

If a graph of the regression residuals – that is, $y_i - \hat{y}_i$ – against the dependent variable (y_i) shows a nonrandom pattern, then we have heteroskedasticity. The Breusch-Pagan test is a popular measure for detecting heteroskedasticity (Wooldridge 2020), which is common, indeed almost standard, in cross-sectional data. While the presence of heteroskedasticity does not alter the coefficient estimates, it gives inaccurate standard errors and tests of statistical significance.

A common solution with economic and financial data such as income or spending, where the underlying distribution tends to be highly skewed to the right, is to use the log of these variables in the model, as done in the Mincer model of earnings reported in Table 2.2. Another solution is to use a robust estimator such as White's estimator, which is easily done in most statistical packages.

2.2.2.3 Outliers

It is not unusual to see a small number of exceptionally high or low values of the Y or X variables. These are outliers and are often identified using box and whisker plots. A single outlier can have a strong *influence* on the estimated regression coefficients, especially in a small dataset. This may be measured by the "difference in fits" (dfits), which is the difference between the predicted values of y_i based on regressions with or without the i-th observation. Some observations also have high *leverage*, meaning they are particularly important drivers of the coefficient estimates.

In some cases, researchers simply drop the outlier observations, or observations at the tails, such as the top or bottom 5%, giving a trimmed (or truncated) estimate. Others replace the identified outliers with the mean value of the variable or censor (Winsorize) the data by bringing (say) the bottom and top values to the 5th and 95th percentiles. Trimmed and Winsorized estimates are more robust to outliers, but in the process, there is a loss of potentially valuable information. Some stock indexes censor extreme values of individual component stocks.

2.2.2.4 Measurement Error

In some contexts, such as survey data, outliers are often viewed as a form of measurement error. But nothing is measured with perfect accuracy, so some

degree of measurement error is inevitable. When the outcome (Y) variable is measured inaccurately, with a random additive error, this reduces the fit of the regression but does not bias the estimates of the coefficients. When the X variables are measured with error, they become noisy and less informative, which biases the estimated coefficients toward zero. There is no easy way to address measurement errors, but careful data cleaning and collection do help.

2.2.2.5 Omitted Variable Bias

Even the best-formulated model will have omissions; variables that one would ideally like to include but are either unobservable (such as a person's inherent "ability") or unobserved (such as a person's weight if data on weight were not collected). The problem arises when an omitted variable such as ability is correlated with an included variable such as years of education, in an earnings equation, for example. The effect of education on earnings will then be overstated because it is, in effect, channeling the effects of both ability and education.

When panel data are available, for instance, showing the education and earnings of a cohort of individuals over time, it may be possible to control for time-invariant unobservables such as ability. Some researchers use stepwise methods to add or subtract variables from a regression model, but this does not address the problem of unobservables, and it is an imperfect substitute for working through the elements of the right model.

2.2.2.6 Simultaneity

In estimating a regression model, we implicitly suppose that changes in the X variable cause changes in Y. For instance, consider a classical demand curve in economics, where

$$Q = f\left(P,\ P_{other},\ income,\ tastes\right)$$

Here, Q is the quantity bought, P is the price of the good, and P_{other} is a vector of prices of other goods. To estimate a demand curve, it would be tempting to estimate

$$Q = a + b\ P + error$$

perhaps with other right-hand variables as well. But the problem is that while a higher price is expected to reduce the quantity bought (Q), causality may run in the other direction, where a high demand (Q) – such as everyone wanting to take Uber on a rainy evening – may lead to a higher price.

We have a lot more to say about handling causality in subsequent chapters, but it is worth noting that the problem is widespread. In some cases, we can tease out the causal effects and analytically or statistically address the issue of simultaneity. We begin the task in Chapter 3.

2.3 Classification: Logistic Regression

When a company sends out offers for a new credit card, or a new phone provider, or asks for money for a worthy cause, people respond either positively (yes, 1) or negatively (no, 0). We would like to be able to predict, and even explain, why some people take up the offer and others do not, based on information that we already have, such as age, location, credit score, gender, and so on.

If we have a causal model in mind – developed perhaps with the help of the graphical methods we discuss in Chapter 3 – then we may be able to estimate the effects of the input variables on the outcome of interest. At issue is how best to do this.

Consider an example where we have a phone provider that is hoping to recruit clients in a developing country. We have information on a sample of 514 individuals who were offered a new service, 19.5% of whom accepted the offer. We also have a set of input variables that we believe influence the outcome. Summary statistics for this (partly hypothetical) example are shown in Table 2.3.

Each point in Figure 2.1 represents a person who either accepted the offer of a new service provider (shown as 1 on the vertical axis) or did not (0 on the vertical axis), graphed against household income. Note how the observations cluster at the top and bottom of the graph.

To measure the effects of household income on the outcome, we could simply estimate a linear regression. The simple case generates the straight line shown in Figure 2.1, and the coefficient estimates for a more complete linear model are reported in column (1) of Table 2.4.

TABLE 2.3

Summary Statistics for Example of Uptake of New Phone Service

	Variable	Mean
Outcome	Accepts a new phone provider	19.5
Inputs	Region: North	31.3
	Region: South	10.7
	Region: East	28.2
	Region: West	29.8
	Gender of household head (% male)	87.2
	Age (years)	47.2
	Education (years)	2.3
	Family size	6.0
	Household income	52.4

Note: Sample size: 514 individuals. Data adapted by authors for illustrative purposes.

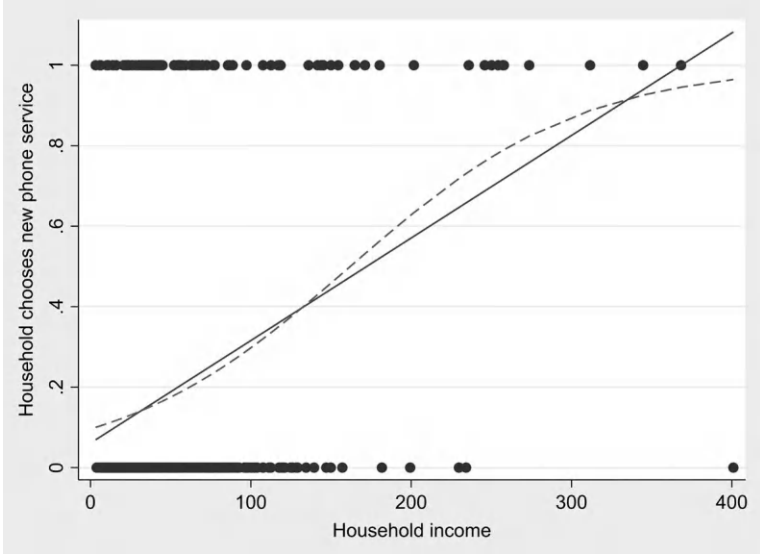


FIGURE 2.1
Income versus household decision to enroll in phone service.

TABLE 2.4

Comparing Linear with Logistic Regression

	Coefficients: Linear Regression	Coefficients: Logistic Regression	Marginal Effects: Logistic Regression
	(1)	(2)	(3)
Dependent variable: household chooses new phone provider			
Region (North is reference)			
South	0.070	0.449	0.072
East	–0.042	–0.311	–0.042
West	–0.117	–1.097	–0.123
Gender (female is reference)			
Male	–0.060	–0.456	–0.060
Continuous variables			
Age of head (years)	0.001	0.010	0.001
Education of head (years)	0.030	0.189	0.023
Size of household	–0.013	–0.091	–0.011
Household income	0.0021	0.0128	0.0016
Intercept	0.135	–2.021	
R-squared	0.193	0.194	

Note: Adjusted R-squared in column (1) and pseudo R-squared in column (2). Data adapted by authors for illustrative purposes.

A problem with this approach is that the linear regression line clearly does not fit the data well. It also could predict a probability of accepting the offer of more than 1, which makes no sense! Note that the marginal effect of additional income, in this simple case, is 0.0021, meaning that if income rises by one unit, the probability of accepting the offer of a new phone service will rise by 0.0021 (or 0.21 percentage points).

A more satisfactory approach would be to estimate a logistic regression. The idea is to transform the estimating equation so the outcome variable is confined to the interval (0,1). The estimating equation for linear regression is

$$Y = X\beta + \varepsilon \quad \text{or} \quad E(Y) = X\beta,$$

where Y is the binary outcome variable, X represents a series of "explanatory" variables, β is a vector of coefficients, and ε is the error term. By way of contrast, the estimating equation for logistic regression, which models the probability that the outcome variable takes on the value (which is the same as the expected value of the binary outcome Y), looks like this:

$$P(Y=1) = \frac{e^{X\beta}}{1+e^{X\beta}} \quad \text{or} \quad E(Y) = \frac{1}{1+e^{-X\beta}}.$$

If $X\beta$ is large, $Y \to 1$, but if $X\beta$ is small, $Y \to 0$. Mechanically, it is easy to estimate. In Stata, for instance, the regression command is

```
reg Y X
```

while the logistic regression command is

```
logit Y X
```

The line produced by the logistic regression in our example, where Y is "change phone provider" and X is household income, is given by the dashed line in Figure 2.1.

The estimates of the raw coefficients for the logistic regression are shown in the middle column of Table 2.4. These estimates are not of much interest in themselves. Usually, we are more interested in the marginal effects, or what happens to $E(Y)$ when we change X (i.e., $\partial E(Y)/\partial X$). In linear regression, this is just the slope, given by the coefficient, but with the logistic curve, the slopes vary – first low (flat), then higher, and then lower again, as we move from left to right in Figure 2.1.

One solution is to calculate the marginal effects at the average values, using the margins command in Stata (or R), as done in the right-hand column in Table 2.4. If we need more detail, we could request the margins at multiple points – for instance, at household income levels of 30 (margin = 0.0014),

150 (margin = 0.0026), and 300 (margin = 0.0020), and compare these with the ordinary least squares regression margin of 0.0021. In our example, the marginal effects from the logistic regression are close to those generated by the linear regression, but this is not always the case.

Logistic regression is one of the workhorses of analytics, but it has limitations. It does not easily handle interactions – for example, when the influence of education also depends on the age of the individual – and non-linearities, and is constrained by its parametric form. For greater flexibility, we may want to use some form of classification or regression tree, which is the topic to which we now turn.

2.4 Classification: Trees

An alternative way to approach the question of how "independent" variables influence an outcome is to build a tree, either singly or with the help of random forests or gradient boosting.

To see how trees work, we start with a simple example, using the data shown in Table 2.5. Each column shows information about an individual. Of the 20 people in this sample, 12 left the firm; 12 had been working overtime; and 11 had been receiving stock options. We would like to know whether overtime work and stock options influence the decision to leave the firm. A simple linear regression gives:

$$\text{Leave} = 0.55 + 0.40\,\text{overtime} - 0.36\,\text{stock option} \qquad R^2 = 0.27$$

If the equation worked perfectly, it would correctly predict who would leave and who would not, but in this case, 6 of the 20 cases are not predicted correctly using a cutoff of 0.6. A logistic model, which may be more appropriate in this case, also misclassifies 6 of the 20 cases.

TABLE 2.5

Hypothetical Data for a Simple Employee Attrition Model

Individual	1	2	3	4	5	6	7	8	9	10
Leave firm (Y = 1)	0	0	0	0	0	0	0	0	1	1
Overtime (Y = 1)	0	0	0	0	0	1	1	1	0	0
Stock option (Y = 1)	1	0	1	0	1	1	1	1	0	1
Individual	11	12	13	14	15	16	17	18	19	20
Leave firm (Y = 1)	1	1	1	1	1	1	1	1	1	1
Overtime (Y = 1)	0	1	1	1	1	1	1	1	1	1
Stock option (Y = 1)	0	0	0	0	0	0	1	1	1	1

2.4.1 Classification and Regression Trees

We might be able to do better by building a classification tree. We start with an initial node – the tree trunk – and then proceed to classify the observations into separate branches (or bins). The idea is to try to create bins for identifiable subsets of the data that have similar outcomes: If all the observations in a bin have the same outcome – for instance, they are all leavers – then that bin is "pure," while a bin that consists of equal numbers of leavers and stayers would be completely "impure" and tell us nothing about what drives the decision to leave.[4]

We may illustrate the idea with our simple dataset. Eight of the 20 employees left, so the Gini impurity measure is 0.48. But now let us split the sample into those who do overtime and those who do not. Of the 12 who do overtime, 9 quit, and the impurity measure is 0.38. At the other side of the split, only 3 of the 8 who did not do overtime left the firm, for an impurity measure of 0.47. The average impurity value for the new model is 0.413 (instead of 0.48), and so the classification has helped us see a pattern in the data: It appears that doing overtime is associated with a greater propensity to quit one's job. If we predict that those who do overtime will quit and those who do not do overtime will stay, then our predictions will be correct 70% of the time, as with the regression model. This basic tree is shown in Figure 2.2.

One could also split the sample based on whether employees have stock options or not. Here too, the split reduces impurity, and having stock options is associated with a lower probability of quitting.

If there are enough independent variables, the splitting process of binary recursive partitioning can continue almost indefinitely, generating a tree with a lot of branches and low impurity. Such a tree would almost certainly overfit the data and predict poorly when applied to a different dataset. The standard solution is to estimate the tree based on a training sample, say two-thirds of the data, and then to use the test (or "holdout") sample to prune the

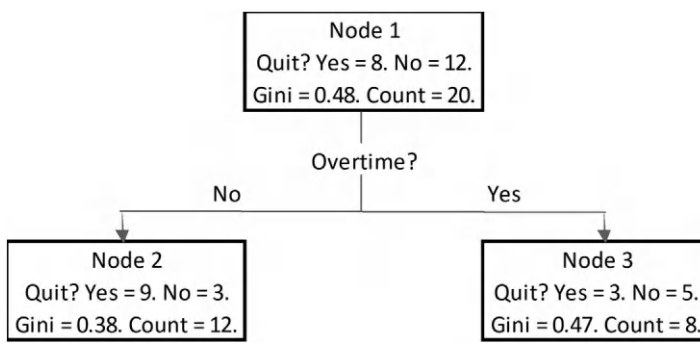


FIGURE 2.2
Classification model for basic example.

tree in such a way as to generate the lowest Gini (or mean square error, in the continuous case) when the model is applied to the testing sample.

CART trees are particularly helpful as pre-processors: They can help show which variables are most relevant because they use a stepwise process to add and then prune nodes. They also show the extent to which interactions and non-linearities are important, and they can handle missing values so that less information is lost than in regression models. On the other hand, trees can be difficult to interpret, can easily sprawl, and are sensitive to the assumptions made about how to handle pruning and the choice of training and validation samples.

To illustrate a more complex tree, we return to the example on employee attrition that was introduced early in this chapter. The trees generated by SAS and Minitab's Salford Predictive Modeler (2023) give essentially the same results; the SAS version is shown in Figure 2.3. The sample is first split according to whether an employee does overtime; very few (10.4%) of those who did not do overtime left the company. Among those who did overtime, the likelihood of quitting was far higher for those who had a monthly income below $2,475 (69.6%) compared to those who made more than that (22.8%). This latter group splits further by job role – sales executive versus research scientist – and the sales executives are then divided into those who have relatively few stock options (59.7% quit) and those who have more (20.0% quit).

This is a typical tree. It is good at pointing to non-linear and interactive effects, but it only uses four variables to make the splits shown in Figure 2.3; if there were many more variables of interest, the tree would quickly become unmanageably large.

2.4.2 Random Forests

While a single decision tree – whether for classification or regression – is straightforward to understand, it may not be very accurate. The building of a tree might get stuck on a single path or fail to recognize the full complexity of interactions and non-linearities. It turns out that greater accuracy may be achieved by growing a random forest. The idea is to build lots of trees independently and then to aggregate the results in a sort of "collective intelligence."

Start with the dataset, with information on the outcome variable of interest, and a set of independent variables. Then draw a random sample (with replacement) of observations from the dataset, and also randomly choose a subset of the independent variables. Build a tree, using the non-chosen observations for validation. Repeat many times. Then use the information on this forest of decision trees for prediction: Use the new information on the independent variables to predict the outcome (e.g., leave the firm or stay) for each tree. If a majority of trees predict that you will leave, then this is the result; otherwise, the forest predicts that you will stay.

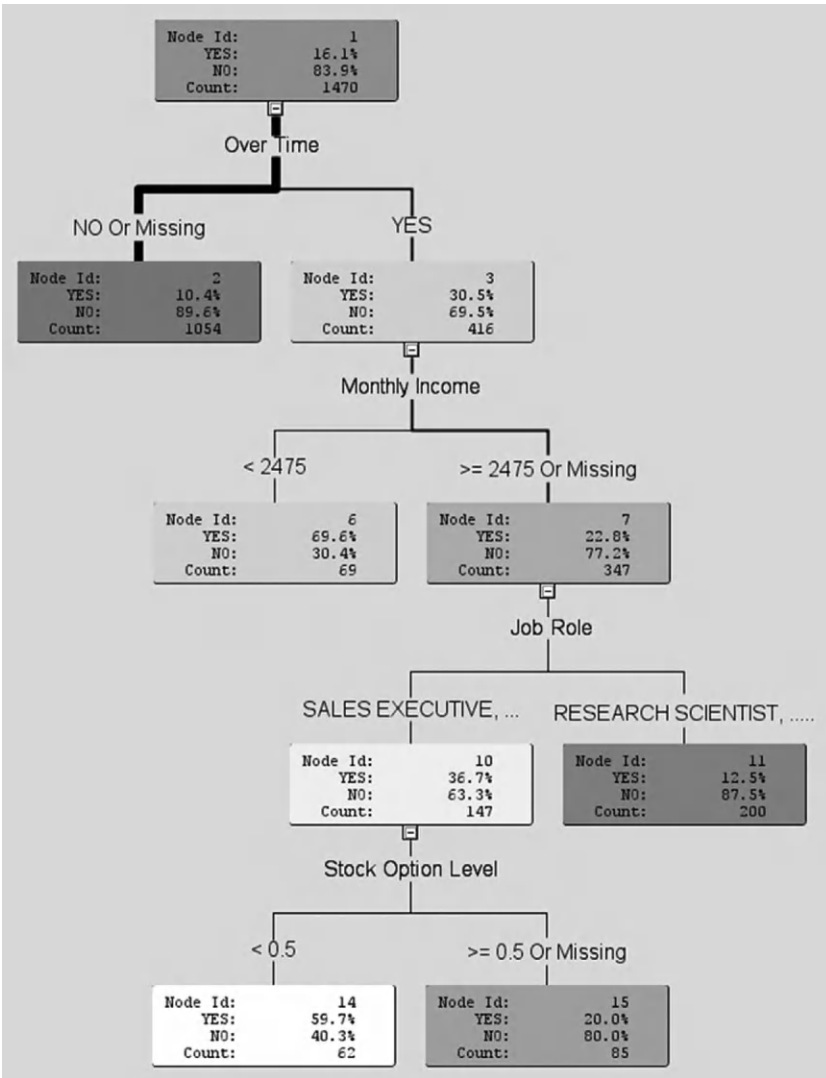


FIGURE 2.3
Descriptive statistics and predictive models of employee attrition.

Random forests are straightforward to run, and the main decisions to make, other than the choice of variables, are how many iterations to run – usually in the hundreds – how many variables to use when building each tree, and how many nodes to allow. A drawback of the random forest method is that it is difficult to interpret – the calculations occur in what seems to be a black box – but the results tend to fit well and to be robust (see for instance Verme 2023).

2.4.3 Gradient Boosted Trees

A single decision tree that relates an outcome (such as whether someone quits their job) to independent variables creates a step function, which inevitably will not fit the data perfectly. The idea behind gradient boosting is that after fitting a first tree, it then fits another tree to the residuals of the first tree, and so on until there is little further improvement in the fit of the overall model, in effect boosting the eventual fit. The technique owes a lot to Friedman (1999), and Salford Systems (now Minitab 2023) has implemented a version of the technique that it calls TreeNet.

In order to avoid the risk of overfitting, the number of trees involved in the approximation is controlled by cross-validation or evaluation of the approximation on a test sample.

There is some evidence that in some circumstances, gradient boosting outperforms random forests in predicting outcomes. As with random forests, gradient-boosted tree models are, like many machine learning techniques, difficult to interpret and hence to explain or evaluate.

2.4.4 Neural Nets

Regression analysis is very powerful, but it can be constraining. Consider the following (overly) simple model of the determinants of monthly earnings:

$$\begin{aligned}\text{Lmthearn} = b_0 &+ b_1 \text{ hrs overtime} + b_2 \text{ yrs experience} \\ &+ b_3 \text{ years in current role} + b_4 \text{ gender}(F = 1).\end{aligned}$$

When we estimate this equation, we are assuming that the relationships are linear, and the right-hand variables do not interact to influence earnings. This parameterization is not always adequate, but often we do not have a clearer idea of what the appropriate functional relationship should look like.

One solution is to model the relationship as a *neural net*. A basic neural net has:

a. An input layer, with information on the (assumed) causal variables – here, the hours of overtime, and so on;
b. One or more hidden layers, each with a series of nodes, as explained below; and
c. An output layer, with the variables that we want to "explain."

The idea is that information from the input layer feeds into the hidden layers, causing the nodes there to "fire" – like neurons in the brain responding to a signal – and these firings then affect the output variables. Schematically, this may be set out as in Figure 2.4, which reflects our example and assumes one hidden layer with three nodes.

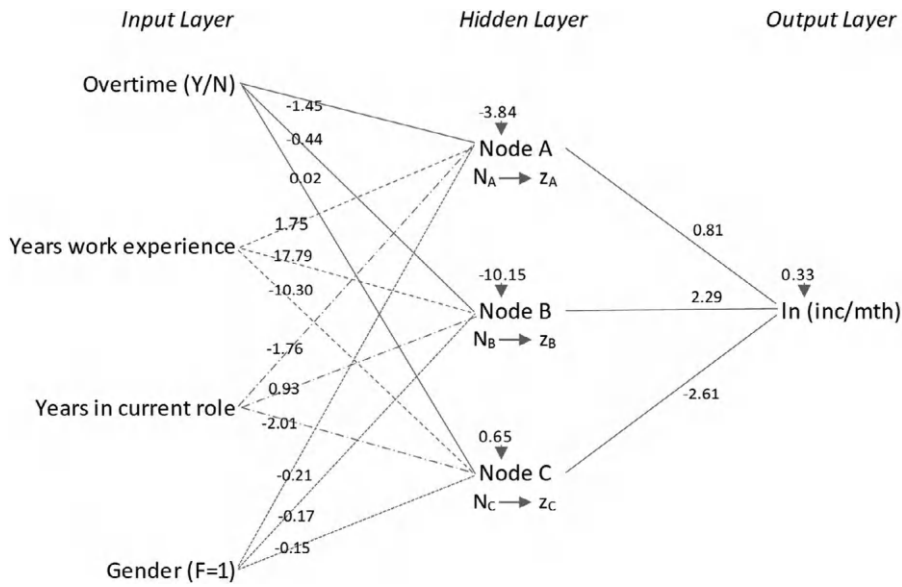


FIGURE 2.4
Schema of neural net model of income. (As for Table 2.1.)

The estimation of a neural network begins by normalizing the observed input and output variables to the closed interval [0,1]. Every relationship in the neural net, shown by lines in Figure 2.4, is represented by a weight. So, for instance, the value of N_A (i.e., the value that is input into Node A) is a weighted average of the values of the input nodes. Thus

$$N_A = w_{0A} + w_{1A}X_1 + w_{2A}X_2 + w_{3A}X_3 + w_{4A}X_4,$$

or, in our example,

$$N_A = -3.84 - 1.45X_1 + 1.75X_2 - 1.76X_3 - 0.21X_4.$$

Within the hidden nodes, the incoming "net" is "squashed" using an activation (or transfer) function that is often a sigmoid function of the form

$$z_A = \frac{1}{1+e^{-N_A}}.$$

Other transfer functions are possible; a "rectifier" or ReLU of the form $f(x) = \max(0, X)$ is reputed to work well, for instance. The sigmoid nonlinear processing magnifies changes in the net if they are close to the mean but dampens them if they are relative outliers. The z_A is typically referred to as a neuron or artificial neuron.

In the final step, the output variable – here the log of earnings per month – is a weighted function of the z_i's that emerge from the sigmoid transformations in the nodes of the hidden layer.

A linear regression is a special case of a neural net, where there is no hidden layer, so the weights linking the input variables with the output variable are just the regression coefficients. When there are multiple hidden layers, we refer to it as a deep learning model.

The challenge in estimating a neural net is to find the best possible set of weights. Because the system is highly non-linear, there is no closed-form solution, and the weights have to be found using a search process. This is largely done by using back propagation (Nielsen 2015), which helps the network "learn" under supervision. Begin with an initial, essentially arbitrary, set of weights, and use these to predict the output variable ($\hat{y}$). Compute the sum of squared prediction errors given by $SSE = \sum_{i\,obs} \left(y_i - \hat{y}_i\right)^2$, which we now want to minimize. Vary the weights one by one to determine the direction they need to move in order to reduce the SSE and iterate until the SSE is close to a minimum.

It is easy to overfit a neural net, so it has become common practice to use a cross-validation procedure. One approach is to designate part – a third, for instance – of the data as a validation sample and build the model on the remaining training sample. At each iteration, apply the revised set of weights to the validation sample, and stop the iterations once the SSE of the validation sample begins to rise. The process can be helped along by the judicious choice of a learning rate and by taking momentum into account. Larose and Larose (2014, Chapter 12) work through a lucid example of the process.

It is instructive to compare the results of estimating a neural net (with sigmoid transfer function, using the brain command in Stata) with those from a regression. Figure 2.4 sets out simple neural net model of the determinants of monthly earnings, showing the optimal weights. These have limited interest in their own right but can be applied in order to predict the values of the dependent variable. We can also measure the effect on the output variable by setting the input variables to zero – one by one – in order to generate "marginal" effects that, once un-normalized, are analogous to regression coefficients. Table 2.6 shows these marginal effects, the coefficients from a regression, and measures of R^2 (for the regression) and its parallel from the neural net.

The neural net fits better, in the sense that it has a higher R^2, as is to be expected given its greater flexibility of functional form, although the difference in this case is not very great. The regression coefficients and marginal effects are quite similar, which suggests that the linear regression may be a reasonable approximation here. The stronger parametric assumptions of the regression model also allow us to test the hypothesis of a gender effect more easily – the p-value is 0.66 – so gender, in this case, does not have a statistically significant effect on wages, although the test is, of course, conditional on the assumptions we have made about the functional form of the regression.

TABLE 2.6

Comparison of Results of Neural Net and Regression Models of the Determinants of the Log of Monthly Earnings

	Mean	Neural Net "Margins"	Regression Coefficients	p-Values
Did overtime? (Y = 1, N = 0)	0.28	0.011	0.003	0.91
Years of work experience	11.3	0.073	0.060	0.00
Years in current role	4.2	0.015	0.015	0.00
Gender (M = 0, F = 1)	0.4	0.015	0.010	0.66
R^2		0.59	0.55	

Source: As for Table 2.1.

We have presented the neural net as a statistical estimator since this is how it is most likely to be used and interpreted in the business context. Computer scientists tend to focus on a neural net as an algorithm, while mathematicians view it as a universal approximator: Hornik et al. (1989) show that it can approximate any theoretical function provided sufficiently many hidden units are available. Two important features of neural nets are worth mentioning: They are feedforward, meaning that effects flow from input notes to hidden nodes to output nodes without any feedback effects, and they represent a completely connected network in that there are no (a priori) blank relationships.

For many users, the biggest problem with neural networks is that they are black boxes – inputs go in, outputs come out, and the intermediate steps lack transparency. This is especially true of deep learning, where there may be multiple hidden layers with many nodes. While the lack of relational clarity may be unsatisfying, it may also represent a useful improvement over restrictive functional forms. As with regression, neural networks do not show causality, and serious thought needs to go into defining the appropriate inputs and outputs – an issue we tackle in other chapters. On the other hand, neural networks are not as easy as regressions to use to reject relationships and so may be less helpful in helping us trim noncausal links.

2.4.5 Other Models

These are not the only techniques that have been used to try to classify outcomes or make predictions. SVMs are widely used in classifying images and in the sciences. The idea is to find a hyperplane (e.g., a line in two-dimensional space) that separates two groups of points defined by the two values of our target variable (such as leave/stay).

In many cases, the points will not be well linearly separated, and finding a suitable hyperplane without transforming the data will prove near impossible. The graphs below, due to Kim (Everything You Wanted to Know about the Kernel Trick, 2013), illustrate the problem and its solution.

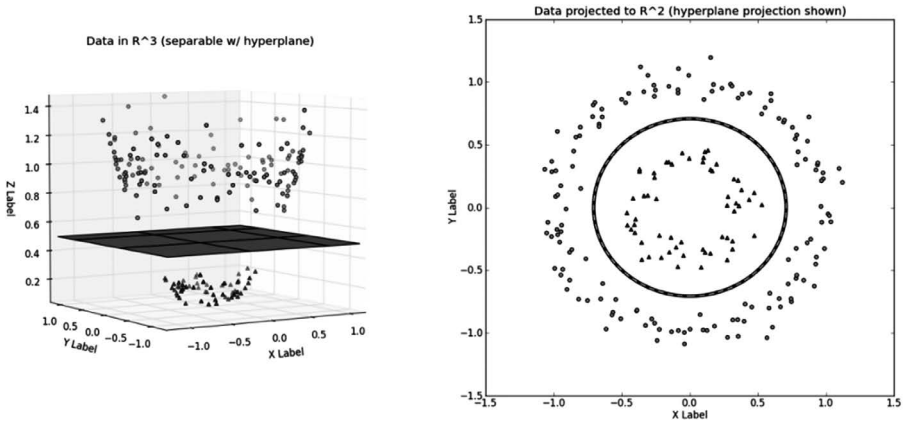


FIGURE 2.5
Illustrations of fitting a support vector machine.

The two groups in the right-hand-side panel can be separated by a circle, but not by a line, whereas transforming the data by the transformation $T(x_1, x_2) = \left[x_1, x_2, x_1^2 + x_2^2\right]$ allows for a separating plane in the three-dimensional space, as shown on the left in Figure 2.5. The art of building SVM models involves finding clever transformations, typically into higher-dimensional spaces, that make the two groups linearly separable, or nearly so.

There is keen interest in stacking, which is a form of model averaging: Random forests average over many trees, but the idea may be generalized, and results from quite different methods can, in principle, be aggregated, yielding predictions that may be more robust (Ahrens et al. 2023).

2.4.6 Which Model to Use?

Given two or more binary classification models (or diagnostic tests) – for instance, linear or logistic regression, neural nets, random forests, or gradient-boosted trees – the issue arises of which model is best, in the sense of being most useful for our purposes. Here we discuss some approaches that help address the issue.

The most traditional approach, largely used to choose between competing linear regression models, is to compare values of adjusted R^2, although the danger here is that analysts will simply tweak their models in their hunting for R^2. It is also fairly common to apply an information criterion such as the BIC or AIC, which favors fit (as measured by the log of the likelihood function), with a penalty for models that have too many variables, or in other words, are not parsimonious.

This approach works less well with highly non-linear models or with classification models. A popular alternative is to measure and compare the area under the receiver operating characteristic (ROC) curve; in business contexts, a common alternative may be to construct lift tables. We now discuss both of these techniques.

2.4.7 AUC

Consider the following issue: We want to predict who is most likely to quit. Using data from past experience, we model this binary outcome – quit or not – using a logistic regression, a neural net, and random forests. Which model should we use?

The problem is that while the underlying variable to be predicted is binary, our models generate a continuous variable that measures the probability that someone will quit. It is possible that one model does well at identifying just a handful of very likely quitters, while another is better at correctly identifying a wider group of potential leavers.

Table 2.7 shows the number of sample observations in four possible states. A model either identifies someone as a quitter (or sick), in which case you "test positive," or as a non-quitter ("test negative"). In reality, in the sample, you are identified as an actual quitter (or sick) or a stayer.

The number of quitters (or sick people) is A + B. A fraction, A/(A + B), is correctly identified by the model as quitter (or being sick). This is a measure of sensitivity, or of the power of the test. Its complement, B/(A + B), is the false negative rate; it shows the proportion of quitters (or sick) who are incorrectly identified by the model as stayers (or well) and is the Type II error.

Similarly, D/(C + D) is a measure of specificity and gives the proportion of non-quitters (or healthy people) who were correctly identified as staying with the firm (or being healthy). Its complement, C/(C + D), gives the proportion of false positives – that is, the proportion of non-quitters (healthy people) who were predicted by the model to be quitters (sick people), measuring the Type I error. The accuracy of the model is given by (A + D)/(A + B + C + D), or the proportion of cases that are classified correctly.

TABLE 2.7

Comparing Actual with Predicted Classifications

	Test Is	
	Positive (Quit/Sick)	**Negative (Stay/Well)**
You actually:		
Quit/Are sick	A	B
Stay/Are well	C	D

A perfect test would have a sensitivity of 1 (so no false negatives) and zero false positives (i.e., a specificity of 1). This test, or model, would exactly identify who will quit and who will not. In practice, this is never achieved: As we tighten the conditions required for identifying positives, the power of the test (sensitivity) falls and specificity rises – the more cautious we are about identifying someone as a quitter, the more likely we are to correctly identify non-quitters as staying.

The tradeoff between sensitivity and specificity is captured by the ROC curve, which shows sensitivity on the vertical axis and (1-specificity) on the horizontal axis. A perfect test would put us at point A in Figure 2.6. A completely noisy test would have us on the diagonal. The curve for an informative model (or test) will typically be curved, as shown by the dashed curve. At a point such as B, we use a rigorous cutoff to determine whether someone is a quitter (or sick); at a point like C, the cutoff is less rigorous, and we get more false positives but fewer false negatives.

The solid curve in Figure 2.6 shows the ROC curve for a different model. As shown, it is a consistently better model because it is closer to the perfect model (point A). The ROCs may intersect, in which case it is harder to choose between them. One solution is to measure the area under the ROC, known as the "area under the curve" (AUC). For a perfect test, the AUC would be 1; for a random test, it would be 0.5, so a higher value of the AUC is typically seen as indicating a better model. In this context, the Gini coefficient is given by

$$\text{Gini} = 2 \times AUC - 1.$$

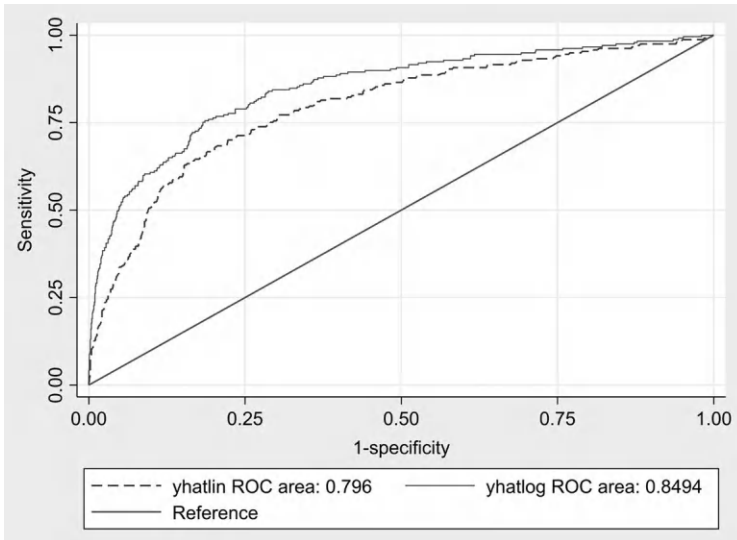


FIGURE 2.6
ROC curves for linear and log models.

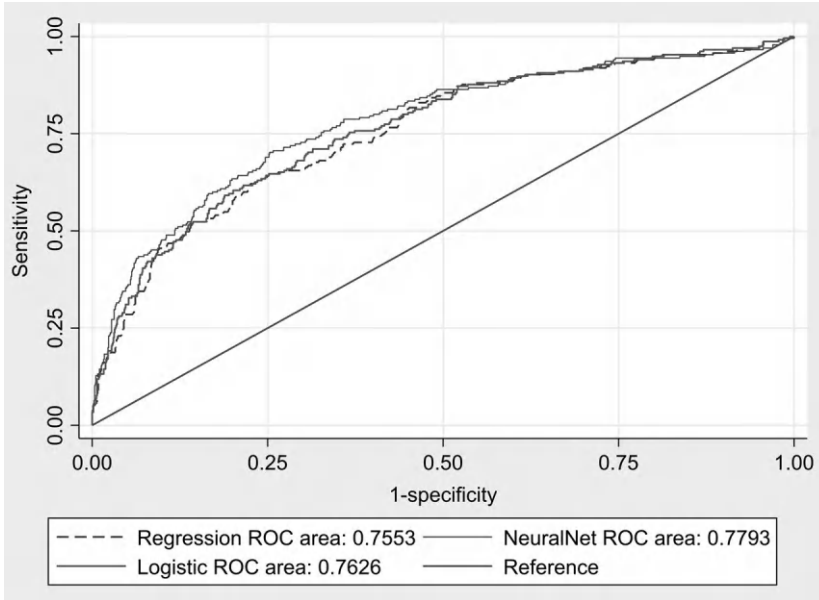


FIGURE 2.7
ROC curves for four different models.

In our attrition example, we estimate linear, log, and neural net models, construct ROC curves, and compute the AUC for each. The results are shown in Figure 2.7, which shows that the neural net has the highest AUC and so is, by this measure, the best model.

The problem with the AUC is that it may not focus on the part of the curve that matters to us. For instance, we may want to find the model that best identifies the 10% of customers most likely to respond to an offer, in which case we want the model to perform well at the upper end of responses but do not care about how well it works elsewhere. Hand (2009) argues that the measure is incoherent and should not be used.

A good way to address this is to compute lift tables. A fuller treatment of lift tables is given in Chapters 6–9, but the essential steps are as follows:

1. Randomly divide the sample data into a training sample – typically 50–70% of the total – and a test sample with the remaining observations.
2. Estimate the model using the training sample.
3. Apply the model to the test sample, getting predicted values for the outcome variable. Sort these predicted values into bins, typically deciles (tenths) or semi-deciles. Then graph the number of actual positive outcomes for these deciles.

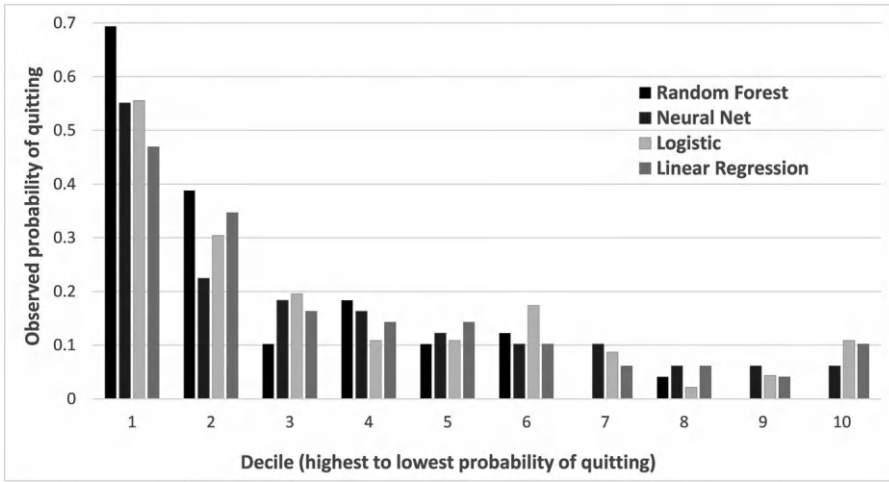


FIGURE 2.8
Lift table for four classification models.

A good model will be successful at separating the positive responses into the top bins. The model can then be used to predict, or at least identify, good prospects. Figure 2.8 illustrates the lift tables for our four models. If the goal is to identify those who are most likely to quit, the random forest performs best, as is often the case. However, neural net and logistic regression models do an adequate job in the top decile, but not elsewhere, and the linear model performs relatively well if the focus is on the top three deciles.

In short, there is not necessarily a single model that always works best – in part because the definition of "best" can be ambiguous, and in part because it may depend on the specific models and data in question.

Notes

1. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/downloads/ibm-hr-analytics-employee-attrition-performance.zip/1
2. When experience rises by 1 from 12 to 13 years, experience squared rises by 25, from 144 to 169. The change in the log of earnings is then $0.025 \times 1 - 0.001 \times 25 = 0$.
3. In the Mincer model in column (3) of Table 2.2, ln(income) = K + 1.877 jjo5. Someone in a level one job would have ln(Y1) = K, while someone in a level five job would have ln(Y5) = K + 1.877. This means that ln(Y5) – ln(Y1) = ln(Y5/Y1) = 1.877; therefore, $Y5/Y1 = e^{1.873} = 6.51$. The raw numbers bear this out, with a mean monthly income for type-one jobs of $2,787 and for type-five jobs of $19,192.

4. Impurity is often measured by the Gini impurity measure, defined as $1 - p_+^2 - p_-^2$ where p_+ and p_- are the probabilities of achieving the target outcome for the right and left child nodes, respectively. In a bin with 8 black balls and 2 white balls, this would be $1 - 0.64 - 0.04 = 0.32$. The maximum value is 0.5 (with 5 each of white and black balls), and the minimum value is 0 (with no black or no white balls).

References

Ahrens, Achim, Christian Hansen, and Mark Schaffer. 2023. "pystacked: Stacking Generalization and Machine Learning in Stata". *The Stata Journal*, in process.

Cameron, Colin, and Pravin Trivedi. 2022. *Microeconometrics Using Stata*, 2nd edition. College Station, TX: Stata Press.

Friedman, Jerome. 1999. "Stochastic Gradient Boosting". Technical Report, Stanford University.

Friedman, Milton. 1953. *Essays in Positive Economics*. Chicago, IL: Chicago University Press.

Hand, David. 2009. "Mismatched Models, Wrong Results, and Dreadful Decisions: On choosing appropriate data mining tools". Working paper. Imperial College.

Haughton, Dominique, and Jonathan Haughton. 2011. *Living Standards Analytics*. New York, NY: Springer.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. "Multilayer Feedforward Networks Are Universal Approximators". *Neural Networks*, 2: 359–366. https://cognitivemedium.com/magic_paper/assets/Hornik.pdf

Kim, Eric. 2013. "Everything You Wanted to Know about the Kernel Trick". Working paper. https://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html

Larose, Daniel, and Chantal Larose. 2014. *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, NJ: Wiley.

Mincer, Jacob. 1974. *Schooling, Experience, and Earnings*. Cambridge, MA: National Bureau of Economic Research.

Minitab. 2023. Introducing Salford Predictive Modeler 8. https://www.minitab.com/en-us/products/spm/

Nielsen, Michael A. 2015. How the Backpropagation Algorithm Works. In *Neural Networks and Deep Learning*. San Francisco, CA: Determination Press. http://neuralnetworksanddeeplearning.com/chap2.html

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society. Series B (Methodology)*, 58(1): 267–288.

Verme, Paolo. 2023. "Predicting Poverty with Missing Incomes". GLO Working Paper No. 1260. Global Labor Organization (GLO).

Wooldridge, Jeffrey. 2020. *Introductory Econometrics: A Modern Approach*. Andover, Hampshire, England: Cengage.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net". *Journal of the Royal Statistical Society. Series B (Methodology)*, 67(2): 301–320.

3

Causality

3.1 Introduction

Among the most important questions that businesses ask are some very simple ones: If I decide to do something, will it work? And if so, how large are the effects?

The important idea here is that we want to evaluate the impact of an intervention or treatment. This situation arises all the time. For instance:

- A business is wondering whether to expand its loyalty program and wants to know whether it would enhance profits;
- A retailer would like to know whether a sales campaign that emphasizes the eco-friendly nature of their clothing would boost sales; where it should advertise; whether it should open new outlets; or what new clothing features might tempt customers;
- An airline is trying to work out what prices to charge on flights from Chicago to Denver in order to maximize revenue;
- A bank is wondering whether greater investment in technical training would be worthwhile, as measured by higher worker productivity;
- A university is debating whether a reduction in tuition will attract significantly more students;
- A mobile phone operator in Latin America is considering whether to introduce a mobile banking facility;
- A car manufacturer is trying to determine whether it should introduce a hybrid pickup.

In every case, we have one or more courses of action ("treatments"), and we want to determine whether they lead to useful results. We want to determine causality.

An important distinction needs to be made here between *observing* and *intervening*. We may observe that customers with loyalty cards are more likely

 DOI: 10.1201/9780429172588-3

to spend on our product than the average customer. Formally, we may write this conditional probability as:

$$P(S \mid L) > P(S)$$

which reads as "the probability of spending (S), given that one has a loyalty card (L), is greater than the probability of spending given that one may or may not have a loyalty card." One might be tempted to conclude that loyalty cards cause customers to be more likely to spend, but this would be premature. Here we have only observed more frequent spending among those who have loyalty cards, but perhaps they were given loyalty cards because they spend more often, which would reverse the direction of causality. What really interests us is whether

$$P(S \mid do(L)) > P(S).$$

Here we introduce Judea Pearl's "do operator" (Pearl 1995), where we intervene to change the number of loyalty cards instead of allowing L to be set in the normal way (Hitchcock 2010, Section 3.6). If we increase L (i.e., "do" L), and sales rise as a result, we can reasonably claim that distributing more loyalty cards does cause people to be more likely to spend on our product. This is a probabilistic causation, in the sense that "causes change the probabilities of their effects" (Hitchcock 2010, Introduction): Intervening typically does not guarantee that there will be an effect, but it raises the likelihood of it.

It is not straightforward to identify causation with a reasonable degree of plausibility, but one can often succeed with the help of the concepts and techniques set out in this chapter and the next. We begin by stressing the importance of specifying an appropriate counterfactual, which allows us to measure treatment effects. Randomized controlled trials (RCTs) (aka experimental design) make this possible in a straightforward way, but this gold standard is rarely achievable. Quasi-experimental designs, including regression adjustment, inverse probability weighting (IPW), and covariate and propensity-score matching, are widely used and often very effective.

3.2 The Counterfactual

Identifying the impact of a treatment is not straightforward because it requires that we compare what actually occurred with a hypothetical *counterfactual* situation, namely what would have occurred in the presence or absence of the treatment. This is often referred to as the "potential outcomes" approach because it seeks to compare two potential outcomes (with or without treatment).

It is easy to go astray. Suppose your sales division introduces an advertising campaign for electric cars, just in the Northeast of the United States, and presents the following information:

	Number of electric cars sold per million population	
	Pre-campaign (2024)	**Post-campaign (2025)**
Northeast US	400	500

The 25% increase in sales is impressive, but we cannot assume that it was due to the advertising campaign unless we can rule out all other potential influences: Perhaps incomes were rising, tastes changing, or tax incentives were due to expire at the end of 2025.

We are now given some additional information about sales in the rest of the United States, where there was no advertising campaign:

	Number of electric cars sold per million population	
	Pre-campaign (2024)	**Post-campaign (2025)**
Rest of the US	200	300

It is now clear that the penetration of electric cars in the Northeast, relative to the rest of the country, has fallen from 2.0 (= 400/200) to 1.67. Has the advertising campaign in the Northeast actually been harmful? Again, we cannot tell unless we can successfully construct a counterfactual that measures the number of cars that would have been sold in the Northeast in the absence of the advertising campaign.

More formally, let the outcome of interest (sales, test scores, customer retention, etc.) be Y_i for individual or household i from a sample of size n. The outcome may be written as Y_i^T under (actual or hypothetical) treatment, and in the non-treated (control or comparison) case, we have Y_i^C. For an individual who has been "treated" – for instance, subject to a marketing intervention, sent to training, and so on – we have $T_i = 1$, otherwise $T_i = 0$.

The gain, if any, from treatment is then given by

$$G_i = Y_i^T - Y_i^C \tag{3.1}$$

and this is the causal effect we want to measure. The "Fundamental Problem of Causal Inference" (Holland 1986) is that for any individual, we either observe Y_i^T (if they were treated) or Y_i^C (if they did not get the treatment), but never both. Because of this missing information, we can never measure the impact of an intervention on a particular individual.

However, our usual interest is in measuring the *average* impact of a treatment or intervention. A popular measure is the average treatment effect on the treated (ATT), given by the expected gain

$$G^{ATT} = E\left[\left(Y_i^T - Y_i^C\right) | \, T_i = 1\right]. \tag{3.2}$$

The focus here is on those who receive the treatment, so in this case, we measure the gain conditional on $T_i = 1$ (i.e., among those in the treated group). Here, we observe $E(Y_i^T \,|\, T_i = 1)$, but the counterfactual $E(Y_i^C \,|\, T_i = 1)$ has to be constructed: It measures the outcome that this group would have experienced if they had not in fact been treated. In our earlier example, it would measure the sales of electric cars in the Northeast of the United States if there had *not* been an advertising campaign there. In effect, we have a missing data problem and need to impute the missing counterfactual, typically with the help of an "auxiliary model" (Drukker 2016, p. 10).

Other measures of impact are possible and often used in practice. The average treatment effect (ATE) on the untreated or controls (ATC) is given by

$$G^{ATC} = E\left[\left(Y_i^T - Y_i^C\right) | \, T_i = 0\right]. \tag{3.3}$$

Here we observe Y_i^C but not Y_i^T. This would be used, for instance, to measure the effect of the electric car advertising campaign in the rest of the United States (where the campaign did not actually occur) if the campaign did happen. The overall ATE is a weighted average of these effects, so

$$G^{ATE} = E(G_i) = E\left(Y_i^T - Y_i^C\right) = \Pr(T = 1).G^{ATT} + \Pr(T = 0).G^{ATC}. \tag{3.4}$$

It is tempting to measure the impact of our hypothetical advertising campaign by comparing before and after results. Using our notation and taking post- (denoted by $T_i = 1$ in this example) minus pre- ($T_i = 0$) as the estimator, this single difference is given by

$$\begin{aligned} D &= E\left(Y_i^T \,|\, T_i = 1\right) - E\left(Y_i^C \,|\, T_i = 0\right) \\ &= \left[\left|E\left(Y_i^T \;\; T_i = 1\right) - E\left(Y_i^C \,|\, T_i = 1\right)\right] + \left[E\left(Y_i^C \,|\, T_i = 1\right) - E\left(Y_i^C \,|\, T_i = 0\right)\right] \\ &= G^{ATT} + B \end{aligned} \tag{3.5}$$

where B is a measure of bias. Only if $B = 0$ will the single difference, D, give a viable measure of impact. The bias measures the difference in outcomes that we would observe, between those who were treated ($T_i = 1$) and those who were not ($T_i = 0$), in the absence of any treatment.

In our example, there is a time dimension: Before the campaign, purchases were 400 per million population, given by $E\left(Y_i^C \,|\, T_i = 0\right)$. Now suppose that the advertising campaign essentially reached those whose spending on

electric cars would have risen anyway – young professionals, for instance – and that their spending would have risen to 460 cars per million population even without the advertising campaign; this means that $E\left(Y_i^C | T_i = 1\right) = 460$. In this case, the bias, B, is 60, and the impact of the campaign (G^{ATT}) is 40 ($= D - B = 100 - 60 = 40$).

This is a common trap. It is easy to compare sales or other variables before and after an intervention, or with or without a treatment (such as a bank loan), but such comparisons rarely measure the causal impact. The above example uses post- and pre-campaign for treated ($T_i = 1$) and untreated ($T_i = 0$) respectively, but the same problem arises when treated and untreated are two groups of individuals that are not similar to each other (e.g., advertising for people in the Northeast versus no advertising for people in the Southeast).

The solution is to get rid of the bias somehow. The key to doing this is to ensure that the assignment of the treatment T, perhaps after conditioning on variables such as age or gender, is independent of the value of the outcomes. Formally, this may be written as

$$\left(Y_i^T, Y_i^C\right) \perp T_i \mid X_i$$

which reads as "the values of potential outcomes for the treated and comparison groups are independent of who is treated, at least conditional on the variables ("covariates") X_i." This is the crucial assumption of *unconfoundedness*, sometimes referred to as the conditional independence assumption, or the assumption of *ignorable* treatment assignment, or the approach of selection on observables. The challenge, then, is that of making the case that unconfoundedness applies; otherwise, it is impossible to identify causal effects.

3.3 Experimental Design

Perhaps the most satisfactory way to avoid sample bias is to assign treatment randomly. With pure randomization, there should be, on average, no difference in outcomes if there were no treatment (i.e., in Y_i^C), between those who receive the treatment and those who do not. In other words,

$$E\left(Y_i^C \mid T_i = 1\right) - E\left(Y_i^C \mid T_i = 0\right) = B = 0.$$

Now we may use the single difference, D, to measure the impact; the only remaining explanation for differences in outcome between the treated and non-treated must be the treatment itself. In other words, with randomization, the two groups $T_i = 1$ and $T_i = 0$ are completely matched on all characteristics, both observed and unobserved, provided the sample is large enough. Randomized experiments are often referred to as Randomized Controlled Trials (RCTs).

As a practical matter, in this case of pure randomization, the impact is often measured using a linear regression of the form

$$Y_i = a + b.T_i + \varepsilon_{i.}$$

Here, $T_i = 1$ if the person is treated, and 0 otherwise, and ε_i is an error term, with (by construction) zero mean. If the sample is large, or the errors may be assumed to be normally distributed, we may test whether the estimated coefficient on the treatment term ($\hat{b}$) is statistically significantly different from zero.

This case, where there is random assignment and the treatment is binary, is often referred to as A/B testing, especially in commercial applications. If we have several treatments to test, we would need to conduct an A/B/n testing (where more than two treatments are compared) or undertake multivariate testing (see Chapter 7.2.5 for further detail). While we have assumed up to now that the outcome Y_i is a continuous variable, such as spending, the same principles apply if the outcome is binary – for instance, if the result is whether an individual signs up for a new credit card or chooses to stay with the firm. In this case, the relevant regression would typically be logit or probit. Some examples of how A/B testing works in practice are given in Box 3.1.

BOX 3.1 A/B TESTING IN PRACTICE

Web pages provide an ideal environment for A/B testing because it is easy to change a page for a randomly determined fraction of viewers. As businesses develop a culture of experimentation, they can successively refine their offerings. In most cases, the goal is to increase the number of viewers who subscribe to an offer. Here are some examples where A/B testing was used to identify improvements (culled from Optimizely 2020).

The 2012 Obama presidential campaign offered an opportunity for donors to win a dinner with the President. The Obama Digital Team found, with experimentation, that adding a photo of the President boosted the number of donations, related to this offer, by 7%.

A few years ago, the BBC ran an experiment on iPlayer – its Internet streaming service – in which, at the end of an episode, it automatically played the next episode. This increased the number of people watching the next episode by 50%.

Hotwire, which is owned by Expedia, provides car and hotel bookings using phones and other apps. The company runs over a hundred experiments annually, mainly aimed at improving the "conversion rate," which is the percentage of viewers who actually book through

Hotwire. In one example, they redesigned the look of the mobile car-rental app to include an image and a clearer navigation button, significantly boosting conversions.

Brooks Running sells shoes online and has a free-return policy that is expensive to operate. It experimented with a pop-up message for shoppers with multiple shoes of different sizes in their cart, offering help with choosing the right size. This modest change reduced returns by 80%.

HP is promoting its Instant Ink service, which can notify the company when toner in a printer is running low, prompting an automatic shipment that seeks to ensure that the user will never actually run out. The key challenge is getting customers to enroll. After some experimentation, the greatest yield was associated with an offer of a free trial and presenting the service as a "printer feature."

3.3.1 Stratified Randomization

Simple random sampling is rare because it is usually inefficient. For instance, if we are trying to measure the impact of a cosmetics campaign geared toward women, it might not make sense to sample men. But we may be interested in whether women in different age groups respond differently to the campaign – for instance, those aged 20–30 or 30–40. These constitute *strata* or blocks, and it is still essential that within these strata, individuals be chosen at random for the intervention. This is known as stratified randomization or randomized block design. Often, we deliberately over-sample some strata in order to have enough observations on people in that block; by using sample weights, we can adjust for over- or under-sampling at the analysis stage.

Formally, we now have *conditional exogeneity of program placement*. It is still possible to estimate the impact of the treatment by estimating a regression model if we are willing to assume that the control variables (i.e., the X_i) operate linearly. Suppose, for the n_T who receive treatment,

$$Y_i^T = \alpha^T + X_i\beta^T + v_i^T,\ i = 1,\ldots,n_T \tag{3.6}$$

and for the n_C who are not treated (i.e., the control group)

$$Y_i^C = \alpha^C + X_i\beta^C + v_i^C,\ i = 1,\ldots,n_C. \tag{3.7}$$

The data from these two groups can be pooled to create Y_i, which is the observed potential outcome, and is given by

$$Y_i = T_iY_i^T + (1-T_i)Y_i^C. \tag{3.8}$$

where, as usual, $T_i = 1$ if the individual is treated and 0 otherwise

Using the pooled data for the n ($= n_T + n_C$) observations, we may estimate a "switching" regression of the form

$$Y_i = \alpha^C + \left(\alpha^T - \alpha^C\right)T_i + X_i\beta^C + X_i\left(\beta^T - \beta^C\right)T_i + \varepsilon_i, \quad i = 1,\ldots,n, \tag{3.9}$$

and the error term is given by $\varepsilon_i = T_i\left(v_i^T - v_i^C\right) + v_i^C$. The error term picks up any "latent" (unobserved or omitted) influences on the continuous outcome variable Y_i. If the errors are uncorrelated with who gets treatment (T_i) or the other control variables (X_i), Eqn. (3.9) may be estimated simply using ordinary least squares. The treatment effect for observation i is given by

$$\begin{aligned} G_i^{ATE} &= \Pr(T=1)E\left(Y_i^T - Y_i^C \mid T_i = 1\right) + \Pr(T=0)E\left(Y_i^T - Y_i^C \mid T_i = 0\right) \\ &= E\left(Y_i^T - Y_i^C\right) = \left(\alpha^T - \alpha^C\right) + X_i\left(\beta^T - \beta^C\right) + E\left(v_i^T - v_i^C\right) \end{aligned}$$

Averaging this over all observations gives the ATE, which is

$$G^{ATE} = \left(\alpha^T - \alpha^C\right) + \bar{X}\left(\beta^T - \beta^C\right). \tag{3.10}$$

If we further assume (heroically) that $\beta^T = \beta^C$, this simplifies to the common impact model, so $G^{ATE} = \alpha^T - \alpha^C$.

Stratified randomization underlies uplift modeling (Chapters 6 to 9), which begins by measuring the causal treatment effect, relates this to characteristics of the individual, such as age, gender, or education, and then uses these results to identify and target good prospects for a treatment.

3.3.2 Randomization: The Gold Standard?

Randomization is often considered to be the gold standard for scientific experimentation. The Nobel Memorial Prize in Economic Sciences in 2019 went to three researchers (Abhijit Banerjee, Esther Duflo, and Michael Kremer) who have popularized the application of RCTs to identify the impact of treatments in developing countries such as deworming, microcredit, sexuality education, and incentives for teachers to turn up for work. In 2020, the vaccines for protecting against COVID-19 first had to be tested using RCTs before they were considered safe enough for public use.

Randomization can be difficult in practice and is, at times, unethical. Haughton and Kelly (2014) wanted to assess whether a flipped hybrid approach to teaching basic statistics would generate better outcomes, as measured by test scores. The most powerful approach would have been to randomly assign students to hybrid and non-hybrid classes. But this would not have been practical because students have to design their schedules to fit with other classes. It would also have been considered unacceptable to force students to enroll in sections of classes that they did not want to take.

Likewise, a study of the effects of smoking that assigned subjects to smoke or not would clearly be unworkable and unethical.

Even when random assignment is possible, there may be selective uptake and nonrandom attrition. Consider the case of a pollster who wants to test attitudes among registered voters toward a political candidate for state governor. The list of voters may not be accurate; some voters may not have (known) phone numbers; most individuals do not respond to opinion survey calls – 94% of those sampled for the Pew Research Center Surveys, for instance (Kennedy and Hartig 2019); and those who do respond may not go to vote. These are only problematic because at each stage there are likely to be biases. For instance, those without phones may be poor or jealous about guarding their privacy; those who respond to a call may be older or lonelier; and those who vote may be more educated. The extent of the biases is unknown, so the results of the poll, which began as a random sample, may be unusable. In fall 2020, for instance, most polls overestimated the number of votes that Joe Biden was expected to receive and underestimated the number of votes that went to Donald Trump (Keeter 2021).

RCTs have other problems. They may be expensive; it may be difficult to get a large-enough sample; there may be leakage (e.g., if a TV advertising campaign aimed at New York gets seen by many from elsewhere); and often there are important questions, such as how a firm should react in times of recession, that they cannot address.

3.4 Quasi-experimental Methods

When randomized experiments are not feasible, what can be done? To repeat, the problem that has to be addressed is that of selection bias.

For instance, suppose we have developed a new variety of soybeans and would like to test market it on a small scale to farmers in several counties in the Midwest of the United States. The sales team has identified some promising areas for the test, but this is likely to lead to a nonrandom program placement, which will affect our results in hard-to-predict ways (due to *unobserved area heterogeneity*). When the product is introduced, there is likely to be self-selection into program participation, so more dynamic or better-informed individuals are the first to sign up. This *unobserved individual heterogeneity* makes it difficult to assess the effects of a stronger push to market soybeans because other farmers may not be so quick to come forward.

Selection bias need not be fatal, and there are a number of approaches to at least attenuating its effects, including regression adjustment and matching, which are discussed in this chapter, and discontinuity design, synthetic samples, and instrumental variables, which are the subjects of Chapter 4.

3.4.1 Regression Adjustment

It is sometimes possible to identify the impact of a treatment by estimating a straightforward common impact regression equation that controls for enough assumed other influences on the outcome. To illustrate this, consider the (hypothetical) numbers shown in Table 3.1: A firm wants to know whether handing out discount coupons boosts sales and has collected information on consumer purchases, whether they were handed a coupon, and gender. A simple comparison shows that the average spending of those with coupons was \$39.5, compared to \$42.3 for those without coupons, a difference of \$2.8. Alternatively, a common-impact regression gives

$$\begin{aligned} purchases &= 42.3 - 2.8\ (gets\ a\ coupon) \\ \text{p-value} &= -0.88 \end{aligned}$$

The coefficient on the treatment variable is –2.8 (although it is also not statistically significant). The coupon campaign looks like a failure.

But is it? The problem here is that the coupons do not seem to have been handed out randomly: two-thirds of those receiving coupons were men, while only half of those who did not receive coupons were men. The path diagram in Figure 3.1 shows the causal links: Gender affects both treatments (i.e., who gets a coupon) and purchases (as women spend more than men). To identify the effect of coupons on spending, one has to "block" the "backdoor"

TABLE 3.1

Hypothetical Data on Sales, Gender, and Coupons

Observation Number	Receives Coupon?	Male?	Purchases (\$)
1	Yes	No	84
2	Yes	No	72
3	Yes	Yes	0
4	Yes	Yes	10
5	Yes	Yes	45
6	Yes	Yes	26
7	No	No	60
8	No	No	77
9	No	No	67
10	No	Yes	0
11	No	Yes	30
12	No	Yes	20
Mean			40.9
Mean, with coupon			39.5
Mean, without coupon			42.3

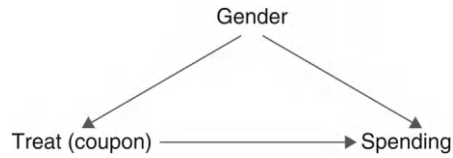


FIGURE 3.1
Path diagram of influences on spending.

path whereby receiving a coupon is associated with gender and, through it, on spending. A simple way to do this is by including gender as a variable in the common-impact equation. This gives

$$purchases = 69.5 - 54.4(male) + 6.2(gets\ a\ coupon)$$
$$\text{p-value} = 0.48$$

This puts the coupon campaign in a different light, and while the p-value is still large, if this sample were eight times larger, the coefficient (6.2) would be considered highly statistically significant.

In this example, we were able to control for observable effects (gender), and we were able to use a linear model. Things are rarely so simple. Often the basis for our selection bias is unobservable, reflecting hidden (latent) influences such as individual motivation or tastes. Sometimes, it is possible to remove the effects of time-invariant observables using panel data, as discussed below.

One of the most difficult decisions that needs to be made in situations such as this is what variables to include as controls. A helpful way to keep one's ideas clear in this context is with the help of path diagrams. In the words of Morgan and Winship (2015, p. 91), such graphs "offer a disciplined framework for expressing causal assumptions for entire systems of causal relationships."

Consider Figure 3.2, where we want to measure the effect of treatment T (coupons) on outcome Y (spending). The problem is that one who gets treatment is affected by confounder C (here, gender), which in turn affects a set of

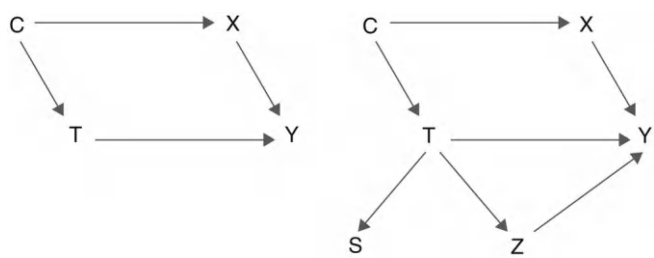


FIGURE 3.2
Path diagrams. (T is the treatment, Y is the outcome of interest, and C, X, S, and Z are other variables)

observable variables (such as "love of fashion," denoted here by X) that influence spending. In the left panel of Figure 3.2, the direct treatment effect of T on Y is confounded by C. For instance, if gender = female, there may be fewer coupons (low T) but more fashion lovers (high X) and spending on fashion (high Y). So we will tend to observe low values of T along with high values of Y, and this gets in the way of ("confounds") identifying the direct effect of T on Y. The trick is to block the backdoor associations between T and Y. In the left panel of Figure 3.2, we have

$T \rightarrow Y$	The causal effect we want to measure
$T \leftarrow C \rightarrow X \rightarrow Y$	The confounding "backdoor" effect.

The solution is to include C (or X, but not necessarily both) in the regression equation.

The situation in the right-hand panel of Figure 3.2 is similar, except that receiving a coupon (the treatment) makes it more likely that the person will visit a store (variable Z), which in turn boosts sales. In this case, Z is a descendent of T, but variable Z should *not* be included in the regression because it risks taking from T some of the influence that is rightly attributable to it.[1]

If, in addition, there is a variable S (e.g., how tall the individual is) that influences who gets a coupon – they catch the eye of the agents handing out the coupons – but does not influence anything else, then it is a parent of T and should not be included in the regression. That would reduce the variation in T, which would make it harder to pick up the causal effects. In this case, *S* is essentially an instrumental variable (IV); as a general rule, it is inappropriate to control for IVs or mediator variables in causal inference.

The use of diagrams such as these can be helpful in trying to identify causality and whether it can even be identified. It also makes it clear that one should not simply dump all variables into a treatment regression in the hope that one is controlling for all possible confounders: Sometimes that will make it harder to measure causal effects. Morgan and Winship (2015, p. 117) have an extended discussion of these issues.

3.4.2 Other Treatment Effects

Suppose the Dean of the College of Arts and Sciences observes that students enrolled in the honors program are more likely to stay after their first year and to earn a higher grade point average (GPA) than those who are not. She would like to know whether this association is accidental or causal: If being in the honors program boosts retention, perhaps the honors program should be expanded, or at least promoted more vigorously.

There is clearly a selection bias here, because honors students are presumably academically stronger, on average, than non-honors students, so a simple comparison of retention rates (or GPA) between the two groups will not tell us much.

Some hypothetical information on 20 students is provided in Table 3.2. In this example, 75% of the honors students stayed for the next academic year (retain = 1) compared to 50% of non-honors students. Honors students had a higher college GPA (3.13 versus 2.40) and entered with higher selective aptitude test (SAT) scores (1,374 versus 1,234).

One visual representation of these data is given in Figure 3.3, where one of the outcomes of interest (GPA) is shown on the vertical axis and the SAT scores on the horizontal axis. Black dots refer to honors students, hollow circles to those who are not in the honors program. The black dots are concentrated in the upper right of the diagram, as we would expect.

Do honors students perform better than non-honors students, as measured by GPA? We see how students actually perform in Figure 3.3, but only with the help of an "auxiliary" model can we estimate their potential outcomes in the alternative (counterfactual) state. There are a number of ways to measure these *treatment effects*.

TABLE 3.2

Data on Student Retention and Performance

ID	Retain	College GPA	Honors	Male	Recruit Score	SAT Score	Propensity Score
14	0	1.44	0	1	5	960	0.036
13	1	2.47	0	1	6	1010	0.068
20	0	1.76	0	0	6	1060	0.142
12	0	2.01	0	1	7	1180	0.168
19	1	2.23	0	0	6	1140	0.173
3	1	3.21	1	1	7	1200	0.177
18	0	2.77	0	0	7	1220	0.309
8	1	3.13	1	0	7	1230	0.316
11	1	2.71	0	1	8	1320	0.340
10	0	2.13	0	1	8	1340	0.354
7	1	3.10	1	0	7	1290	0.355
17	1	2.72	0	0	7	1330	0.382
2	0	3.17	1	1	9	1450	0.561
6	0	3.02	1	0	8	1410	0.570
9	1	3.06	0	1	9	1470	0.576
16	0	3.00	0	0	8	1430	0.584
15	1	2.45	0	0	9	1350	0.653
5	1	2.88	1	0	9	1400	0.685
1	1	3.43	1	1	10	1510	0.721
4	1	3.11	1	0	10	1500	0.832

Retain = 1 if the student stays another year; GPA is grade point average, on a scale of 1 through 5; Honors = 1 if the student is in the honors program; Male = 1 for male, otherwise female; Recruit Score is assigned by admissions officers on a scale of 1 (poor prospect) through 10. The propensity score is discussed below.

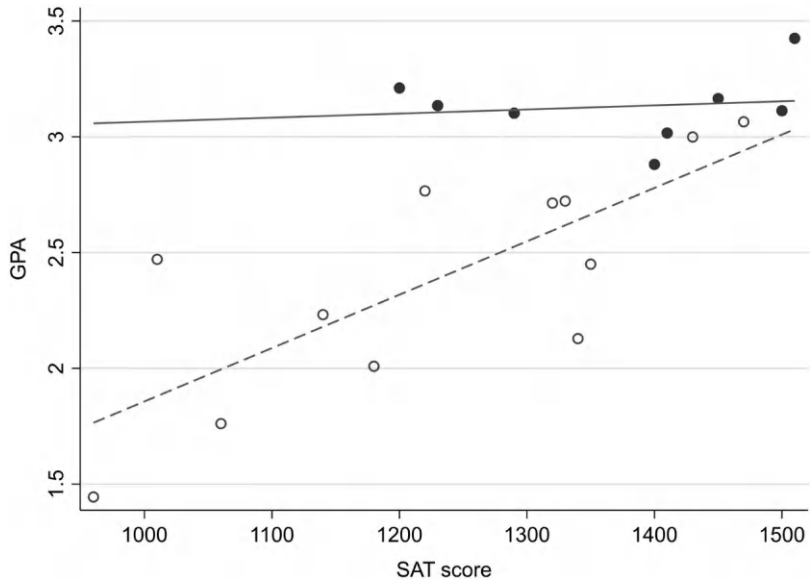


FIGURE 3.3
GPA and SAT scores for a sample of students. (Honors students are marked by black dots, and non-honors students by hollow circles.)

3.4.3 Regression Adjustment (Again)

The idea here is to estimate one equation that predicts performance for the treated (here, honors students) and another for the non-treated. Assuming linearity, these are shown in Figure 3.3 as the solid-black and dashed lines, respectively. Now, for each observation, find the difference between the potential outcome in the honors program ($\hat{Y}_i \,|\, T_i = 1$), from the solid line, and the potential outcome if not in the honors program ($\hat{Y}_i \,|\, T_i = 0$) from the dashed line. The average of these differences gives the ATE.

Equivalently, we may estimate the switching equation (Eqn. 3.9) directly, using the small invented dataset from Table 3.2, and this gives the estimates shown in Table 3.3. To measure the ATE, we apply Eqn. (3.10), which gives

$$G = 3.349 + 0.292\ (\text{Male}) + 0.0435\ (\text{Recruit score}) - 0.0025\ (\text{SAT})$$

Computing this for each observation, and averaging over *all* observations, gives

$$G^{ATE} = 0.616$$

while averaging just over *the honors ("treated") students* yields

$$G^{ATT} = 0.418.$$

TABLE 3.3

Estimates of Regression Adjustment "Switching" Equation for GPA

	Coefficient	Standard Error	Memo: Mean Value
Outcome variable: GPA			2.69
Right-hand variables:			
In honors program (treatment)	3.349	2.269	0.40
Male (Y = 1)	–0.075	0.191	0.45
Recruitment score at admission	–0.046	0.201	7.65
SAT score at admission	0.0026	0.0015	1,290
Honors × Male	0.292	0.307	0.15
Honors × Recruitment score	0.0435	0.349	3.35
Honors × SAT score	–0.0025	0.0034	549.5
Intercept	–0.442	0.815	
Memo items			
Adjusted R^2	0.664		
Number of observations	20		

Note: Based on data in Table 3.2. The estimation equation is of the form shown in Eqn. (3.9).

The method is flexible. For instance, if we limit our sample to those students who entered with at least 1,200 SAT score and re-run the switching regression and the ensuing computations, we find that in this case $G^{ATE} = 0.436$ and $G^{ATT} = 0.425$.

3.4.4 Propensity Scores

Before going further, we need to introduce the notion of the *propensity score*. This is the estimated probability that a person will be treated. In a simple random sample, this would be the same for everyone, since every person would have the same probability of being chosen (although this probability does not need to be 0.5). However, in a quasi-experimental setting, the probability that someone will be treated may vary widely and systematically.

A propensity score results from an *assignment model*, which seeks to determine who is treated and who is not. A common and straightforward approach is to estimate a logit (or probit) regression, where the dependent variable is binary (1 if treated, 0 otherwise), and the right-hand variables are those believed to influence the probability of treatment – for instance, age, gender, location, and so on. Then the propensity score is the predicted value of the dependent variable: $\hat{p}_i = P(T_i = 1 \mid X_i)$.

In Table 3.4, we show the estimates of a logit "propensity score equation" using the data on university retention presented in Table 3.2. In this case, "treatment" consists of being enrolled in the honors program, and we assume that this is driven by the student's gender, SAT score, and recruitment score (as determined by the university's admissions office). The predicted probabilities of being in the honors program – the propensity scores – are then

TABLE 3.4

Estimates of Logit Propensity Score Equation for Participation in Honors Program

	Coefficient	Standard Error	Memo: Mean Value
Outcome variable:			0.40
In honors program (Y = 1)			
Right-hand variables:			
Male (Y = 1)	–0.678	1.111	0.45
Recruitment score at admission	0.527	1.068	7.65
SAT score at admission	0.0029	0.0096	1,290
Intercept	–8.061	6.339	
Memo items			
Pseudo R^2	0.177		
Number of observations	20		

Note: Based on data in Table 3.2.

calculated for each observation using this estimated equation and are shown in the final column of Table 3.2. Although a student is either in the honors program (honors = 1) or non-honors (honors = 0), the probabilities range from 0 to 1, as would be expected. As usual, the estimated propensity scores are only as sound as the underlying model. In this small model, we keep predictors that are not statistically significant, but many researchers reduce the predictors (or features) through regular significance testing.

The estimation of propensity scores is an area of active research. Diamond (2005) uses a robust logit, which reduces the influence of outliers. Imbens (2004) favors non-parametric binary response models, which can better handle non-linearities. Maciel (2020) reviews a number of approaches, including the use of Random Forests, Gradient Boosting, Support Vector Machines, and Neural Networks, which we summarize in Chapter 2.

3.4.5 Matching Methods

Let us return for a moment to the data on the honors program that are set out in Table 3.2 and Figure 3.3. As noted above, honors students have, on average, higher SAT scores and higher GPAs than non-honors students. But enrollment in the honors program has a random element, so some non-honors students have comparable characteristics to some non-honors students, and vice versa. This opens up the possibility that we could match some honors students with otherwise comparable non-honors students, and then, after preparing the data in this way, use a model on the matched data to measure whether retention is higher in one group. Presumably, any differences in the outcome – here retention – could mainly be attributed to the one obvious difference between the groups, which is participation in the honors program. King and Nielsen (2019) express the idea well: Matching "amounts to a search for a data set that might have resulted from a randomized experiment but is hidden in an observational data set" (p. 1).

Consider the raw (hypothetical) data set out in Table 3.2. Of the eight students in the honors program, six (75%) stayed on after the first year, compared to six of the 12 students (50%) of the non-honors students. But, in trying to match each honors student with an otherwise identical non-honors student, we note that there are no exact matches, in that no two students have the same gender and recruitment score and SAT. So any matching or pairing of students will necessarily be inexact.

There are, in fact, a number of ways to do this (imprecise) matching, which we describe and discuss in the following subsections. Perhaps the most widely used approach is propensity score matching, so we begin there before moving on to a number of methods that are becoming more popular.

3.4.5.1 Propensity Score Matching

A popular solution to the matching problem is to use propensity scores. In our example, the propensity score is the estimated probability that a student will be in the honors program. If we use nearest-neighbor matching, we would match students using the propensity scores, essentially by pairing each student in the honors program with the non-honors student with the closest propensity score, and vice versa. So, for instance, referring to the propensity scores in Table 3.2, student 5 (honors, propensity score of 0.685) is matched with student 16 (non-honors, score of 0.584); and in turn, student 16 is matched with student 9 (honors, score of 0.576). For each pair of students, we calculate the difference in the outcome variable (retention), and the average of these differences is a measure of the ATE. In this particular example, the ATE is 0.200, which tells us that the retention rate is increased by 0.200 as a result of enrollment in the honors program. This is a causal conclusion.

Rosenbaum and Rubin (1983) prove that unconfoundedness is preserved when using the propensity score. Formally, this means that

$$\left(y_i^T, y_i^C\right) \perp T_i \mid x_i \Rightarrow \left(y_i^T, y_i^C\right) \perp T_i | p(x_i)$$

where $p(x_i)$ is the propensity score. This means that matching on the propensity score should allow us to control for bias and so get to the underlying causal effects.

There are many possible flavors of propensity score matching, which is both a strength and a weakness. On the one hand, it holds the promise of flexibility; on the other, researchers inevitably end up trying out different variants of propensity score matching, typically reporting only the more successful results, and thereby overstating its effectiveness.

Perhaps the commonest approach is to use nearest-neighbor matching, where (for the ATT) each treated case is matched with the m closest non-treated case. In the example above, we used $m = 1$, which is the most popular choice.

One alternative is to use caliper matching, which compares each treated case with all the non-treated cases whose propensity scores are within a given (modest) distance of the treated case, often set at quarter of a standard deviation of the propensity score. Some researchers use kernel or Gaussian matching, techniques that compare each treated case with all the non-treated cases but put greater weight on observations whose propensity scores are close to that of the treated case. Dehejia and Wahba (2002) argue that the choice of matching method is less important than the careful estimation of the underlying propensity scores.

Propensity score matching only works effectively if comparisons (between the treated and comparison cases) are confined to the *region of common support*. Consider Figure 3.4, where we have graphed the distribution of propensity scores for the treated (the right-hand distribution) and the comparison group. For treated cases with propensity scores between A and B (or 0.1–0.7 in this illustration), there is *overlap*, so we can find examples of non-treated cases with similar propensity scores. But this is not the case above B, where there are no good comparators, or below A, where there are no treated cases. Cases in the region of common support are those with propensity scores between A and B; all cases with propensity scores outside this range should be discarded. Ho et al. (2006) emphasize the importance of this trimming or "preprocessing" of the data. Diamond (2005, p. 9) notes that the discarded cases "cannot support causal inferences about missing potential outcomes."

This trimming comes at a cost: It limits the applicability of the results (their "external validity"), which now only hold in the range of propensity scores between A and B, and not universally.

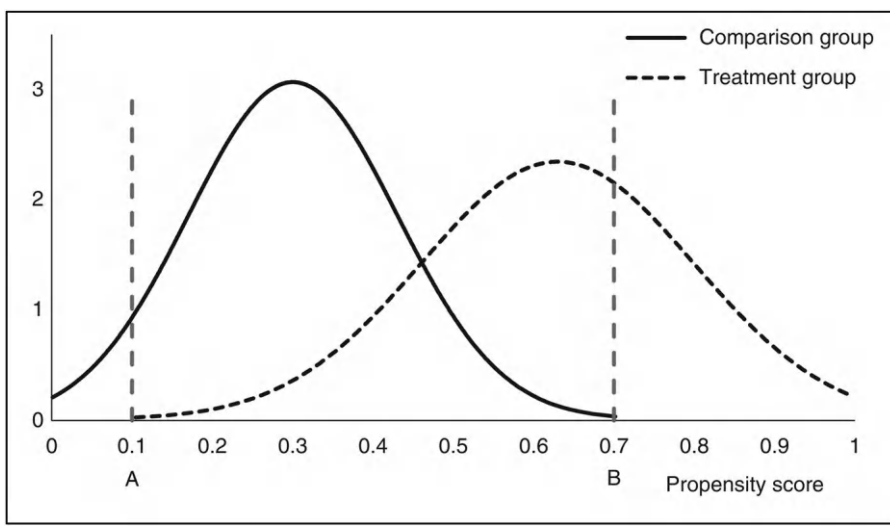


FIGURE 3.4
Illustration of the region of common support.

When matching treated cases with their comparators, the two sets of data should be "balanced." This means that they should have comparable distributional characteristics for the dimensions that can be measured. For instance, the treated and non-treated with, say, a propensity score of between 0.6 and 0.7 should have similar proportions of old people, women, health workers, and the like. In theory, propensity score matching ensures balance, but in reality, it may not occur. Thus, we may end up comparing a treated individual with a propensity score of 0.65 who is a 70-year-old retired man with a non-treated individual, also with a propensity score of 0.65, who is a 30-year-old female lawyer. This would be a surprising way to match supposedly "otherwise similar" individuals. It is important to check for balance; where it is lacking, the usual advice is to re-specify the propensity score model, which can often be helpful.

In an article, King and Nielsen (2019) argue that propensity scores should not be used for matching. They show that the more balanced the initial data are, the greater the danger that propensity score matching will "degrade inferences" (p. 2), a phenomenon they refer to as the propensity score paradox. Their other argument is that researchers are likely to report the results from versions of propensity score matching that reflect, perhaps unconsciously, their own biases, so the model chosen itself depends on the data ("model dependence"); in their words, "human choice turns model dependence into bias" (p. 5). This weakens the ability of propensity score matching to deliver unbiased judgments.

There are a number of other, increasingly popular, methods for inferring causality via matching, starting with IPW.

3.4.6 Inverse Probability Weighting (IPW)

An elegant method for measuring the treatment effect is to estimate a regression of the form

$$Y_i = \alpha + \beta T_i \tag{3.11}$$

where Y_i is the outcome of interest (e.g., GPA), and T_i is the treatment, for instance, being in the honors program (see Hernán and Robins 2006). However, the twist here is that this needs to be a weighted regression, where the weights are the inverse probabilities from the propensity scores. More specifically, we have:

$$w_i = \begin{cases} 1/\hat{p}_1 & \text{for treated case} \\ 1/(1-\hat{p}_1) & \text{for non-treated case} \end{cases}$$

Then the estimated coefficient $\hat{\beta}$ in Eqn. (3.11) gives the ATE.

In effect, this estimator puts higher weights on rare occurrences. This is analogous to the use of survey weights, where more weight is put on observations that are less likely to have been sampled because they are standing in for larger numbers of unsampled items. To measure the ATT, the weights needed are:

$$w_i = \begin{cases} 1 & \text{for treated cases} \\ \hat{p}_i \big/ \left(1-\hat{p}_i\right) & \text{for non-treated cases} \end{cases}$$

and for the ATC, we have:

$$w_i = \begin{cases} \left(1-\hat{p}_i\right) \big/ \hat{p}_i & \text{for treated cases} \\ 1 & \text{for non-treated cases} \end{cases}$$

When the propensity scores are either very low or very high, the weights can become rather large, and this method becomes less robust.

One solution to the problem of extreme weights is to trim them. For instance, we may cap ("windsorize") the weights at the 5th and 95th percentiles. There is no generally recognized standard for the cut-off percentiles (Crump et al. 2009; Morgan and Winship 2015), so the procedure used will be decidedly ad hoc. However, Lee et al. (2011), using a logistic-regression-based propensity score model, find that the trimming does lead to more robust results.

The other potential problem with the "original" weights shown here is that they are known to increase the sample size artificially, which is likely to affect the standard errors of the estimators. Hernán and Robins (2025) make the case for using "stabilized" weights, which ensure the weighted sample is the same size as the original sample while leaving the coefficients unchanged (see Chapter 9, Eqn. 9.5). This is achieved by replacing $\hat{p}_i$ with $\hat{p}_i / \overline{p}$, and $(1-\hat{p}_i)$ with $(1-\hat{p}_i)/\left(1-\overline{p}\right)$, where $\overline{p}$ is the average propensity (or the predicted value of a logistic regression with only a constant term). So, for example, the stabilized weights for the ATE would be given by:

$$w_i = \begin{cases} \overline{p} \big/ \hat{p}_i & \text{for treated cases} \\ \left(1-\overline{p}\right) \big/ \left(1-\hat{p}_i\right) & \text{for non-treated cases} \end{cases}$$

There is a growing literature advocating for stabilized, rather than ordinary, weights.

As with propensity score matching, the inverse probability weighted regression should only be estimated using those observations that are in the region of common support.

3.4.7 Doubly-robust Methods

It is possible to combine the regression adjustment and the IPW approaches in a rather natural manner: Estimate the regression adjustment equations using the inverse probability weights.

Another possible extension is the *augmented inverse probability weighting* approach (Glynn and Quinn 2010), which estimates Eqn. (3.11) but includes a number of additional variables, such as those used in the regression adjustment model.

It is also possible to combine matching and regression methods. Matching could be used first to trim (or take a "strategic subsample" from) the data, and then regression techniques could be applied to the dataset that should now look more like a dataset from a natural experiment.

All of these are examples of "doubly-robust" methods. The idea is that they hold greater promise of yielding good results, even if one of the sets of assumptions underlying (for instance) regression adjustment or IPW does not entirely hold. For instance, regression adjustment requires that the treatment be independent of the error in Eqn. (3.9), the causal effect does not vary with the other regressors, and the estimated equation is fully flexible (Morgan and Winship 2015, p. 205). The IPW also requires the assumption of ignorable treatment and overlap between the treated and non-treated samples.

3.4.7.1 Covariate ("Nearest Neighbor") Matching

There are also ways to match someone who is treated with someone similar who is not, but without using propensity scores. A basic form of *nearest-neighbor matching* proceeds as follows:

a. Normalize all the relevant variables ("covariates"), such as age, SAT score, and so on, so they have zero means and standard deviations of 1, using

$$u_i = \frac{X_i - \overline{X}}{s_i},$$

where X_i is the variable, $\overline{X}$ is its mean, and s_i is the standard deviation of the X_i.

b. For every pair of treated and untreated (comparison) individuals, with vectors of normalized variables U_i^T and U_i^C, define the distance between them as

$$\left\|U_i^T - U_i^N\right\|_V \equiv \sqrt{\left(U^T - U^C\right)' V \left(U^T - U^C\right)}$$

where V is a weight matrix. If V is the identity matrix, we have a standard Euclidean distance, which gives equal weights to each variable when comparing two individuals. Note that there can be a lot of pairs for which this calculation must be done.

c. Match each treated case with the closest non-treated case and find the average difference in outcome to get the Average Treatment Effect on the Treated (ATT); or match every treated case with the closest non-treated case, *and vice versa*, to obtain the average treatment effect (ATE).

In practice, the most common choice of weight matrix V is the inverse of the variance-covariance matrix of U^T and U^C, which gives the Mahalanobis distance. This method is non-parametric and flexible but converges relatively slowly to the true measure. In our example, this measure shows that enrollment in the honors program raised retention by 0.294 (see Table 3.6).

There are many possible variations. For instance, one can insist that some variables be matched exactly: Maybe men can only be matched with men and women with women. Or an iterative method may be employed, such as the genetic matching proposed by Diamond and Sekhon (2013), wherein the weight matrix V is adjusted to achieve the best possible balance – although Morgan and Winship (2015) dismiss this as "error-prone sausage making" (p. 167). In this context, balance is obtained when the average value of covariates, such as age or SAT score, does not differ between the treated group and the sample with which they are matched (see Table 3.2).

3.4.7.2 Coarsened Exact Matching

Suppose we want to match subjects from one dataset (the treated, perhaps) with those from another (the non-treated). Let us also suppose that we only have very simple data, such as information on the gender of the head of household and whether the head is over 65 years old or not. The exact matching is straightforward: Match old males from dataset 1 with old males from dataset 2, young females with young females, and so on.

In reality, the data we work with are rarely so simple. Suppose we have everyone's age, and not just whether they are old or young. Now we might try to match 51-year-old men in dataset 1 with men of the same age in dataset 2, and so on. However, if the datasets are relatively small, there may be a lot of non-matches: Perhaps dataset 2 has men who are 50 or 52 but none who are 51. When matches cannot be made, we potentially lose information.

Gary King and his coauthors (Iacus et al. 2008; Blackwell et al. 2009) propose that, in such cases, the variables be "coarsened" – in effect assigning them to a limited number of bins – before being exactly matched. In our example, this might mean classifying people's ages into ten-year intervals (e.g., 40–49, 50–59), which then raises the probability of "exact" matches. The coarsening might be done using an algorithm, in the same manner as is often done when constructing histograms, or rest on the researcher's judgment.

Not every treated case will be matched with a non-treated comparison case, and the non-matched cases will be dropped. It is helpful to think of this as a way to preprocess the data, thereby confining the observation to a region of common support or overlap: The outcomes of the remaining matched cases could then simply be compared or be subject to a regression-adjustment or other model. An advantage of coarsened exact matching (CEM) over propensity score matching (PSM; discussed above) is that CEM seeks to balance the data in advance, while PSM needs to check for balance after matching and, if necessary, go back to adjusting the propensity score model until the balance is achieved.

To illustrate how CEM might work, consider the data on retention from Table 3.2. First, we create a limited number of bins; for instance, gender crossed with SAT scores, where the SAT scores are assigned to three equal-width classes. The treatment consists of being enrolled in an honors program, so we may assign the observations to bins as shown in Table 3.5.

The non-honors students whose SAT scores fall in the range 960–1142 do not overlap with any honors students and so cannot usefully serve as comparators (or, equivalently, we assign them a weight of 0). Each honors student is assigned a weight of 1 since we want to estimate the ATT effect. There are two female honors students with SAT scores in the range 1143–1325, and they will be compared with the one non-honors student in that bin. Since this non-honors student has to serve in two comparisons, she is assigned a weight of 2. There is one male honors student in the same SAT category and two non-honors males, so the latter comparators are each assigned a weight of 0.5.

TABLE 3.5

Classification of Observations into "Coarsened" Bins

	Female			Male		
SAT	**960-**	**1143-**	**1326-**	**960-**	**1143-**	**1326-**
Observations:						
In honors program	0	2	3	0	1	2
Not in honors program	2	1	3	2	2	2
Weights:						
In honors program		1	1		1	1
Not in honors program	0	2	1	0	0.5	1

We may then use these weights in a treatment regression. In our example, our estimates give:

$$\text{GPA} = 2.657 + 0.434 \text{ Honors} \qquad R^2 = 0.43.$$

So, we may conclude that being in the honors program boosts one's GPA by 0.434 points. However, this result is sensitive to the decisions made in the coarsening process. If the number of bins is different, or the cutoffs are different, then the result will be different too. This approach is at least as susceptible to manipulation by the researcher as the other approaches to matching.

3.4.7.3 Which Matching?

How important is the choice of method in measuring the impact of a treatment? For each of the methods discussed above, we have used the data shown in Table 3.2 to estimate the effect of being in an honors program on (i) GPA and (ii) retention (i.e., the probability of staying at the university for another year).

The results are gathered together in Table 3.6, and although they are based on a small hypothetical example, they suggest that different methods can yield quite different results. For example, the measured impact on retention for students in an honors program (the ATT) is zero using propensity score matching, but 0.11 if one applies the regression adjustment method and 0.25 based on CEM.

If our linear regression adjustment model of causality is plausible and complete, then the results in Table 3.6 show that being in the honors program boosts retention by 36 percentage points and raises GPA by 0.62. If the differences are averaged over just the honors students, then we get a measure of the Average Treatment Effects on the Treated (ATT), which in this case are somewhat smaller than the ATE effects.

An important implication is that the choice of method matters. This leaves researchers and practitioners with a lot of discretion, so the answers they provide will reflect the methodological choices they make. There is no royal road to "the truth," but recent practice has moved away from propensity score matching. The abundance of methodological choices does require us to caution, once again, against simply reporting the results of the method that seems to give the "most sensible" results, because this results in potentially serious bias, thereby weakening our ability to infer causality.

The techniques described in this chapter are widely used, especially in the social sciences, but do not exhaust the possibilities. In the next chapter, we consider discontinuity designs, the creation of synthetic controls, and the important method of instrumental variables.

TABLE 3.6

Measures of Impact of Being in an Honors Program (Treatment Effects) on Retention and GPA

Outcome Variable and Method	ATE	p-Value	ATT	p-Value
Retention				
Regression methods				
Regression adjustment, linear	0.361	0.08	0.106	0.64
Regression adjustment, logit	0.195	0.27	0.114	0.60
Common impact version, linear	0.176	0.53		
Common impact version, logit	0.165	0.46		
Inverse probability weighting				
With original weights	0.269	0.22	0.126	0.57
With stabilized weights	0.269	0.24	0.126	0.58
Doubly robust:				
IPW on regression adjustment	0.348	0.07	0.074	0.73
IPW, reg. adj., logit	0.188	0.28	0.094	0.65
Augmented IPW	0.358	0.08	n.a.	n.a.
Matching				
Nearest neighbor/Mahalanobis	0.294	0.31	0.036	0.89
Propensity score matching	0.200	0.25	0.000	1.00
GPA				
Regression adjustment	0.616		0.418	
RA if SAT ≥ 1,200	0.436		0.425	
Coarsened exact matching			0.474	0.003

Note: Based on the (hypothetical) data displayed in Table 3.1. Twenty observations in all cases except for the final row, where there are 15 observations. Since "retention" is a binary variable, linear models (which are widely used) are generally more approximate than logit models. Logit model effects are the marginal effects at the mean values of the covariates.

Most of the matching methods considered here, including covariate matching, can be estimated in Stata using the teffects command. CEM may be done with the cem command.

Note

1. Z is also known as a mediator variable, as it lies between the treatment variable T and the outcome variable Y. There is another subfield of causal inference known as mediator analysis that seeks to understand the direct and indirect effects of T on Y and is beyond the scope of this book; see Vanderweele (2015) for details of this topic.

References

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. 2009. "Cem: Coarsened Exact Matching in Stata". *The Stata Journal*, 9(4): 524–546.

Crump, Richard, Joseph Hotz, Guido Imbens, and Oscar Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects". *Biometrika*, 96(1): 187–199.

Dehejia, Rajeev, and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies". *Review of Economics and Statistics*, 84(1): 151–161.

Diamond, Alexis. 2005. Reliable Estimation of Average and Quantile Causal Effects in Non-Experimental Settings. Working draft. Cambridge, MA: Harvard University.

Diamond, Alexis, and Jasjeet Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies". *Review of Economics and Statistics*, 95(3): 932–945.

Drukker, David. 2016. "Estimating Treatment Effects from Observational Data using Teffects, Stteffects, and Eteffects. Presented at UK Stata Users' Group Meeting 2016. London: Stata Users Group.

Glynn, Adam, and Kevin Quinn. 2010. "An Introduction to the Augmented Inverse Propensity Weighted Estimator". *Political Analysis*, 18: 36–56.

Haughton, Jonathan, and Alison Kelly. 2014. "Student Performance in an Introductory Business Statistics Course: Does Delivery Mode Matter?" *Journal of Education for Business*, 90(1): 31–43.

Hernán, Miguel, and James Robins. 2006. "Estimating Causal Effects from Epidemiological Data". *Journal of Epidemiology and Community Health*, 60: 578–596.

Hernán, Miguel, and James Robins. 2025. *Causal Inference: What If*. Boca Raton, FL: CRC Press.

Hitchcock, Christopher. 2010. "Probabilistic Causation". In *Stanford Encyclopedia of Philosophy*, https://stanford.library.sydney.edu.au/archives/sum2010/entries/causation-probabilistic/#Int [Accessed June 11, 2020]

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stewart. 2006. "Matching as Non-parametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference". http://gking.harvard.edu/files/matchp.pdf [An excellent guide for practitioners.]

Holland, Paul W. 1986. "Statistics and Causal Inference". *Journal of the American Statistical Association*, 81(396): 945–960. https://doi.org/10.2307/2289064

Iacus, Stefano Maria, Gary King, and Giuseppe Porro. 2008. "Matching for Causal Inference Without Balance Checking". https://doi.org/10.2139/ssrn.1152391

Imbens, Guido. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review". *Review of Economics and Statistics*, 86(1): 4–29.

Keeter, Scott. 2021. "Q&A: A Conversation about U.S. Election Polling Problems in 2020". *Short Reads*, Pew Research Center. https://www.pewresearch.org/short-reads/2021/07/21/a-conversation-about-u-s-election-polling-problems-in-2020/

Kennedy, Courtney, and Hannah Hartig. 2019. Response Rates in Telephone Surveys Have Resumed Their Decline. In *Short Reads*, Washington, DC: Pew Research Center. https://www.pewresearch.org/short-reads/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching". *Political Analysis*, 27: 4. https://j.mp/2ovYGsW

Lee, Brian, Justin Lessler, and Elizabeth Stuart. 2011. "Weight Trimming and Propensity Score Weighting," *PLOS ONE*, 6(3): e18174.

Maciel, Fernanda. 2020. The Analytics of Vulnerable Populations in Brazil. *PhD dissertation*, Bentley University.

Morgan, Stephen, and Christopher Winship. 2015. *Counterfactuals and Causal Inference Methods and Principles for Social Research*, 2nd edition. New York, NY: Cambridge University Press.

Optimizely. 2020. *The Big Book of Experimentation*. https://www.optimizely.com/resources/experimentation-case-studies/ [Accessed June 11, 2020]

Pearl, J. 1995. "Causal Diagrams for Empirical Research". *Biometrika*, 82(4): 669–710.

Rosenbaum, P., and D. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika*, 70(1): 41–55.

Vanderweele, Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York, NY: Oxford University Press.

4

Causality: Synthetic Control, Regression Discontinuity, and Instrumental Variables

4.1 Introduction

In Chapter 3, we discussed a number of techniques for estimating causal effects. But there is more to be said on the subject, and in this chapter, we begin with a discussion of double differencing and follow up with an explanation of the methods of synthetic controls, regression discontinuity, and instrumental variables.

A good starting point is to return to a modified version of the example, from Chapter 3, of sales of electric cars. We imagine there was an advertising campaign in the Northeast of the United States in 2020, but not in the rest of the United States, and we would like to know whether the campaign was effective. The data are summarized in Table 4.1.

Based on these numbers, a single difference shows a rise of sales in the Northeast of 160, but as we discussed in Chapter 3, from this information alone we cannot conclude that the campaign was a success; perhaps sales would have risen anyway.

TABLE 4.1

Sales of Electric Cars per Million Population

	2019	**2020**	**Difference**
Northeastern US	400	560	+160
Rest of the US	200	300	+100
Difference	+200	+260	**+60**

Note: The numbers are invented by the authors for illustrative purposes.

DOI: 10.1201/9780429172588-4

4.2 Double Differencing

A common way to address the problem is by using a *double difference* (aka difference in difference). We note that sales of cars rose by 160 in the Northeast and by 100 in the Rest of the United States. The rise was larger by 60 in the Northeast, and we could reasonably attribute this to the advertising campaign in the Northeast. This conclusion crucially depends on the *parallel trends assumption*, which is the idea that in the absence of the advertising campaign, the trend in sales in the Northeast, and in the Rest of the United States would have been the same.

The idea is captured graphically in Figure 4.1. The vertical axis (Y) shows sales; the lower solid line shows the evolution of sales in the Rest of the United States between 2019 and 2020, and the upper solid line shows the trajectory of sales in the Northeast. The dashed line shows the counterfactual, based on the assumption of parallel trends. The difference between point A and point B, here 60, represents the double-difference estimate and is our measure of the impact of the advertising campaign.

It helps to formalize the presentation somewhat. We want to measure the average treatment effect on the treated (ATT), which is given by

$$ATT = E\left[Y_{it}^{T} - Y_{it}^{C} \mid T_i = 1\right] \tag{4.1}$$

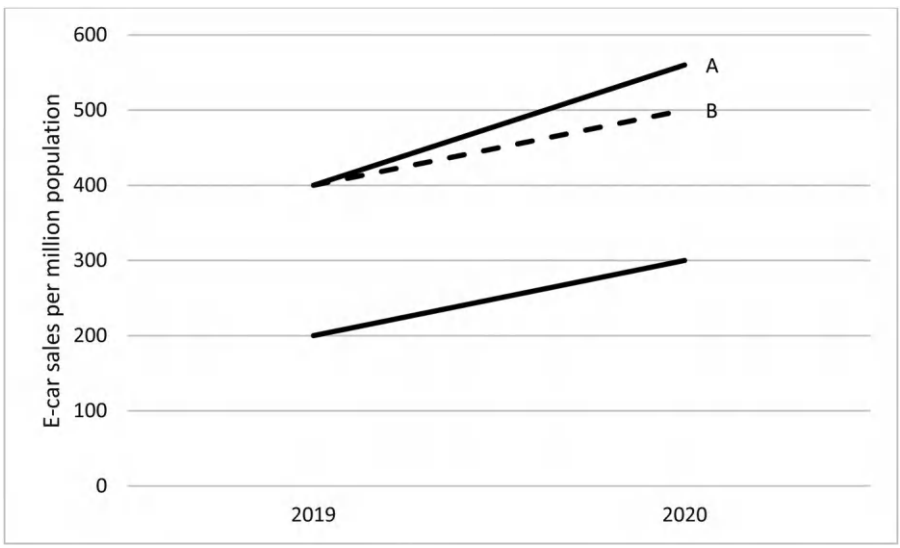


FIGURE 4.1
E-car sales in 2019 per million people for the Northeast of the United States (top solid line) and the rest of the United States (bottom solid line). Dotted line is a counterfactual (if there were no advertising campaigns). (Data from Table 4.1.)

In words, this is the expected value for the treated ($T_i = 1$) of the outcome in time t when they are treated (Y_{it}^T) compared to the outcome that would have been observed had they not been treated (Y_{it}^C). The latter term is unobservable.

If we assume parallel trends, we are supposing that

$$E\left[\left(Y_{it}^C - Y_{i\,t-}^C\right) \middle| T_i = 1\right] = E\left[\left(Y_{it}^C - Y_{i\,t-}^C\right) \middle| T_i = 0\right] \tag{4.2}$$

This says that, in the absence of the advertising campaign, sales would have risen by the same amount in the treated area (left-hand side) as in the reference area (right-hand side). Substituting Eqn. (4.2) into Eqn. (4.1) gives

$$ATT = E\left[\left(Y_{it}^T - Y_{i\,t-1}^C\right) \middle| T_i = 1\right] - E\left[\left(Y_{it}^C - Y_{i\,t-1}^C\right) \middle| T_i = 0\right] \tag{4.3}$$

This is the double-difference estimator. In our example, it gives

$$ATT = [560 - 400] - [300 - 200] = 60. \tag{4.4}$$

With data on many units, this is typically estimated using an ordinary least squares (OLS) regression of the form

$$Y_{it} = \beta_0 + \beta_1 \text{Time}_i + \beta_2 \text{Treatment}_i + \beta_3 \text{Time}_i \times \text{Treatment}_i + \varepsilon_i \tag{4.5}$$

in the canonical case where there are just two time periods and a treatment that applies only in the second period and to only some of the units. Then the estimate of β_3 gives the impact.

Equation (4.5) is often generalized to the case of more than two time periods, in which case we get the *two-way fixed effects* (TWFE or 2FE) model:

$$Y_{it} = \theta_t + \eta_i + \alpha T_{it} + \varepsilon_{it} \tag{4.6}$$

This has become "the default method for estimating causal effects from panel data" (Imai and Kim 2020), and we summarize a celebrated example related to the impact of minimum wages in Box 4.1. However, the use of Eqn. (4.6), when there are multiple time periods, is problematic. Consider Figure 4.2, which traces the effects of staggered treatments – perhaps an early and ongoing advertising campaign in the Northeast (top solid line), a later one in the Midwest (middle solid line), and no campaign elsewhere (bottom solid line). We also show a slowing trend in the second interval. The parallel trends assumption needs to apply to all groupings of observations (Northeast, Midwest, rest of United States), and the appropriate measure of impact applies double differences relative to the untreated baseline, shown on the graph as the differences between the solid lines and their dotted counterfactuals. Unfortunately, Eqn. (4.6) does not calculate this, although recent work by Callaway and Sant'Anna (2021b) and Wooldridge (2021) suggests

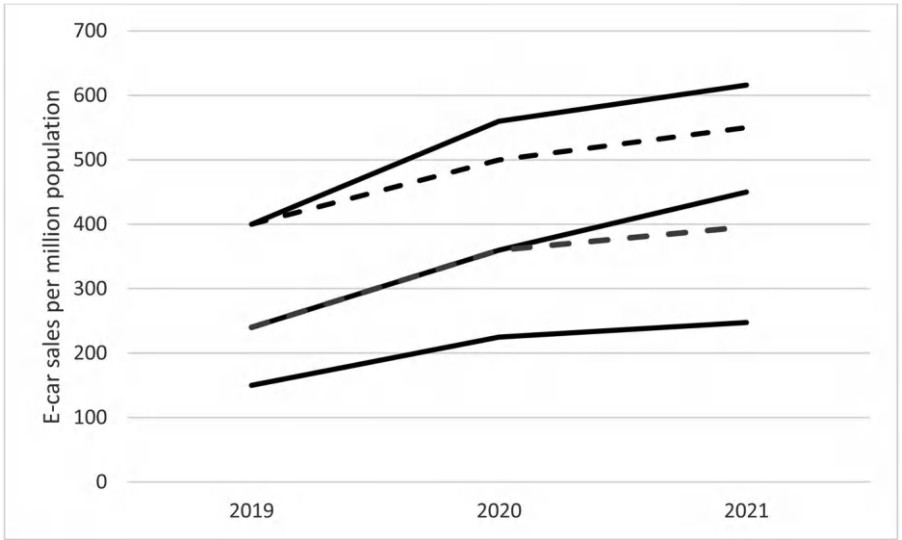


FIGURE 4.2
E-car sales in 2019 per million people for the Northeast of the United States (top solid line), the Midwest of the United States (middle solid line), and the rest of the United States (bottom solid line). Dotted lines are counterfactuals if there were no advertising campaigns. (Illustrative data created by authors.)

some work-arounds. Where interventions are not staggered – that is, they come and go – it is especially hard to measure impact consistently. This is an active area of research.

It is important to note that the parallel trends assumption is sensitive to the units used. Suppose we were to express the numbers in Table 4.1 in log form, as shown in Table 4.2.

The (log) growth rate of sales is 33.6% in the Northeast and 40.5% in the rest of the United States. Now the parallel trends assumption is that the growth rate (i.e., the change in the log value) would have been the same for the treated and non-treated groups in the absence of the advertising campaign. In this case, we may conclude that the advertising campaign in the Northeast may

TABLE 4.2
Log of Sales of Electric Cars per Million Population

	2019	2020	Difference
Northeastern US	5.991	6.328	+0.336
Rest of the US	5.298	5.704	+0.405
Difference	+0.693	+0.624	**−0.069**

Note: The numbers are invented by the authors for illustrative purposes and are derived from Table 4.1.

BOX 4.1 EXAMPLE

A celebrated application of double differencing in economics is the study by Card and Krueger (1994) of the effects of raising the minimum wage on employment in the fast-food industry in New Jersey and Pennsylvania. In April 1992, New Jersey raised its minimum wage by 18%, from $4.25 to $5.05 per hour, but there was no change in the minimum wage in neighboring Pennsylvania. Traditional economic theory predicts that the increase in the minimum wage in New Jersey would reduce employment there relative to Pennsylvania.

Card and Krueger collected information on employment in standard fast-food outlets (Burger King, KFC, Roy Rogers, and Wendy's) in New Jersey and Eastern Pennsylvania in February 1992 (pre-treatment) and November 1992. The number of full-time equivalent employees per outlet is shown in Table 4.3.

The results were completely unexpected: Outlets in New Jersey *increased* the average number of full-time equivalent employees by 2.75 (13%) relative to Pennsylvania, as the double-difference calculation shows in Table 4.3. The explanation for this result is not entirely clear but may reflect the noncompetitive nature of the market for low-skilled workers or that the higher minimum wage in New Jersey put more purchasing power in the hands of fast-food clients. Card and Krueger's study, which was done with great care, led to a flurry of subsequent work. The applicability of the findings in other contexts ("external validity") may be limited: Nobody is seriously suggesting, for instance, that a tripling of the minimum wage would lead to more employment! However, it is a good illustration of how meticulous empirical work can force us to think more deeply about how a market or economy actually behaves.

TABLE 4.3

Full-time Equivalent Employment per Fast-Food Establishment

	February 1992	November 1992	Difference
Eastern Pennsylvania	23.33	21.17	–2.16
New Jersey	20.44	21.03	+0.59
Difference	+2.89	+0.14	**+2.75**

Source: Based on information from 410 restaurants (Card and Krueger 1994).

actually have hurt the growth of sales: The ATT double-difference estimate is –0.069, signifying that the growth of sales in the Northeast was 6.9 percentage points lower than in the rest of the United States.

The parallel trends assumption, whether absolute or in logs, is not testable. The implication is that our measure of causal impact is, once again, dependent on the reasonableness of our underlying model.

Stata has two commands that allow the computation of double differences, namely didregress for repeated cross-sectional data, and xtdidregress for panel data (Stata 2021). Callaway and Sant'Anna (2021a) have created a Difference-In-Difference (DID) package for R that has an updated treatment of the case where there are multiple time periods.

The double-difference model can often be improved by matching observations. The choice of the untreated reference group is also important and is the key element in the synthetic control method, to which we now turn.

4.3 Synthetic Control

In 2008, when the United States was in the grip of a deep recession, sales of vehicles fell sharply – by 18% compared to 2007 – threatening the solvency of some of the major automakers. Under the Troubled Asset Relief Program (TARP), the Federal government lent over $80 billion to the industry, mainly to General Motors and Chrysler. A loan of $12.5 billion was extended to Chrysler in January 2009 on the condition that the company make a number of managerial and strategic changes.

Fremeth et al. (2016) ask a simple question: What impact did the government support have on the sales of vehicles made by Chrysler?

Figure 4.3 shows the evolution of sales of Chrysler vehicles. The first vertical line shows January 2009, when the loan was made, and the second

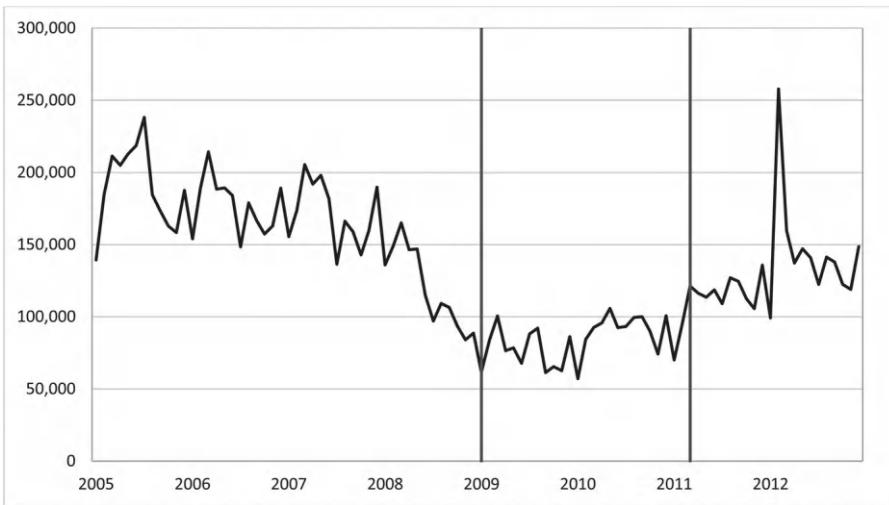


FIGURE 4.3
Monthly sales of Chrysler vehicles in the United States, 2005–2012. (Fremeth et al. 2016.)

indicates May 2011, when it was repaid and Chrysler became free to manage its own affairs once again. From the time series alone, it is difficult to assess the impact of the Federal loan on Chrysler sales, because it was possible that Chrysler sales were just mirroring industry experience.

Until relatively recently, a question like this would have been addressed using a straightforward comparative case study approach. Ford did not take any TARP funds, so one could, for example, compare the trajectories of Ford and Chrysler before and after the bailout to see if there was evidence of divergence in their paths after January 2009. This would be the DID approach discussed in Section 4.2.

In a widely cited study, Card (1990) compared the evolution of wages and unemployment in Miami, before and right after the influx of 125,000 Cubans associated with the 1980 Mariel boatlift, with the experience of four other cities that he deemed similar to Miami, namely Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg. He found, somewhat unexpectedly, that the inflow of Cubans did not lower wages or raise unemployment, relative to the experience of the comparator cities. This is certainly a reasonable approach to the problem, but the choice of comparator cities was essentially ad hoc. Could one do better?

The traditional weakness of comparative case studies is the improvised nature of the choice of comparators (as well as the parallel trends assumption). This has now been addressed by the method of *synthetic control*, hailed by Athey and Imbens (2017, p. 9) as "arguably the most important innovation in the policy evaluation literature in the last 15 years." The method is now widely used in the social sciences and by business analysts. The key innovation is that the synthetic control method provides a systematic way to identify the relevant comparators.

The synthetic control technique is useful when one wishes to estimate the impact of an intervention on an outcome, and potentially only one, or a few, items are targeted by the intervention. A classic example is the study by Hsiao et al. (2012) of the impact of the 1997 change of sovereignty in Hong Kong (from Britain to the People's Republic of China); they concluded that the change of sovereignty had essentially no impact on the growth rate of Hong Kong's GDP.

The idea is to compare the outcome over time of the item targeted by the intervention (the treated item) with the outcome of a control group not targeted by the intervention. The synthetic control thus complements the DID design and potentially improves on any *comparative interrupted time series design* (see Bernal et al. 2019). The synthetic control approach attempts to remove the arbitrariness in the choice of a control group by proposing a "synthetic" control item equal to a weighted average of a set of control items. The weights, which are typically positive or zero, are selected in such a way as to best approximate the characteristics of the targeted item prior to the intervention.

More rigorously, to apply synthetic controls, we need panel data with information for the "focal" unit (e.g., Chrysler) and some potential comparator units (the "donor pool") over T time periods. At some time T_0, there is a "treatment" or shock or intervention that affects the focal unit but not the donor pool. At a minimum, the variables must include an outcome measure Y_{it} (such as vehicle sales), but typically there will be a number of other predictor variables ("covariates"). In the Chrysler example, these might include such variables as the price of autos, miles per gallon, number of company employees, and so on.

Following Abadie (2021), let the focal unit be indexed by 1, and the units in the donor pool be indexed from 2 through $J+1$. The superscript T refers to an intervention ("treatment") and C to the case of no intervention ("controls" or "comparators"). Then the treatment effect on the treated is given by

$$ATT = Y_{it}^{T} - Y_{1t}^{C},\ t > T_0. \tag{4.7}$$

We actually observe the potential outcome Y_{1t}^{T} but not the counterfactual Y_{1t}^{C}.

The key idea in synthetic controls is to replace the unobserved Y_{1t}^{C} with one or more untreated units that look similar to the treated unit as of time T_0. In the Chrysler example, this could be just the sales of Ford vehicles, or a weighted average of sales by several automakers (Ford, Toyota, Subaru, and so on). The synthetic control comparator is given by a weighted average of outcomes from the donor pool:

$$\hat{Y}_{1t}^{C} = \sum_{j=2}^{J+1} w_j Y_{jt}. \tag{4.8}$$

A straightforward regression of, say, Chrysler vehicle sales on the vehicle sales of all the other automakers would generate a set of coefficients that could serve as weights – indeed this is often done – but some of the weights would likely be negative, which implies some extrapolation. So, it is typical to constrain the vector of weights, W, to be between 0 and 1 ($0 \le w_j \le 1$), and for them to sum to 1 if they are well-scaled, so $\sum w_j = 1$. The weights are obtained by minimizing the weighted squared prediction error:

$$\sum_{m=1}^{k} v_m \left(X_{1m} - X_{0m}W\right)^2 \tag{4.9}$$

Here, X_{1m} is the value of the m-th predictor or outcome for the focal unit, X_{0m} is a vector of predictors and/or outcomes for the j units in the donor pool, and the v_m is the weight on the attributes (i.e., covariates and/or outcomes). As a practical matter, this process tends to yield a sparse set of weights that are transparent to use, as we illustrate below.

The central improvement of using synthetic controls is that it gives us a satisfactory method for choosing the weights. Instead of comparing unemployment in Miami with four handpicked (and hence potentially biased) comparators, Card could now use a more formal method to identify, from a larger donor pool, what weighted average of unemployment in other cities best replicated the experience of Miami prior to the Mariel boatlift. It turns out that Peri and Yasenov (2019) did exactly this and found similar results, based on their synthetic control analysis, to Card's more informal approach.

Abadie and Gardeazabal (2003), who first developed the synthetic control method, argue that the weights should be chosen such that the resulting synthetic control "best resembles the pre-intervention values, for the treated unit, of *predictors* of the outcome variable," but some authors instead try to match on the *outcome* variables prior to the intervention. All that is then needed is to examine the post-intervention outcome of this synthetic control and compare it with that of the focal unit.

4.3.1 Chrysler Example

Many of these ideas will be clearer if we examine the study by Fremeth et al. (2016), who wanted to estimate whether the government bailout and control of Chrysler, after the 2008 financial crisis, had an impact on sales of Chrysler vehicles. They got monthly data on vehicle sales (from Ward's Automotive Reports) and other variables from January 2005 through December 2012 for each of the 19 major automobile companies selling in the United States. The treatment period ran from January 2009, when the government loan was disbursed, through May 2011, when Chrysler finished repaying the loan. General Motors was excluded from the donor pool of potential comparators because it too received a government bailout loan, and Jaguar Land Rover was excluded because of missing data. The donor pool thus consisted of 16 companies.

The results of a synthetic control analysis are typically shown visually, as in Figure 4.4. The solid line shows the 12-month moving average of Chrysler sales by month both prior to and during the government intervention. The dashed line tracks the sales of the synthetic control (based on the moving-average data) and is based on applying the weights estimated based on outcome data (using Stata code created by Fremeth et al. 2016) that are reported in Table 4.4. Prior to January 2009, the two series were relatively close, but during the period of the government loan, and associated government restrictions on Chrysler's actions, the two series diverged. This is even easier to see in Figure 4.5, which simply graphs the difference between the two series shown in Figure 4.4. The reduction in Chrysler sales, compared to what would have been expected had the company's sales tracked the synthetic control, is striking. One interpretation is that government involvement lowered the company's sales by about 40,000 vehicles per month.

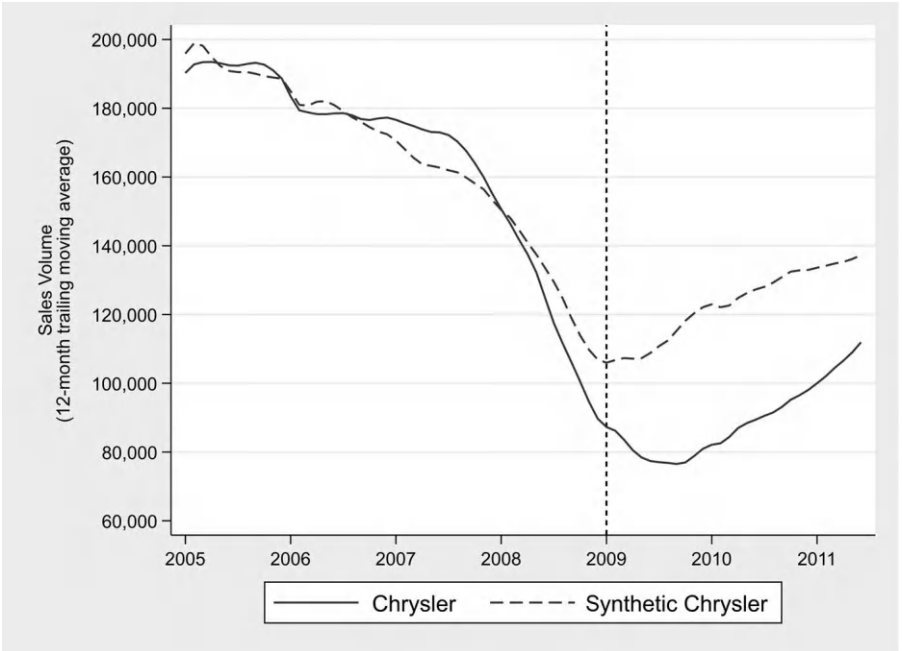


FIGURE 4.4
Sales volume for Chrysler and for Synthetic Chrysler.

TABLE 4.4

Weights Estimated in Synthetic Control Model of Impact of Government Bailout on Chrysler

Company	Weight	Company	Weight	Company	Weight
BMW	0	Isuzu	0.068	Saab	0
Daimler	0.022	Jaguar Land Rover	**	Subaru	0
Ford	0.664	Mazda	0	Suzuki	0
General Motors	*	Mitsubishi	0	Toyota	0.077
Honda	0	Nissan	0.169	Volkswagen-Audi	0
Hyundai-Kia	0	Porsche	0	Volvo	0

Source: Fremeth et al. (2016).
Notes:
* GM was excluded because it was also subject to government intervention.
** JLR was excluded because some data were missing.

This is not the end of the story, however. Perhaps a number of other companies experienced similar patterns, post-January 2009, to that seen by Chrysler, in which case the Chrysler case would not be exceptional. A clever way to address this, in a way that is somewhat similar to testing for significance, is to apply an **"across-unit" placebo test**: Estimate a synthetic

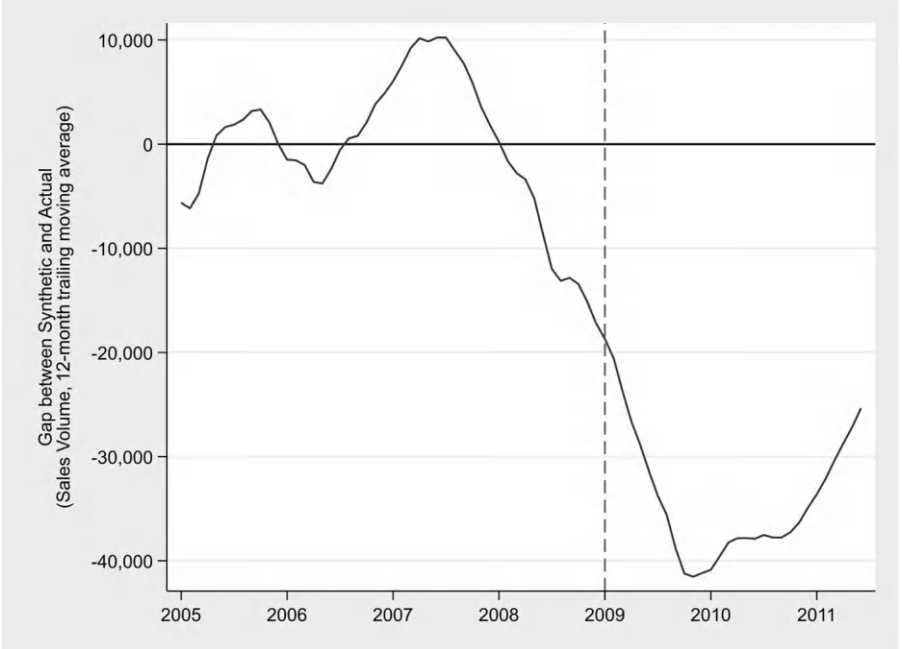


FIGURE 4.5
Difference between Chrysler and Synthetic Chrysler.

control model for each of the other car companies that did not have bailouts, and use those models to measure the gap between each company's sales and its corresponding synthetic control. This is done in Figure 4.6, where each line represents a different car company. It is clear that Chrysler is an outlier; the probability that it would end up at the bottom of the graph is less than 10% (given that there are time series for 17 companies), if the effect were random, so we can be pretty confident that the bailout did have an effect on Chrysler.

If there are enough time periods prior to the intervention, it is often possible to use an "in-time" placebo test. Let us imagine, hypothetically, that there was an intervention in, say, January 2005 instead of January 2009. Construct a synthetic control model for Chrysler using the January 2005 date: If, subsequently, the sales of Chrysler cars closely track the synthetic control, and the placebo intervention had no discernible effect (which is what Fremeth et al. 2016 found), then we have greater confidence that the deviation of Chrysler sales after the real intervention of January 2009 was not accidental.

There are other useful checks on the robustness of the results. For instance, the synthetic control for Chrysler may be sensitive to the weights, as shown in Table 4.4, that are used. A leave-one-out test would recalculate the synthetic

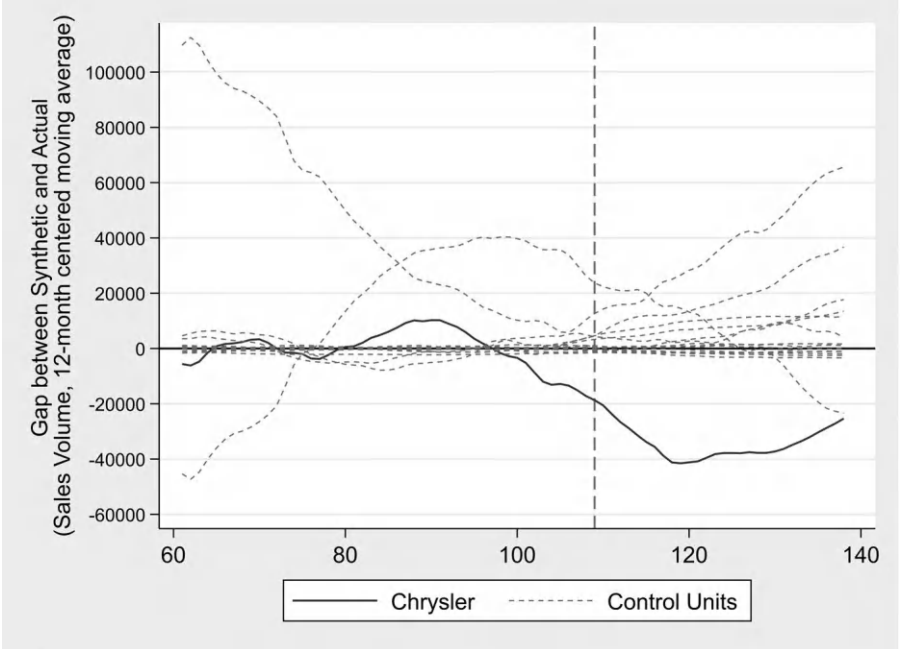


FIGURE 4.6
Placebo tests, among untreated units.

control after leaving out one of the weights (e.g., Daimler, Ford) in turn and examining whether each of the resulting series still shows Chrysler's sales deviating from the synthetic measure. One could also base the analysis on data on outcomes and covariates that run through, for instance, 2007 rather than 2008, in case the latter data are contaminated by being too close to the intervention of January 2009. Fremeth et al. (2016) provide further details.

Mechanically, synthetic controls can be implemented using the synth commands in Stata or R. Cunningham (2021) works through some useful examples.

The synthetic control method works well if the donor pool has untreated units that are otherwise sufficiently similar to the treated unit. It follows that the approach will not be successful if the treated unit is unusual or extreme in some dimension. It is important that there be a stable structure of weights so that the synthetic controls serve as good predictors (under normal circumstances). If either the focal unit or controls in the donor pool have been affected by shocks in the pre-treatment period, this makes it harder to establish a stable underlying relationship. We also need to assume that the treatment of the focal unit – for instance, Chrysler – does not have spillover effects on the control units. There is no formal way to test this, but the case needs to be made that spillovers can be ignored.

4.4 Regression Discontinuity

The median annual wage of a 40-year-old U.S. worker with an undergraduate degree in economics was $90,000 in 2018, compared to $66,000 for any major other than economics (Bleemer and Mehta 2022). Is it reasonable to claim that the choice of an economics major causes one to earn at least $20,000 more?

The problem with the simple comparison of salaries is that the individuals who pick economics may be different in some fundamental, and perhaps unobservable, way from those who opt for other majors. Perhaps only unusual people select themselves into economics, in which case the salary difference might reflect the atypical nature of economists, or perhaps an economics major really does lead to higher salaries, for instance, by imparting technical skills or attitudinal changes.

A recent study by Bleemer and Mehta (2022) tries to isolate the effect of the choice of economics major by making use of a *discontinuity*. In 2008, the University of California Santa Cruz, a large public university, introduced a requirement that anyone wishing to major in economics had to get a grade point average (GPA) of at least 2.8 on the two economics principles courses (microeconomics and macroeconomics). Most of those with a GPA of 2.8 or higher chose to major in economics, while those with a lower GPA did not have that choice. It is reasonable to assume that students who had a GPA of, say, 2.75 were not very different from those with a GPA of 2.85, except that only the latter could major in economics. By exploiting this discontinuity and using some additional tools (including regression) that we consider below, Bleemer and Mehta found that those who just made the cutoff and majored in economics had early-career salaries that were $22,000 higher than those who just failed to make the cutoff, and half of this differential was because economics majors gravitated toward higher-paying industries. This use of discontinuity creates a natural experiment and provides a plausible strategy for identifying a causal effect. Because it offers a clear identification strategy, the technique of regression discontinuity design has become very popular in the last decade.

The starting point of any regression discontinuity design is, naturally, the jump or discontinuity. Such cases are surprisingly common. For instance, in the United States, people become eligible for Medicare at the age of 65; young people can drink when they reach the age of 21; postulants are admitted to university if they have at least some threshold GPA; and candidates are elected to office if they get at least one more vote than anyone else. Many government programs have eligibility criteria, which represent discontinuities that may often be exploited in order to measure the causal impact of these programs.

To develop the ideas, we begin with a simulated example and then consider some real-world cases, including a study of whether companies perform better if their corporate compensation policy has a long-run, rather than short-run, orientation.

Consider the case of a university that automatically tracks students into an honors program if they have a GPA of at least 3.33 (a B+ in most places) on a scale that runs from 0 to 4. If the policy is strictly enforced, we have a sharp discontinuity. We would like to know whether the honors program has any causal impact on subsequent academic performance. In this case, GPA is the *running variable*, and we assume that the outcome of interest is the final exam score (Y), which is on a scale of 0–100. We simulated 1,000 observations of GPA based on a normal distribution (censored at 4), and in the baseline case assumed that the final exam score is unrelated to whether a student is in the honors program.[1] The association between the exam score and GPA is shown in Figure 4.7, where the cutoff of a GPA of 3.33 is marked. Rather than showing all thousand underlying data points, the dots in Figure 4.7 are the average values of the exam score for bins that have equal numbers of observations. The regression lines in Figure 4.7 are fitted to the underlying data using quadratics, separately on each side of the vertical discontinuity. While the two lines do not meet exactly at the discontinuity, they are very close, so a first visual inspection suggests that there is no measurable effect. In other words, the honors program does not appear to boost the final exam scores relative to what we would have expected based on anyone's GPA. This is not surprising, because we constructed the data this way.

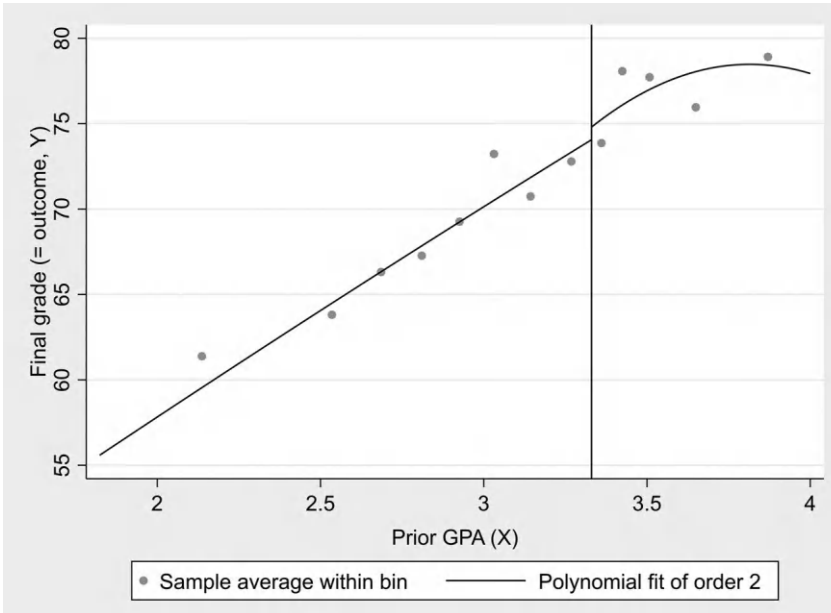


FIGURE 4.7
Relationship of final exam score to GPA and participation in the honors program. (*Note:* Honors program participation requires a GPA of 3.33 or higher. Based on simulated data (see text) that assumes that participation has no impact on exam scores.)

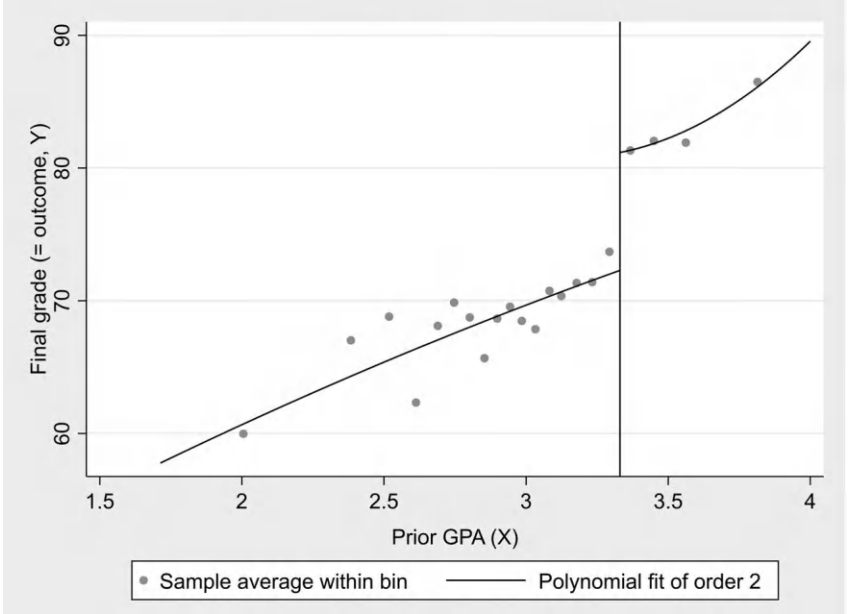


FIGURE 4.8
Relationship of final exam score to GPA and participation in the honors program. (*Note:* Honors program participation requires a GPA of 3.33 or higher. Based on simulated data (see text) that assumes that participation raises exam scores.)

Now let us assume that the honors program raises final exam scores by seven percentage points.[2] We simulated a dataset that only differs from the base case in this respect. The data are summarized visually in Figure 4.8, and it is clear that there seems to be a jump in exam scores for those who are in the honors program relative to those who are not.

These graphs are useful, but they are only the starting point in our analysis. We are interested in how outcomes vary close to the discontinuity. Often, we may think of the points near the cutoff as being essentially randomly scattered a bit above or a bit below the line, with the position being determined substantially by chance – a lucky guess on an exam question, an inconsistency in grading, and the like. For this group, it is as if they were randomly assigned to treatment or not. So, one method of measuring the impact of the treatment is to compare the average outcomes of those just below the cutoff with those just above it, using only the observations that fall within modest bandwidths h^- and h^+ on each side of the cutoff. One problem with this is that there may be a trend in outcomes around the cutoff, in which case it would be better to fit a regression line below and another above the cutoff, using only the observations near the cutoff. The gap between the lines, if any, would measure the impact. This is the local linear (or sometimes quadratic) approximation favored by many researchers (Gelman and Imbens 2019).

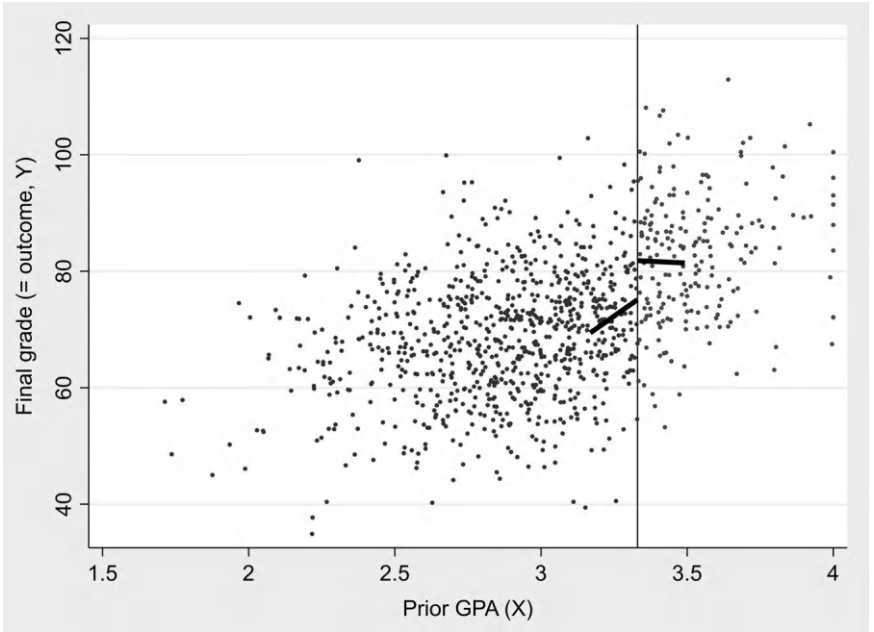


FIGURE 4.9
Relationship of final exam score to GPA and participation in the honors program. (*Note:* Honors program participation requires a GPA of 3.33 or higher. Based on simulated data (see text) that assumes that participation raises exam scores. Dots represent observations; thick black lines show linear regression lines based on observations in the vicinity of the cutoff point.)

The situation is illustrated in Figure 4.9, which shows all one thousand points from our simulated dataset in which the honors program boosts scores. The solid black lines are linear regression lines, estimated using the data points found within 0.168 units below and then above the GPA cutoff of 3.33. There are a number of ways to measure the bandwidth, here 0.168, and the results of regression discontinuity studies are somewhat sensitive to the choice of bandwidth.

To measure the impact of joining the honors program, we measure the distance between the intercepts of the regression lines at the cutoff. Here, the estimated difference is 7.22, and the associated p-value is 0.05, indicating a statistically significant difference from zero; we designed the simulation to have a difference of 7 points.

This measure is a Local Average Treatment Effect (LATE), and strictly speaking, it only applies at the cutoff. In practice, such measures tend to have good internal validity because the identification strategy is usually clear and robust, but there may be limited applicability to other situations ("external validity"). The measure of impact relies on successful extrapolation; in contrast to matching methods, there is no area of common support

in sharp regression discontinuity analysis, although the situation is a bit different in the case of a fuzzy regression discontinuity, discussed further below.

More formally, let Y_i^C be the outcome (such as exam score) if individual i is not treated, and Y_i^T be the outcome if they are treated. We would like to measure $Y_i^T - Y_i^C$, but since an individual is either treated or not, we cannot observe this directly: This is the Fundamental Problem of Causal Inference that we discussed in Chapter 3. Let the treatment (T_i) be associated with the running (or "forcing" or "treatment-determining") variable X_i so $T_i = 1\{X_i \geq c\}$, where c is the cutoff value, or point of discontinuity. Then

$$LATE = \lim_{x \downarrow c} E\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow c} E\left[Y_i \mid X_i = x\right]. \tag{4.10}$$

The terms may be obtained by estimating regressions, the first above the cutoff and the second below it. These regressions may include additional covariates, denoted by the vector Z_i, although in practice this is unlikely to alter the estimates substantially (Imbens and Lemieux 2008). Researchers frequently fit high-order polynomial regressions, but Gelman and Imbens (2019) warn against this practice and make a strong case for using a linear, or at most a quadratic, approximation. One reason is that the estimate of the LATE requires extrapolation to the point of the cutoff, which works best if the estimated equations are fairly robust.

Some researchers graph these functions by rescaling the running variable so that it takes on a value of zero at the discontinuity. The conclusion remains the same, but this emphasizes the change at point zero. In Box 4.2, we provide short illustrations of some further examples of regression discontinuity studies.

BOX 4.2 EXAMPLES

Many applications of regression discontinuity have been in the context of education, and the earliest use of the method was by Thistlethwaite and Campbell in 1960. They sought to measure whether awarding students' certificates of merit had an impact on student attitudes or career plans and compared students who just got an award with those who just missed the cutoff. They found that the public recognition of academic achievement had no observable effect on attitudes or career plans.

An interesting business-related example comes from Flammer and Bansal (2017). They ask an important question: If a corporation designs its compensation package to give executives an incentive to take a long-run view (the treatment), does this add to the value of the firm?

Proposals for long-term executive compensation generally need to be approved at a company's annual general meeting (AGM), and the clever idea here is to compare companies where long-term executive compensation proposals were barely approved at the AGM with companies where such proposals just barely failed. The key to this example is that shareholder proposals for long-term executive compensation that pass by a small margin can be considered as a close-to-random assignment to a long-term orientation treatment for a company.

Based on data for S&P 1500 companies and some additional widely held U.S. companies for the period 1997–2011, Flammer and Bansal identify a total of 808 long-term executive compensation proposals, of which 65 were found to have passed within 5% of the majority threshold (of 50% of the votes) and 152 within 10% of the majority threshold. The outcome of interest (Y) is the rate of abnormal returns on the day of the shareholder meeting.

They then estimate

$$Y_{it} = \beta\, Pass_{it} + P_l\left(v_{it}, \gamma_l\right) + P_r\left(v_{it}, \gamma_r\right) + \varepsilon_{it}, \tag{4.11}$$

where Y_{it} is the outcome variable for company i on the day of the proposal vote, $Pass_{it}$ is a dummy variable equal to 1 if a long-run compensation proposal passed, and the other two terms are polynomial functions that include the vote shares as well as other covariates, one for firms where the proposal did not pass ($P_l(.)$) and the other for proposals that did pass ($P_r(.)$). The estimated value of β, given by $\hat{\beta}$, measures the effect of the compensation plan on the value of the firm and is the difference in the intercept of the two functions at the cutoff point.

Equation (4.11) may be used with or without the polynomial terms. If the sample is restricted to close calls, say proposals that pass within 5% of the majority threshold, the polynomial terms might be omitted (so we simply compare averages on each side of the cutoff) or linear functions used. Flammer and Bansal argue that including the polynomial terms, and using all 808 data points, allows for a more efficient estimate of the effect of passing the proposal, but not all researchers would agree.

4.4.1 Extensions and Further Considerations

We note that the Regression Discontinuity technique relies on the idea that the only jump is the discontinuity under consideration. For this to be reasonable, it is important that the discontinuity not coincide with other jumps. More formally, we require continuity in the running variable, outcome, and covariates (if any) at the cutoff in the absence of treatment. Continuity is not assured a priori: For instance, many people retire at 65, but other relevant

series jump at the same time, such as Medicare or senior discounts. These risks confound our efforts to measure the impact of retirement on such outcomes as the pattern of household spending.

It is good practice to check for continuity. A popular technique is to graph the distribution of the running variable to see whether there are any jumps near the cutoff, and to graph covariates against the running variable, again looking for any possible jumps. It is not possible to check the continuity of the outcome variable because, by construction, we expect a jump at the cutoff. But it may be useful to do a placebo test: Pick one or more cutoffs that are *not* at the point of interest and do a regression discontinuity analysis. There should be no effect, but if there is, then the model may be in doubt.

The discussion so far has assumed a *sharp* discontinuity, so at the cutoff, the probability of treatment suddenly went from 0 to 1 and is not subject to manipulation. In reality, many discontinuities are *fuzzy*, so the probability of treatment jumps by less. For instance, at the age of 65, the probability of being retired might rise from 30% to 70%. This means that close to the cutoff there may be treated as well as non-treated individuals, as Figure 4.10 shows. The techniques needed to measure the impact of treatment (retirement) on, for instance, household consumption spending will be somewhat different. The most common approach is first to regress treatment (i.e., retirement) on an exogenous variable such as age – either in a linear or logistic regression – and then use the predicted probability of treatment on the right-hand side of a consumption regression. This two-stage least squares (TSLS) approach is tantamount to treating age as an instrument for retirement, and retirement is closer to an "intention to treat" rather than a definitive treatment per se.

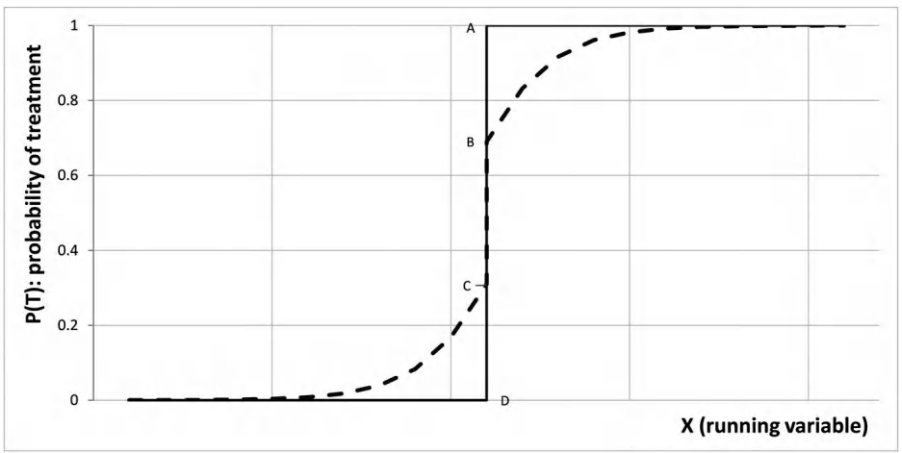


FIGURE 4.10
Sharp versus Fuzzy discontinuity. (The black line shows the probability of treatment under a sharp discontinuity design, while the dashed line traces the probability of treatment if the discontinuity is "fuzzy".)

We return to the technique of instrumental variables later in this chapter. A useful way to think of this is that we are trying to measure $\frac{\Delta Y}{\Delta P}$, which is the response of the outcome variable to a change in the probability. When the discontinuity is sharp, $\Delta P = 1$, given by the distance AD in Figure 4.10. This is no longer the case with a fuzzy discontinuity, where ΔP is given by the distance BC. The probability of retirement may depend on variables, often unobserved and potentially subject to manipulation, other than a person's age. By isolating the change in the probability of retiring that is attributable to age, we purge the probability of retirement of influences that can be altered by the individual and can then identify the effects of exogenously driven retirement on the outcome of interest.

There are variations on the theme. A *regression kink design* looks not for a jump but for a kink in the outcome response function at some cutoff. David Card et al. (2015) used such a design to estimate the effect of unemployment benefits on the duration of unemployment in Austria.

4.4.1.1 Implementation

Regression discontinuity analyses may be performed easily enough in Stata and R, thanks to the efforts of Calonico et al. (2017). In Stata, the `rdplot` command generates graphs such as those in Figures 4.11 and 4.12; the `rdbwselect` command provides flexibility in determining the optimal bandwidths (on each side of the cutoff) for local regressions; and the `rdrobust` command formally measures, and tests the significance of, the LATE. There are equivalent commands in Python. The relevant packages for R and Python may be found at https://rdpackages.github.io/.

4.5 Instrumental Variables

Imagine that you run a bank and decide to offer credit cards to some of your customers. You would like people to use your credit card for their spending (Y_i), thereby generating fee revenue for the bank. To measure the impact of your credit card campaign, you offer credit cards to a random sample of your clients – this is the treatment (R_i) equal to 1 for those who get the offer and 0 otherwise. Spending is also influenced by other "control" variables, such as the client's age or income (the variables A_{ji}) as well as random effects that we cannot or do not measure (ε_i).

The directed acyclic graph (DAG) for this case is shown in Figure 4.11. Assuming the effects to be linear, we may construct a regression of the form.

$$Y_i = \alpha + \beta_0 R_i + \sum_j \beta_j A_{ji} + \varepsilon_i \tag{4.12}$$

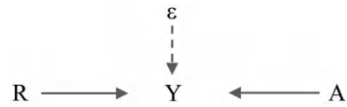


FIGURE 4.11
The directed acyclic graph (DAG) for the effect of offer of treatment (R) on outcome (Y). (*Note:* ε is an unobserved "error" variable, hence the dotted arrow. Other observable influences on Y are represented by A.)

which may be estimated easily enough. The effect of the credit card campaign on spending is given by the estimated value of β_0.

Unfortunately, the problem is never this simple, because not everyone accepts the offer of a credit card, so we have a randomized trial with noncompliance. If R_i is defined as an *offer* of a credit card, then $\widehat{\beta_0}$ measures the effect of the *intention to treat*. This may not be what we need. It combines two distinct effects – the proportion of people who opt for a card and the spending behavior of those who accept a card. Often, we are interested in analyzing these effects separately so that we may seek ways to (i) increase uptake and (ii) increase spending, given uptake.

In this case, we might be tempted to measure the treatment as whether or not the client *accepts* a new credit card (T_i) rather than just whether they were offered one. Then we might want to estimate

$$Y_i = \alpha + \beta_0 T_i + \sum_j \beta_j A_{ji} + \varepsilon_i \tag{4.13}$$

The difficulty here is that those who accept a credit card are unlikely to be typical. Perhaps they are more needy, enjoy spending, or are unusual in ways that we cannot easily measure. So, if we compare those who accept a card with either those who did not accept a card or those who were not offered a card, we do not have a quasi-random assignment of treatment. The unobserved characteristics of the acceptors may make them more prone to spending and more likely to accept the credit card offer. In this case, our measure of the effect of treatment $\widehat{\beta_0}$ from Eqn. (4.13) may be biased: The treatment (T_i) may be picking up the effects of other variables that simultaneously raise the probability of accepting a credit card and spending.

The problem may be clarified with the help of the DAG in Figure 4.12. The unobserved characteristics (U) affect the treatment (T), and outcome (Y), so some of the observed influence of the treatment – getting a credit card – on the outcome is attributable to the effect of the unobserved U; here U is a confounder.

The solution is to find an *instrumental variable* (Z) that causes variation in T but has no direct effect on Y, meaning it only affects Y via its effect on T.

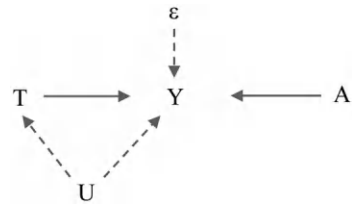


FIGURE 4.12
The directed acyclic graph (DAG) for the effect of treatment (T) on outcome (Y) with an unobserved confounder (U).

Then we construct a new version of treatment, call it $\hat{T}$, that represents variation in treatment that is purged of the influence of the unobservables that are correlated with Y.

The situation is illustrated in the DAGs in Figure 4.13. On the left-hand side, we see how the instrumental variable affects treatment and hence Y. By creating a variable $\hat{T}$ that varies with Z, but not with U, we get the right-hand DAG, and it now becomes possible to get an unbiased measure of the influence of treatment on outcome.

In our example, the offer of treatment (T) would potentially be a suitable instrument because acceptance of a new credit card is unlikely unless preceded by an offer of one.

Mechanically, the commonest technique is to use TSLS. First, regress the treatment on the instrumental variable(s) and other exogenous controls in a first-stage equation of the form:

$$T_i = a + bZ_i + \sum_j c_j A_{ji} + e_i \tag{4.14}$$

From this, get the predicted values of T_i (i.e., $\widehat{T_i}$), and use these instead of T_i in Eqn. (4.13). These predicted values will no longer be influenced by the unobserved Us. In practice, most software packages allow one to estimate

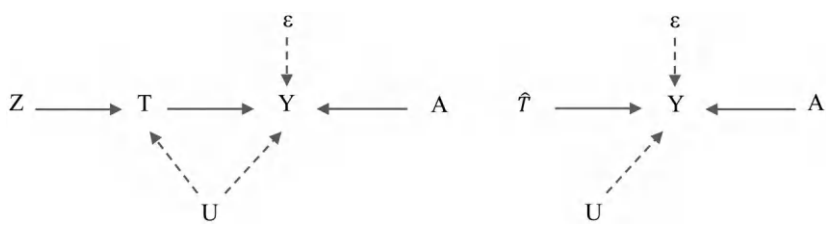


FIGURE 4.13
The directed acyclic graph (DAG) for the effect of treatment (T) on outcome (Y) with an instrument (Z).

TSLS with a single command, such as ivregress in Stata, and extract the effect of treatment in $\widehat{\beta_0}$. Other estimation methods are also possible.

Instrumental variables estimation is widely used by economists (e.g., Boonperm et al. 2013), but it does not always work well. The main challenge is finding a suitable instrumental variable (or variables). A good instrument (Z) needs to be strongly correlated with the treatment indicator (T) – instrument relevance – while at the same time being uncorrelated to the error term in Eqn. (4.2) – instrument exogeneity, or the "exclusion restriction." It is possible to test for instrument relevance, using Eqn. (4.14), but the case for instrument exogeneity has to be made using logic, theory, or common sense. Murray (2005, p. 18), in his review of IV estimation, notes that

> all instruments arrive on the scene with a dark cloud of invalidity hanging overhead … [and] the credibility of IV estimates rests on the arguments offered for the instruments.

4.5.1 Simulation Example

To give a sense of how IV estimation works, consider the following example that is simulated with 1,000 observations. Suppose that a consumer's credit card spending (Y) depends on whether they accept another credit card (the treatment T), their income (A), an unobservable factor (U), and further random elements (u_Y), as follows:

$$Y = 3 + 1.5\,T + 4\,U + A + u_Y \tag{4.15}$$

All variables are normalized, and U, A, and u_Y are drawn from standard normal distributions. The true influence of treatment – which is set to 1 if the consumer has accepted another credit card and to zero otherwise – on spending is 1.5. Let us further suppose that acceptance of another credit card is influenced by income, the unobserved factor, and some exogenous outside influence (Z), giving

$$\begin{aligned} T &= 1 \quad \text{if } 0.05 + 0.1\,U + 0.8\,Z + u_u > 0 \\ T &= 0 \quad \text{otherwise} \end{aligned} \tag{4.16}$$

We are not able to estimate Eqn. (4.16) because U is unobservable. If we omit U, our regression estimate (from Eqn. 4.15) gives

$$\begin{aligned} Y = \quad & 2.00 \quad + \quad 3.50\,T \quad + \quad 3.85\,A \\ & p < .001 \qquad p < .001 \qquad p < .001 \end{aligned} \tag{4.17}$$

The estimate of the treatment effect, at 3.5, is too large – it should be 1.5 – because it is picking up not just the effect of T but of U (working via T).

An estimate of the first-stage ("participation") Eqn. (4.16) gives

$$\begin{array}{rccccc} \hat{T} = & 0.52 & + & 0.21\ Z & - & 0.001\ A \\ & p < .001 & & p < .001 & & p < .95 \end{array} \tag{4.18}$$

Using the predicted values ($\hat{T}$) from this equation in the second-stage Eqn. (4.15) gives

$$\begin{array}{rccccc} Y = & 2.07 & + & 1.46\ \hat{T} & + & 3.87\ A \\ & p < .001 & & p < .001 & & p < .001 \end{array} \tag{4.19}$$

The coefficient here on $\hat{T}$, at 1.46, is very close to the theoretical value of 1.5 established in the simulation. The t-statistic on the coefficient of T in Eqn. (4.17) is 28.1, compared to 3.6 in Eqn. (4.19). Ignoring the endogeneity issue may give estimates that are precisely wrong, while IV estimates are approximately right.

Despite the difficulty in identifying compelling instruments, IV is popular because it is frequently the only identification technique available. We have used the approach to help measure the impact of microcredit in Thailand: Borrowers may be different from non-borrowers, but the Thailand Village Fund makes a fixed amount of credit available per village. In smaller villages, the probability of a loan (T) is higher, so village size can serve as an instrument for microcredit borrowing (Haughton and Khandker 2009). We have also looked at the impact of different scripts in a fundraising telemarketing campaign. Not everyone picks up the phone, but calling someone is an instrument that helps drive whether one picks up the phone, while not influencing whether they donate, given that they pick up the phone (Lo and Li 2021).

If we assume that everyone responds to treatment (such as a credit card offer) in the same way, we have the homogeneous response case, and the results may plausibly be generalized to the wider population, giving the results external validity. However, it is often the case that we have heterogeneous responses to treatment – women may respond differently compared to men, for instance – in which case we only have a LATE on the outcome, and there may be low external validity.

Consider again the campaign to offer credit cards. Using the terminology of Angrist et al. (1996), some people may sign up for a credit card even without the offer ("always takers"), some may never accept another credit card ("never takers"), a few might give up credit cards when they get another offer ("defiers"), and the remaining subpopulation responds to the offer ("compliers"). When responses are heterogeneous and a number of assumptions are met (Imbens 2014), the IV estimates only measure the effect of treatment *on the compliers*. This might be a special group, not necessarily representative of the population as a whole, and it may be difficult to identify, in which case the scope of application of the measure of impact is hard to determine (Imbens 2014).

The field of practical causal inference is rapidly expanding. The techniques discussed in Chapters 3 and 4 provide an introduction to the field and should be enough to allow the user to get started with the estimation of causal effects.

Notes

1. We assumed $GPA \sim N(3, 0.4)$, and $Y = 40 + 10 \times GPA + N(0, 10)$.
2. Here $Y = 40 + 7 \times T + 10 \times GPA + N(0, 10)$, where T is the "treatment," here induction into the honors program.

References

Abadie, Alberto. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects". *Journal of Economic Literature*, 59(2): 391–425.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country". *American Economic Review*, 93(1): 113–132.

Angrist, Joshua, Guido Imbens, and Donald Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables". *Journal of the American Statistical Association*, 91: 444–472.

Athey, Susan, and Guido Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation". *Journal of Economic Perspectives*, 31(2): 3–32.

Bernal, James Lopez, Steven Cummins, and Antonio Gasparrini. 2019. "Difference in Difference, Controlled Interrupted Time Series and Synthetic Controls". *International Journal of Epidemiology*, 48(6): 2062–2063.

Bleemer, Zachary, and Aashish Mehta. 2022. "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major". *American Economic Journal: Applied Economics*, 14(2): 1–22.

Boonperm, Jirawan, Jonathan Haughton, and Shahidur Khandker. 2013. "Does the Village Fund Matter in Thailand? Evaluating the Impact on Incomes and Spending". *Journal of Asian Economics*, 25: 3–16.

Callaway, Brantly, and Petro Sant'Anna. 2021a. "Getting Started with the DID Package". https://www.stata.com/new-in-stata/difference-in-differences-DID-DDD/ [Accessed Jan 6, 2022.]

Callaway, Brantly, and Pedro Sant'Anna. 2021b. "Difference-in-Differences with Multiple Time Periods". *Journal of Econometrics*, 225(2): 200–230.

Calonico, Sebastian, Matias Cattaneo, Max Farrell, and Rocio Titiunik. 2017. "Rdrobust: Software for Regression-Discontinuity Designs". *Stata Journal*, 17: 372–404.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market". *Industrial and Labor Relations Review*, 43(2): 245–257.

Card, David, and Alan Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". *American Economic Review*, 84(4): 772–793.

Card, David, David Lee, Zhuan Pei, and Andrea Weber. 2015. "Inference on Causal Effects in a Generalized Regression Kink Design". *Econometrica*, 83(6): 2453–2483.

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. New Haven, CT: Yale University Press.

Flammer, Caroline, and Pratima Bansal. 2017. "Does a Long-Term Orientation Create Value? Evidence from a Regression Discontinuity". *Strategic Management Journal*, 38(9): 1837–1847.

Fremeth, Adam, Guy Holburn, and Brian Richter. 2016. "Bridging Qualitative and Quantitative Methods in Organizational Research: Applications of Synthetic Control Methodology in the U.S. Automobile Industry". *Organization Science*, 27(2): 462–482.

Gelman, Andrew, and Guido Imbens. 2019. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs". *Journal of Business and Economic Statistics*, 37(3): 447–456.

Haughton, Jonathan, and Shahidur Khandker. 2009. *Handbook of Poverty and Inequality*. Washington, DC: World Bank.

Hsiao, C., H. S. Ching, and S. K. Wan. 2012. "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China". *Journal of Applied Econometrics*, 27: 705–740.

Imai, Kosuke, and In Song Kim. 2020. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data". *Political Analysis*, 29: 405–415.

Imbens, Guido. 2014. "Instrumental Variables: An Econometrician's Perspective". *Statistical Science*, 29(3): 323–358.

Imbens, Guido, and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice". *Journal of Econometrics*, 142(2): 615–635.

Lo, Victor S. Y., and Zhuang Li. 2021. "Causal Inference and Machine Learning for Outbound Strategy". *Joint Statistical Meetings (JSM) Proceedings*. American Statistical Association.

Murray, Michael. 2005. *The Bad, the Weak, and the Ugly: Avoiding the Pitfalls of Instrumental Variables Estimation*. Lewiston ME. Bates College.

Peri, Giovanni, and Vasil Yasenov. 2019. "The Labor Market Effects of a Refugee Wave: Synthetic Control Method Meets the Mariel Boatlift". *Journal of Human Resources*, 54(2): 267–309.

Stata. 2021. "Difference-in-differences (DID) and DDD Models". https://www.stata.com/new-in-stata/difference-in-differences-DID-DDD/

Thistlethwaite, Donald, and Donald Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment". *Journal of Educational Psychology*, 51(6): 309–317.

Wooldridge, Jeffrey. 2021. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators". https://doi.org/10.2139/ssrn.3906345

5

Directed Acyclic Graphs

At the heart of all business decision-making is a desire to know whether A causes B. Does an ad campaign boost sales? Does green packaging increase customer loyalty? Do pay raises boost worker productivity?

Sometimes we may be able to infer causality by running an experiment – for instance, by piloting green packaging in just one area and tracking the outcomes relative to other areas. We examine randomized trials in more detail in Chapter 3, but in practice, we may not have the opportunity, resources, or time to conduct complete experiments. Most of the time we have to make do with imperfect data.

It turns out that under certain circumstances, it is actually possible to "discover causal structures in raw statistical data" (Conrady and Jouffe 2013, p. 17), but as we will see, this is not always easy to do, requires special techniques, and sometimes fails. One of the foremost exponents of this approach is Judea Pearl, who won the Turing Prize in 2011 "for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning" (http://amturing.acm.org/award_winners/pearl_2658896.cfm), and whose book *Causality* (Pearl 2000) remains perhaps the most important in the field, along with a more concise survey (Pearl 2009) and the more accessible *The Book of Why* (Pearl and Mackenzie 2020). The approach uses a mixture of logic and statistical correlations to infer (where possible) the directions of causality among a set of variables.

Practically, this approach is implemented through the estimation of *directed acyclic graphs* (DAGs), also referred to as *Bayesian Networks*, a term coined by Pearl. Causal networks are Bayesian Networks that require the relationships to be causal.

In this chapter, we first show how causality can sometimes be inferred from data; explain the logic behind, and main components of, DAGs; discuss the software and algorithms needed to estimate DAGs in practice; and develop an extended example related to marketing mix that serves to illustrate the application of these and related techniques to an important practical problem.

5.1 Can Causality Be Inferred from Data?

Suppose we have extensive survey data that show a clear relationship between variables A and B: For instance, how thin a person is (A) may be related to

DOI: 10.1201/9780429172588-5

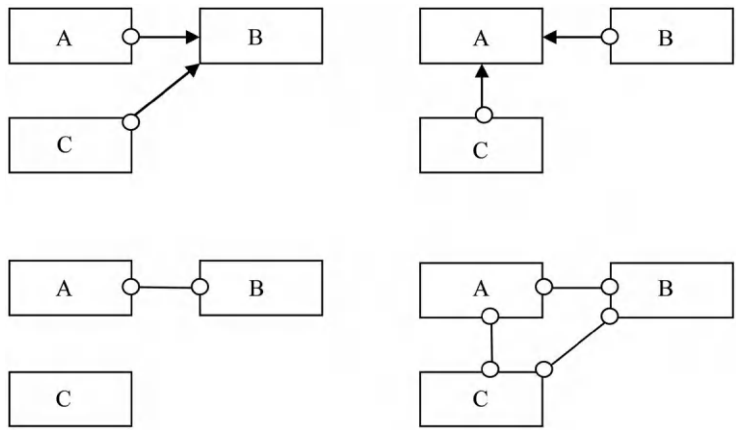


FIGURE 5.1
Graphs of three variables (A, B, and C) with varying causal structures. (Based on Bryant et al. 2009, Figure 2.)

the quantity of diet soda that they drink (B). We would like to know whether they drink diet soda because they are thin or are thin because they drink diet soda. If this is the only information we have, then it is not practically possible to infer the direction of causality.

Surprisingly enough, if we have some additional information, we may be able to say something about causality. Suppose, for instance, we have information about the age of individuals (variable C). We now set out the possible relationships between A, B, and C in the path diagrams shown in Figure 5.1 (which draws on Bryant et al. 2009). Here, the symbol A $\circ\!\!\rightarrow$ B represents a directed "edge" indicating that either A causes B, or they share a common latent cause, or both; and the edge A $\circ\!\!-\!\!\circ$ B means that either variable causes the other, or they share a common latent cause, or both.

In our example, we know that there is some relationship between A (thinness) and B (diet soda consumption): Formally, A and B are not independent (i.e., $A \not\perp B$). However, we do not yet know anything about the causal relationship between the two.

Given our information on variable C – age in our example – there are now four possible cases of interest:

1. The first possibility is that A and C are independent, but B and C are not (i.e., $A \perp C$ and $B \not\perp C$). This is the situation shown in the top left panel of Figure 5.1. In this case, we cannot reject the possibility that A causes B. This is the only logical possibility; if C caused B and B caused A, then C and A would not be independent; and if B caused A and C, then A and C would not be independent.

   In our example, if age (C) is unrelated to how thin someone is (A), but young people are more likely to drink diet soda (B), then we infer

that thin people drink diet soda, but drinking diet soda does not affect how thin you are. If diet soda kept you slender, and is drunk mainly by the young, then there would be an association between age and how thin you are, but we have ruled this out.

2. A second possibility, shown in the top right panel of Figure 5.1, is that C is independent of B but not of A (i.e., $A \not\perp C$ and $B \perp C$). If A caused B and is not independent of C, then B would not be independent of C; yet we are told it is. So we infer that *A cannot cause B* in this case, and we reject the null hypothesis that A causes B. In our example, this is the case where young people are thinner but are not more likely to drink diet soda. Here we conclude that diet soda keeps you slim.
3. The next possibility (bottom left panel of Figure 5.1) is that C is independent of both A and B (i.e., $A \perp C$ and $B \perp C$). We have assumed that there is some relationship between A and B, but the information on C is of no help in determining the direction of causality.
4. Finally, C may not be independent of either A or B (i.e., $A \not\perp C$ and $B \not\perp C$), as shown in the bottom right panel of Figure 5.1. We cannot reject the possibility that A causes B, but it is also possible that B causes A, or that neither causes the other, but both are simply buffeted by a common cause (such as C, in this case).

Bryant et al. (2009) illustrate the logic of this situation with a nice example. It is well known that greater alcohol consumption is associated with higher road fatalities, but is the alcohol causing the deaths, or are the deaths inducing people to drink away their sorrows?

Using nationwide time series data for the United States from 1947 through 1993, they find that, as expected, there is indeed a correlation between road deaths and alcohol consumption per head in the population at large. After filtering the data to remove the effects of non-stationarity (Bryant et al. 2009, pp. 370–371), they get a correlation coefficient of 0.341, which we mark with an asterisk because it is statistically significant: Thus ρ(death, alcohol) = 0.341*. However, without further information, we cannot be sure of the direction of causality in this relationship.

They also have information on the average speed on highways and are able to estimate the following correlation coefficients:

ρ(death, speed) = 0.386*

ρ(alcohol, speed) = 0.136.

Following our earlier logic, there is only one way to organize this information, which is as shown in Figure 5.2.

In other words, it is alcohol that causes road deaths, and not deaths that lead to more drinking. We now have made a clear causal inference. Geiger et al. (1990) further discusse how to identify independencies in DAGs.

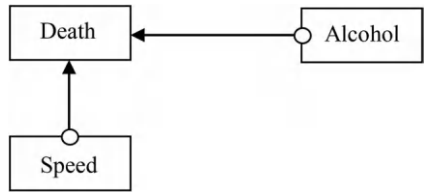


FIGURE 5.2
DAG identifying the causes of road deaths. (Based on Bryant et al. (2009).

The introduction of additional information in order to help make sense of causality is somewhat analogous to the use, in regression analysis, of instrumental variables to identify the impact of one variable on another, a method that we discussed in more detail in Chapter 4.

5.2 Using DAGs to Infer Causality

Six months ago your company launched a campaign to sell more backpacks. Prior to the campaign, 2% of your customers bought backpacks. Now, six months after the campaign, 2.2% of customers are buying backpacks, and the marketing department is trumpeting its achievement, arguing that the campaign has worked. We assume here that a randomized control group is not available for a "Gold Standard" causal measurement. You are almost persuaded until your data analyst points out that purchases of backpacks among male customers fell from 4% to 3.5% and among female customers from 1% to 0.7%. Perhaps the campaign was a failure. At a minimum, you want to understand more about what effects were caused by the campaign.

A DAG can often be created in order to make sense of this situation (which is an illustration of Simpson's Paradox). The process of building a DAG begins with the construction of a path diagram, which summarizes the qualitative links between variables. As before, variables are shown in nodes, but now they are either joined by a directed arc ($A \rightarrow B$) that specifies an assumed causal direction, or there is no arc, in which case there is, by assumption, no direct causal connection between A and B.

DAGs are built up of sets of just three basic components, as shown in Figure 5.3: Indirect connections, common causes, and colliders. Let us introduce each in turn:

a. An indirect connection is given by $A \rightarrow C \rightarrow B$.
   Here, A causes B, but only via its effect on C. Variables A and B are unconditionally dependent, but if C is pre-set, the path between A and B is "blocked," and A and B are now conditionally independent, meaning that they are independent, given C. Formally, we have $A \not\perp B$ and $A \perp B \,|\, C$.

Indirect connection	*Common cause*	*Collider*
$A \rightarrow C \rightarrow B$	$A \leftarrow C \rightarrow B$	$A \rightarrow C \leftarrow B$

FIGURE 5.3
The three basic component graphs of DAGs.

b. A common cause occurs when we have $A \leftarrow C \rightarrow B$
   In this case, variable C causes both A and B. Again, A and B are unconditionally dependent, but if C is pre-set, any further variation in A is independent of that of B. Here too $A \not\perp B$ and $A \perp B \,|\, C$.

c. The third building block of DAGs is a *collider*, where $A \rightarrow C \leftarrow B$.
   This interesting situation shows that both A and B cause C, and A and B are unconditionally independent. However, for any given value of the collider C, A and B are (conditionally) dependent. Formally, we say that A and B are *d-connected* given C, and we write this as $A \perp B$ and $A \not\perp B \,|$C.

We are now ready to return to the backpack advertising campaign. The issue here is that gender appears to interact with "treatment" (i.e., with the campaign). We may depict this situation with a graph, indeed with a DAG, because the graph does not show cyclic behavior. This is shown in Figure 5.4 and illustrates all three of the basic components of a DAG: There is an indirect connection between gender and the sales of backpacks ($G \rightarrow C \rightarrow S$); gender is a common cause of the sales campaign and the outcome ($S \leftarrow G \rightarrow C$); and gender and the campaign collide to produce the result on the outcome of backpack sales ($G \rightarrow S \leftarrow C$).

Our interest is in measuring the effect of the ad campaign (C) on sales (S) in order to identify the extent to which the campaign caused a change in sales. But there are noncausal paths, including $G \rightarrow C$ that need to be "blocked" so that we can concentrate just on the $C \rightarrow S$ link, without any contamination.

Often we use regression to try to measure the strength of the effects. If we assume that the effects of the ad campaign (C) on backpack sales (S) are linear, we might control for gender and estimate

$$\mathrm{E}(S) = \beta_0 + \beta_1 G + \beta_2 C \tag{5.1}$$

where we expect that β_2 will isolate the effect of the ad campaign.

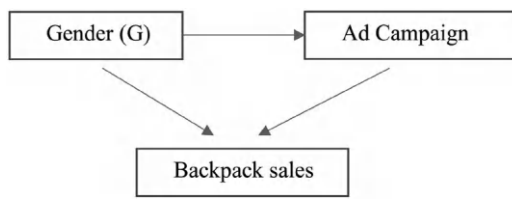


FIGURE 5.4
DAG of backpack example.

In passing, we note that this only measures the *direct* or proximate causal effect; in more complicated models, there may also be indirect effects that need to be taken into account. For instance, if our interest is in the effect of gender on backpack sales (in addition to the campaign effect on sales), there are two effects to take into account: A direct effect $G \circ\!\!-\!\!\rightarrow S$ and an indirect effect $G \rightarrow C \rightarrow S$. To measure both of these effects, we do *not* want to condition on C, so in this case we would need to estimate

$$E(Z) = \gamma_0 + \gamma_1 G. \tag{5.2}$$

Put another way, the coefficient on G in Eqn. 5.1 does not measure the *total* causal effect of G on S. In epidemiology, this is known as *The Table 2 Fallacy* (Westreich and Greenland 2013). In general, if we need to measure multiple causal effects, we need several regression models (see also Baron and Kenny 1986 on measuring direct versus indirect effects).

A broader implication of this discussion is that regression coefficients have no causal meaning without an explicitly stated causal structure (Conrady and Jouffe 2013, p. 22).

5.3 When Can DAGs Be Created?

The identification of causal effects, even with the help of a DAG, can be fragile. Following Conrady and Jouffre (p. 342), suppose that there is an unobserved variable U that influences both the treatment and the outcome. We would want to include this in our regression equation (if the effects may be assumed to be linear, or some variation thereon), but are unable to do this since we do not observe U. Now we cannot measure the causal effect of treatment on the result in an unbiased way. This may also be thought of as a case of omitted variable bias, which would only be unimportant if the omitted variables were orthogonal to the other right-hand-side variables in the regression – which (by assumption) is not true here and is rare in practice.

More generally, when can we use DAGs with a reasonable degree of hope that they might illuminate the direction of causality? It turns out that three major conditions need to hold (Spirtes et al. 2000).

The most important of these is that the set of variables be *causally sufficient*. This means that there are no omitted variables that cause any two of the variables included in the model that we have built. Taken literally, this would imply that we can never infer causality because it is impossible to be sure that we have taken every possible influence into account. No wonder Democritus despaired of finding any causal laws! In practice, the analyst needs to be able to make a persuasive case that nothing that is likely to be important has been left out of the model.

The second condition is that the joint distribution of all the variables in the set satisfies a *causal Markov condition*. That means that one only needs to condition on parents (i.e., parent variables in the DAG) and not on grandparents, uncles, aunts, or siblings to fully capture the probability distribution for any variable. There may be a chain of causes, but they all work their effects through the proximate determinants of the values of a variable.

The third assumption is that of *faithfulness*, which means that for any pair X, Y, X and Y are dependent if and only if there is an edge (i.e., a link) between X and Y.

There is no need to be paralyzed because we do not have a perfect mechanism for inferring causality. The logic of DAGs is powerful and can often be very helpful as we work to identify those variables that we can actually use in order to have effects on outcomes.

5.4 Estimating DAGs

How can we measure causal effects in practice? Clearly, we have to assume that our model is complete, in the sense that we have information on all relevant variables, and there are no relevant omitted variables. This still does not guarantee that we will be able to identify causality, but it opens the possibility.

The estimation of DAGs requires the use of specialized software and may be attempted using one of a large number of possible algorithms. There are a number of software packages that can then be employed, the main ones being TETRAD, GeNIE, bnlearn in R, and BayesiaLab.

In order to construct a DAG, all one needs is multiple observations – from a data survey, for instance – for some variables. The software then seeks to identify the relationships among these variables.

The best-established solution mechanism is the *Partial Correlation (PC) algorithm*, due to Scheines et al. (1994). It first appeared as part of the Tetrad project and is maintained by the philosophy department at Carnegie Mellon University (2025). It is included as an option in other software packages such as GeNIE and bnlearn. The PC algorithm is one of a class of constraint-based algorithms that also include the greedy search (GS) algorithm. These construct an undirected network and then identify colliders and other directed links.

The first step is to lay out all the variables (nodes or vertices) and create edges joining every node to every other node in an undirected graph. Then edges are removed from pairs of variables that are unconditionally independent or are independent after conditioning on a subset of the remaining variables. The tests for independence are based on correlation tests for continuous data and contingency table tests for categorical data.[1] Tetrad allows

one to build DAGs provided all the variables of interest are continuous, or are categorical, but does not allow one to mix the two.

The second step is to orient the remaining links between variables, starting with colliders. If X and Y are linked, and Y and Z are linked, but X and Z are independent (unconditionally, or conditionally on the variables other than Y), then we may conclude that $X \rightarrow Y \leftarrow Z$. For the cases where $X \rightarrow Y$, Y and Z are linked, but $Y \not\leftrightarrow Z$ and X and Z are not correlated (conditional on Y), we may infer that we have an indirect connection of the form $X \rightarrow Y \rightarrow Z$. Further details are given in the Tetrad manual (Carnegie Mellon University 2025); a discussion of the choice of algorithm, in the context of Tetrad and other statistical packages that create DAGs, may be found in Haughton et al. (2006). Statistical packages that construct DAGs typically allow the user to impose some a priori known links. If the entire DAG is known a priori, its estimation reduces to the estimation of a structural equation model (SEM; see Chapter 11).

There are many other possible algorithms that can be used, and they often generate somewhat different DAGs. Score-based algorithms search among all possible DAGs for the one that fits best, measured typically using the minimum description length (MDL), which is an implementation of the BIC criterion. This turns out to be a difficult optimization problem, in part because equivalences may create many local optima. More recent approaches include efforts based on sparsest permutations (Raskutti and Uhler 2014). Larranaga et al. (1996) have tried to use genetic algorithms, and Larranaga et al. (2013) have written a survey of the subsequent literature. The hunt for good algorithms remains an active research field. Bessler and Loper (2001) estimate an interesting model of the causes of economic development.

5.4.1 DAGs and Theory: Publishing Productivity

Some advocates of DAGs and other "discovery algorithms" argue that these may largely substitute for theory and that common sense, coupled with attention to the underlying statistical assumptions, suffices in much of the social sciences (Spirtes et al. 2000). In an interesting example, Spirtes and his collaborators examine a study by Rodgers and Maranto (1989) that seeks to explain the determinants of publishing productivity, as measured by the rate at which publications are cited. The data come from 162 responses to a survey of academic psychologists who obtained doctoral degrees between 1966 and 1976, and the variables include measures of "ability" (based on undergraduate performance), the quality of the graduate program (GPQ), the number of early publications, the quality of the first job (QFJ), gender, and publication rate.

There are several possible theories that seek to explain publishing productivity. A standard human capital model has the ability (ABILITY) to influence both the GPQ and QFJ, through these channels driving the publication rate (PROD), as shown in the top left panel of Figure 5.5. A version of the

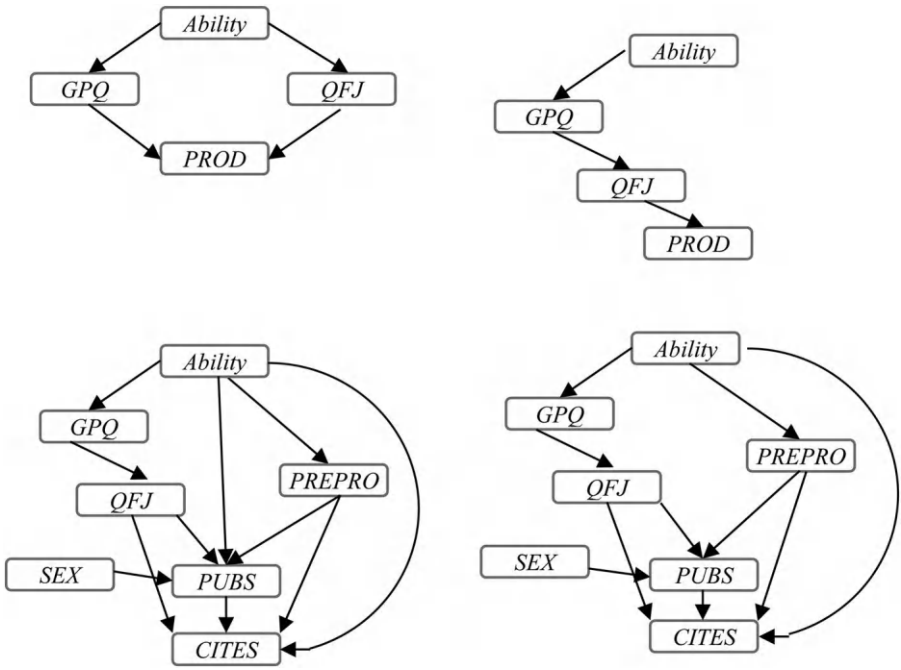


FIGURE 5.5
Graphs showing causes of research productivity in psychology. (*Note:* Top left panel shows human capital model; top right panel shows "screening" model; bottom left panel shows Rodgers and Maranto (1989) graph; bottom right panel has DAG from Spirtes et al. (2000). GPQ is quality of graduate program attended; QFJ is quality of first job; PREPRO is pre-doctoral publications; PUBS measures publication rate; and CITES measures citation rate.)

"screening" hypothesis supposes that individuals with ability need to signal their capability by attending a high-quality graduate program, and this in turn propels them into a high-quality first job, which provides the setting for scholarly productivity, as shown in the top right panel of Figure 5.5. There are a number of other possible models, and after reviewing the literature and trying several models, Rodgers and Maranto eventually settle on the pathways shown in the bottom left panel of Figure 5.5.

Using the same data, Spirtes et al. (2000) use TETRAD to create a very similar graph – shown in the bottom right panel of Figure 5.5 – using a process that they say "takes a few minutes." Their conclusion is provocative: "Any claim that social scientific theory – other than common sense – is required to find the essentials of the Rodgers and Maranto model is clearly false" (p. 102).

The challenge posed here goes well beyond the specifics of this particular case – it is odd, for instance, that in this example "ability" has no apparent direct effect on PUBS – to the issue of how to divide time and effort between theory and numbers, and even to the nature of what constitutes useful knowledge.

5.5 Case Study: Marketing Mix

A central problem for marketers is how to allocate a promotional budget across different media – whether to print ads, social media, TV spots, free samples, coupons, and so on. The standard approach is to vary the allocation of promotional spending from area to area and then use this variation to identify the contribution of the different activities to the objective (such as sales or profit). The identification is typically done using time series regression techniques, along the lines outlined in Chapter 2 and Chapter 6.

Consider the case of a pharmaceutical company that wishes to increase the number of prescriptions that are to be written for a relatively new antibiotic drug. The company's strategy is to try to market directly to physicians, using a variety of methods including personal visits ("calls"), other interventions with physicians ("contacts"), advertising in medical journals, and the provision of free drug samples. However, it is not known which of these strategies has the highest payoff, as measured by the number of new prescriptions written in the period following the promotional activities.

To illustrate how one might proceed, we construct a synthetic dataset based on the real-world case reported by Lim et al. (2008). They have monthly data on the number of new prescriptions (*new_Rx*), and 11 other variables, collected over a period of 71 months from a promotional campaign for a new antibiotic. The original data are not available, so we have created a simulated dataset based on the correlation matrix and summary statistics. The essential information on the variables, including their definitions, is shown in Table 5.1.

TABLE 5.1

Summary Data on Variables Related to a Marketing Campaign for a New Antibiotic

Variable	Mean	Std. Dev.	Description
new_Rx (nrx)	1,080,544	512,580	Number of new prescriptions dispensed
Calls (cal)	38,965	17,163	Visits by pharma reps to physicians
contacts_n (con)	54,114	21,607	Contacts with physicians for a given product
contacts_totcost (coc)	4,357,318	1,740,648	Cost related to contacts with physicians
contacts_unitcost (cpc)	87.9	10.7	Cost related to contacts, per contact
minutes_withphys (min)	182,602	72,862	Time spent with physicians
ads_n (ads)	14.1	6.6	Number of distinct ads
ads_pages (adp)	39.2	20.2	Number of ad pages
ads_totcost (jas)	217,927	112,178	Cost of ads in medical journals
samples_n (sam)	1,094,355	627,397	Samples provided to physicians
samples_units (eus)	3,556,611	1,743,432	Samples provided to physicians, weighted by sample size
samples_retailval (rvs)	8,582,260	3,820,827	Retail value of samples provided to physicians

Note: Observations cover 71 months. Each variable refers to the quantity per month.

TABLE 5.2

Regression Results for Basic Regression Model of New Prescriptions

Variable	Coefficient	p-value	Coefficient	p-value
Calls (cal)	12.4	0.47		
contacts_n (con)	–31.3	0.04		
contacts_totcost (coc)	0.19	0.12		
Contacts_unitcost*(cpc)	–11.3	0.02		
minutes_withphys (min)	1.23	0.46		
ads_n* (ads)	34.0	0.02	33.0	0.00
ads_pages* (adp)	2.71	0.61		
ads_totcost (jas)	–1.56	0.01	–1.21	0.01
samples_n (sam)	0.48	0.05	0.23	0.01
samples_units (eus)	–0.14	0.08		
samples_retailval (rvs)	0.08	0.00	0.07	0.00
R squared	0.93		0.92	

Note: * coefficients are per thousand units. Dependent variable is new_Rx (number of new prescriptions written). Descriptions of variables are shown in Table 5.1.

A naïve approach to measuring the effect of the available variables on the outcomes would be to estimate a linear regression, where new_Rx is the dependent variable and the other 11 variables are the regressors (note that the data were not differenced or otherwise transformed). This substitutes data for thought, so it should not be surprising that the results – shown in Table 5.2 – are hard to interpret. For example, higher spending on journal ads is associated with fewer new prescriptions, other things being equal, which seems odd. The high degree of multicollinearity among the right-hand-side variables contributes to the lack of statistical significance: Four of the variables are not statistically significant at the 10% level (i.e., they have p-values greater than 0.1).

At this point, some researchers might use a stepwise or lasso procedure to trim the variables (features) in the model. The results of a forward stepwise regression, which included only variables that are significant at the 10% level or better, are shown in the right-hand columns of Table 5.2. The apparently good news is that the fit is almost as high as for the larger equation, while the model is more parsimonious. The problem here is that the use of stepwise regression may be incorrect (see Chapter 2), and in the current example does not eliminate the surprising negative correlation between spending on journal advertising and the number of new prescriptions, which undermines the credibility of the model.

An alternative, and tempting, route is to reduce the number of variables to just a few factors (see Chapter 11). In Table 5.3, we show the varimax-rotated factor loadings that apply to the 11 variables; the underlying variables fairly naturally fall into two groups, a first factor that reflects the influence of

TABLE 5.3

Factor Loadings and Coefficients for Marketing Mix Model

	Factor Loading		Coefficients	
	Factor 1	**Factor 2**	**Factor 1**	**Factor 2**
Variable				
new_Rx (nrx)				
Calls (cal)	0.940		0.149	–0.105
contacts_n (con)	0.945		0.151	–0.109
contacts_totcost (coc)	0.918		0.144	–0.094
contacts_unitcost* (cpc)	–0.769		–0.135	0.162
minutes_withphys (min)	0.941		0.151	–0.117
ads_n* (ads)		0.910	–0.014	0.375
ads_pages* (adp)		0.936	–0.028	0.406
ads_totcost (jas)		0.954	–0.051	0.446
samples_n (sam)	0.886	0.402	0.129	–0.034
samples_units (eus)	0.845	0.416	0.120	–0.015
samples_retailval (rvs)	0.887		0.132	–0.051

Note: Only factor loadings greater than 0.4 are shown here. These are rotated factor loadings (using the varimax method).

contacts and samples, and a second factor that might be labeled "journal advertising." The coefficients applicable to the variables, which then produce the factors, are shown on the right-hand side of Table 5.3. We can now regress the number of new prescriptions on these two factors, as shown in Table 5.4. The overall fit is still good, but we have lost some information along the way and have certainly lost sight of the causal connections.

A more defensible approach is to try to build a DAG, and the result of this effort is shown in Figure 5.6. This particular graph was built using a modified PC algorithm called PC Pattern, using the Tetrad IV package. There is a node for each of the 13 variables, including the number of new prescriptions, which is the variable of most interest to us. The values attached to each node are simply the mean values and are the same as in Table 5.1.

TABLE 5.4

Regression Estimates for Two-Factor Marketing Mix Model

	Coefficients	**p-Values**
Variable		
Factor 1 ("contacts and samples")	0.916	0.00
Factor 2 ("journal advertising")	0.200	0.00
R squared	0.88	

Note: Coefficients are standardized.

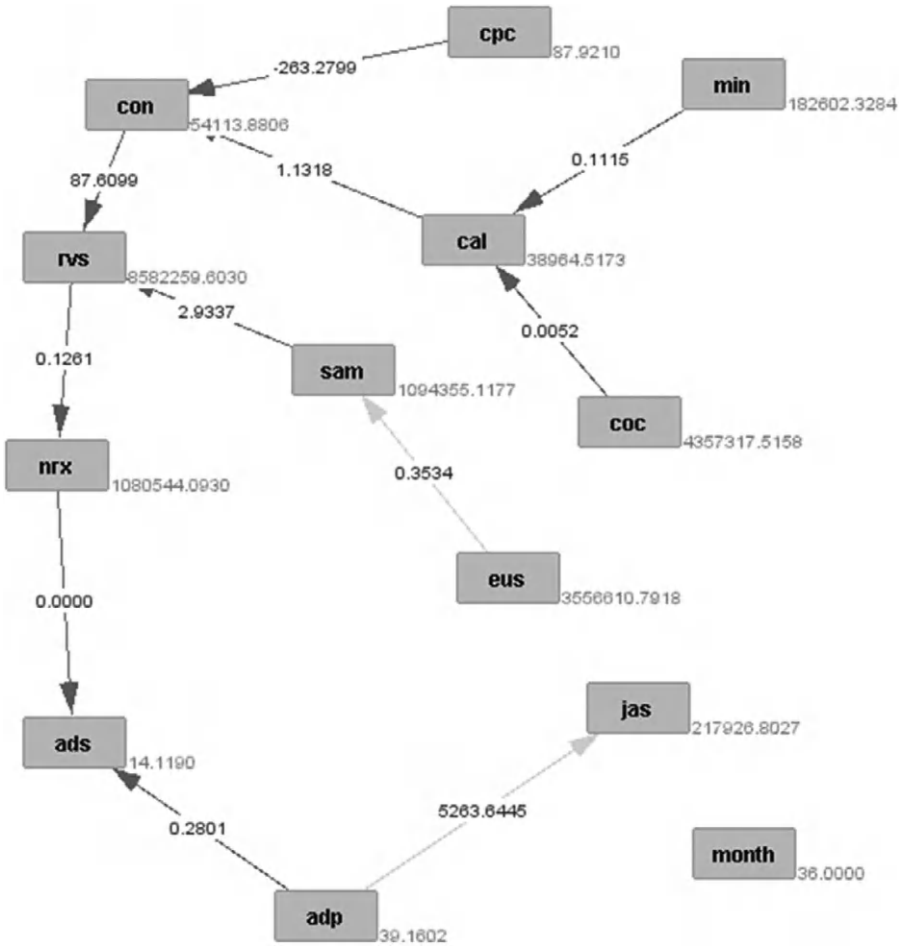


FIGURE 5.6
DAG for marketing mix model.

The DAG has a series of directed links ("edges") that indicate the direction of causality (as inferred from the data). The numbers on each of the edges measure the effect of each parent node on its child and are the coefficients from linear regressions of the value of the child on the values of its parents. So, for instance, when the number of contacts (con) rises by one, the retail value of prescriptions (rvs) goes up by $87.6.

In this example, only one variable has a direct effect on the number of new prescriptions written, and that is the retail value of samples provided to physicians. Seven variables have an indirect effect in that they first influence the value of samples and, through this indirect route, change the number of new prescriptions written.

Indeed only four "primary" variables influence the number of new prescriptions, and these are the ancestors that do not themselves have any parents. In our example, these are cpc, min, coc, and eus. A regression of nrx on these variables would measure the impact of these causes, and one could then apply linear programming to optimize the way in which a limited market budget should be spent, given the goal of selling more of the drug.

The variables that measure the extent of advertising in medical journals are shown in the model as being the *result* of additional prescriptions, not a cause. It is possible that the more was prescribed, the more advertising in journals was expanded.

Our results are specific to the situation being studied here and do not have "external validity" in the sense of being immediately applicable to other contexts. Every situation calls for its own careful analysis. But as studies of this nature are repeated, practitioners will gradually build up a better sense of where the lines of causality run, and hence what does and does not work when one is trying to market a new product.

5.6 Estimating DAGs: Practical Considerations

In estimating DAGs, one first has to choose which software package to use. The main choices are one of the free packages (Tetrad, GeNIe, bnlearn in R) or a commercial product such as BayesiaLab. All offer a variety of solution algorithms, but they differ substantially in the quality of the graphics and in the user interfaces and options available.

The choice of appropriate solution algorithm can be daunting. Figure 5.7, from Tetrad, shows a box with data in the center of a wheel, with spokes pointing to 17 distinct possible algorithms, five of which are variants of the PC algorithm discussed above.

Figure 5.8 shows the result of applying Tetrad's PC algorithm to data related to churn. The data come from the customer database of a wireless provider, which was interested in determining which variables influence churn (measured as a binary variable that is set to one if the customer switches to another provider). Many of the variables are perceptual, while others measure sociodemographic characteristics such as income or household size, or variables related to the quality of service. Rather surprisingly, the likelihood of switching to another provider ("lswitchinglik") is only influenced by the customer's satisfaction with the service ("lsatisfaction"), and there are no clear influences on what drives satisfaction. This does not point the way toward clear solutions aimed at reducing churn but may be a useful reality check nonetheless.

Figure 5.9 shows the result of the same exercise, again using the PC algorithm, but using GeNIe rather than Tetrad. The result is substantively the same. One link could not be directed by the algorithm (between the number

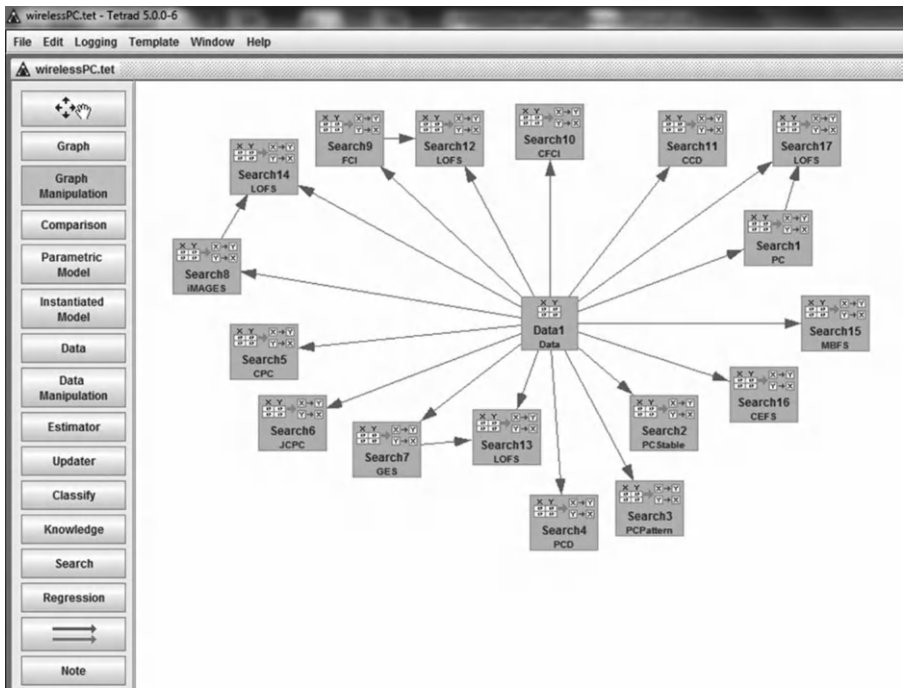


FIGURE 5.7
The spectrum of DAG solution algorithms in Tetrad.

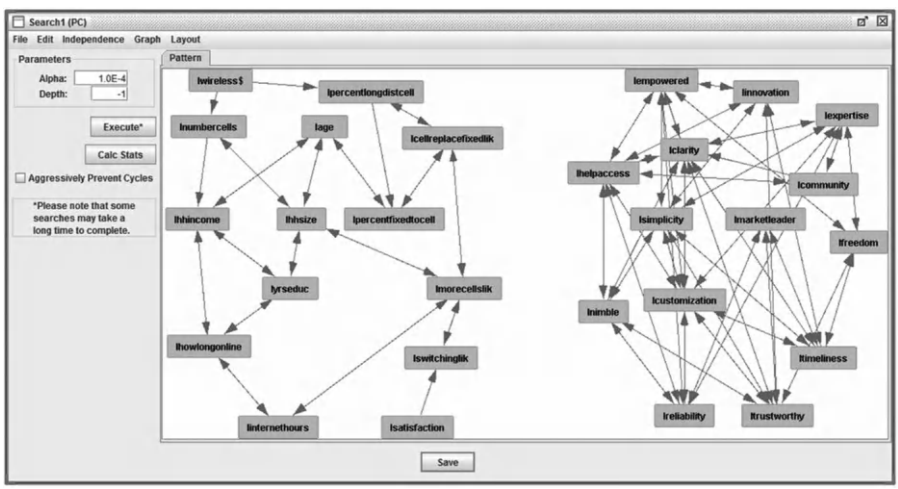


FIGURE 5.8
Tetrad PC algorithm applied to churn data.

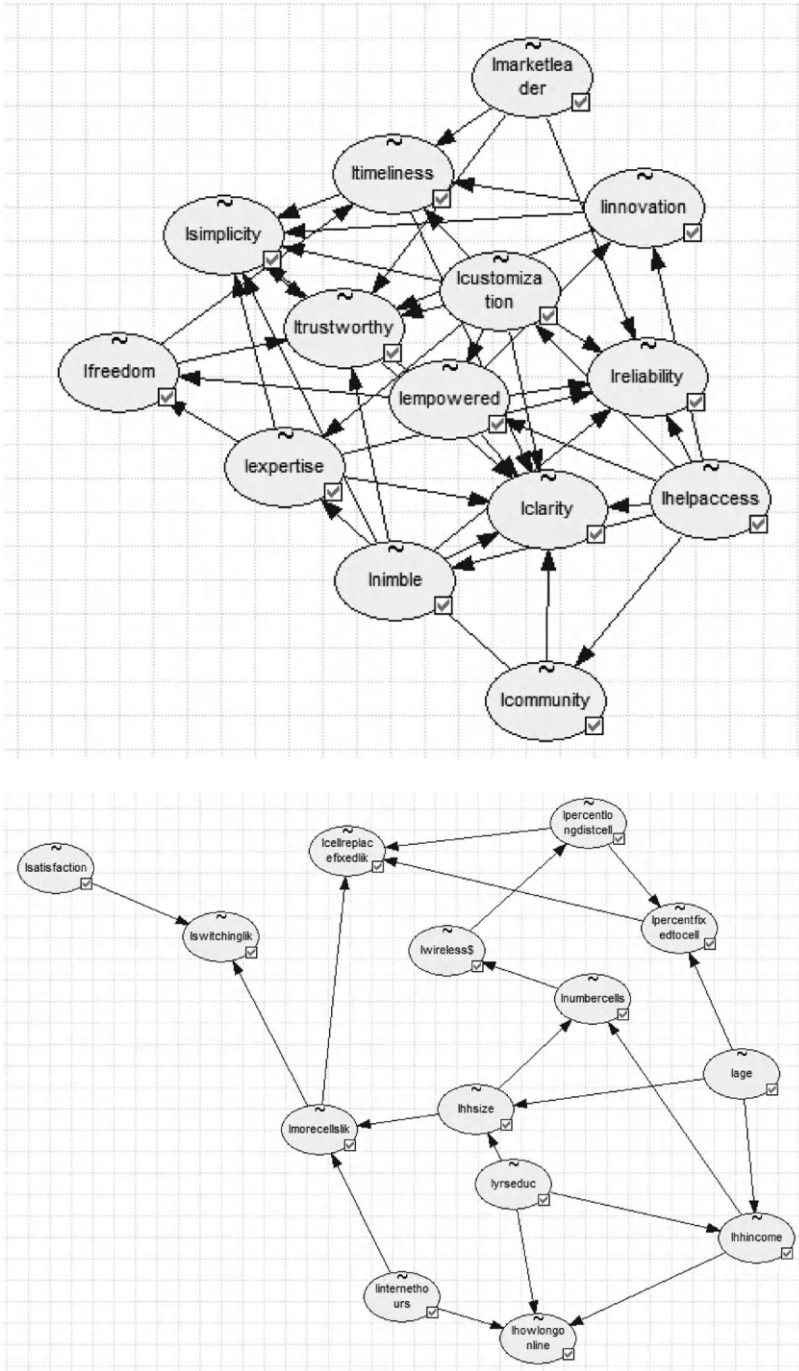


FIGURE 5.9
GeNIe PC algorithm applied to churn data.

of cell phones in the household and spending on wireless service) and had to be determined manually. One nice feature of GeNIe is that it is straightforward for the analyst to impose a direction of causality. For instance, age might influence spending, but spending cannot change your age (which is clearly exogenous), so it makes sense to specify this before the software is asked to search for causal links. This and some other sensible restrictions have been imposed prior to constructing the DAG shown in Figure 5.9. This graph is somewhat more helpful than the Tetrad version: Not only does satisfaction affect the amount of churn, but so does the likelihood of the household acquiring more cell phones ("lmorecellslik"). Here too the perceptual variables appear to have no effect on churn.

The Blearn package in R (Scutari 2010) offers a very wide range of algorithms, including the grow-shrink (GS) constraint-based method, hill-climbing approaches, and hybrids of the two.[2]

The BayesiaLab approach to constructing DAGs begins by discretizing all variables before applying any algorithms. It then relies on score-based algorithms published by its authors (Conrady and Jouffe 2013). The package is marketed to business analysts (and academics who talk to businesspeople). One particular strength is that it has an implementation of Pearl's "do" operator, which allows one to observe what happens when a variable is not only observed but also changed (by the business, for instance). This allows one to see the effects of an intervention, such as spending more money on a particular marketing mode. Figure 5.10 shows the BayesiaLab tabu algorithm,

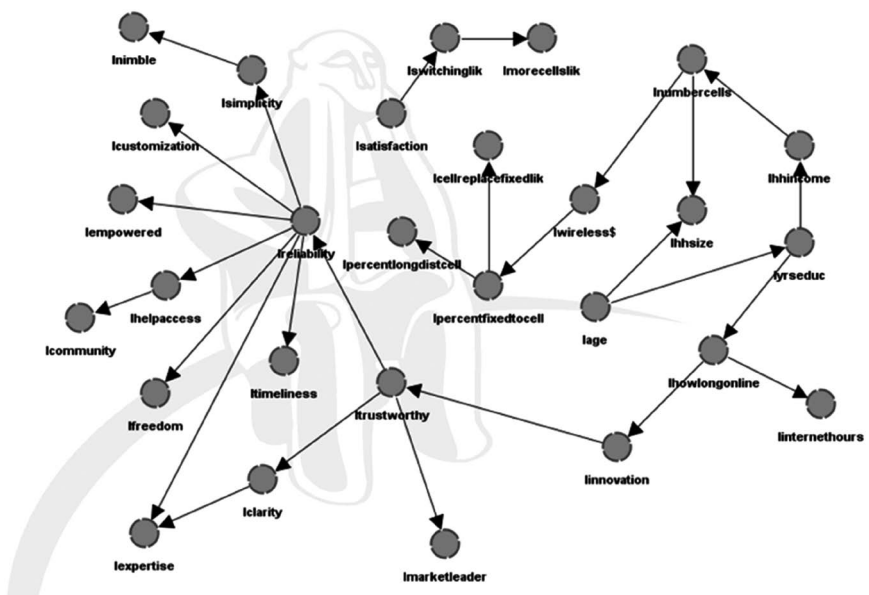


FIGURE 5.10
The BayesiaLab tabu order algorithm applied to the churn data.

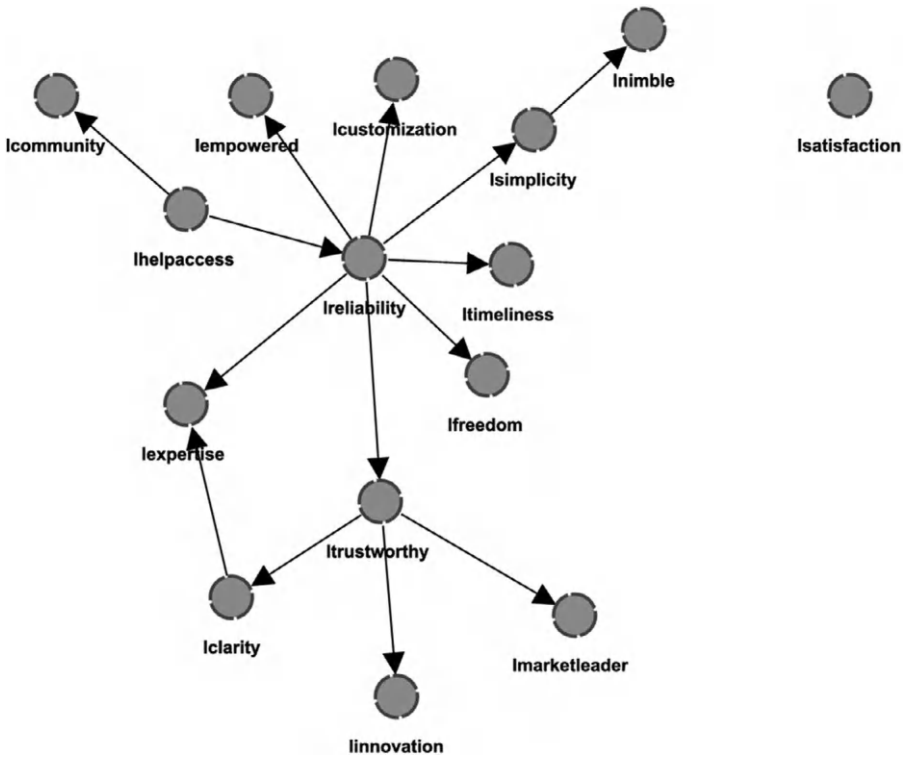


FIGURE 5.11
The BayesiaLab EQ algorithm applied to the perception variables (on features of a wireless service) in the churn data.

applied to the churn problem, while Figure 5.11 shows the EQ algorithm. As with the other implementations of DAGs to this problem, the perception variables are linked with one another and with a few of the demographic variables, but satisfaction is still the only variable that appears to influence the likelihood of churning.

5.7 Conclusion

The use of DAGs is still relatively rare among business analytics or healthcare analytics professionals; perhaps they are less well known, and perhaps because standard regression techniques can be applied so quickly and easily. This is unfortunate, because DAGs are relatively straightforward to construct and interpret, and as long as the number of variables is not too large, can be very useful in helping to identify causal paths.

As with all techniques, there is some art involved in getting DAGs to "speak," and they cannot be used in a purely mechanical way. But at a minimum they can offer a check against received wisdom: In the marketing example presented in Section 5.4, the DAG-based analysis helped make it clearer that print advertising likely followed, rather than led, the growth rate of new prescriptions. This may call for the marketing department to change the way it spends its budget.

Not all business problems are amenable to DAGs, and nor are DAGs always needed. In the next four chapters, we examine the role of uplift analytics, which explores the ways in which individual data on sales (experimental or observational) may be used to determine how best to target spending on such items as advertising or how best to change the type of product that the firm offers.

Notes

1. This typically assumes that the data, possibly transformed, are multivariate normal.
2. The Blearn package includes the following algorithms:

   C*onstraint-based structure learning algorithms:* Grow-Shrink (GS); Incremental Association Markov Blanket (IAMB); Fast Incremental Association (Fast-IAMB); Interleaved Incremental Association (Inter-IAMB).

   Score-based structure learning algorithms: Hill Climbing (HC); Tabu Search (Tabu).

   Hybrid structure learning algorithms: Max-Min Hill Climbing (MMHC); General 2-Phase Restricted Maximization (RSMAX2).

   Local discovery algorithms: Chow-Liu; ARACNE; Max-Min Parents & Children (MMPC); Semi-Interleaved Hiton-PC.

   Bayesian network classifiers: Naive Bayes; Tree-Augmented naive Bayes (TAN).

References

Baron, R. M., and D. A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research – Conceptual, Strategic, and Statistical Considerations". *Journal of Personality and Social Psychology*, 51(6): 1173–1182.

Bessler, D. A., and N. Loper. 2001. "Economic Development: Evidence from Directed Acyclic Graphs". *The Manchester School*, 69: 457–476.

Bryant, Henry, David Bessler, and Michael Haigh. 2009. "Disproving Causal Relationships Using Observational Data". *Oxford Bulletin of Economics and Statistics*, 71(3): 357–374.

Conrady, Stefan, and Lionel Jouffe. 2013. *Introduction to Bayesian Networks & BayesiaLab*. Laval, France: Bayesia S.A.S.

Geiger, D., T. Verma, and J. Pearl. 1990. "Identifying Independencies in Bayesian Networks". *Networks*, 20: 507–534.

Haughton, D., A. Kamis, and P. Scholten. 2006. "A Review of Three Directed Acyclic Graphs Software Packages: MIM, Tetrad, and WinMine". *The American Statistician*, 60(3): 272–286.

Larranaga, P., C. M. H. Kuijpers, R. H. Murga, and Y. Yurramendi. 1996. "Learning Bayesian Network Sructures by Searching for the Best Ordering with Genetic Algorithm". *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 26(4): 487–493.

Larrañaga, Pedro, Hossein Karshenas, Concha Bielza, and Roberto Santana. 2013. "A Review on Evolutionary Algorithms in Bayesian Network Learning and Inference Tasks". *Information Sciences*, 223: 109–125.

Lim, Chee, Wooi Lim, and Toru Kirikoshi. 2008. "Understanding the Effects of Pharmaceutical Promotion: A Neural Network Approach Guided by Genetic Algorithm-Partial Least Squares". *Health Care Management Science*, 11(4): 359–372.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press.

Pearl, Judea. 2009. "Causal Inference in Statistics: An Overview". *Statistics Surveys*, 3: 96–146.

Pearl, Judea, and Dana Mackenzie. 2020. *The Book of Why*. New York NY: Basic Books.

Raskutti, G., and C. Uhler 2014. "Learning Directed Acyclic Graphs Based on Sparsest Permutations". https://doi.org/10.48550/arXiv.1307.0366

Rodgers, Robert, and Cheryl Maranto. 1989. "Causal Models of Publishing Productivity in Psychology". *Journal of Applied Psychology*, 74(4): 636–649.

Scheines, R., P. Spirtes, C. Glymour, and C. Meek. 1994. *TETRAD II: Tools for Discovery*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Scutari, Marco. 2010. "Learning Bayesian Networks with the Blearn R Package". *Journal of Statistical Software*, VV(II): 1–22.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*, 2nd edition. Cambridge, MA: MIT Press.

Carnegie Mellon Univeersity. 2025. *Tetrad*. http://www.phil.cmu.edu/projects/tetrad/tetrad4.html.

Westreich, Daniel, and Sander Greenland. 2013. "The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients". *American Journal of Epidemiology*, 177(4): 292–298.

6

Uplift Analytics I: Mining for the Truly Responsive Customers and Prospects

6.1 Introduction to Target Marketing

The previous two chapters discussed causal measurement at the population or group level. This chapter introduces uplift modeling, also known as true-lift or net lift modeling, a key technique for target marketing and other applications of personalization.

A long time ago, when one of the authors introduced uplift/true-lift modeling at a data mining conference, a prominent data mining expert who had written multiple popular technical books was in the audience, and he openly rejected the idea and believed the traditional modeling approach would work just fine. About a decade later, the same expert spoke at another conference. To our surprise, this expert was actually "converted." He not only mentioned uplift modeling but also described it as an important emerging field and acknowledged the contribution of our work. In fact, over the past several years, many academic scholars and industrial practitioners have increasingly been involved in this emerging field. To our knowledge, however, very few books been written to date on the technical aspects of uplift modeling. Thus, the first purpose of this and the subsequent chapter is to introduce the technical details in a readable fashion. Siegel (2013b) provides a general nontechnical description of this approach, but not the technical aspects.

Uplift modeling is essentially the same as measuring the causal treatment effect at the individual (customer) level, relating such effect to individual attributes such as age and income, and then applying estimated individual-level effects to other individuals for target marketing. Before discussing uplift, it is important to define target marketing. Business analytics (or data mining or data science) has been widely applied to marketing since the 1980s. Many corporations collect large amounts of customer data in order to understand their needs, predict their future behavior, and optimize future contacts. Such target marketing methods are part of a more general concept in marketing called Customer Relationship Management (CRM), where customer contact data are collected and analyzed to optimize future contacts in order

DOI: 10.1201/9780429172588-6

to improve acquisition, development, and retention. CRM, and its predecessor Database Marketing, are key areas where analytic methods are utilized to understand customers so that companies can provide the best offers and messages to the right customers through the right channels (see, for instance, Jackson and Wang 1996, Roberts and Berger 1999, and Maex and Brown 2012).

Section 6.2 describes the traditional predictive modeling approach for target marketing, and you may skip this section if you are familiar with it. Section 6.3 describes the concept of uplift/true-lift modeling in detail. If you already have some knowledge of the uplift concept and understand why it is important, you may jump to Section 6.5 directly.

6.2 Traditional Predictive Modeling for Target Marketing

Consider email marketing where the goal is to get more individuals (customers or prospects) to respond (e.g., click or buy). If there is a historical marketing campaign data set that has both a treatment group (where the individuals received a marketing message in an email) and a control group* (to which no marketing email was mailed) and assuming the treatment and control groups were randomly split so they look alike in terms of their individual characteristics, i.e., it is a randomized experiment, could we learn about who are likely to respond based on historical data so as to benefit future marketing efforts? To maximize return on investment in marketing campaigns, predictive modeling through statistical or machine learning techniques has been routinely applied to uncover the characteristics of customers or prospects who are likely to respond. The model, which is based on data from a previous campaign, can then be used to identify likely responders to similar future campaigns. This improves efficiency by increasing the proportion of responders within the contacted group. Such an approach has been used very widely since the late 1980s. In addition to typical customer analytics in for-profit organizations, it can be applied to non-profit organizations in analyzing donor data to identify appropriate donors to contact (see Examples 6.1 and 6.2 in a later section of this chapter).

Note that our discussion focuses on predictive modeling or supervised learning methods for marketing. Customer segmentation through cluster analysis is another widely used method; it is an unsupervised learning technique for customer grouping, typically for more strategic objectives rather than maximizing response rates directly, and is beyond the scope of this book (see Chapter 2 for a short discussion of clustering).

* The control group may also receive a "business-as-usual" (BAU) email as opposed to a newer or more creative design for the treatment group email.

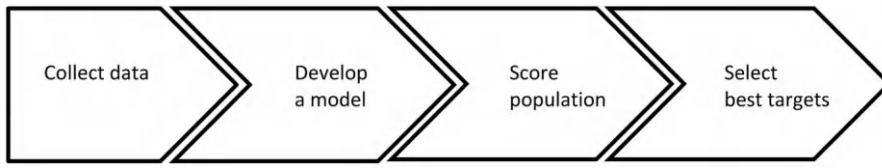


FIGURE 6.1
Response modeling process.

A typical application of predictive modeling is developing response models to identify likely campaign responders. As summarized in Figure 6.1, a previous campaign provides data on the "dependent" or "target" variable (responded or not), which is merged with individual characteristics, including behavioral and demographic variables, to form an analyzable data set. A response model is then developed to predict the response rate given the individual characteristics. The model is then used to score the population to predict response rates for all individuals. Finally, the best list of individuals will be targeted in the next campaign in order to maximize effectiveness and minimize expense.

More generally, response modeling can be applied in the following key initiatives in marketing (see Peppers and Rogers 1997 and Peppers et al. 1999):

1. **Acquisition:** Which prospects are most likely to buy your product and become a customer? Acquiring customers may be the most exciting but is often difficult, especially for certain industries (e.g., getting someone to buy a movie for online viewing is easier than selling an SUV). Acquisition is sometimes decomposed into two steps: (a) Converting a pure prospect to a lead and (b) converting a lead to a customer.
2. **Development:** Which customers are most likely to purchase additional products (cross-selling) or to increase monetary value (up-selling)?
3. **Retention:** Which customers are most retainable? This can be relationship or value retention: Relationship retention is to minimize complete attrition (e.g., closing an account), while value retention is to minimize the chance of buying less or reducing value.

Standard response modeling typically addresses a binary question: Among the treated population, who responded and who did not? Commonly used techniques include the most classical statistical methods, such as logistic regression and discriminant analysis, and more recent machine-learning and advanced-nonlinear techniques, such as decision trees (CART, CHAID, C4.5, TreeNet/MART), MARS, neural networks, and support vector machines (many of these were reviewed in Chapter 2). The objective of this approach

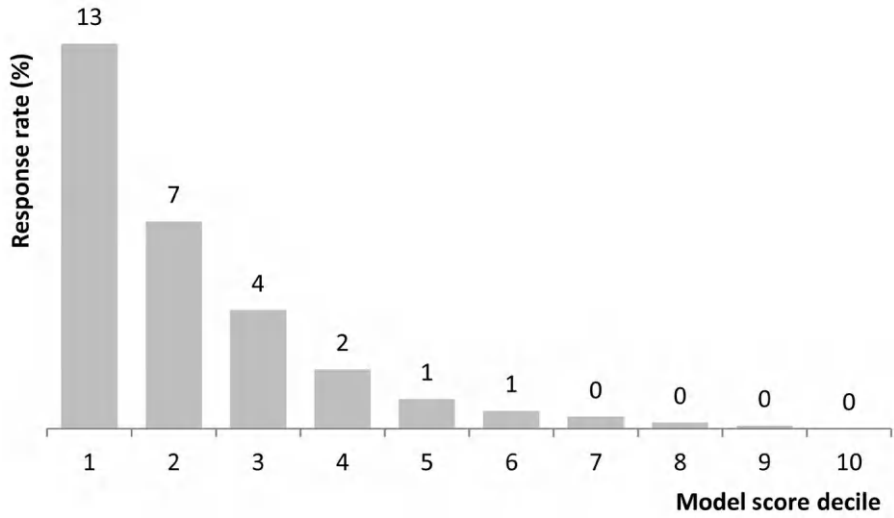


FIGURE 6.2
A successful standard response model concentrates treatment responders in the top model deciles. (Illustrative example created by the authors.)

is to concentrate treatment responders in the top deciles, as ranked by model scores. Figure 6.2 illustrates how a successful model would expect to have higher response rates in the top deciles, so the top decile has the highest score values. More specifically, the standard approach aims to differentiate between the responders and nonresponders in the treated group while ignoring behavior in the control group. These models are then used to identify future likely responders in order to improve the success of future marketing campaigns. Typically, the first few deciles, say top 2 or 3 deciles, are used for targeting. The reason is that the marketing budget is limited and marketers would want the highest probability of success from its targeting effort. Using Figure 6.2 as an example, the average response rate is 2.9%, and if marketers want to focus on the top 3 deciles with a cumulative response rate (average of deciles 1–3) of 8%, the modeling or data mining "lift over random" would be 8%/2.9% = 2.8, i.e., the top 3 deciles have a response rate 2.8 times of that of random, which is typically considered pretty good. (Another way is to target those deciles that have response rates higher than random, and in Figure 6.2, they happen to be the first 3 deciles as well.)

The standard approach is flawed. Scientific measurement practice in marketing typically measures "lift over control" by comparing the results to a control group (where treatment is NOT given, e.g., no-mail control group in the case of direct mails) to causally address whether a campaign is successful. In contrast, the standard response model does not address how the success or failure of a treatment is measured, as the objective is to maximize "lift over random" rather than "lift over control." Response measurement and

response modeling should instead share the same measure of success, so the most appropriate modeling strategy should also maximize the treatment lift over control. In this chapter and the next three chapters of this book where individual targeting using uplift analytics is the focus, "lift" is defined as lift over control, i.e., treatment response rate minus control response rate, and "uplift" refers to a set of methods to "up" or maximize lift over control.

6.3 Uplift Modeling: What and Why?

In this section, we describe the concept of uplift modeling in detail. In case you already have some familiarity with the uplift concept and would love to get to the math right away, you may skip to Section 6.5.

Once upon a time – and this is based on a real story – a team of sophisticated PhD-level modelers in a large organization developed marketing response models using the traditional approach (including advanced methods such as decision trees). As in the approach described above in Section 6.2, data from previous marketing campaigns were used to uncover the characteristics of individuals who were likely to respond. The models developed were then applied to new campaigns designed by the marketing group, which initiated and sponsored the whole effort. This marketing team had highly experienced marketing MBAs who understood customer needs through past experience, market research, and behavioral analysis.

In addition to the modeling team and the marketing team, there was also a separate measurement team of quantitative MBA-type business analysts who were responsible for measuring the success of such a model against a random control group. Specifically, the model-based selected targets (from the top two model deciles) were randomly split into treatment and control groups. Then the success of a modeled campaign was assessed by the difference in response rate between the modeled treatment and the modeled control groups. Table 6.1 provides an (imagined) example. In this table, "Model" is the group of customers identified as "good" (e.g., likely responders) by the

TABLE 6.1

Campaign Measurement of Model Effectiveness: Example

	Treatment (E.g., Mail)	**Control (E.g., No Mail)**	**Increment (Treatment Minus Control)**
		% Response Rate	
Model	1.0	1.0	**0.0**
Random	0.3	0.3	**0.0**
Model minus random	**0.7**	**0.7**	0.0

model; in our example, these customers are in the top two deciles, as predicted by the model. "Random" refers to customers who may be targeted randomly. "Treatment" is the group that received a treatment (an offer) through direct mail, email, web, or a live channel, and "Control" refers to groups to which no treatment is offered.

Do the numbers in Table 6.1 indicate that this campaign was successful? The measurement group said "no," as it focused on lift over control (which was 0%); yet the modeling group said the model no doubt had a great lift (over random), which was 1% – 0.3% = 0.7%. Moreover, the modeling group believed it was the marketing group's responsibility to generate lift over control, but on this, the marketing group did not agree, as they expected the modeling group would produce a model to help generate value. The measurement group suggested that perhaps the models were not designed in the best way, but the modeling group cited existing literature and industry "best practice" at the time to support their work. The marketing group, being the sponsor of the campaign and all the related work, was frustrated but was not sure what to do. The three teams were pointing fingers at each other, and the debate went on endlessly.

Rather than joining the debate, the newly appointed modeling manager believed it was the model that was not designed correctly or appropriately. He realized that the approach taken was not meant to maximize lift over control, but rather only lift over random, so he proposed a new way of modeling, which is now called uplift or true-lift modeling (from Radcliffe and Surry 1999 and Lo 2002).

Why was "lift over control" the right measurement instead of "lift over random"? We explained in Chapter 3 of this book that the gold standard of causal measurement is through randomization. In other words, a random split between treatment and control can ensure homogeneity of individual characteristics between the treatment and control groups, so any observed difference in measurement between the two groups can be attributed to the marketing campaign itself. When the campaign uses a model, the "lift over control" measurement truly assesses whether the model works in generating responses in campaign selection. The modeled control group represents the counterfactual that shows what would happen to the modeled treatment group if its members had not received the treatment.

Traditional response models are designed to identify likely responders and often target customers who will take the desirable action anyway, whether they receive the marketing interaction or not. Lo (2002) proposed the true-lift method to find the customers whose decisions will be *positively influenced* by marketing interaction. The methodology is easy to implement and can be used in conjunction with commonly used supervised learning algorithms. It is also known as uplift modeling by Radcliffe and Surry (1999) and net lift by Lund (2012) and Kubiak (2012). The early articles sparked interest in this new field, leading to numerous discussion sessions in recent data mining and analytics conferences. The data mining usage report by Rexer (2012) mentions

that uplift/true-lift modeling is more popular among corporate practitioners and consultants than in academic or government agencies. Rexer (2012) and Rexer et al (2016) also report the growing usage of uplift modeling, followed by expanded academic and industry focus on this topic in conferences and journals. Since the term "uplift" appears to be more appealing and popular, we will call this approach uplift modeling in this book, which includes all techniques that are designed to generate lift over control.

Uplift modeling applies to experiments in which a randomly selected control group is withheld, as described in Lo (2002, 2009), Radcliffe (2007a), and Radcliffe and Surry (2011). In marketing, this type of experiment is often used for direct mail, email, and outbound phone "treatments" in which a target prospect is invited to make a product inquiry or purchase (the "response"). Control prospects are randomly selected from the same population to receive either no invitation or a typical business-as-usual "champion" invitation (for champion-challenger comparisons). In medical experiments, a parallel approach would be giving the treatment group a trial medication in order to measure its effectiveness relative to a placebo or standard care in the control group (see Cai et al. 2011). Siegel (2011) and Chapter 7 of Siegel (2013b) provide further nontechnical introductions to this field. The 2012 presidential campaign also utilized similar techniques for targeting campaign contributors and voters, significantly affecting the election results, according to Scherer (2012), Samuelson (2013), and Siegel (2013a). The presentation by Potter (2013) also describes their uplift approach (aka persuasion modeling) in the presidential campaign.

The business benefits of true-lift modeling are rooted in its ability to identify four types of targets and the most efficient use of marketing budgets for each type. This is shown schematically in Figure 6.3, where we may identify:

- **Sure Things,** who will purchase the product whether they are contacted or not. The marketing budget applied to these contacts is wasted because it has no effect on their action.

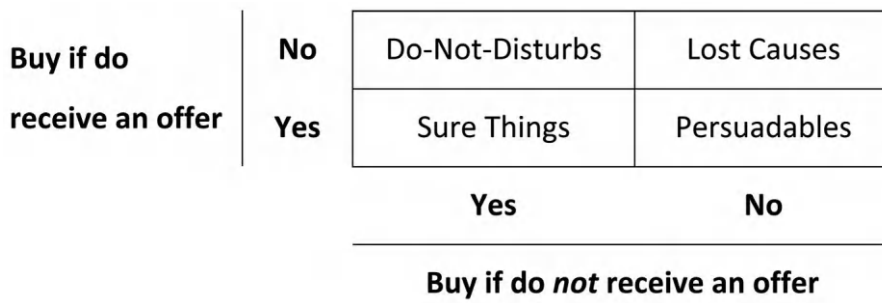


FIGURE 6.3
Four types of targets. (Radcliffe 2007b, Siegel 2011, and Kane et al. 2014.)

- **Lost Causes,** who will not purchase the product, whether they are contacted or not. The marketing budget applied to these contacts is wasted because it has no effect on their action.
- **Do-Not-Disturbs,** who have a negative reaction to the marketing contact. Perversely, they will purchase if not contacted but will not purchase if contacted. The marketing budget applied to these contacts is not only wasted but it also has a negative impact on the results. For example, populations targeted for retention efforts by standard response models could result in withdrawing from current products or services.
- **Persuadables,** who have a positive reaction to the marketing contact. They purchase only if contacted (or sometimes they purchase more or earlier, but only if contacted). ***They are the only efficient targets.*** Modeling techniques discussed in this chapter and Chapter 8 will focus on finding the likely persuadables**.**

The purpose of campaign-specific response modeling is, of course, to identify the customers (or prospects) who are most likely to respond, where a response can be accepting an offer or increasing revenue or sales.

After a model has been used for a campaign, analysts may measure the model's effectiveness. Table 6.2, which is a generalized version of Table 6.1, shows a typical example: A, B, C, and D are the observed response rates of a campaign. Alternatively, these values could represent average sales or revenue generated, and so they can be continuous or proportions. If A is statistically significantly larger than C, it means that the model is in the right direction of targeting customers who are likely to respond, but it is not sufficient to determine whether the model is effective. For that, we also need to ask what would have happened if the customers had not received the treatment, and that is the role of the control group.

Consider the difference between "Treatment" and "Control." Assuming the effects are statistically significant, if A > C and B > D, but A = B, it means that while the model is able to pick the likely responders, it does not add any "value" to the campaign. That is because A = B implies that the customers in the model-treatment cell (who received a mail or other solicitation)

TABLE 6.2

Campaign Measurement of Model Effectiveness: Generalization

	Treatment (E.g., Mail)	**Control (E.g., No Mail)**	**Increment (Treatment Minus Control)**
		Response Rate	
Model	A	B	**A – B**
Random	C	D	**C – D**
Model minus random	**A – C**	**B – D**	**(A – B) – (C – D)**

responded in the same way as those in the model-control cell (who did not receive the treatment). To claim that the model "works" (i.e., "Model" is better than "Random"), we not only require that A – B > 0 but also A – B > C – D. As a result, the appropriate measure of the campaign gain (in response rate, revenue, or sales) due to the treatment is the double difference [(A – B) – (C – D)]. Hence, the quantitative business objective of a model, measured by its effectiveness for campaigns, is to maximize [(A – B) – (C – D)] rather than the (A – C) that is used in the traditional approach.

The *ideal* response model pushes treatment responders into the top deciles and control responders into the bottom deciles, as illustrated in Figure 6.4, where the dark bars represent the treatment responders and the light bards represent the control responders. A standard treatment-only response model may often concentrate the control responders in each decile at the same rate as treatment responders, and the model is not expected to generate any lift in the top deciles. This happens most often when the lift of the overall campaign is very small compared to the baseline response. We define the true-lift "Signal-to-Noise" (S/N) ratio as lift divided by control response rate, which measures the incremental contribution from the treatment (signal) relative to the control response rate (noise). For example, if the control response is 5.00% and the treatment response is 5.01%, S/N = 0.01/5 = 0.002, then most of the treatment responders are very similar to baseline (natural) responders. As a result, there is very little lift signal from the data, so baseline response will dominate the standard model.

The difficulty with lift as a model's objective is that observations cannot be classified as "lift" responses versus baseline responses versus nonresponses. Any treatment responder could be either a baseline responder (i.e., someone

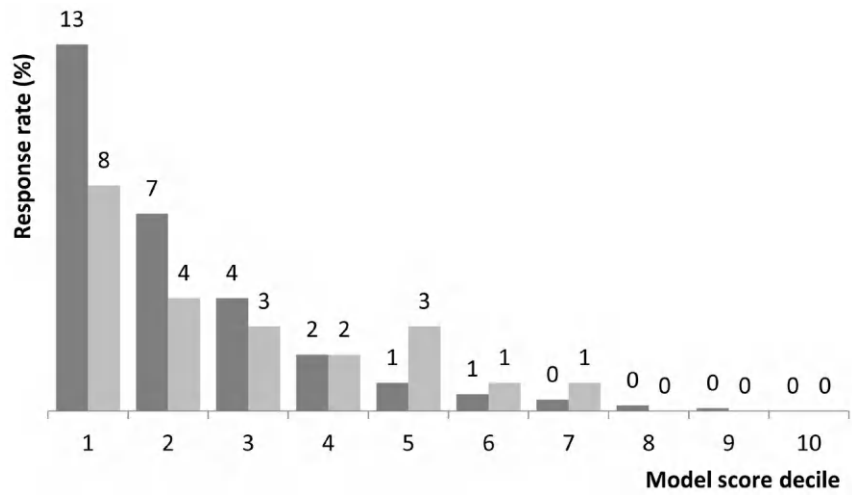


FIGURE 6.4
Illustration of an *ideal* uplift model.

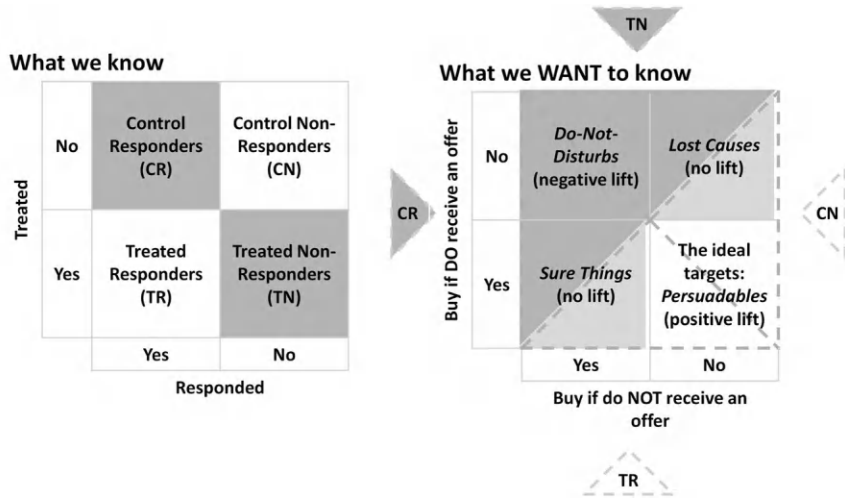


FIGURE 6.5
A conceptual 2 × 2 table.

who would have responded anyway) or a lift responder (i.e., someone who responded purely because of the treatment). The objective is to take the data we know, such as who responded and who did not, plus who were treated and who were not, to get to the information we want to know: Who are the lift responders, in other words, those who responded because of the campaign.

Based on our initial experiment, we have the information shown in the left panel of Figure 6.5, and we may distinguish the following groups:

- **Control Responders (CR)** are Sure Things, or Do-Not-Disturbs, in unknown proportion; they responded without a treatment but might *not* respond if treated. We want to avoid treating them in the future to save the cost of unnecessary and possibly deleterious treatment.
- **Control Nonresponders (CN)** are Lost Causes or Persuadables in unknown proportions. We want to find the Persuadables in this group to treat them in the future.
- **Treatment Responders (TR)** are Sure Things or Persuadables, also in unknown proportions. We want to continue treating the Persuadables in this group but avoid wasting the cost of treatment on the Sure Things.
- **Treatment Nonresponders (TN)** are Lost Causes or Do-Not-Disturbs in unknown proportions; they did not respond in spite of being treated. Some of them might have responded if they had not been treated (the Do-Not-Disturbs). We want to avoid treating them in the future since the cost of treatment is wasted on them and might even have a negative impact.

The underlying objective of uplift modeling is to figure out the identities of the Sure Things, Do-Not-Disturbs, Lost Causes, and Persuadables from the known identities of Control Responders, Control Nonresponders, Treatment Responders, and Treatment Nonresponders.

6.4 When Is the Traditional Response Modeling Approach Sufficient?

In response to a presentation on uplift modeling at an analytics conference, an industrial practitioner did not agree there was a need for uplift modeling, since their traditional approach of finding responsive customers (using only the treatment group) performed very well (in terms of lift over control). Armed with years of successful experience from a team of world-class modelers, the practitioner was highly confident that the traditional approach was fine, and the uplift/true-lift approach would not be necessary. The presenter then realized that this practitioner was affiliated with a large credit card company and agreed that uplift would likely not be useful for the credit card industry, as the control response rate (given by B and D in Table 6.2) would be close to zero, at least in the United States, so the lift would be close to the treatment response rate (A – C in Table 6.2). In other words, potential customers very rarely apply for a credit card by themselves but rather respond to a credit card invitation in the mail. This section discusses under what scenarios the traditional approach is sufficient.

Since uplift modeling seeks to maximize the treatment response rate minus the control response rate, while the traditional response modeling approach is to maximize only the treatment response rate, the two approaches would be identical if the control response rate is close to zero. When there are very few control responders, so customers would not buy the product by themselves without being pushed, uplift modeling would not be necessary. This situation can arise in the following scenarios:

1. Certain mature products have a very low response rate without initiation of marketers – for example, credit cards in the United States – possibly due to strong competition, and product maturity (e.g., most individuals have a credit card already in the United States).
2. Brand-new products, where the response would be close to none without marketing, since consumers are not yet familiar with the product.
3. Products that are only sold by invitation, such as acceptance of an email invitation to a webinar.
4. Unsought products due to low frequency or fear, such as funeral services or certain health screening services, where most consumers do not actively seek consumption without targeted promotion.

6.5 Uplift Model Development Methods

This section provides more technical details on how to implement uplift modeling. We discuss three simple methods in this chapter – a "baseline" model and two true uplift models – followed by some advanced methods in the next chapter.

6.5.1 Method 0: The Baseline Model

The standard "traditional" approach is set out in Figure 6.6. First, treatment data from a previous campaign are used for model development. With the traditional supervised learning methodology, a holdout ("validation") sample is set aside for model validation. The response model is developed and estimated using the training dataset and is then used to predict the response to treatment ("scored") using the validation sample. The responses are ordered from highest to lowest and assigned to deciles. The observed response rate for each of these deciles is then graphed to create a traditional lift table, as in Figure 6.2.

In the uplift modeling approach, however, both treatment *and* control data from the previous campaign are used, as illustrated in Figure 6.7. The combined treatment and control data are split into training and validation samples. As in the traditional approach, the model developed using the training data is then scored in the holdout sample, sorted into deciles, and graphed to show the observed response rates for each of these deciles. Alternative models may be developed through this process, and the same holdout sample can be used for model comparison.

The difficult bit here is that there is more than one way to develop appropriate models using the training sample of treatment and control data, so

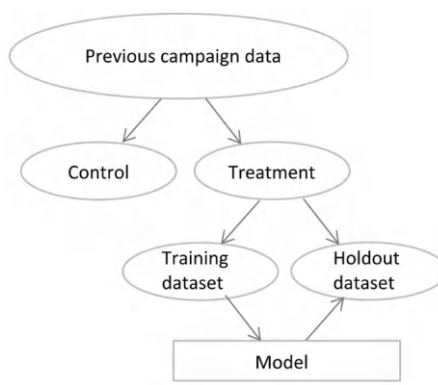


FIGURE 6.6
Traditional approach: Baseline treatment only response model.

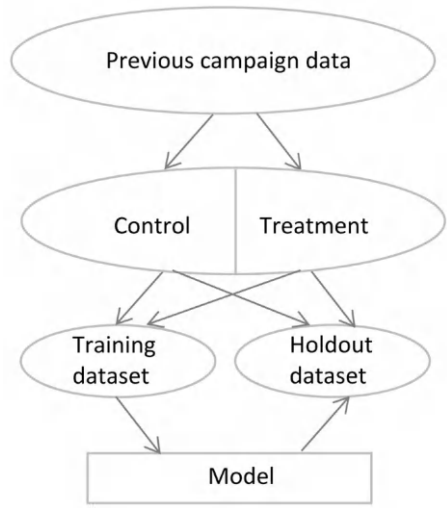


FIGURE 6.7
Uplift modeling approach: Both treatment and control groups are used.

we now describe two approaches to uplift modeling in more detail.[1] Let R be the binary response variable, set equal to 1 if the client responded and 0 otherwise; T is a binary treatment variable (= 1 for treatment and 0 for control); and x is a set of observable individual characteristics such as age and income.

6.5.2 Method 1: Two Model Approach

This is the most straightforward approach and consists of developing treatment and control response models separately. The necessary steps are as follows:

1. Estimate the probability of response among the treated population, $p(R \mid T, x)$ as a function of individual characteristics, x. This is the same as a standard response model. If logistic regression is used, the standard stepwise or a lasso procedure can be used to select variables. In many practical applications where there are a large number of potential predictors, other pre-modeling variable-selection methods can be adopted, such as graphical methods and correlation analysis. Other statistical or machine-learning procedures, such as decision trees and neural networks (see Chapter 2), are also common techniques for predicting probability of response.
2. Estimate the probability of response among the untreated (control) population, $p(R \mid C, x)$, using a procedure similar to 1 above.

3. For each individual in the treatment and control groups, calculate the estimated scores using the two models above. The lift score combines the two model scores to estimate lift:

$$\text{lift}(x) = p(R \mid T, x) - p(R \mid C, x)$$

This is a familiar approach for modelers, except that two models are developed: First, a standard supervised learning technique to predict response among the treatment population; then the same method to predict response among the control population. The lift score is the difference between the two models – the expected response, or response probability if treated, minus the expected response if *not* treated – and is shown as the shaded area in Figure 6.8. The most popular method is to use a standard logistic regression, in which nonlinear and interaction terms may be included if desired, but other common methods such as decision trees, support vector machines, neural networks, or Naïve Bayes algorithms may also be applied.

Compared to the standard response model approach, the drawbacks of the two-model method are:

- It is twice the work because two models have to be created rather than one (though they may be straightforward to develop).
- The models aim to estimate gross response, not lift. Therefore, variable reduction for each model is geared toward response and may not capture some of the variables that are correlated with lift.

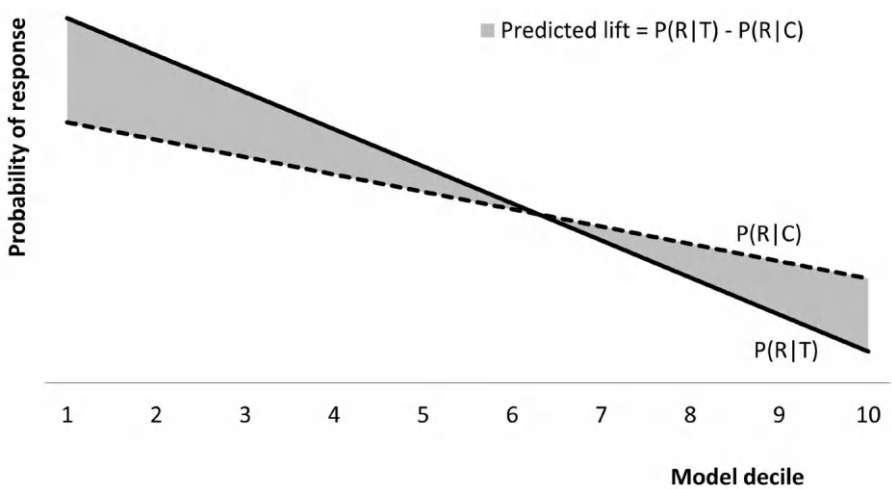


FIGURE 6.8
In Method 1, the lift is the difference between probability of response given treatment $P(R \mid T)$ minus probability of response given NO treatment $P(R \mid C)$.

- The scales of the two models may not be compatible, causing the lift estimate to be noisy or meaningless – see Swait and Louviere (1993) for detail.
- Taking the difference between two models as the score will capture all the errors of both models. If the S/N ratio of the lift is low, then the model scores may be all noise and therefore meaningless.

6.5.3 Method 2: A Single-Model Approach with a Treatment Dummy

Lo (2002) proposes to model the treatment and control groups together in a single model, capturing lift through terms that interact with the treatment variable. This method was also mentioned by Potter (2013) as the chosen approach for a recent presidential election. In this method, the model population includes both the treatment and control observations, and as before, the dependent variable is a binary measure of response. But here the independent variables comprise two sets: first are the variables that capture baseline response for both treatment and control populations (the "main" effects), while the second set captures lift response by interacting the control variables with a dummy variable that is set to 1 for treatment and takes on values of 0 for all control observations.

Formally, the probability of response, P_i, is a function of the treatment dummy variable, T_i, and a vector of independent variables or predictors, X_i:

$$P_i = \frac{\exp\left(\alpha + \beta' X_i + \gamma T_i + \delta' X_i T_i\right)}{1 + \exp\left(\alpha + \beta' X_i + \gamma T_i + \delta' X_i T_i\right)}$$

where α, β, γ, and δ are parameters to be estimated. Therefore,

$$\begin{aligned} P_i|\text{treatment} - P_i|\text{control} &= P_i|(T_i = 1) - P_i|(T_i = 0) \\ &= \frac{\exp\left(\alpha + \gamma + \beta' X_i + \delta' X_i\right)}{1 + \exp\left(\alpha + \gamma + \beta' X_i + \delta' X_i\right)} \\ &= \frac{\exp\left(\alpha + \beta' X_i\right)}{1 + \exp\left(\alpha + \beta' X_i\right)} \end{aligned} \tag{6.1}$$

The final prediction of true lift is then equal to the difference of two scores, given by

$$\text{Prob}\left(\text{response if Treated}\right) - \text{Prob}\left(\text{response if not Treated}\right)$$

$$= \text{score with treatment dummy } \textit{set} \text{ to } 1 - \text{score with treatment dummy } \textit{set} \text{ to } 0$$

Since this model is fit to the data only once, one is guaranteed that the two scores will be on the same scale.

The process is summarized in detail as follows:

1. Include the response and covariate data, $\{Y_i, X_i\}$ from both the treatment and control groups in the analysis data set and assign a dummy variable T_i to 1 for the treatment group and 0 for the control group.
2. Randomly divide the data set into training and holdout samples;
3. Further divide the training sample into two sub-samples by T_i, that is, one is treatment and the other is control.
4. Multiply all independent variables, X_i, by T_i to form the interaction effects, $X_i{*}T_i$.
5. Fit a stepwise logistic model using Y_i as the dependent variable and X_i, T_i, and $X_i{*}T_i$ as independent variables.

Method 2 is not without its own disadvantages. Including the interaction terms, there may be a large number of variables, and multicollinearity is likely. When the correlation between a baseline and its interaction variable is very high (say, over 95%), some judgment call may be needed to select one of them for model development. As with Method 1, taking the difference between two model scores may compound errors.

Method 2 proposed for marketing is actually similar to measuring heterogeneous treatment effects (also known as effect modification) in fields such as epidemiology and social sciences where interaction effects are used for analysis. The subtle difference is that Method 2 focuses on predicting individual level lift for future targeting while those other fields typically study group-level differences (e.g., Greenland et al. 2008; Brand and Xie 2010).

Note that both Methods 1 and 2 can be easily applied with other supervised-learning techniques such as MARS, neural network, and decision trees (e.g., Haughton and Oulabi 1997).

6.6 Uplift Model Evaluation Methods

In addition to the model development method, Lo (2002) has proposed a procedure for validation using a holdout sample. Because treatment and control groups are both involved, and the objective is to measure the "lift over control," this procedure is different from validating a standard supervised learning model, such as the standard logistic regression approach, where only

the treatment group is used. Given that such a procedure is used by other industrial practitioners, we adopt the same validation procedure, which may be summarized as follows:

1. For every individual in the holdout sample (regardless of whether they were originally in the treatment or control group), compute the predicted values of response probabilities for both the treatment and control by assigning $T_i = 1$ or 0, respectively, and then create the predicted uplift score by taking the difference between the two estimated probabilities, as in Eqn. 6.1.
2. Rank the entire holdout sample by this score and derive the score-driven semideciles (i.e., 5% intervals) to provide sufficient granularity.
3. In each semidecile, calculate the observed response rates in the treatment group and control group respectively, and then take the observed difference (this is the *actual or observed* lift in each decile).
4. Plot the observed difference between treatment and control by decile to validate the model and graphically assess the model.
5. Compute relevant metrics to support model evaluation; this is discussed more fully below.

Step 4 is a simple graphical method proposed by Lo (2002). Step 5 builds on suggestions made by Kane et al. (2014), who used three metrics, obtained from holdout samples, related to how the models would likely perform in the next iteration of the marketing campaign. These metrics are (1) a "Gini" coefficient, (2) a "Gini top 15%," and (3) a "Gini repeatability metric" (or R^2 of the lift chart). Each calls for some further explanation.

The *Gini coefficient* gives an estimate of the overall model fit. It measures the area between the gains curve – this measures the cumulative percent of responders and is the top curve in Figure 6.9 – and the population curve, shown by the straight line that gives the cumulative percent of population, sorted by decreasing size of uplift. For standard response models, the gains curve always rises. For lift models, however, the gains curve can rise (for positive lift) and fall (for negative lift, i.e., suppressed response). The Gini coefficient computation is described in Appendix 6.2.

The *Gini top 15%* is another useful measure. For a campaign with an unlimited budget, the target population would include all candidates up to the point where the gains curve begins to fall (and uplift goes from positive to negative). However, many campaigns have limited budgets, in which case the gains for the initial deciles (say, deciles 1 and 2) are more important than the gains for the higher deciles. For this evaluation, we chose the Gini area for the top 15% of scorers as an example, since many campaigns target just

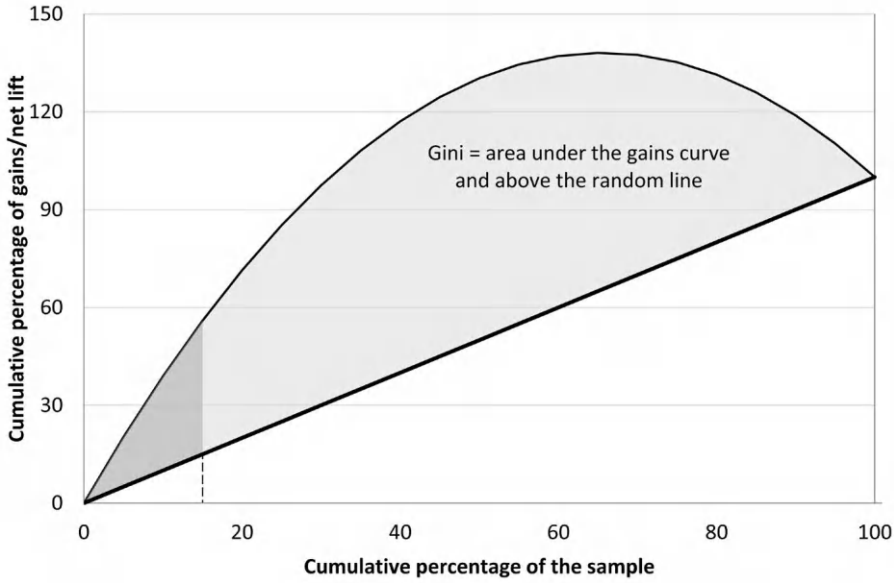


FIGURE 6.9
Gains chart for uplift/true-lift models.

the top 1 or 2 deciles in practice; this is shown as the dark-shaded area to the left of Figure 6.9. One can use similar metrics such as Gini top 30% as a metric if the next campaign expects to target the top 3 deciles (30%) instead of the top 15%. The computational formula of Gini Top 15% is in Appendix 6.2.

Our final measure of model performance is the *Gini repeatability metric*. Lift models can be unstable, so the third metric is a simple measure of the model's overall stability, given by the R-squared (R^2) of a straight line fit to the lift chart on the holdout sample. Ideally, a model would have a very high lift at the top semideciles and no positive lift afterwards, but is practically unrealistic to achieve. More realistically, a model with a relatively smoothly declining lift pattern from the top semidecile to the bottom would be considered good; however, most uplift models have a much more ragged fit on the holdout sample, e.g., the top few semideciles can flip from positive to negative lift values. This metric should be considered secondarily to Gini and Gini top 15%, because the lift could sometimes be curved or non-linear in a desirable way. R^2 evaluation should include a visual assessment of the lift chart for desirable non-linear results. Figure 6.10 shows a hypothetical lift table with a perfect fit, while Figure 6.11 shows a much more realistic lift chart where the trend line fits more loosely. Note also that this metric becomes meaningful if the regression slope is positive (which would mean the lift has a tendency to go up in supposedly worse deciles), or equivalently, the Gini coefficient is negative.

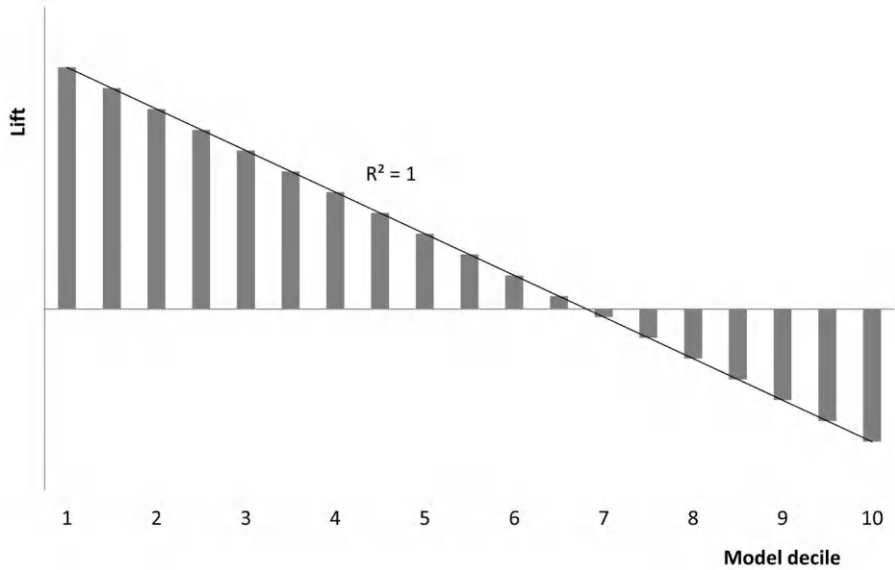


FIGURE 6.10
Lift chart on holdout sample (*ideal* case).

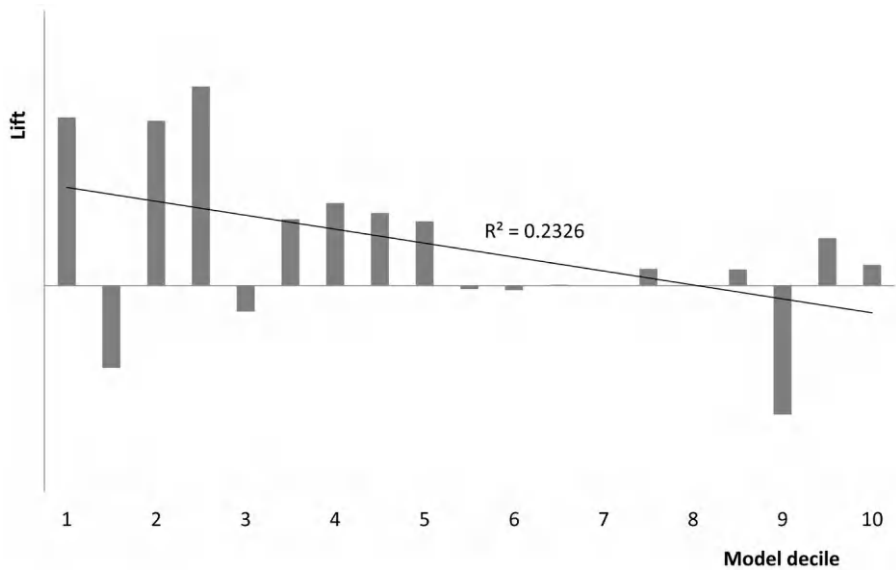


FIGURE 6.11
Lift chart on holdout sample (*realistic* case).

6.7 Illustrated Examples

We will now illustrate the above procedures of model development and validation, and associated metrics, with two examples: A simulated data set, and a real data set.

Example 6.1: Nonprofit Organization Donation Analysis – A Simulated Example

Our first example is about ACE (disguised name), a nonprofit school serving mostly low-income students in a city; although the data are simulated, the example is based on a real case. The school mostly relies on donations from individuals and organizations. The workforce includes permanent staff, part-time staff, and many volunteers. Their fundraising effort requires staff and volunteers to make personal calls, or in-person visits, to potential donors. Due to limited resources, they need to prioritize their fundraising treatment effort. Concretely, the school needs to come up with a targeted list of donors to contact through outbound calls or in-person visits. A donor database with multiple years of data is available for analysis. The response is defined as a monetary donation in the last year (*"donated"*). The treatment is a contact (call or visit). The potential predictors are obtained from history prior to the last year. We have an 80-20% split between treatment (those who received a contact) and control (those who did not). The whole data set is also randomly split into training (n = 300 K) and holdout (n = 200 K) samples. The potential variables that drive donations, and that are available in the dataset, include:

- ***Age*** of donor;
- ***Frequency***, defined as the number of times (years) a donation was made in the past;
- ***Spent***, which is the average amount donated in the past;
- ***Recency***, which measures the number of years that have elapsed since the last donation (so 1 = made a donation last year, 2 = last donation was made two years ago, etc.);
- ***Income***, given by estimated annual income;
- ***Wealth***, measured by estimated wealth.

Summary statistics of the response variable (*donated*) are shown in Table 6.3 using the training sample. The lift over control response rate is 19.3% – 6.3% = 13%, and the S/N is 13%/6.3% = 207%. We now apply our two uplift methods to these data.

As noted above, with the Two-Model Approach (Method 1), we fit separate logistic models to the treatment and control data (using `proc logistic` in SAS), with stepwise inclusion of variables. For the Treatment Dummy Approach (Method 2), we again use `proc logistic` in SAS, but with a single

TABLE 6.3

Summary Statistics of Simulated Example

	Response	No Response	Total
Treatment: Number	46,327	193,696	240,023
Row %	19.30	80.70	
Column %	92.47	77.51	
Control: Number	3,774	56,203	59,977
Row %	6.29	93.71	
Column %	7.53	22.49	
Total	**50,101**	**249,899**	**300,000**
Row %	16.7	83.3	100.0

equation that includes a treatment dummy (whether or not they received a contact) as well as all treatment interaction variables (treatment dummy times each variable outlined above).

The standard stepwise regression with 5% level is used for both methods. One can try other significance levels or other standard variable selection methods, such as forward selection or backward elimination, for testing alternative models. The estimation results are summarized in Table 6.4. Note that the Treatment Only model is simply the standard baseline model that

TABLE 6.4

Model Results from Methods 1 and 2 for the Simulated Example

	Treatment Only		Control Only		Dummy Var. Model		
Variable	**Estimate**	**p-Value**	**Estimate**	**p-Value**	**Estimate**	**p-Value**	**Actual**
Intercept	–8.0693	<0.0001	–10.0986	<0.0001		<0.0001	–10
Age	0.0068	<0.0001	n.s.		n.s.		0.001
Frequency	0.7057	<0.0001	0.6861	<0.0001	0.7041	<0.0001	0.7
Spent	0.0010	<0.0001	0.0010	<0.0001	0.0010	<0.0001	0.001
Recency	–0.0409	<0.0001	n.s.		–0.0411	<0.0001	–0.07
Income	n.s.		n.s.		n.s.		0
Wealth	n.s.		n.s.		n.s.		0
Treatment (main effect)	n/a		n/a		1.9662	<0.0001	2
Treatment × age	n/a		n/a		0.0067	<0.0001	0.005
Treatment × freq	n/a		n/a		n.s.		
Treatment × spent	n/a		n/a		n.s.		
Treatment × recency	n/a		n/a		n.s.		0.03
Treatment × income	n/a		n/a		n.s.		
Treatment × wealth	n/a		n/a		n.s.		

Note: n.s. is not statistically significant (at 1% level or better), and n/a means not applicable.

most modelers would use without the knowledge of uplift modeling. The Two Model Approach (Model 1) requires one to develop two models – the Treatment Only and Control Only, respectively, and then take the difference between the two model scores. The Treatment Dummy Approach (Model 2), which relies on the treatment dummy and interaction effects, is in the third column of figures in Table 6.4; note that only one interaction effect is picked up.

So how good are the two models? The last column of Table 6.4 has the actual values used to simulate the data – a big advantage of running a simulation exercise is that we know the actual model and can compare the estimated models to the actual. To understand how to read this, for example, the age effect without treatment (i.e., control) is 0.001, while the age effect with treatment is 0.001 + 0.005 = 0.006, while Method 1 (the Two Model Approach) results in an age effect of 0 in control and 0.00675 in treatment, and Method 2 finds an age effect of 0 in control and 0.00672 in treatment. So both models result in estimates very close to the actual values for the age effect. Another example is the main effect of treatment, which has an actual value of 2 and a model estimate of –8.069 – (–10.099) = 2.030 from Method 1 and an estimate of 1.956 according to Method 2. So once again, both models result in the main treatment effects very close to the actual value. We now evaluate the models using the holdout sample.

Figure 6.12 shows the lift chart by semi-decile (1 = semi-decile with the highest lift, 20 = lowest). The baseline model uses data from the treatment sample only, and it has a peak shifted to the second semi-decile, clearly not showing a desired decreasing pattern.[2] Both Methods 1 and 2 produce a smooth decreasing pattern by semi-decile and, in this sense, outperform the baseline model.

There is hardly any performance difference between Methods 1 and 2 in this example. Using Gini, or Gini Top 15%, which measure the predictive accuracy, it is clear from Table 6.5 that Methods 1 and 2 are better than the

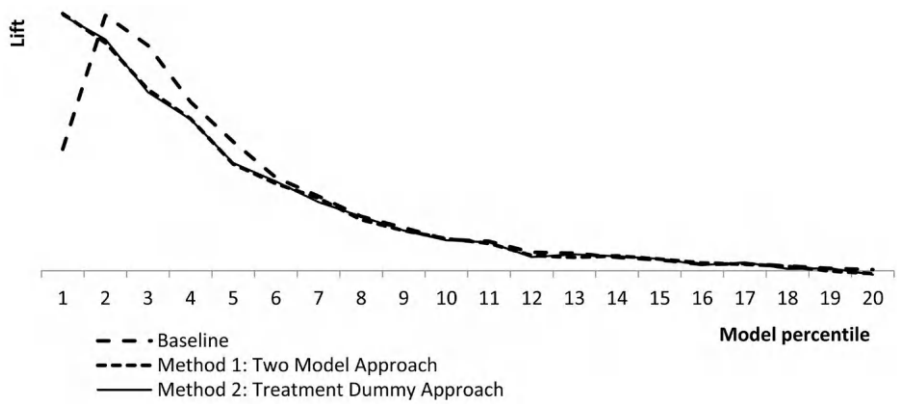


FIGURE 6.12
Lift graph of the simulated example.

TABLE 6.5

Holdout Sample Model Validation Statistics for the Simulated Example

	Gini	Gini 15%	Gini Repeatability R^2
Baseline	5.6420	0.5412	0.731
Method 1 (Two Model Approach)	6.0384	0.7779	0.783
Method 2 (Treatment Dummy Approach)	6.0353	0.7766	0.784

standard ("baseline") approach. In terms of Gini repeatability (R^2), which measures the linearity of the lift graph, Methods 1 and 2 are also better than the baseline model and very similar to each other. Note that the Gini repeatability (R^2) is generated from a linear regression applied to the lift charts in Figure 6.13.

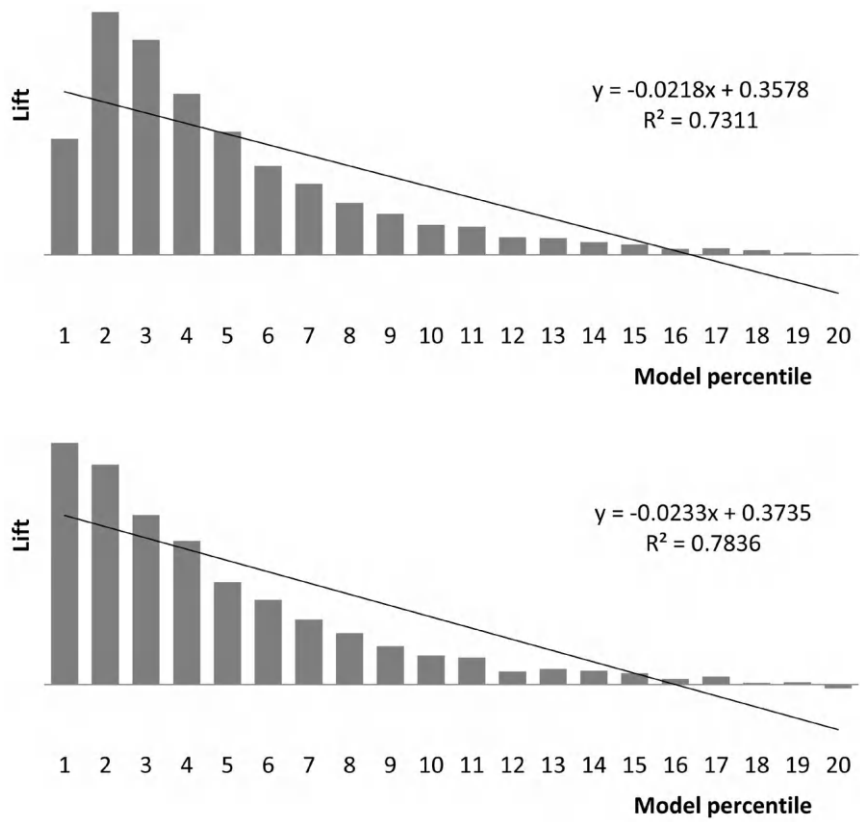


FIGURE 6.13

Lift charts from the baseline method and methods 1 and 2. (a) Baseline method (standard treatment only response model); (b) Method 1: Two model approach; and (c) Method 2: Treatment dummy approach. *(Continued)*

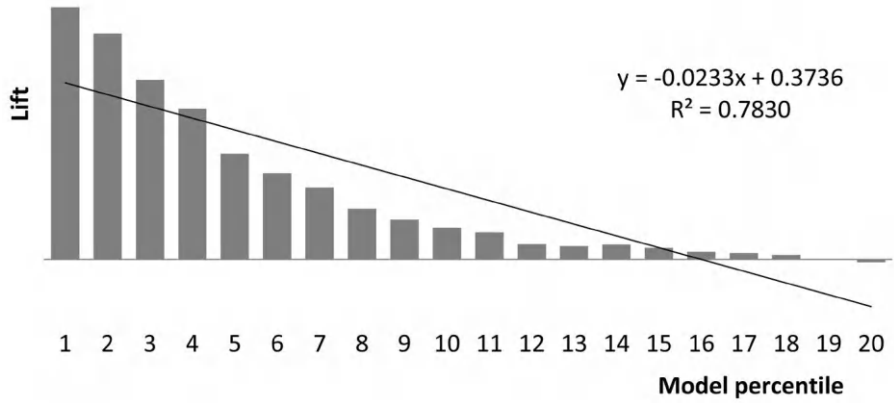


FIGURE 6.13 (Continued)

Example 6.2: An Email Marketing Example

Kevin Hillstrom of MineThatData has posted an excellent example of uplift data on his blog.[3] The data have been analyzed by Radcliffe (2008), Kane et al. (2014), Lo and Pachamanova (2015), Pachamanova et al. (2020), and Lo and Pachamanova (2023). This email marketing campaign for men's and women's clothing contains 64,000 customers who last purchased within the past 12 months:

- 1/3 of the sample were randomly chosen to receive emails featuring men's merchandise;
- 1/3 of the sample were randomly chosen to receive emails featuring women's merchandise;
- 1/3 of the sample were randomly chosen not to receive emails; this is the no-email control group.

Additionally, the data are randomly split into training and holdout samples in the proportions of 70% to 30%.

To illustrate the techniques introduced in the last section, we focus on the campaign featuring women's merchandise using "*visit*" (i.e., whether or not the customer visited the website) as our response variable. As a result, we have a 50-50% breakdown between treatment (those who received emails) and control (those who did not). The potential drivers available include:

- ***Recency,*** as measured by months since last purchase;
- ***History,*** which measures the total number of dollars spent in the last year;
- ***Men's*** is a binary variable set to 1 if men's merchandise was purchased in the last year and to 0 otherwise. We create a similar variable called ***women's***.

- We also construct a variable called ***bothgenders*** set to 1 if both *men's* and *women's* are equal to 1 and to 0 otherwise.
- ***Zip_rural_flag*** is coded to 1 if the area is rural and to 0 otherwise (i.e., if the area is urban).
- ***Newbie*** is a binary variable set to 1 if the customer was new within the past 12 months and to 0 otherwise.
- Two variables are used to indicate whether a customer used the Web, or a phone, or both, to purchase merchandise in the past year: We set ***channel_phone_flag*** = 1 if the phone was used and 0 otherwise, and ***channel_multi_flag*** = 1 if both Web and phone were used but is 0 otherwise.
- Squared terms are also created for the continuous ***recency*** and ***history*** variables.

Summary statistics of the response variable (*visit*) are shown in Table 6.6 for the training sample. The lift over control response rate is 15.3% – 10.7% = 4.6%, and the S/N is 4.6%/10.7% = 43%.

Once again we fit separate logistic models to the treatment and control data (for the Two-Model Approach) and use a single equation that includes a treatment dummy and interaction terms (for the Treatment Dummy method). As before, we apply forward stepwise regression with an inclusion threshold of 5% significance. Only the coefficient estimations that are statistically significant are reported in Table 6.7. While Table 6.7 shows which variables are important, it is not meaningful for targeting, and in order to validate the model, we need to apply the scoring equations to the holdout sample; the lift charts are shown in Figure 6.14.

In Figures 6.14a–c, all three models show a general trend of declining lift by semi-decile. While this is quite commonly seen in regular supervised learning, it cannot be taken for granted for uplift modeling. In particular, the baseline model in Figure 6.14a shows that the first semi-decile is

TABLE 6.6

Summary Statistics of Marketing Data

	Response	No Response	Total
Treatment: Number	2,292	12,691	14,983
Row %	15.3	84.7	
Column %	58.8	48.7	
Control: Number	1,605	13,375	14,980
Row %	10.7	89.3	
Column %	41.2	51.3	
Total	**3,897**	**26,066**	**29,963**
Row %	13.0	87.0	100.0

TABLE 6.7

Model Results from Methods 1 and 2 for the Marketing Example

	Treatment Only		Control Only		Dummy Var. Model	
Variable	**Estimate**	**p-Value**	**Estimate**	**p-Value**	**Estimate**	**p-Value**
Intercept	–1.91	<0.001	–1.61	<0.001	–1.62	<0.001
Recency	–0.10	<0.0001	–0.07	<0.001	–0.07	<0.001
Recency squared	0.00	<0.0001	n.s.		n.s.	
History	0.00	<0.001	0.00	<0.001	0.00	<0.001
Men's	0.38	<0.001	n.s.		n.s.	
Women's	0.91	<0.001	n.s.		n.s.	
Both genders	n.s.		0.51	<0.001	0.43	<0.001
Rural (=1)	0.25	<0.001	0.56	<0.001	0.55	<0.001
Newbie	–0.45	<0.001	–0.77	<0.001	–0.78	<0.001
Used phone? (Y=1)	–0.24	<0.001	–0.38	<0.001	–0.30	<0.001
Used Web+phone? (Y=1)	–0.20	0.01	n.s.		n.s.	
Treatment (main effect)	n/a		n/a		n.s.	
Treatment × recencysq	n/a		n/a		0.00	0.00
Treatment × historysq	n/a		n/a		0.00	0.02
Treatment × women's	n/a		n/a		0.53	<0.001
Treatment × Rural	n/a		n/a		–0.30	0.00
Treatment × newbie	n/a		n/a		0.33	<0.001
Treatment × Web+phone	n/a		n/a		–0.24	0.00

Note: n.s. is not statistically significant (at 1% level or better), and n/a means not applicable.

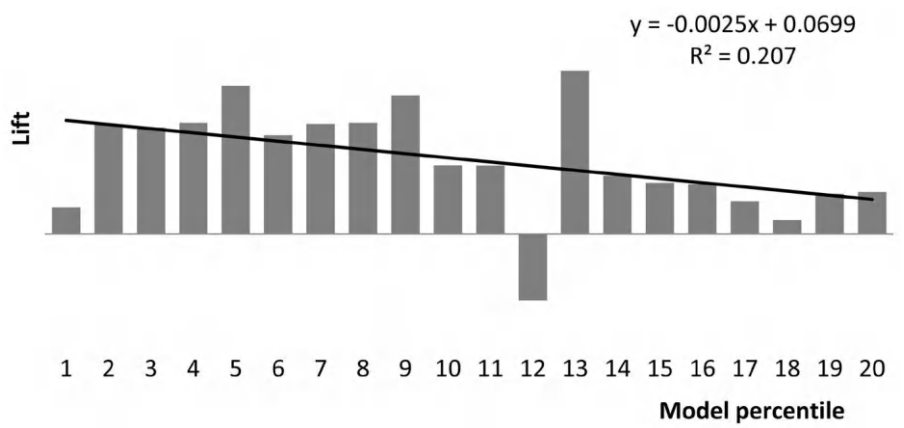


FIGURE 6.14

Lift charts from the baseline method and Methods 1 and 2. (a. Baseline method (standard treatment only response model); b. Method 1: Two model approach; c. Method 2: Treatment dummy approach; and d. Lift chart of all the three methods.) *(Continued)*

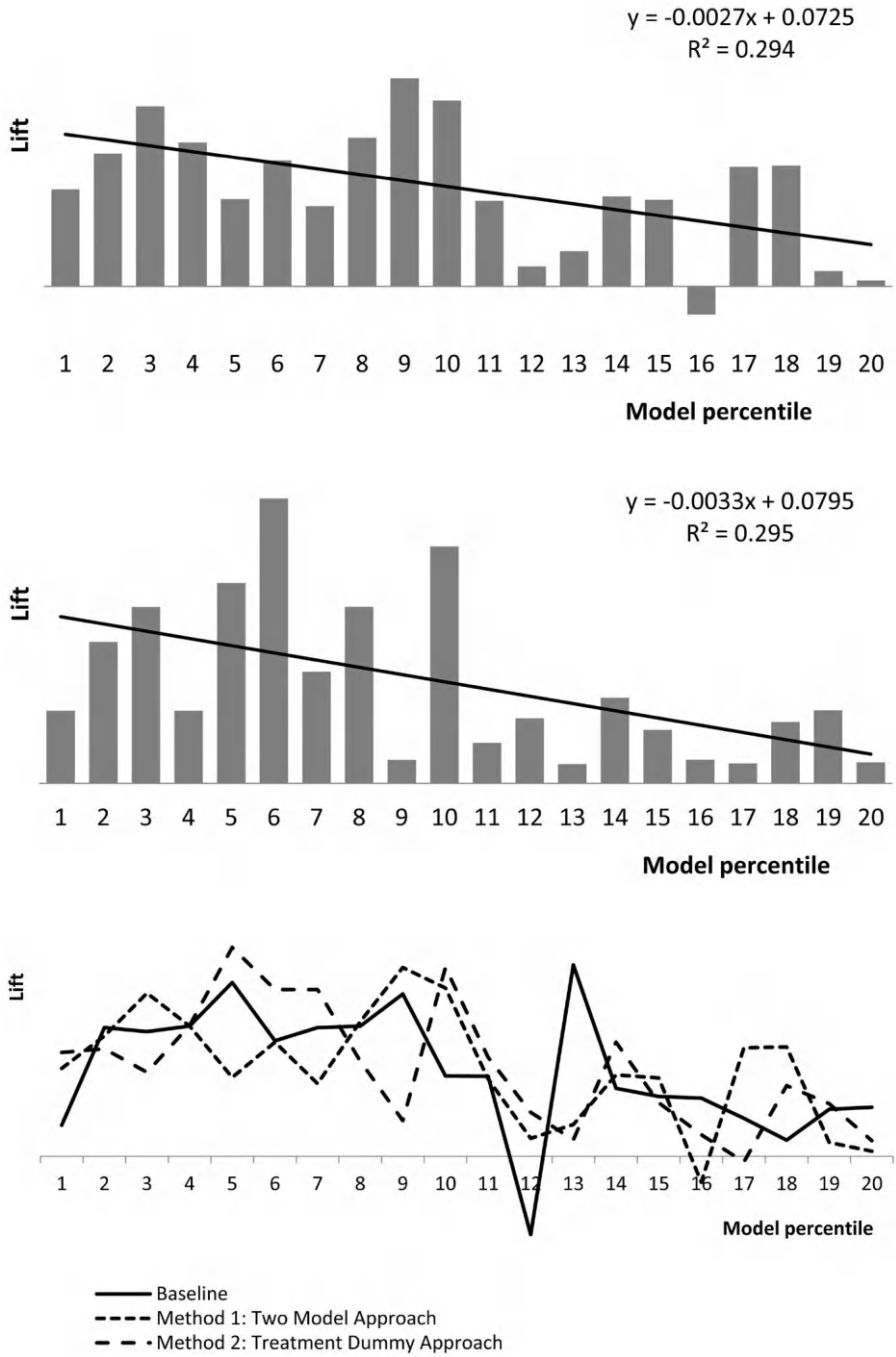


FIGURE 6.14 (Continued)

TABLE 6.8

Holdout Sample Model Validation Statistics for the Marketing Example

	Gini	Gini 15%	Gini Repeatability R^2
Baseline	1.8556	–0.0240	0.2071
Method 1 (Two Model Approach)	2.0074	0.0786	0.2941
Method 2 (Treatment Dummy Approach)	2.4392	0.0431	0.2945

particularly low in lift, but it has some value in predicting lift at the next several semi-deciles (2–9), although the differentiation within these semi-deciles seems minimal. Note that Figure 6.14c shows that Method 2 has a deeper negative slope (a stronger declining trend) but does not perform very well in the first few semi-deciles, which often is the business focus. When putting all three methods together in Figure 6.14d, it is not visually clear which method is the best, other than that the baseline method appears slightly worse than others in the 1st semi-decile, and Method 2 has a higher peak at the 6th semi-decile. When we turn to the model validation statistics in Table 6.8, all three metrics show consistently that Methods 1 and 2 outperform the baseline approach in these data; this is especially true of the Gini 15% measure, which is negative in the baseline case, meaning it is not useful at all.

6.8 Concluding Remarks

We are now in a position to summarize the contribution of uplift modeling. Just like any supervised learning approach, it has two parts: Model development and model validation. Without careful model validation (as recommended in this chapter), one would not even know if the baseline model can do a decent job at all. Whether the baseline model does well is very data-specific. In an extreme situation such as credit card acquisition, where the natural (control) response rate is nearly zero (at least in the United States), the baseline treatment-only model may be all one needs to identify promising new customers. In general, however, the baseline model's objective only meets half of the real objective because it is not meant to maximize uplift. The two model development methods for uplift in this chapter have the right objective, and whether they are better than the baseline can be easily tested empirically. In reality, the authors have seen situations where the uplift model beats the baseline model and, less commonly, cases where the baseline model beats the uplift model. In either case, having a validation test can help the modeler select the most appropriate model.

We will revisit uplift modeling techniques in Chapter 8 and apply them to more complex situations. Since the best modeling technique is often data specific as in the traditional supervised learning world (e.g., Wolpert 1995, 2001), having more alternative methods would give us a better chance of finding a good model for targeting "persuadables."

Additionally, we include a technical counterfactual explanation behind uplift modeling in Appendix 6.1 and some guidelines on variable selection for uplift modeling in Appendix 6.3.

Appendix 6.1: Counterfactual Framework for Uplift/True-lift Modeling

A customer can only be assigned to either treatment ($T = 1$) or control ($T = 0$) at a specific time, but not to both. This makes it necessary to construct a counterfactual, and this appendix provides a theoretical explanation for why Uplift/True-lift modeling can do this.

Consider a counterfactual or potential outcome framework where a customer can be assigned to treatment or control, as set out in Figure A6.1. If she is indeed assigned to treatment, she could choose to respond or not. Similarly, if she is assigned to the control group, she could also respond or not. While she cannot be assigned to both treatment and control at the same time, the question is this: If she receives treatment, can we use others (assumed to look like her) to represent her potential outcome if she had been assigned to the control group? The answer is yes if she and the others are "exchangeable" on all relevant pre-treatment characteristics or variables, x.

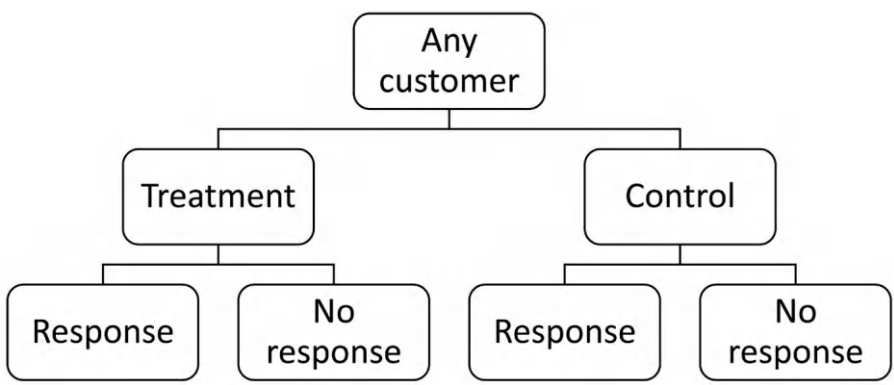


FIGURE A6.1
Counterfactual framework.

More formally, let $Y(1)$ and $Y(0)$ be the potential outcomes under treatment and control, respectively, for a particular customer. Then the treatment effect is simply $Y(1)-Y(0)$, and its expected value is $E\big(Y(1)-Y(0)\big)$. To measure the expected value, given a set of characteristics x, we need to be able to compute $E\big(Y(1)-Y(0)\,|\,x\big)$. To measure the average treatment value of the treated, we need to compute $E(Y(1)-Y(0)\,|\,T=1,x)$. This gives the treatment effect among the treated group with characteristics x. While no one can receive both treatment and control at the same time, we may "borrow" information from others with the same characteristics provided that they are "exchangeable." The exchangeability (or ignorability) condition is a necessary condition for measuring the causal effect of treatment and is formally expressed as $\big(Y(1),Y(0)\big)\perp\!\!\!\perp T\,|\,x$; in other words, neither the treatment potential outcome nor the control potential outcome depends on the treatment assignment mechanism given x, as explained by Rosenbaum and Rubin (1983) and Morgan and Winship (2007). Ignorability, or exchangeability, is satisfied with randomized experiments, but not as a general rule with non-randomized experiments or observational studies. We have:

$$
\begin{aligned}
&E\big(Y(1)-Y(0)|\,T=1,x\big)\\
&\quad=E\big(Y(1)|\,T=1,x\big)-E\big(Y(0)|\,T=1,x\big)\\
&\quad=E\big(Y(1)|\,T=1,x\big)-E\big(Y(0)|\,T=0,x\big),\ \text{with the exchangeability or ignorability condition}\\
&\quad=E\big(Y|\,T=1,x\big)-E\big(Y|\,T=0,x\big),
\end{aligned}
$$

using the consistency assumption that the observed outcome under a treatment condition is the same as the potential outcome under the same treatment condition.

Hence, *we can model the treatment effect given a common set of characteristics, using data for both treatment and control groups.* While this chapter focuses on binary outcomes (i.e., Y = 1 or 0), so the expected value is the same as the response probability, the previous statement applies to both continuous and discrete outcomes.

Appendix 6.2: Computations of Gini and Gini Top 15%

Assume we rank the holdout sample by semidecile, i.e., 20 groups with 5% in each group. Define the average lift at group j as:

$$lift\big(j\big)=P\big(R|\,T,j\big)-P\big(R|\,C,j\big),$$

where $P(R|T,j)$ and $P(R|C,j)$ represent the response probabilities in semidecile subgroup j in the treatment and control groups, respectively, and can be estimated by the relative frequencies of response in the holdout sample. Then,

$$\text{Gini coefficient} = \sum_{g=1}^{20} \left(cum\%lift(g) - cum\%sam(g) \right),$$

where,

$$cum\%lift(g) = \text{cumulative \% lift up to semidecile group } g$$

$$= \frac{\sum_{j=1}^{g} lift(j) n_{tj}}{\sum_{j=1}^{20} lift(j) n_{tj}},$$

$$cum\%sam(g) = \text{cumulative \% sample up to semidecile group } g$$

$$= \frac{\sum_{j=1}^{g} n_{tj}}{\sum_{j=1}^{20} n_{tj}} = \frac{\sum_{j=1}^{g} n_{tj}}{n_t},$$

n_{tj} = treatment sample size in semidecile j in the holdout sample,

n_t = total treatment sample size in the holdout sample.

In the common Gini formula for regular supervised learning, there is a denominator representing the maximum possible value of the numerator, i.e., the gap between the best possible model (horizontal line at 100%) and the diagonal random line, which can be approximated by (1 − 0.05) + (1 − 0.1) + … + (1 − 0.95) + (1 − 1) = 9.5. However, for uplift modeling, the maximum value is data dependent and much more complicated and can be greater than the traditional maximum value. Hence, we choose not to use a constant denominator as model comparisons within the same data set remain valid.

Similarly, Top 15% Gini is simply focused on the top 15%, or the top 3 semideciles, of the Gini coefficient formula:

$$\text{Top 15\% Gini} = \sum_{g=1}^{3} \left(cum\%lift(g) - cum\%sam(g) \right).$$

Appendix 6.3: Variable Selection for Uplift/True-lift Modeling

The uplift/true-lift modeling methodologies introduced in this chapter are relatively straightforward. A natural question is how to pre-select variables prior to model development. This appendix provides some alternative ideas.

1. **No pre-selection:** If the number of variables is not enormous, one can feed all available variables into the modeling procedure and allow the modeling procedure to pick up the important ones, through the standard stepwise procedure, forward selection, backward elimination, or the more recent Lasso method.
2. **Union of the treatment and control lists of variables:** Standard variable selection procedures, e.g., picking variables that are more correlated with the dependent variable, running a decision tree model to pre-select important variables, or performing a principal component analysis or some form of variable clustering to find groupings of variables can all be applied *separately* to treatment and control samples, respectively. The combined (union) of the two selected lists of variables can be used as an input to uplift/true-lift modeling.
3. **One variable at a time using exploratory analysis:** The above two methods are not designed to find heterogeneous treatment effects directly, i.e., the interaction effects in an uplift/true-lift model. The most classical way to display a picture of heterogeneous treatment effect is by graphing a plot of the dependent variable against an independent variable (continuous, binned continuous, or categorical) by treatment and control. If the two lines are parallel, it means the independent variable is related to the dependent variable consistently for treatment and control groups, meaning there is no heterogeneous treatment effect for the given independent variable. If the two lines are not parallel, it shows a potential heterogeneous treatment effect, which can be used as a variable selection tool (see Zink et al. 2015, for example). However, if there is a long list of variables, graphical methods like this may not be practical to go through.
4. **One variable at a time by fitting a simple model:** A more automated variation of method 3 above is to detect whether the treatment and control lines are parallel statistically, using a simple model. One can fit a model with just the main treatment effect, the given independent variable, and their interaction (product term). Statistical significance of the interaction term shows there may be a heterogeneous treatment effect. After automatically detecting the variables that show a heterogeneous treatment effect, one can then plot the line (as in method 3) for the selected list of variables to gain visual insights on the pattern.

5. **Use one model as an input to another:** As in standard supervised learning, one may choose a relatively simple uplift/true-lift modeling algorithm to pre-select a set of variables and use them as an input to a more sophisticated modeling algorithm. For instance, one may use an uplift decision tree method (available from Quadstone, JMP, R, or KNIME) for variable selection and use the most important variables as an input to another algorithm. One may also use another method such as the one described in Section 8.2 (from Kane et al. 2014) for variable selection.

Notes

1. Appendix 6.1 provides the theoretical framework to explain why true-lift modeling can work in reality, where a customer is assigned to either treatment or control group at a given time but not both.
2. In practice, the baseline model's peak in the lift chart can appear almost anywhere, depending on the data.
3. Data is available at http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html in Excel format.

References

Brand, Jennie E. and Yu Xie. 2010. "Who Benefits Most from College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education". *American Sociological Review*, 75(2): 273–302.

Cai, Tianxi, Lu Tian, Peggy H. Wong, and L. J. Wei. 2011. "Analysis of Randomized Comparative Clinical Trial Data for Personalized Treatment Selections". *Biostatistics*, 12(2): 270–282.

Greenland, Sander, Timothy L. Lash, and Kenneth J. Rothman. 2008. Concepts of Interaction. In Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash (eds.), *Modern Epidemiology*, 3rd edition. Philadelphia, PA: Lippincott Williams & Wilkins.

Haughton, Dominique, and Samer Oulabi. 1997. "Direct Marketing Modeling with CART and CHAID". *Journal of Direct Marketing*, 11(4): 42–52.

Jackson, Robert, and Paul Wang. 1996. *Strategic Database Marketing*. Chicago, IL: NTC Publishing.

Kane, Kathleen, Victor S. Y. Lo, and Jane Zheng. 2014. "Mining for the Truly Responsive Customers and Prospects Using True-Lift Modeling: Comparison of New and Existing Methods". *Journal of Marketing Analytics*, 2(4): 218–238.

Kubiak, Roman. 2012. "Net Lift Model for Effective Direct Marketing Campaigns at 1800flowers.com". SAS Global Forum, Paper 108-2012.

Lo, V. S. 2002. "The True Lift Model – A Novel Data Mining Approach to Response Modeling in Database Marketing". *SIGKDD Explorations*, 4(2): 78–86.

Lo, V. S. 2009. New Opportunities in Marketing Data Mining. In J. Wang (ed.), *Encyclopedia of Data Warehousing and Mining*. Hershey, PA: IGI Global, 1409–1505.

Lo, Victor S. Y., and Dessislava Pachamanova. 2023. "From Meaningful Data Science to Impactful Decisions: The Importance of Being Causally Prescriptive". *Data Science Journal*, 22. https://doi.org/10.5334/dsj-2023-008

Lund, Bruce. 2012. Direct Marketing Profit Model. In *Proceedings of Midwest SAS Users Group*, Paper CI-04.

Maex, Dimitri, and Paul B. Brown. 2012. *Sexy Little Numbers: How to Grow Your Business Using the Data Your Already Have*. New York, NY: Crown Business.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference*. New York, NY: Cambridge University Press.

Pachamanova, Dessislava, Victor S. Y. Lo, and Nalan Gulpinar. 2020. Uncertainty Representation and Risk Management for Direct Segmented Marketing. *Journal of Marketing Management*, 36: 149–175. https://doi.org/10.1080/0267257X.2019.1707265

Peppers, Don, and Martha Rogers. 1997. *Enterprise One-to-One*. New York, NY: Doubleday.

Peppers, Don, Martha Rogers, and Bob Dorf. 1999. *The One-to-One Fieldbook*. New York, NY: Doubleday.

Potter, Daniel. 2013. *Pinpointing the Persuadables: Convincing the Right Voters to Support Barack Obama*. Boston, MA: Predictive Analytics World. http://www.predictiveanalyticsworld.com/patimes/pinpointing-the-persuadables-convincing-the-right-voters-to-support-barack-obama/ (available with free subscription)

Radcliffe, Nicholas J. 2007a. "Using Control Groups to Target on Predicted Lift". *DMA Analytic Annual Journal*, 14–21.

Radcliffe, Nicholas J. 2007b. *Generating Incremental Sales: Maximizing the Incremental Impact of Cross-selling, Up-selling and Deep-selling through Uplift Modelling*. Edinburgh: Stochastic Solutions Limited.

Radcliffe, Nicholas J. 2008. *Hillstrom's MineThatData Email Analytics Challenge: An Approach Using Uplift Modeling*. Edinburgh: Stochastic Solutions Limited.

Radcliffe, Nicholas J., and Patrick D. Surry. 1999. Differential Response Analysis: Modeling True Response by Isolating the Effect of a Single Action. In *Proceedings of Credit Scoring and Credit Control VI*, Credit Research Centre, University of Edinburgh Management School.

Radcliffe, Nicholas J., and Patrick D. Surry. 2011. *Real-World Uplift Modelling with Significance-Based Uplift Trees*. Portrait Technical Report TR-2011-1 and Stochastic Solutions White Paper. http://stochasticsolutions.com/pdf/sig-based-up-trees.pdf

Rexer, Karl. 2012. *5th Annual Data Mining Survey – 2011 Survey Summary Report*. Rexer Analytics.

Rexer, Karl, Paul Gearan, and Heather Allen. 2016. *2015 Annual Data Mining Survey*. Rexer Analytics.

Roberts, Mary Lou, and Paul D. Berger. 1999. *Direct Marketing Management*. Saddle River, NJ: Prentice-Hall.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika*, 70(1): 41–55.

Samuelson, Douglas A. 2013. "Analytics: Key to Obama's Victory". *OR/MS Today*, Feb: 20–24.

Scherer, Michael. 2012. "How Obama's Data Crunchers Helped Him Win". CNN News. http://www.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team/index.html?hpt=hp_bn5

Siegel, E. 2011. *Upilft Modeling: Predictive Analytics Can't Optimize Marketing Decisions Without It*. Prediction Impact white paper sponsored by Pitney Bowes Business Insight.

Siegel, E. 2013a. "The Real Story Behind Obama's Election Victory." *The Fiscal Times* 01/21/2013, at https://www.thefiscaltimes.com/2013/01/21/Real-Story-Behind-Obamas-Election-Victory

Siegel, E. 2013b. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken, NJ: Wiley.

Swait, Joffre, and Jordan Louviere. 1993. "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models". *Journal of Marketing Research*, 30: 305–314.

Wolpert, David H. 1995. The Relationship Between PAC, the Statistical Physics Framework, the Bayesian Framework, and the VC Framework. In D. H. Wolpert (ed.), *The Mathematics of Generalization*. Reading, MA: Addison-Wesley, pp. 117–214.

Wolpert, David H. 2002. The Supervised Learning No-Free-Lunch Theorems. In R. Roy, M. Köppen, S. Ovaska, T. Furuhashi, and F. Hoffmann (eds.), *Soft Computing and Industry*. London: Springer. https://doi.org/10.1007/978-1-4471-0123-9_3

Zink, Richard C., Lei Shen, Russell D. Wolfinger, and H.D. Hollins Showalter. 2015. Assessment of Methods to Identify Patient Subgroups with Enhanced Treatment Response in Randomized Clinical Trials. In *Applied Statistics in Biomedicine and Clinical Trials Design*. Switzerland: Springer International Publishing.

7

Uplift Analytics II: Test and Learn for Uplift

7.1 Introduction

The previous chapter introduced the concept of uplift/true-lift, two model development methods, and the model evaluation procedure. Since uplift analytics cannot be achieved without data, and its power can be enhanced with a well-designed experiment, we will discuss a broader topic in this chapter – how to test and learn using experimental design in the context of uplift analytics.

We introduce the approach in Section 7.2, with a discussion of the general strategy and process in Sections 7.2.1 and 7.2.2, sample size determination in Section 7.2.3, A/B testing in Section 7.2.4, and Multivariate Testing or Experimental Design in Section 7.2.5. As elsewhere in this book, we focus on the practical and methodological aspects rather than a purely theoretical discussion. Additionally, we discuss several measurement and modeling metrics for Test and Learn in Section 7.3, followed by opportunities for continuous improvement in Section 7.4.

7.2 Test and Learn for Uplift Analytics

7.2.1 Test and Learn Strategy

In this section, we describe a highly important business strategy that is not frequently mentioned in academic literature but is frequently employed and discussed by marketers – Test and Learn – to come up with a set of value propositions and to refine them in order to meet business goals. It is an *information-driven* strategy to drive marketing decisions and to maximize the return on investment. The strategy consists of market (survey) research, competitive analysis, and in-market testing (where the latter is also known as database marketing or sometimes customer relationship management).

 DOI: 10.1201/9780429172588-7

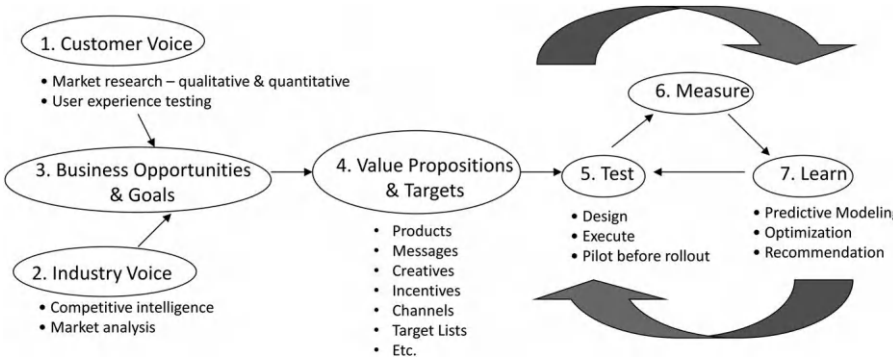


FIGURE 7.1
Test and learn strategy.

Figure 7.1 describes the process of the Test and Learn Strategy, explained below:

1. **Customer Voice:** The first step is to collect customer inputs from market research through qualitative focus groups and quantitative surveys. The goal is to come up with a solid list of value propositions to start with (such as new products, features, messages, etc.). Some corporations employ advanced quantitative market research methods such as conjoint analysis and discrete choice analysis to scientifically derive a list of value propositions, and such methods typically employ advanced experimental designs, similar to those described in Section 7.2.5. Others rely on usability testing or user-experience testing of value propositions, recruiting test participants to provide feedback in a lab environment.
2. **Industry Voice:** The next step is to understand what is going on in the industry. The tools include competitive analysis to uncover the strengths and weaknesses of your competitors versus yours (e.g., the BCG matrix and GE/McKinsey matrix) and market analysis to identify the opportunities in various market segments as well as the trends in the industry. Practical tools that are commonly employed in this area include the SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis; see, for example, Rao and Steckel (1998), Luecke (2005), and Campbell et al. (2014).
3. **Business Opportunities and Goals:** Customer Voice and Industry Voice can provide solid inputs to uncover opportunities or customer needs in market segments that can be met by your products, which can lead to certain business goals. This step also translates to what exactly you are trying to achieve.

4. **Value Propositions:** Based on the above steps, a business can list a set of potential value propositions to meet customer needs. In other words, this is the step to come up with a prioritized list of things to be tested. We should also review what has been done in the past and what we have already learned from past experience. The list of potential value propositions developed becomes the "treatments" for testing.
5. **Test:** Conduct actual in-market tests (as opposed to customer surveys) to test value propositions through randomized experiments (treatment and control). At this stage, some Call-To-Action (CTA) metrics should be established for goal measurement.
6. **Measure:** Measure the results of the in-market tests through statistical measurement related to the CTA metrics. Multiple metrics can be used for measurement, from inquiry rates to purchase rates to Lifetime Value; see Section 7.3 for a discussion of measurement metrics.
7. **Learn:** Apply Uplift modeling to refine targets for the next campaign. When multiple treatments are available, optimization methods can be employed to find the right treatment for the right customer (to be described in Chapter 8).

The Test and Learn procedures in steps 5–7 are explained in detail in Section 7.2.2.

To illustrate the above process with an example, suppose you manage a credit card business line. Steps 1 and 2 help you identify the opportunities: Is the product already popular in certain areas (country/state/province)? If not, would customers be receptive to your product? Step 3 allows you to identify your goals – acquisition, retention, or cross-selling/up-selling – and Step 4 identifies what value propositions may work; for instance, should we be testing different colors (black/blue/green/silver/gold/titanium/diamond), different co-branding strategies (are you partnering with VISA/MasterCard/Amex or another brand, and how to name your co-branded product), various pricing points (introductory rates (aka teaser rates) and long-term annual percentage rates), and different benefits (frequent flyer, rebate, etc.) for some selected market segments (would you want to cover higher-risk groups, are you more interested in younger customers, etc.)? Steps 5–7 test your value propositions and help you refine them in in-marketing tests. Additionally, risk consideration should also be included in this analysis – for example, a higher price (interest rate) may be required for a higher-risk group, a diamond card may be best for a more affluent lower-risk group, and so on.

7.2.2 Test and Learn (Database Marketing Campaign) Process for In-Market Testing

This section describes the core Test and Learn campaign process, which is the *in-market* testing or the *Test-Measure-Learn* portion of Figure 7.1. The process

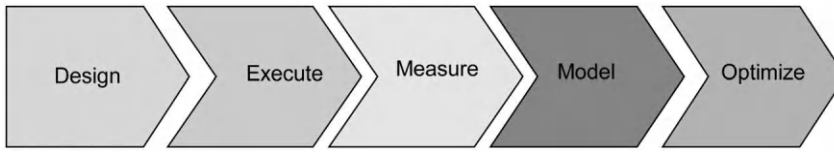


FIGURE 7.2
Test and learn database marketing campaign process.

comprises various components, from campaign design to response modeling to campaign optimization, as set out in Figure 7.2.

The various components of the campaign process are explained below:

1. **Design:** Designing a campaign requires selecting targets (whether randomly, by business rules, segmentations, or predictive models) and splitting them into treatment and control randomly so that each individual in the treatment group has the same probability of being selected and similarly for individuals in the control group. One has to handle issues such as sampling and sample size determination (to be explained in Section 7.2.3). A/B testing and advanced experimental design methods are quite commonly employed (to be introduced in Sections 7.2.4 and 7.2.5).
2. **Execute:** This involves physically selecting a list of targets and distributing them to the contact channel (e.g., direct mail, email, telemarketing, online, in-person visits).
3. **Measure:** The objective here is statistically to measure the campaign results in a way that is meaningful to your business and so to determine whether the campaign has generated any success, that is, lift over control. Marketers are often not only interested in the overall campaign success but also whether it has worked by customer segment or subgroup. Additionally, there are many possible measurement metrics; see Section 7.3 for a discussion of various metrics.
4. **Model:** Develop uplift models to identify characteristics of individuals who responded due to the treatment. Details of uplift modeling are described in Chapter 6 and also in Chapter 9.
5. **Optimize:** Apply the developed model to select appropriate targets for the next campaign. Sophisticated campaigns may include more than one treatment, resulting in an opportunity to optimize treatment at the individual level (to be discussed in Chapter 8).

The simplest and perhaps most well-known Test and Learn process is *A/B Testing*, where two treatments (treatment versus control or treatment A versus B) are compared in a test. For example, A/B Testing is frequently used in online advertising, where two methods of promotion are compared in a randomized test (see, e.g., Kohavi et al. 2012). Measuring whether A or B is

a better treatment is the fundamental step. Modeling (Uplift Modeling) will take it to the next level to determine the right treatment for the right targets. A/B Testing compares only two treatments at a time, and if more treatments are available, the winner of each comparison can compete with other treatments, and so on in a sequential fashion (see, e.g., Goward 2013, McFarland 2013, or Siroker and Koomen 2013 in the marketing context). More discussion on A/B Testing is presented in Section 7.2.4, and advanced methods of testing multiple treatments simultaneously will be discussed in Section 7.2.5.

7.2.3 Sampling and Sample Size Determination

In order to develop uplift models, we need to have random treatment and random control groups, and ideally, they are a random sample or a good representative sample of the larger relevant population. The most commonly used sampling schemes in business are simple random sampling and stratified random sampling (see classical texts such as Scheaffer et al. 1990, or Cochran 1977, for details). Both are straightforward to implement using common campaign software packages or statistical software. Stratified random sampling is helpful in guaranteeing enough subsamples for each subpopulation and is particularly useful when the subpopulations are relatively small. In the following Subsections 7.2.3.1 and 7.2.3.2, we mainly focus on sample size determination in the context of uplift analytics. On a similar topic, Section 7.2.4 discusses randomization *between* treatment and control.

7.2.3.1 Standard Sample Size Determination

We revisit the standard method for sample size determination for two proportions in this subsection. Assume we have two population proportions p_1 and p_2 (for two target groups 1 versus 2, or for treatment versus control), and the pair of hypotheses[1] are: H_0: $p_1 = p_2$ and H_1: $p_1 > p_2$. Our goal here is to determine the appropriate sample sizes for the two groups, n_1 and n_2, such that the Type I and Type II errors (and their probabilities of occurrence, α and β, respectively) are well balanced; see Table 7.1 for their definitions and further discussion in Chapter 2.

TABLE 7.1

Definitions of Type I and Type II Errors (and Their Probabilities of Occurrence, α and β in Hypothesis Testing

	Do Not Reject H_0	**Reject H_0**
If H_0 is true	Correct decision $(1-\alpha)$ ☺	Type I error (α) ☹ [false negative]
If H_0 is not true	Type II error (β) ☹ [false positive]	Correct decision $(1-\beta)$ ☺

It can be shown that, given the values of α and β the sample sizes $n1$ and $n2$, would satisfy:

$$\frac{\left(p_1 - p_2\right)^2}{\left(z_\alpha + z_\beta\right)^2} = \frac{p_1\left(1-p_1\right)}{n_1} + \frac{p_2\left(1-p_2\right)}{n_2}, \tag{7.1}$$

where z_α and z_β are $\Phi^{-1}\left(1-\alpha\right)$ and $\Phi^{-1}\left(1-\beta\right)$, respectively, representing the standard normal critical values associated with the Type I and Type II errors. For example, if $\alpha = 0.05$, $z_\alpha = 1.645$, and if $\beta = 0.20$, $z_\beta = 0.842$.

If $n_1 = n_2 = n$ then with some algebraic manipulation of Eqn. (7.1), the minimum sample size would follow the standard formula:

$$n = \frac{\left(z_\alpha + z_\beta\right)^2\left[p_1\left(1-p_1\right) + p_2\left(1-p_2\right)\right]}{\left(p_1 - p_2\right)^2} \tag{7.2a}$$

Alternatively, if n_1 is given, we can determine the minimum sample size for the second group, again from Eqn. (7.1):

$$n_2 = \frac{p_2\left(1-p_2\right)}{\frac{\left(p_1 - p_2\right)^2}{\left(z_\alpha + z_\beta\right)^2} - \frac{p_1\left(1-p_1\right)}{n_1}} \tag{7.2b}$$

The derivation of Eqns. (7.1), (7.2a), and (7.2b) can be found in most standard statistics books and will be omitted here. Equation (7.2a) is easy to apply provided that some reasonable estimates of the two population proportions, p_1 and p_2, are available. It is also implemented in standard sample-size tools such as nQueryAdvisor. To estimate the proportions, if historical data are not available, the usual practice is to use sensitivity analysis with a reasonable range of values to arrive at a range of minimum sample sizes and select the largest sample size from the range. See Mathews (2010) for determining sample sizes in a variety of situations.

7.2.3.2 Sample Size Determination for Uplift Analytics

For Uplift/True-lift analytics, we need to address the difference between treatment and control. Additionally, we may want to know if there is a difference in lift (treatment minus control) by some demographic group. As in Section 7.2.3.1, let's suppose we have two population groups (e.g., older and younger age groups, male and female, or higher versus lower income groups) and a treatment and control split within each group, which means we have four groups to handle. We now generalize the formulas in Section 7.2.3.1 to handle a four-group proportion comparison.

Our objective is to determine appropriate sample sizes that allow us to statistically test whether the (treatment minus control) lift in group 1 is equal to or better than the lift in group 2, which leads to the following pair of hypotheses:

$$H_0\colon p_{1t} - p_{1c} = p_{2t} - p_{2c} \quad \text{and} \quad H_1\colon p_{1t} - p_{1c} > p_{2t} - p_{2c}$$

where p_{1t}, p_{1c}, p_{2t}, and p_{2c} are response rates in group 1's treatment subgroup, group 1's control subgroup, group 2's treatment subgroup, and group 2's control subgroup, respectively. Rejection of H_0 would indicate that the demographic split between the two groups is effective in differentiating between higher and lower lift. For simpler notation, we can rewrite the above hypotheses as:

$$H_0\colon \Delta p_1 = \Delta p_2 \quad \text{and} \quad H_1\colon \Delta p_1 > \Delta p_2,$$

where Δ denotes the lift (incremental difference) between treatment and control. It can be shown (in Appendix 7.1) that:

$$\text{Power} = \text{Probability of detecting a difference given that } H_1 \text{ is true}$$

$$= 1 - \text{P}\left(\text{Type II Error}\right) = 1 - \Phi\left(z_\alpha - \frac{\left(\Delta p_1 - \Delta p_2\right)}{I}\right) \tag{7.3}$$

where $(\Delta p_1 - \Delta p_2)$ is estimated by its sample estimate, $(\Delta \hat{p}_1 - \Delta \hat{p}_2)$.

$$\text{Define } J = \hat{p}_{1t}\left(1-\hat{p}_{1t}\right) + \hat{p}_{1c}\left(1-\hat{p}_{1c}\right)R_1 + \hat{p}_{2t}\left(1-\hat{p}_{2t}\right)R_t + \hat{p}_{2c}\left(1-\hat{p}_{2c}\right)R_tR_2 \tag{7.4}$$

where the following ratios chosen by the analyst are input parameters to the sample design:

$$R_1 = \frac{n_{1t}}{n_{1c}}, \quad R_2 = \frac{n_{2t}}{n_{2c}}, \quad \text{and} \quad R_t = \frac{n_{1t}}{n_{2t}} \tag{7.5}.$$

In Eqn. (7.5), R_1 is the ratio of treatment to control in Group 1, R_2 is the ratio of treatment to control in Group 2, and R_t is the ratio of Group 1 to Group 2 among those in treatment.

Then n_{1t}, n_{1c}, n_{2t}, and n_{2c} can be obtained as follows:

$$n_{1t} = \frac{J\left(z_\alpha + z_\beta\right)^2}{\left(\Delta p_1 - \Delta p_2\right)^2},$$

$$n_{1c} = \frac{n_{1t}}{R_1}, \quad n_{2t} = \frac{n_{1t}}{R_t}, \quad \text{and} \quad n_{2c} = \frac{n_{2t}}{R_2} = \frac{n_{1t}}{R_tR_2}, \tag{7.6}$$

where J is defined in Eqn. (7.4), and R_1, R_2, and R_t are defined in Eqn. (7.5). Appendix 7.1 proves the sample size determination formulas in Eqns. (7.3) and (7.6). Although Eqns. (7.3) and (7.6) may look a bit tedious to use, they are implemented in an Excel spreadsheet for easy computation, as illustrated below.

Example 7.1a Minimum Sample Sizes for Two Groups (with Treatment minus Control Lift)

To prepare for an upcoming marketing program, you are asked to provide the minimum sample size to test the response rate resulting from contacting two target groups (1 and 2) versus their corresponding no-contact control groups. That is, we would like to test the following pair of hypotheses:

$$H_0\text{: } p_{1t} - p_{1c} = p_{2t} - p_{2c} \text{ and } H_1 \; p_{1t} - p_{1c} > p_{2t} - p_{2c}$$

Or equivalently, H_0: $\Delta p_1 = \Delta p_2$ and H_1: $\Delta p_1 > \Delta p_2$, where again Δ denotes the incremental difference (lift) between treatment and control. Your hypothesis is that target group 1 may have a higher response rate than group 2.

You have found that historically, the response rate for target group 2 without contacting customers (i.e., control response rate) is about 0.5%. While there was insufficient historical data on response rate with a contact (i.e., treatment response rate), you are willing to assume that with a contact the response rate will increase by about 30%, that is, from 0.5% to 0.65%, a 0.15% absolute increase. For the better target group (group 1), you have found that the response rate is about 20% higher than group 2's. Also, target group 1 is about half the size of target group 2 in the population, so you may choose to sample them with the same ratio. Additionally, you will need to have some idea about the significance level (α) and the power desired, and let's assume you are fine with $\alpha = 0.15$ and power ≥ 0.75. So you enter these input numbers into a spreadsheet that has Eqn. (7.6) implemented; see Table 7.2a.[2]

The shaded cells in the INPUT BOX in Table 7.2a are to gather user inputs. The shaded area in the OUTPUT BOX provides the output, which are the minimum sample sizes for target groups 1 and 2, by treatment and control. Suppose that after examining the numbers, you feel the quantities may be too high. You are now thinking that a 20% higher treatment response rate over control may be too low. Or if indeed the difference is that small, you may find it acceptable if statistical significance is not detected. So you try a 50% higher treatment response rate over control in Table 7.2b; that is, what are the sample sizes required so that H_0 can be rejected with a 50% higher treatment response rate over control (and target group 1's response rate remains 20% higher than target group 2's)? As expected, the minimum sample sizes required are now lower and may be considered reasonable.

TABLE 7.2a

Sample Size Determination Example: With a 20% Increase in Treatment Response Rate Over Control

Sample Size Requirement for Comparing Two Pairs of Treatment and Control Groups Find Sizes Given Power – One-Tailed Test			
INPUT BOX	group 1 (the better one)	est. treatment rate	0.7800%
		est. control rate	0.6000%
	group 2	est. treatment rate	0.6500%
		est. control rate	0.5000%
	Stat limits	significance level (alpha)	0.15
		Power	0.75
	Ratios	ratio of treatment to control for group 1	1
		ratio of treatment to control for group 2	1
		ratio of group 1 treatment to group 2 treatment	0.5
OUTPUT BOX	computations	critical value of alpha	1.036
		critical value of beta	0.674
		expected 4-way difference	0.03%
		product of ps and qs (Intermediate calculation)	0.0194195
	size required	group 1 treatment	631,622
		group 1 control	631,622
		group 2 treatment	1,263,244
		group 2 control	1,263,244

After proposing these quantities to your marketing partners, you are now told that the quantities are reasonable, but they would be happier if you could lower the quantities slightly to 200 K for each of target group 1's treatment and control and 400 K for each of target group 2's treatment and control. Now your task is to check the power associated with these numbers, which is a reverse calculation as before, and Eqn. (7.3) would come in handy. Similar to Eqn. (7.6), Eqn. (7.3) is also implemented in the same spreadsheet, and the result with these sample sizes is in Table 7.2c, which shows a power of 69.3%, a slight decrease from the original 75% in Table 7.2b that you feel is acceptable.

Recall that Eqn. (7.6) specifies the requirement to ensure the sample sizes are sufficiently large for H_0: $\Delta p_1 = \Delta p_2$ and H_1: $\Delta p_1 > \Delta p_2$, which concerns the lift of two groups. In fact, Eqn. (7.6) can be adapted to the situation for Uplift Modeling. Before we build an uplift model, we would typically like to see that the top decile[3] exhibits a higher lift than the overall average lift. Our objective now is to determine appropriate sample sizes such that the

TABLE 7.2b

Sample Size Determination Example: With a 50% Increase in Treatment Response Rate Over Control

Sample Size Requirement for comparing Two Pairs of Treatment and Control Groups Find Sizes Given Power – One-Tailed Test			
INPUT BOX	group 1 (the better one)	est. treatment rate	0.9000%
		est. control rate	0.6000%
	group 2	est. treatment rate	0.7500%
		est. control rate	0.5000%
	Stat limits	significance level (alpha)	0.15
		Power	0.75
	Ratios	ratio of treatment to control for group 1	1
		ratio of treatment to control for group 2	1
		ratio of group 1 treatment to group 2 treatment	0.5
OUTPUT BOX	computations	critical value of alpha	1.036
		critical value of beta	0.674
		expected 4-way difference	0.05%
		product of ps and qs (Intermediate calculation)	0.0210924
	size required	group 1 treatment	246,971
		group 1 control	246,971
		group 2 treatment	493,943
		group 2 control	493,943

TABLE 7.2c

Sample Size Determination Example: Determine Power Given Sample Sizes

Find Power Given Sizes – One-Tailed Test (i.e., Testing If One Group Is Better Than the Other)			
INPUT BOX	group 1 (the better one)	est. treatment rate est. control rate treatment size control size	0.9000% 0.6000% 200,000 200,000
	group 2	est. treatment rate	0.7500%
		est. control rate	0.5000%
		treatment size	400,000
		control size	400,000
		significance level (alpha)	0.15
OUPUT BOX	computations	critical value of alpha	1.036
		expected 4-way difference	0.05%
		4-way s.d.	0.000324749
	Result	power	69.3%

data will allow us to statistically detect the difference in lift between the top decile and the overall average:

H_0: $\Delta p_1 = \Delta p$ and H_1: $\Delta p_1 > \Delta p$, where in this case Δp_1 represents the lift in the top decile and Δp is the overall average lift across the entire sample. This pair of hypotheses is equivalent to the following pair:

H_0: $\Delta p_1 = \Delta p_2$ and H_1: $\Delta p_1 > \Delta p_2$, where again Δp_1 represents the lift in the top decile and Δp_2 *now* represents the lift in the rest of the deciles (i.e., deciles 2–10).

Assuming the treatment size, n_t is given, our aim is to find the appropriate value of the control size, n_c. Appendix 7.1 shows that:

$$n_c = \frac{\frac{\hat{p}_{1c}\left(1-\hat{p}_{1c}\right)}{0.1} + \frac{\hat{p}_{2c}\left(1-\hat{p}_{2c}\right)}{0.9}}{\left(\frac{\Delta p_1 - \Delta p_2}{z \quad + z}\right)^2 - \frac{1}{n_t}\left[\frac{\hat{p}_{1t}\left(1-\hat{p}_{1t}\right)}{0.1} + \frac{\hat{p}_{2t}\left(1-\hat{p}_{2t}\right)}{0.9}\right]} \tag{7.7}.$$

In Eqn. (7.7), the user inputs include estimates of treatment and control response rates in the top decile and the rest of the deciles.

In addition to making sure the top decile can be statistically tested to be stronger than the overall sample in terms of response rate, one may also want to make sure that the top decile itself has enough sample size to detect statistical significance between treatment and control (within the top decile). The pair of hypotheses in this case is: H_0: $p_{1t} = p_{1c}$ and H_1: $p_{1t} > p_{1c}$, where p_{1t} and p_{1c} are treatment and control response rates in the top decile. Again, assuming the overall treatment size, n_t is given, this can be accomplished by applying Eqn. (7.2b) to the top decile:

$$n_{1c} = \frac{p_{1c}\left(1-p_{1c}\right)}{\frac{\left(p_{1t}-p_{1c}\right)^2}{\left(z_\alpha + z_\beta\right)^2} - \frac{p_{1t}\left(1-p_{1t}\right)}{10n_t}}. \tag{7.8a}$$

Equivalently, the overall minimum control size, which by definition of decile is 10 times the size of the top decile, becomes:

$$n_c = 10n_{1c} = \frac{10\,p_{1c}\left(\quad -p_{1c}\right)}{\frac{\left(p_{1t}-p_{1c}\right)^2}{\left(z_\alpha + z_\beta\right)^2} - \frac{p_{1t}\left(\quad -p_{1t}\right)}{10n_t}}. \tag{7.8b}$$

If both Eqns. (7.7) and (7.8b) need to be satisfied, the minimum overall control sample size will be the greater of the two values determined by Eqns. (7.7) and (7.8b).

Example 7.1b (Continuation of Example 7.1a)

In Example 7.1a, we have determined the sample sizes required for target groups 1 and 2 by treatment and control in order to detect that target group 1 is better than target group 2 in terms of lift (treatment over control response rate). Suppose target group 2 is close to random targeting, and the plan with this group is that after campaign response data are back, you would like to use group 2's data to develop an uplift model to improve future targeting. The question is now: What is the control size required in group 2 to make sure that (1) the top decile of an uplift model has enough sample size to detect statistical significance between treatment and control response rates, and (2) the top decile response rate can be statistically detected to be stronger than the overall sample's response rate? These two questions, which are associated with two pairs of hypotheses, can be answered by Eqns. (7.8b) and (7.7), respectively. As in the previous example, these equations have been implemented in a spreadsheet for ease of computations. Table 7.3 shows the spreadsheet where the INPUT BOX gathers the overall (baseline) treatment and control rates as well as the assumed ratio of top decile response rate to

TABLE 7.3

Determining Sample Size That Can Meet Two Lift Criteria

True Lift/Uplift Modeling Requirement for Marketing Programs			
INPUT BOX	overall treatment rate	0.750%	estimated value
	overall control rate	0.500%	estimated value
	overall treatment size	400,000	normally known
	ratio of top decile response rate to random	3	estimated value
	alpha	0.05	
	power	0.75	
	critical value of alpha	1.645	
	critical value of beta	0.674	
OUTPUT BOX 1: Top decile lift > 0	est treatment rate in top decile	2.250%	
	est control rate in top decile	1.500%	
	numerator	0.01478	
	denominator	9.9068E-06	
	overall control size required	14,914	
OUTPUT BOX 2: Top decile lift > bottom 9 deciles lift	est treatment rate in bottom 9 deciles	0.583%	
	est control rate in bottom 9 deciles	0.389%	
	4-way difference (lift in top – lift in bottom 9)	0.556%	
	numerator	0.15205	
	denominator	5.172E-06	
	overall control size required	29,402	

random. The OUTPUT BOX 1 and 2 provide the minimum control size for the two goals above in the shaded areas. In this example, they both happen to be relatively mild compared to the size requirement in Example 7.1a. In practice, one may want to try different sets of input values to make sure that the sample size will be sufficient for a range of possible inputs.

The above methods are technically only applicable for a single treatment and a single control case. For more general situations where there are many treatment groups, the above are still valid when we consider all treatments collectively as one treatment group. We will discuss estimability[4] for individual treatment effects using simulation later in Subsection 7.2.5.2.2.

7.2.4 A/B Testing and A/B/n Testing for Campaign Design

7.2.4.1 Randomization and A/B Testing

All experimental design methods rely on randomization, that is, randomly assigning experimental units to treatment and control groups or several treatment groups. Randomization can be done with simple random number generation on a computer (see Chapter 3 for an introduction). The biggest advantage of randomization for causal measurement is that it ensures that the characteristics between the treatment and control groups are the same prior to the design, that is, balanced in both observable and unobservable covariates.

The most popular method associated with experimental design in modern-day business is A/B Testing, where only two options are compared – treatment versus control, challenger versus champion, or simply A versus B. When more than two options are compared, it is called A/B/n Testing.

Consider a retail company selling shoes that can be promoted with two different messages (treatment versus control) on their website, where:

- Control (current way) emphasizes their durability and all-weather features, and
- Treatment (new idea) focuses on their multi-functional features for both business and leisure.

One can also package other differences in the two options. For example, the current price in the control group may range from $89.99 to $109.99; the treatment price can be lowered to $79.99–$99.99. Obviously, compounding multiple factors in only two groups would not allow us to separate the price effect from the promotional message. So A/B testing (or A/B/n testing for more than two attribute levels) only allows us to test one attribute at a time (or a combined set of attributes without the ability of separating their effects). For testing three attributes or more, more sophisticated designs are required, and we will cover this more advanced topic (Multivariate Testing or MVT) in Section 7.2.5.

7.2.4.2 Randomized Block Design

A classical method that is highly practical but may be less known in business applications is Randomized Block Design (also known as Stratified Randomization), where treatment and control (or A versus B) are randomly assigned to individuals (experimental subjects) *within* each block (or stratum or group). Instead of assigning treatment and control to the entire study population completely randomly, random assignment of treatment and control under the Randomized Block Design is done at the block level – for instance, separately for men and women. These blocks are preselected by experimental designers with the prior knowledge that the block factor has a high impact on the outcomes. With very large samples, it is more likely that individuals or experimental units assigned to treatment and control share similar characteristics. However, in smaller samples, it is possible that the treatment and control groups have differences in characteristics; thus, the measurement can be confounded by their differences. The Randomized Block Design would then be an effective technique to minimize the potential differences due to confounding.[5]

Consider a retail chain that is interested in knowing whether blue color (treatment) should be used to replace the current yellow color (control) as the theme color used in stores (e.g., for walls and employee uniforms). In this case, our experimental units are stores as opposed to people. Suppose there are 1,000 stores that are unevenly distributed across the 50 states of the United States. The study is to select 20 stores in the pilot to test whether blue is better than yellow, which will lead to the ultimate decision of possibly changing the theme color for the entire chain. In this example, we will have 10 stores assigned to treatment (blue) and 10 stores assigned to control (yellow). One method is the traditional way of completely random assignment. With only 10 stores randomly assigned to treatment, it is possible that many of them are concentrated in certain parts of the country (e.g., too many of them on the West Coast), and similarly the 10 stores assigned to control could be concentrated in another part of the country. Thus, the treatment and control stores can be confounded with geography (e.g., West Coast versus East Coast). Using the Randomized Block Design, we would first divide the country geographically, say by region, and then randomly assign the stores to treatment and control within each region. Table 7.4 shows a possible assignment with 5 stores to be tested in each region.

TABLE 7.4

Example of Randomized Block Design by Geographic Region

Block	Geographic Region[6]	Random Assignment of Stores to Treatment and Control within Region
1	Northeast	Blue, Yellow, Blue, Yellow, Blue
2	Midwest	Blue, Yellow, Blue, Blue, Yellow
3	South	Blue, Yellow, Yellow, Blue, Yellow
4	West	Yellow, Blue, Blue, Yellow, Yellow

Blocks can also be based on more than one variable. For instance, in the same example, in the situation where not all stores can be tested at the same time, and the total testing period has to be spread across three months, one may include month as an additional variable for blocking (because of possible seasonal effects on sales), so each block represents the combination of a specific geographic region and a particular month. There would be 3 months × 4 regions = 12 blocks in total. Obviously, this can get further complicated if more variables are involved.[7] Again, this method is very useful when the total sample size is relatively small but is less important if one has a huge sample.[8]

7.2.5 Multivariate Testing/Experimental Design for Campaign Design

The design of a marketing campaign is the key starting point of a campaign process. However, it often does not receive enough attention in data science, data mining, and the machine learning literature, where the focus is often on supervised and unsupervised learning techniques on observational data (also sometimes known as "found data") or relatively simple experimental data. It is known that a poorly designed campaign could make learning infeasible, while a scientifically designed one can not only make learning feasible but also maximize learning opportunities. Experimental design has been discussed in statistics and the clinical trial literature for decades, for example, Box et al. (1978), Kirk (1982), Fleiss (1986), Montgomery (1991), Clarke and Kempson (1997), Box (2006), and Friedman et al. (2010). For introductions to experimental design in the business literature, see Almquist and Wyner (2001), Davenport (2009), and Manzi (2012).

The design process includes activities such as sample size determination techniques, which have been covered in Section 7.2.3. We focus on cell design structure – testing various offers along with age, income, or other demographic-liked variables – in this section.

Classical designs often test one attribute (also called factor or variable) at a time, as discussed in Subsection 7.2.4.1. For example, in a smartphone email campaign, they may test a few price levels of the phone. After launching the campaign and waiting for a week or two of measurement period, analysis can tell which price level is associated with the highest sales level. Then another email campaign can be launched to test monthly fees, and a third campaign can test the mail content, and so on. The idea is to test only a SINGLE attribute at a time. While this traditional method is simple to process, it tends to take a longer time (and would be even longer for direct mail campaigns) to realize the ultimate best combination, and, in fact, technically one can never know what the best combination is because only the best level from the first test (for the first attribute) will be incorporated in the second test (for the second attribute), etc. A more efficient way is to structure the cell design such that all these attributes are testable in one campaign. Such design is called MVT in the business literature; see, for example, Holland (2005). We describe the methodology below.

Since at least the 1990s, large banks have done many experimental in-market tests for financial products such as credit cards, mortgage refinancing, home equity lines of credit, and home equity loans. Many offers that regular households receive in their mailboxes are part of large-scale experiments by banks seeking to fine-tune their offers. The financial services firm Capital One was founded by two former management consultants with a strong mindset of experimentation (see Paige 2001, Brunner and Kirchoff 2012, and Manzi 2012). More recently, technology firms such as Google and Amazon are known to routinely carry out experiments. The following are two realistic examples from two different industries.

Example 7.2 Retailer Coupon Campaign Design

Many large retailers are experts in coupon strategies and some primarily rely on couponing for their consumer business. Suppose a large retailer selling mainly clothes would like to test a marketing campaign, as set out in Table 7.5. There are three attributes (also called factors or variables) and two attribute levels for each attribute. The total number of combinations is $2 \times 2 \times 2 = 2^3 = 8$, and all the 8 possible combinations are listed in Table 7.6. Such a design with all possible combinations is called the *Full Factorial Design*.

TABLE 7.5

Retailer Coupon Campaign Design: Attributes and Attribute Levels

Attribute	Attribute Level
Coupon Discount	20% (0), 30% (1)
Frequency	Once a month (0), Twice a month (1)
Mail Design	Postcard (0), Letter (1)

Note: Numbers in parentheses are coded levels.

TABLE 7.6

Full Factorial Design of Retailer Coupon Campaign

Cell	Discount	Frequency	Design
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	1	0	0
6	1	0	1
7	1	1	0
8	1	1	1

To use the design in Table 7.6, the analysts at the Retailer will need to randomly assign the target customers into 8 cells, with each cell receiving one of the 8 possible combinations in the mail. The design is straightforward, but when the number of attributes is larger, such design can get very large. Let's consider another realistic example in the credit card industry.

Example 7.3 Credit Card Marketing Campaign Design

A large bank selling a credit card would like to determine the best combination of treatments for each prospect, and the treatment attributes and attribute levels are summarized in Table 7.7 (note that channel is also incorporated as an attribute in this example). To accomplish this goal, the bank plans to conduct an in-market experiment trying lots of combinations on millions of prospects (noncustomers). The number of all possible combinations in this example with four 4-level attributes and two 2-level attributes = $4^4 \times 2^2 = 1{,}024$. This means, in order to test all possible combinations, we would need 1,024 cells. This would require the bank to divide the targets into 1,024 groups, with each group (cell) receiving a unique treatment combination. For example, the first cell may be APR = 4.95, Credit Limit = $2,500, Color = Green, Rebate = None, Brand = SmartCard, and Channel = Direct mail; the second is the same except that Channel = Email, and so on. While such a test is theoretically feasible, most marketers would consider it administratively too difficult because one would have to have lots of card designs and then randomly put millions of prospects into 1,024 cells. Such design with all possible combinations is called the *Full Factorial Design*.

This full factorial design is conceptually simple and mathematically straightforward to handle (simply multiply the numbers of levels of all attributes together). In practice, however, when the number of attributes is large or even moderate like in this example, the total number of possible combinations can be large, which makes it operationally difficult or infeasible. An alternative is to use a *Fractional Factorial Design*. By definition, the full factorial refers to the design that includes all possible combinations, while a fractional factorial design includes only a subset of the full factorial design (see Almquist and Wyner 2001 for an introduction).

TABLE 7.7

Credit Card Campaign Design: Attributes and Attribute Levels

Attribute	Attribute Level
Annual Percentage Rate (APR)	4.9% (0), 6.9% (1), 9.9% (2), 11.9% (3)
Credit limit	$2,500 (0), $5,000 (1), $8,000 (2), $12,000 (3)
Color	Green (0), Platinum (1), Gold (2), Diamond (3)
Rebate	None (0), 0.5% (1), 1% (2), 1.5% (3)
Brand	SmartCard (0), SuperAdvantage (1)
Channel	Direct mail (0), Email (1)

Note: Numbers in parentheses are coded levels.

There are two types of fractional factorial design:

1. **Orthogonal design,** where all attributes are made orthogonal (uncorrelated) with each other, and
2. **Optimal design,** a more flexible method that can handle complex business requirements, where some criterion related to the covariance matrix of parameter estimates is optimized (see Kuhfeld 1997, 2010) for the applications of SAS PROC FACTEX and PROC OPTEX in market research; Kuhfeld's market research applications are applicable to database marketing).

We discuss both designs in the next section. This example will be continued in Subsections 7.2.5.2 and 7.2.5.3. Before going into the technicalities of the design, we need to introduce some fundamental terminology below.

7.2.5.1 Main and Interaction Effects

We first introduce the concepts of main and interaction effects, which are essential features of any experimental design, starting with an example.

Example 7.4 Airline Premium Membership Experiment

Consider an airline premium membership, which provides premium club services at many airports. The current membership charges an annual fee of $99 and allows customers to have a quicker check-in time, with an average time reduction of 5 minutes. The airline is interested in testing whether customers are willing to pay more for a shorter check-in time, with two levels of annual fee and two levels of check-in time reduction. The full factorial design involves $2 \times 2 = 4$ cells, with customers randomly selected into each of the test cells for promotion (see Table 7.8a). The "control" group, or the current base case in this example, is cell 2 (higher fee and lower time reduction). The metric of interest is customer acceptance (buy or not), with test results summarized in Table 7.8b.

TABLE 7.8a

Design of a Simple Airline Membership Experiment for Two Factors

Cell/Run	Annual Fee	Reduce Average Check-in Time by
1	$99 (+)	15 min (+)
2 (base case)	$99 (+)	5 min (–)
3	$69 (–)	15 min (+)
4	$69 (–)	5 min (–)

Note: +/– indicates the high/low level of each of the two 2-level attributes.

TABLE 7.8b

Test Results of the Airline Membership Experiment: Response (Acceptance) Rate by Fee and Reduction in Time

	Annual Fee = $69	**Annual Fee = $99**
Reduce average check-in time by 5 minutes	3% (Cell 4)	2% (base case, Cell 2)
Reduce average check-in time by 15 minutes	5.2% (Cell 3)	4% (Cell 1)

Table 7.8b shows that both a reduced annual fee and an increase in average time reduction would increase the response rate from the base case. However, the time reduction has a stronger effect than a reduced fee, which implies that if only one factor can be changed, a time reduction would lead to a higher customer acceptance. It is also clear that when both factors are improved, that is, lower fee AND further time reduction, the response rate is the highest (5.2%). The data in Table 7.8b are plotted in Figure 7.3, which clearly shows the benefits of both fee (the lower the better) and time reduction (the more the better). The fee effect and the time reduction effect can be calculated by taking the difference in average response rates (see Table 7.9 for the computations). Note that the almost parallel lines in Figure 7.3 indicate that

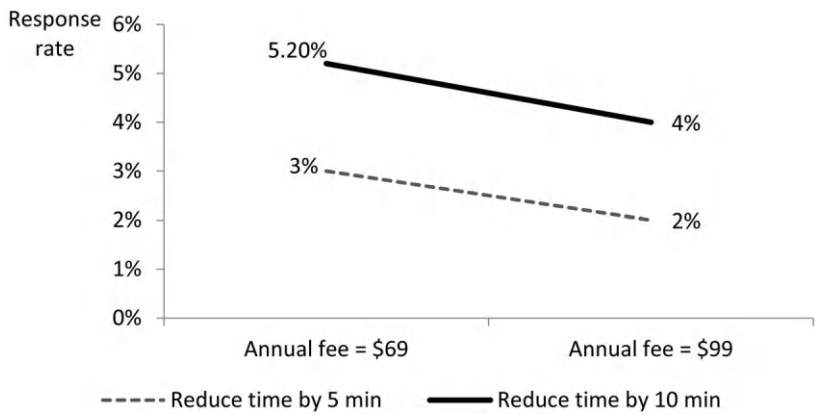


FIGURE 7.3
Response rate by fee and time reduction.

TABLE 7.9

Attribute Effect Computation of the Airline Membership Example

$$\textbf{Fee Effect} = \frac{4\% + 2\%}{2} - \frac{5.2\% + 3\%}{2} = -1.1\%$$

$$\textbf{Time Reduction Effect} = \frac{5.2\% + 4\%}{2} - \frac{3\% + 2\%}{2} = 2.1\%$$

TABLE 7.10

New Test Results of the Airline Membership Experiment: Response (Acceptance) Rate by Fee and Reduction in Time

	Annual Fee = $69	**Annual Fee = $99**
Reduce average check-in time by 5 minutes	3% (y_4)	2% (y_3)
Reduce average check-in time by 15 minutes	6% (y_2)	3% (y_1)

the fee effect does not depend much on the level of time reduction, and also the time reduction effect does not depend much on the level of fee. In other words, there is no or little interaction effect between fee and time reduction.

Let us consider the same example, but the data are now different, as shown in Table 7.10 and its graphical representation in Figure 7.4.

Figure 7.4 shows that the two lines are not parallel, which means the effect of the fee depends on the time reduction level; that is, when the time reduction is low (5 min, dotted line), the effect of fee is relatively low (3% versus 2%, a 1% difference), but when the time reduction is high (15 min, solid line), the effect of fee is stronger (6% versus 3%, a 3% difference). Similarly, the effect of time reduction also depends on the level of fee. Such dependence of one factor's effect on another factor is called the interaction effect, which is defined as the difference between a factor's effects at two different levels of the other factor, and its calculation is shown in Table 7.11. Table 7.12 displays all the possible combinations (also called runs or cells) for this example, with the +/− signs indicating the high/low level of each of the two-level attributes that correspond to the signs in the effect equations in Table 7.11. For example, for the main effect of time reduction in Table 7.11, y_1 and y_2 have a positive sign while y_3 and y_4 have a negative sign, which also shows up in the same

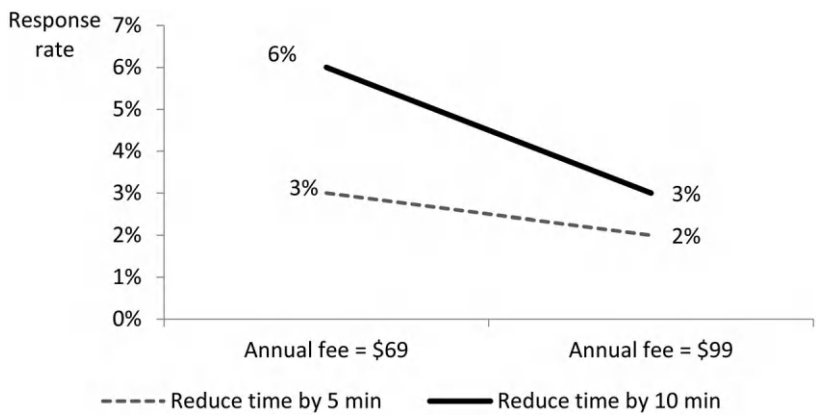


FIGURE 7.4
Response rate by fee and time reduction (from Table 7.10).

TABLE 7.11

Attribute Effect Computation of the Airline Membership Example (from Table 7.10)

$$\textbf{Fee Effect} = \frac{y_1 + y_3}{2} - \frac{y_2 + y_4}{2} = \frac{3\% + 2\%}{2} - \frac{6\% + 3\%}{2} = -2\%$$

$$\textbf{Time Reduction Effect} = \frac{y_1 + y_2}{2} - \frac{y_3 + y_4}{2} = \frac{3\% + 6\%}{2} - \frac{2\% + 3\%}{2} = 2\%$$

$$\textbf{Interaction Effect} = \frac{(y_1 - y_3) - (y_2 - y_4)}{2} = \frac{(y_1 - y_2) - (y_3 - y_4)}{2} = \frac{(3\% - 6\%) - (2\% - 3\%)}{2} = -1\%$$

TABLE 7.12

Full Factorial Design, Including Interaction, for the Airline Membership Example

Run	Time Reduction	Fee	Interaction: Time Reduction × Fee	Outcome
1	+	+	+	y_1
2	+	−	−	y_2
3	−	+	−	y_3
4	−	−	+	y_4

way in Table 7.12. Note that the signs for the interaction column are exactly the product of the time reduction and the fee columns using the common multiplication rule (i.e., − × − = +, + × + = +, and + × − = −). The full factorial design (in this case, only 4 runs or cells) supports estimation of all main and interaction effects as demonstrated in Table 7.11. This can be verified by the signs in Table 7.12, where all three columns (main and interaction) are different and uncorrelated.[9]

To illustrate using a slightly more sophisticated case: If the airline is now interested in testing one more 2-level factor, namely, the style of the membership card: Gold versus Diamond, we will have three factors in total. With a 3-factor full factorial design, we have 2 × 2 × 2 = 23 = 8 runs. Table 7.13 displays the signs of all the main and interaction effects (two-way and three-way). Again, because this is a full factorial design that covers all possible combinations, all the main and interaction effects are estimable as all sign columns are different and uncorrelated with each other.

7.2.5.2 Generating Orthogonal Fractional Factorial Design

Full factorial designs can be made orthogonal,[10] which means that all main and interaction effects are uncorrelated with each other (by definition of orthogonality). However, only carefully constructed fractional factorial designs are orthogonal.

TABLE 7.13

A 3-Factor Full Factorial Design for the Airline Membership Example

	Main Effects			Two-way Interactions			Three-way Interaction
Run	Time Reduction	Fee	Card	Time Reduction × Fee	Time Reduction × Card	Fee × Card	Time Reduction × Fee × Card
1	+	+	+	+	+	+	+
2	+	+	−	+	−	−	−
3	+	−	+	−	+	−	−
4	+	−	−	−	−	+	+
5	−	+	+	−	−	+	−
6	−	+	−	−	+	−	+
7	−	−	+	+	−	−	+
8	−	−	−	+	+	+	−

Let us answer this question first – why do people want an orthogonal design? Because of the advantages listed below:

1. Estimability is completely guaranteed. An orthogonal design guarantees that all main effects and selected desirable interaction effects are estimable, which may be the most important advantage.
2. There is no multicollinearity problem. A classical issue that arises when running a predictive model on correlated variables is that the estimated effects may not necessarily represent the true causal effects (see Chapter 5). Having a full factorial design enables orthogonality of the variables so all parameters can be "cleanly estimated" without interference of other variables. In other words, adding or dropping a variable will not affect the estimated parameter values of other variables.
3. Minimum variance (i.e., maximum precision). Mathematically, it can be proven that having an orthogonal design will achieve the minimum variance of the estimated parameters (or, equivalently, maximum precision). Consider linear models, $E(Y) = \boldsymbol{X\beta}$, where $\boldsymbol{X}$ is now the design matrix (with columns representing the intercept [1's] and independent variables from the design) and $\boldsymbol{\beta}$ is a vector of parameters associated with the intercept and the independent variables. With an orthogonal design, the variance of the estimated parameters can be shown as[11]: $Var\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\boldsymbol{X'X}\right)^{-1} = \sigma^2\boldsymbol{I}$ (where σ is the standard deviation of the error term) because $\boldsymbol{X'X} = \boldsymbol{I}$. As a result, the variance of the estimated parameters is minimized.

In general, nonlinear models (including logistic regression), however, the variance form is less straightforward but can still be written as in the following form: $Var\left(\hat{\beta}\right) = \sigma^2 \left(X'WX\right)^{-1} + O\left(\frac{1}{n}\right)$, where W can be interpreted as a set of weights and n is the total sample size. Some numerical evidence suggests that orthogonal designs work reasonably well even for nonlinear models; see Kuhfeld et al. (1994).

So how are we going to generate a fractional factorial design? There are generally three ways:

1. **By Hand:** This is only feasible for the simplest designs. We will illustrate this in the next Subsection 7.2.5.2.1 in order to introduce the general concept.
2. **By a Computer:** This is now the most practical and common method, as experimental designs are available in many common software packages, including SAS, JMP, and R. We discuss this further in Subsection 7.2.5.2.2.
3. **By Tables:** There are canned tables available for selected designs as in, for instance, Box et al. (1978) or Montgomery (1991).[12] These are rarely used anymore.

Although experimenters typically use computer algorithms for most fractional factorial designs, it is still instructive to illustrate how simple designs can be generated manually.

7.2.5.2.1 Generation by Hand and the Concept of Resolution

Although experimenters typically would use computer algorithms for most fractional factorial designs, it is still instructive to illustrate how simple designs can be generated manually.

Example 7.4 Airline Premium Membership Experiment – Continued

Example 7.4 above generates a full factorial design in Table 7.13, with 3 factors leading to 8 (= 2^3) total combinations. If we want to create a fractional factorial design by taking half of the 8 combinations, that is, reducing 2^3 to $2^3 - 1 = 4$ combinations, we can simply discard 4 combinations and only keep 4 of them. The question is: Which 4 combinations should we keep?

Let's say we are mostly interested in the first two factors: Time Reduction and Fee, and we want to construct a "full" factorial design using just the first two factors first; this would take us back to Table 7.12. Fractional factorial design is typically generated by sacrificing some interaction effects, so in this example, we can generate a fractional design by aliasing the interaction effect

of Time Reduction and Fee with the main effect of the third attribute, Card. The procedure goes as follows:

Step 1: Choose two attributes and write down the full 2^2 factorial as the initial design for the two attributes, for example, Table 7.12.

Step 2: Create aliasing relationships. For a 2^{3-1} design, we can equate the levels of the third attribute to either the positive interaction of the first two attributes or the negative interaction of the first two attributes, that is,, set Card = Time Reduction × Fee or Card = –Time Reduction × Fee, see Table 7.14.

Note that the first half of Table 7.14 has a clear alias, which, by construction, is Card = Time Reduction × Fee, which means we cannot tell the difference between the main effect of Card or the interaction effect of the other two attributes in this design. One can also verify the following additional aliasing relationships (also known as design generators):

$$\text{Fee} = \text{Time Reduction} \times \text{Card}, \text{ and}$$
$$\text{Time Reduction} = \text{Fee} \times \text{Card}.$$

TABLE 7.14

A 3-Factor Full Factorial Design = Sum of Two ½ (Fractional) Factorial Designs

First Half of the Full Factorial Design (Corresponding to Runs 1, 4, 6, 7 in Table 7.13)

Run	Time Reduction	Fee	Card = Time Reduction × Fee
1	+	+	+
2	+	–	–
3	–	+	–
4	–	–	+

Second Half of the Full Factorial Design (Corresponding to Runs 2, 3, 5, 8 of Table 7.13)

Run	Time Reduction	Fee	Card = – Time Reduction × Fee
1	+	+	–
2	+	–	+
3	–	+	+
4	–	–	–

This indicates that all two-way interactions are aliased (completely confounded) with a main effect. What it means is that while the main effects are estimable, the two-way interactions are not, and also the main effects and two-way interactions are confounded with each other. In other words, if the interaction effects are all close to zero (or negligible), all the main effects can be estimated correctly. In fact, if we consider the identity I column, defined as + for all runs, we can use this *defining relation* as a *design generator*: I = Time Reduction × Fee × Card. Then one can multiply both the left and right by Card to achieve: Card = Time Reduction × Fee, because Card × Card = I, since any attribute level multiplied by itself becomes + (for + × + = + and − × − = +). Likewise, we also have Fee = Time Reduction × Card, if we multiply both left and right by Fee, and Time Reduction = Fee × Card, if we multiply both left and right by Time Reduction. This way finds all possible aliases associated with the first half of Table 7.14.

Similarly, in the second half of Table 7.14, we have Card = – Time Reduction × Fee by construction, as well as the following additional aliasing relationships:

$$\text{Fee} = -\text{Time Reduction} \times \text{Card}, \text{ and}$$
$$\text{Time Reduction} = -\text{Fee} \times \text{Card}.$$

The corresponding *defining relation* is: I = –Time Reduction × Fee × Card, which will lead to all three aliasing relationships. This means all two-way interactions are (negatively) aliased with a main effect.

7.2.5.2.1.1 Rule of Resolution

To assess the quality of a fractional factorial design, statisticians use a term called Resolution, *R*. We should understand the *Rule of Resolution* because it is a common term used in tables and computer software (in both user input and computer output).

The rule goes as follows: To check if a *p*-factor effect is aliased with a *q*-factor effect of a resolution *R* design (e.g., a 1-factor effect is simply a main effect and a 2-factor effect is a two-way interaction):

- If $p + q < R$, the *p*-factor and *q*-factor effects are not aliased;
- If $p + q \geq R$, the *p*-factor and *q*-factor effects can be aliased.

Let's take a look at some examples. The previous example in Table 7.14 is a Resolution III design (either we take the first half or the second half of the original 2^3 full factorial design) because all main effects are NOT aliased with any other main effect (1 + 1 < 3), but all two-way interaction effects are aliased with a main effect (2 + 1 = 3), and the two-way interactions may also be aliased with each other (2 + 2 > 3). As a result, it is a Resolution III design and can be denoted as 2^{3-1}_{III}.

Let's consider a Resolution IV design, which means, according to the above rule of Resolution:

- All main effects are estimable as no main effect is aliased with any other main effect (because 1 + 1 < 4).
- No main effect is aliased with any two-way interaction (because 1 + 2 < 4).
- Two-way interactions can be aliased with each other (because 2 + 2 = 4).
- Main effects can be aliased with three-way interactions (because 1 + 3 = 4).

The above example for three attributes is relatively simple, and, in reality, there can be many more attributes, sometimes with constraints, which is why generating fractional factorial designs with a computer program is a more common method, which will be described below.

7.2.5.2.2 Generation by a Computer Program

As mentioned, a computer program is typically required for relatively large designs. To illustrate the capability in SAS/QC,[13] we revisit the Credit Card Campaign Design in Example 7.3. The problem with the design in Table 7.7 is that a full factorial design of this credit card marketing example would result in $44 \times 22 = 1{,}024$ cells. A fractional factorial design requires a tradeoff between size (i.e., number of runs or cells) and the number of interaction effects that can be taken into account. Quite often, it is reasonable to assume that some high-order interaction effects are negligible. If one is interested in only the main effects, all possible two-way interaction effects, and quadratic effects on continuous variables (equivalently, assuming all three-way interaction effects or above are negligible), the following program[14] using SAS PROC FACTEX results in 256 runs[15,16]:

```
/* Efficient way: code four-level vars as two binary vars */
/* All two-way interaction effects as well as quadratic
effects are estimable */
proc factex;
factors apr1 apr2 limit1 limit2 color1 color2 rebate1 rebate2
brand chan;
size design=minimum;
model est=(apr1|apr2 limit1|limit2 color1|color2
rebate1|rebate2 brand chan
apr1|apr2|limit1|limit2 apr1|apr2|color1|color2
apr1|apr2|rebate1|rebate2
apr1|apr2|brand apr1|apr2|chan
limit1|limit2|color1|color2 limit1|limit2|rebate1|rebate2
limit1|limit2|brand
```

```
limit1|limit2|chan
color1|color2|rebate1|rebate2 color1|color2|brand
color1|color2|chan
rebate1|rebate2|brand rebate1|rebate2|chan
brand|chan
);
examine design aliasing;
output out=twoway
      [apr1 apr2]=apr nvals=(0 1 2 3)
      [limit1 limit2]=limit nvals=(0 1 2 3)
      [color1 color2]=color nvals=(0 1 2 3)
      [rebate1 rebate2]=rebate nvals=(0 1 2 3)
      brand nvals=(0 1)
      chan nvals=(0 1);
run;

proc print data=twoway;
run;

/* Completely balanced for EACH variable */
proc freq data=twoway;
table apr limit color rebate brand chan;
run;

/* Also balanced at the multivariate level */
proc freq data=twoway;
table apr*limit*color*rebate*brand*chan/list;
run;
```

Table 7.15a summarizes part of the aliasing structure from the PROC FACTEX output, indicating that some high-order interaction effects are aliased (or confounded) with other interaction effects. The first equation shows what are aliased with the identity column (denoted as 0 as opposed to I in SAS), that is, a defining relation. The truncated design is shown in Table 7.15b for the first 50 runs. The PROC FREQs (frequency distributions) in the above program are used to check and confirm that the attributes are balanced (i.e., all levels of each attribute are equally represented), as expected from an orthogonal design.

Depending on the experience of applying fractional factorial design, some companies may feel fine with 256 combinations, while others may feel it is still too large. It is exactly the situation where optimal design can come to rescue if there is a need to reduce the combinations.

Optimal design aims at optimizing certain statistical criteria such that the variances of estimated parameters are as small as possible (subject to constraints) or, in other words, the estimated parameters are as "precise" as possible. Since the variance-covariance matrix of estimated parameters in a linear model is[17]: $Var\left(\hat{\boldsymbol{\beta}}\right)=\sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, it is reasonable to have a design such

TABLE 7.15a

Aliasing Structure of an Orthogonal Fractional Factorial Design Example, Truncated

Aliasing Structure
0 = apr1*limit1*color1*brand*chan
apr1 = limit1*color1*brand*chan
apr2 = limit2*color2*rebate1*rebate2*chan
limit1 = apr1*color1*brand*chan
limit2 = apr2*color2*rebate1*rebate2*chan
color1 = apr1*limit1*brand*chan
color2 = apr2*limit2*rebate1*rebate2*chan
rebate1 = apr2*limit2*color2*rebate2*chan
rebate2 = apr2*limit2*color2*rebate1*chan
brand = apr1*limit1*color1*chan
chan = apr1*limit1*color1*brand = apr2*limit2*color2*rebate1*rebate2
apr1*apr2 = apr2*limit1*color1*brand*chan
apr1*limit1 = color1*brand*chan
apr1*limit2 = limit1*limit2*color1*brand*chan
apr1*color1 = limit1*brand*chan
apr1*color2 = limit1*color1*color2*brand*chan
apr1*rebate1 = limit1*color1*rebate1*brand*chan
apr1*rebate2 = limit1*color1*rebate2*brand*chan
apr1*brand = limit1*color1*chan
apr1*chan = limit1*color1*brand

TABLE 7.15b

An Orthogonal Fractional Factorial Design Example, Truncated

Run	brand	chan	apr	Limit	color	rebate
1	1	0	0	0	0	0
2	0	1	0	0	0	2
3	0	1	0	0	0	1
4	1	0	0	0	0	3
5	0	1	0	0	2	0
6	1	0	0	0	2	2
7	1	0	0	0	2	1
8	0	1	0	0	2	3
9	0	0	0	0	1	0
10	1	1	0	0	1	2
11	1	1	0	0	1	1
12	0	0	0	0	1	3
13	1	1	0	0	3	0
14	0	0	0	0	3	2

(Continued)

TABLE 7.15b (Continued)

Run	brand	chan	apr	Limit	color	rebate
15	0	0	0	0	3	1
16	1	1	0	0	3	3
17	0	1	0	2	0	0
18	1	0	0	2	0	2
19	1	0	0	2	0	1
20	0	1	0	2	0	3
21	1	0	0	2	2	0
22	0	1	0	2	2	2
23	0	1	0	2	2	1
24	1	0	0	2	2	3
25	1	1	0	2	1	0
26	0	0	0	2	1	2
27	0	0	0	2	1	1
28	1	1	0	2	1	3
29	0	0	0	2	3	0
30	1	1	0	2	3	2
31	1	1	0	2	3	1
32	0	0	0	2	3	3
33	0	0	0	1	0	0
34	1	1	0	1	0	2
35	1	1	0	1	0	1
36	0	0	0	1	0	3
37	1	1	0	1	2	0
38	0	0	0	1	2	2
39	0	0	0	1	2	1
40	1	1	0	1	2	3
41	1	0	0	1	1	0
42	0	1	0	1	1	2
43	0	1	0	1	1	1
44	1	0	0	1	1	3
45	0	1	0	1	3	0
46	1	0	0	1	3	2
47	1	0	0	1	3	1
48	0	1	0	1	3	3
49	1	1	0	3	0	0
50	0	0	0	3	0	2

Note: See Table 7.7 for attribute level definitions.

that $\left(\boldsymbol{X'X}\right)^{-1}$ is small or $\boldsymbol{X'X}$ is large. (Note that the inverse of $Var\left(\hat{\boldsymbol{\beta}}\right)$ is known as the Fisher Information Matrix, which, by definition of "information," is the larger the better.) That said, since $\boldsymbol{X'X}$ is a matrix, not a scalar, how do we determine a matrix is large? A common way is to calculate its determinant: $\left|\boldsymbol{X'X}\right|$. Maximizing $\left|\boldsymbol{X'X}\right|$ subject to a candidate set of design runs is called the

D-optimal design.[18] To construct a D-optimal design, we first need to provide a starting set of runs, known as the candidate set of design runs. A common way is to start with the full factorial design as the candidate set, or an orthogonal fractional factorial design, which is already smaller. Kuhfeld (1997, 2010) gives details of optimal design and its implementation using PROC OPTEX in SAS/QC.

To continue with the credit card design example in Table 7.7, let's start with Table 7.15b as the candidate set for the D-optimal design.[19] Running the following program using PROC OPTEX[20] in SAS/QC results in only 37 runs, a substantial reduction from the orthogonal design of 256 runs or the original full factorial design of 1,024 runs:

```
proc optex data=twoway;
class color brand chan; /* nominal variables */
model apr|limit|color|rebate|brand|chan@2 apr*apr limit*limit
rebate*rebate;
generate n=saturated method=m_fedorov;
output out=optdesign;
run;

proc print data=optdesign;
run;
```

The output of this optimal design in Table 7.16a shows various efficiency measures (the higher the better) – the best D-efficiency design (with a D-efficiency measure of 56.8%) in Design Number 1 is selected as the final design, which only has 37 runs as printed in Table 7.16b.

Once a design is obtained, the next step is to check whether the required effects are truly estimable. The following SAS code first transforms the coded levels of continuous attributes from 0-3 to the actual values, followed by centering to reduce correlations,[21] and the creation of all two-way interaction terms. Then, correlation analysis (PROC CORR) and a linear model procedure (PROC GLM) are used to make sure the correlation coefficients are reasonable[22] and the required effects are really estimable. The beginning part of PROC GLM output is shown in Table 7.17,

TABLE 7.16a

Statistical Output of an Optimal Design Example (Truncated)

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	56.7902	18.9995	45.7578	1.3367
2	55.9370	16.4208	45.8300	1.4018
3	55.7041	13.9121	40.3855	1.4856
4	55.6898	18.1791	47.2150	1.3757
5	55.6537	17.8955	49.3404	1.3982

TABLE 7.16b

An Optimal Design Example

Run	apr	limit	rebate	color	brand	chan
1	0	0	2	3	0	0
2	0	0	3	3	1	1
3	0	0	2	2	1	0
4	0	0	3	2	0	1
5	0	0	3	1	0	0
6	0	0	1	0	0	1
7	0	0	3	0	1	0
8	0	1	0	2	1	1
9	0	3	0	3	1	0
10	0	3	2	3	0	1
11	0	3	0	2	0	0
12	0	3	0	1	0	1
13	0	3	2	1	1	0
14	0	3	0	0	1	1
15	1	0	0	3	0	1
16	1	0	0	1	1	0
17	1	3	0	2	1	0
18	1	3	3	0	0	1
19	2	1	3	3	1	0
20	2	2	0	1	0	0
21	2	3	3	2	1	1
22	2	3	2	1	0	1
23	2	3	0	0	0	0
24	3	0	0	3	1	0
25	3	0	0	2	0	0
26	3	0	2	2	1	1
27	3	0	0	1	0	1
28	3	0	0	0	1	1
29	3	0	2	0	0	0
30	3	1	3	1	1	1
31	3	2	3	3	0	1
32	3	2	2	2	0	0
33	3	3	0	3	1	1
34	3	3	1	3	0	0
35	3	3	0	2	0	1
36	3	3	3	1	0	0
37	3	3	3	0	1	0

Note: See Table 7.7 for attribute level definitions.

TABLE 7.17

PROC GLM Output(Beginning Part) for Checking Estimability of the D-Optimal Design (Lack of Equations Below Indicates No Effects Are Linked and All Parameters Are Estimable)

General Form of Estimable Functions	
Effect	**Coefficients**
aprn	L1
limitn	L2
color1	L3
color2	L4
color3	L5
rebaten	L6
brand	L7
chan	L8
aprnsq	L9
limitnsq	L10
rebatensq	L11
aprnlimitn	L12
aprncolor1	L13
aprncolor2	L14
aprncolor3	L15
aprnrebaten	L16
aprnbrand	L17
aprnchan	L18
limitncolor1	L19
limitncolor2	L20
limitncolor3	L21
limitnrebaten	L22
limitnbrand	L23
limitnchan	L24
rebatencolor1	L25
rebatencolor2	L26
rebatencolor3	L27
brandcolor1	L28
brandcolor2	L29
brandcolor3	L30
chancolor1	L31
chancolor2	L32
chancolor3	L33
rebatenbrand	L34
rebatenchan	L35
brandchan	L36

which indicates that all effects are estimable (otherwise, one would see linear equations between some of the variables).

```
data optdesign; set optdesign;
cell=_n_;
run;

data try;
set
optdesign;

y=rannor(-1);

select (apr);
      when (0) aprn=4.9;
      when (1) aprn=6.9;
      when (2) aprn=9.9;
      otherwise aprn=11.9;
end;

/* limitn = actual limit/1000 */
select (limit);
      when (0) limitn=2.5;
      when (1) limitn=5;
      when (2) limitn=8;
      otherwise limitn=12;
end;

select (rebate);
      when (0) rebaten=0;
      when (1) rebaten=0.5;
      when (2) rebaten=1;
      otherwise rebaten=1.5;
end;

/* For color: color1=platinum, color2=gold, color3=diamond,
base color = green */
select (color);
      when (0) do; color1=0; color2=0; color3=0; end;
      when (1) do; color1=1; color2=0; color3=0; end;
      when (2) do; color1=0; color2=1; color3=0; end;
      otherwise do; color1=0; color2=0; color3=1; end;
end;

/* Centering to reduce correlations */
aprn=aprn-mean(4.9,6.9,9.9,11.9);
limitn=limitn-mean(2.5,5,8,12);
rebaten=rebaten-mean(0,0.5,1,1.5);
```

```
/* Squared terms for testing quadratic effects */
aprnsq=aprn*aprn;
limitnsq=limitn*limitn;
rebatensq=rebaten*rebaten;

/* Two-way interaction terms */
/* Consider all the following interaction terms:
apr*limit apr*color apr*rebate apr*brand apr*chan
limit*color limit*rebate limit*brand limit*chan
color*rebate color*brand color*chan
rebate*brand rebate*chan
brand*chan */

aprnlimitn=aprn*limitn;
aprncolor1=aprn*color1; aprncolor2=aprn*color2;
aprncolor3=aprn*color3;
aprnrebaten=aprn*rebaten;
aprnbrand=aprn*brand;
aprnchan=aprn*chan;

limitncolor1=limitn*color1; limitncolor2=limitn*color2;
limitncolor3=limitn*color3;
limitnrebaten=limitn*rebaten;
limitnbrand=limitn*brand;
limitnchan=limitn*chan;

rebatencolor1=rebaten*color1; rebatencolor2=rebaten*color2;
rebatencolor3=rebaten*color3;
brandcolor1=brand*color1; brandcolor2=brand*color2;
brandcolor3=brand*color3;
chancolor1=chan*color1; chancolor2=chan*color2;
chancolor3=chan*color3;

rebatenbrand=rebaten*brand;
rebatenchan=rebaten*chan;

brandchan=brand*chan;

run;


/* Checking correlations of all main and two-way interactions
for:
aprn limitn color rebaten brand chan; */
proc corr data=try;
var
aprn limitn color1-color3 rebaten brand chan
aprnsq limitnsq rebatensq
aprnlimitn aprncolor1-aprncolor3 aprnrebaten aprnbrand aprnchan
```

```
limitncolor1-limitncolor3 limitnrebaten limitnbrand limitnchan
rebatencolor1-rebatencolor3 brandcolor1-brandcolor3
chancolor1-chancolor3
rebatenbrand rebatenchan
brandchan
;
run;

/* Using GLM to check that all required effects are estimable */
proc glm data=try;
model y=
aprn limitn color1-color3 rebaten brand chan
aprnsq limitnsq rebatensq
aprnlimitn aprncolor1-aprncolor3 aprnrebaten aprnbrand
aprnchan
limitncolor1-limitncolor3 limitnrebaten limitnbrand limitnchan
rebatencolor1-rebatencolor3 brandcolor1-brandcolor3
chancolor1-chancolor3
rebatenbrand rebatenchan
brandchan
/ noint e solution;
run;
```

To further ensure estimability when a non-contact control group is included, we can mimic the process of model estimation assuming data are available, with the following steps:

1. Combining the treatment group (with all the attribute combinations) with the control group;
2. Adding demographic variables and replicating the design m times (m often in thousands or tens of thousands);
3. Simulating the outcome response variable with an assumed theoretical model;
4. Adding interaction terms between design and demographic variables; and finally,
5. Checking for estimability as well as statistical significance and values of estimated parameters in a statistical model using simulated data.

Note that replicating the design m times in Step 2 essentially mimics the actual campaign where each cell/run has m replicates, while each replicate corresponds to an individual with certain demographic variables. Step 5 allows us to check not only whether the parameters are estimable but also whether they are statistically significant. If Step 5 fails in capturing some key parameters, one can go back to Step 2 to increase m in order to increase the likelihood of capturing the key parameters significantly at a reasonable level. Additionally, as a sensitivity analysis, one can also try different

sets of parameter values in Step 3, and then repeat Steps 4 and 5 to ensure estimability and statistical significance in a reasonable range of parameter values.[23]

The following sample code illustrates these steps for this credit card marketing example.

```
/* Now, combine the treatment group and the no-contact control
group */
data ch6.try2; set try;
trt=1; /* treatment group */
run;

data control;
cell=38;
trt=0; /* control group */
array all_design_var[*]
aprn limitn color1-color3 rebaten brand chan
aprnsq limitnsq rebatensq
aprnlimitn aprncolor1-aprncolor3 aprnrebaten aprnbrand
aprnchan
limitncolor1-limitncolor3 limitnrebaten limitnbrand limitnchan
rebatencolor1-rebatencolor3 brandcolor1-brandcolor3
chancolor1-chancolor3
rebatenbrand rebatenchan
brandchan;
do i=1 to dim(all_design_var); all_design_var[i]=0; end;
run;

data ch6.try3;
set ch6.try2 control;

seed=150;
m=30000; /* number of replicates in treatment */
ctrl_size=100000;

/* Make some assumptions on demographic variables,
assuming m replicates for each cell plus control */
if trt=1 then do;

do i=1 to m;

age=45+13*rannor(seed);
wealth=800+150*rannor(seed)+3*age;
balance=400+150*rannor(seed)+wealth*0.3;
hvalue=wealth*0.7+70*rannor(seed);

if age<=18 then age=18;
if balance<0 then balance=0;
if hvalue<0 then hvalue=0;
```

```
lnodd=-10+(20*age+5*wealth+1.65*balance)/1000;
lnodd=lnodd

+0.0*(color=1)
+0.08*(color=2)
+0.12*(color=3)
+0.03*aprn
+0.02*rebaten
+0.01*limitn

+0.01*age
+0.01*age*(color=3)
+0.01*age*rebaten;
prob=1/(1+exp(-lnodd));

res=ranbin(seed,1,prob);
output;
end;

end; else do;

do i=1 to ctrl_size;

age=45+13*rannor(seed);
wealth=800+150*rannor(seed)+3*age;
balance=400+150*rannor(seed)+wealth*0.3;
hvalue=wealth*0.7+70*rannor(seed);

if age<=18 then age=18;
if balance<0 then balance=0;
if hvalue<0 then hvalue=0;

lnodd=-10+(20*age+5*wealth+1.65*balance)/1000;

prob=1/(1+exp(-lnodd));

res=ranbin(seed,1,prob);
output;
end;

end;

run;

proc means data=ch6.try3;
class trt;
var res;
run;
```

```
data ch6.try3; set ch6.try3;
array demo_var[4] age wealth balance hvalue;
array design_var[9] trt aprn limitn color1-color3 rebaten
brand chan;
array demo_design_int[4,9]
agetrt ageaprn agelimitn agecolor1-agecolor3 agerebaten
agebrand agechan
wealthtrt wealthaprn wealthlimitn wealthcolor1-wealthcolor3
wealthrebaten wealthbrand wealthchan
balancetrt balanceaprn balancelimitn
balancecolor1-balancecolor3
balancerebaten balancebrand balancechan
hvaluetrt hvalueaprn hvaluelimitn hvaluecolor1-hvaluecolor3
hvaluerebaten hvaluebrand hvaluechan;
do i=1 to dim(demo_var);
      do j=1 to dim(design_var);
            demo_design_int[i,j]=demo_var[i]*design_var[j];
      end;
end;
run;


proc logistic data=ch6.try3 descending;
model res=
age wealth balance hvalue

trt aprn limitn color1-color3 rebaten brand chan
aprnsq limitnsq rebatensq
aprnlimitn aprncolor1-aprncolor3 aprnrebaten aprnbrand
aprnchan
limitncolor1-limitncolor3 limitnrebaten limitnbrand limitnchan
rebatencolor1-rebatencolor3 brandcolor1-brandcolor3
chancolor1-chancolor3
rebatenbrand rebatenchan
brandchan

agetrt ageaprn agelimitn agecolor1-agecolor3 agerebaten
agebrand agechan
wealthtrt wealthaprn wealthlimitn wealthcolor1-wealthcolor3
wealthrebaten wealthbrand wealthchan
balancetrt balanceaprn balancelimitn
balancecolor1-balancecolor3
balancerebaten balancebrand balancechan
hvaluetrt hvalueaprn hvaluelimitn hvaluecolor1-hvaluecolor3
hvaluerebaten hvaluebrand hvaluechan
/selection=stepwise;
run;

quit;
```

Examination of the logistic regression output of the above code (not shown) indicates that most of the key parameters are statistically significant (and the estimated values are quite consistent with the actual values from the theoretical model).

7.2.5.3 Handling Constraints in Optimal Factorial Design

The previous section employs fractional factorial design to generate an orthogonal design. In practical situations, however, there are often business constraints imposed. For example, you may not want to offer a product that has the highest quality with the lowest price, or marketers may really want to test a new product with certain attribute combinations that may not exist in your current fractional factorial design. The typical way to handle this situation is to first try incorporating these constraints into the design and then test the design if the required effects are still estimable. If not estimable, one can return to running another optimal design with these constraints explicitly incorporated. We resume Example 7.3 below with a realistic scenario.

Example 7.3 Credit Card Marketing Campaign Design – Handling Constraints

Two constraints are imposed by your colleagues:

1. Suppose you are told by the Finance Department that the following combination of attributes should not exist because it is not financially sustainable: APR = 4.9% (level 0), credit limit = $12,000 (level 3), and rebate = 1.5% (level 3), that is, the lowest price and the highest credit limit and rebate levels. After examining the current design in Table 7.16, you are happy to report that such a scenario does not exist, so you feel you are all set. Your Chief Financial Officer (CFO), however, is also concerned about the next low-price and high-quality scenario, APR at 6.9% (level 1), along with the highest credit limit and rebate levels. This scenario does exist at Run 18 in Table 7.16 and thus needs to be removed.
2. The Chief Marketing Officer (CMO) wants you to test what would happen when the Diamond card (Color at level 3) goes with the highest APR (11.9%, level 3), the highest credit limit ($12,000, level 3), the highest rebate (1.5%, level 3), the SuperAdvantage brand (level 1 of brand), and the email channel (level 1 of channel). Such a scenario does not exist in the current design, and you are interested in the estimability of the design if that scenario is added.

Your natural action now is to replace the current Run 18, which your CFO does not like, with the scenario from the CMO. The new design is in Table 7.18, where the changed run is shaded.

TABLE 7.18

Current Design with Run 18 Replaced by the CMO Scenario

Run	apr	limit	rebate	color	brand	chan
1	0	0	2	3	0	0
2	0	0	3	3	1	1
3	0	0	2	2	1	0
4	0	0	3	2	0	1
5	0	0	3	1	0	0
6	0	0	1	0	0	1
7	0	0	3	0	1	0
8	0	1	0	2	1	1
9	0	3	0	3	1	0
10	0	3	2	3	0	1
11	0	3	0	2	0	0
12	0	3	0	1	0	1
13	0	3	2	1	1	0
14	0	3	0	0	1	1
15	1	0	0	3	0	1
16	1	0	0	1	1	0
17	1	3	0	2	1	0
18	**3**	**3**	**3**	**3**	**1**	**1**
19	2	1	3	3	1	0
20	2	2	0	1	0	0
21	2	3	3	2	1	1
22	2	3	2	1	0	1
23	2	3	0	0	0	0
24	3	0	0	3	1	0
25	3	0	0	2	0	0
26	3	0	2	2	1	1
27	3	0	0	1	0	1
28	3	0	0	0	1	1
29	3	0	2	0	0	0
30	3	1	3	1	1	1
31	3	2	3	3	0	1
32	3	2	2	2	0	0
33	3	3	0	3	1	1
34	3	3	1	3	0	0
35	3	3	0	2	0	1
36	3	3	3	1	0	0
37	3	3	3	0	1	0

Re-running correlation analysis and PROC GLM with this modified design in Table 7.18 shows that the required effects are all still estimable and the correlation coefficients do not seem to have much change (not shown). As a result, the design in Table 7.18 is reported and can be used for actual campaign execution.

In case this manually constrained design is considered not as good (in the sense of variables being linked in the PROC GLM output or much higher correlation coefficients), one can remove the unwanted scenarios and add the desired one to the candidate set and re-run the optimal design procedure (PROC OPTEX) and then make sure the desired scenario does show up (if not, one can always add it as an additional scenario to the optimal design). The final design has to be checked using correlation analysis and PROC GLM. For our example, if we had a doubt on the latest design, the following code could be used to re-run the optimal design with the constraints incorporated. The statistical output is then shown in Table 7.19a (where the CMO desired scenario is in Run 32) with the output design in Table 7.19b. Note that the D-efficiency measure is only marginally reduced from 56.8% with the original design in Table 7.16a to 56.0% with the constraints incorporated in Table 7.19a.

```
/* Generate an optimal design with the constraints explictly
handled */

/* twoway is the original orthogonal design generated by PROC
FACTEX */
/* twoway_constr has the constraints incorporated */
data twoway_constr;
set twoway;
/* Delete the scenario CFO does not like */
if apr in (0,1) and limit=3 and rebate=3 then delete;
/* Check if the CMO scenario exists */
if apr=3 and limit=3 and rebate=3 and color=3 and brand=1 and
chan=1 then
      CMO_scenario=1;
else CMO_scenario=0;
run;
```

TABLE 7.19a

Statistical Output of an Optimal Design with Constraints Incorporated

Design Number	D-Efficiency	A-Efficiency	G-Efficiency	Average Prediction Standard Error
1	56.0460	22.3305	46.3830	1.3351
2	56.0460	22.3305	46.3830	1.3351
3	55.6233	16.8085	43.1266	1.3927
4	55.5979	18.4516	39.8424	1.3701
5	55.4959	18.0044	45.5846	1.3894

TABLE 7.19b

An Optimal Design with Constraints Incorporated

Run	apr	limit	rebate	color	brand	chan
1	0	0	3	3	1	1
2	0	0	2	2	1	0
3	0	0	3	2	0	1
4	0	0	2	1	1	1
5	0	0	3	1	0	0
6	0	0	3	0	1	0
7	0	1	0	3	0	1
8	0	1	3	1	1	0
9	0	2	3	3	0	0
10	0	3	0	3	1	0
11	0	3	0	2	0	0
12	0	3	2	2	1	1
13	0	3	0	1	0	1
14	0	3	0	0	1	1
15	0	3	2	0	0	0
16	1	0	0	2	1	1
17	1	1	0	1	0	0
18	1	2	3	1	0	1
19	2	0	0	0	0	1
20	2	1	3	2	0	0
21	2	3	0	3	0	1
22	3	0	0	3	1	0
23	3	0	2	3	0	1
24	3	0	0	2	0	0
25	3	0	0	1	0	1
26	3	0	2	1	1	0
27	3	0	0	0	1	1
28	3	1	1	2	0	1
29	3	1	3	0	0	1
30	3	2	3	2	1	1
31	3	2	0	0	0	0
32	3	3	3	3	1	1
33	3	3	1	2	1	0
34	3	3	3	2	0	1
35	3	3	1	1	1	1
36	3	3	3	1	0	0
37	3	3	3	0	1	0

```
/* Verify that the CMO scenario does exist */
proc freq data=twoway_constr;
tables CMO_scenario;
run;

proc optex data=twoway_constr;
class color brand chan; /* nominal variables */
model apr|limit|color|rebate|brand|chan@2 apr*apr limit*limit
rebate*rebate;
generate n=saturated method=m_fedorov;
output out=optdesign_constr;
run;

/* The CMO scenario is confirmed to be in Run 32 */
proc print data=optdesign_constr;
run;
```

7.2.5.4 *Putting Them Together: End-to-End Computer-Based Experimental Design Process*

Having described the various components of experimental design in previous subsections, it is useful to put them together in a flow diagram in Figure 7.5 as a general guideline.

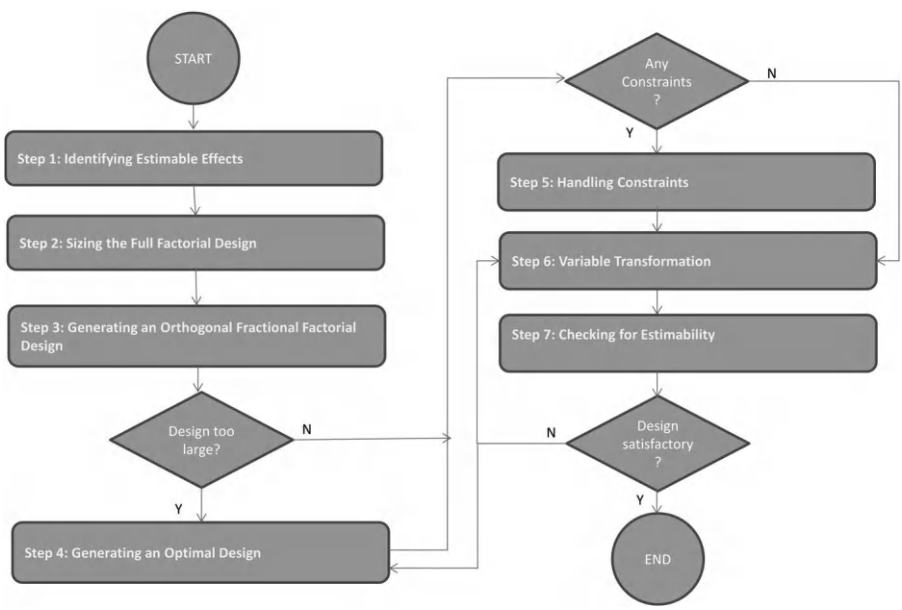


FIGURE 7.5
Flow diagram for the end-to-end experimental design process.

Let's describe the various steps shown in Figure 7.5:

Step 1: Identifying Estimable Effects – Identify all attributes in the design and all effects to be estimable such as main effects, two-way interactions, squared terms, etc. It is hard to know in advance which two-way or three-way interactions would be significant, but if there is a good business reason for attribute linkage (e.g., price and quality), and there is at least a slight expectation that certain interaction effects may exist, you would rather make them estimable in the design.

Step 2: Sizing the Full Factorial Design – Calculate the total number of combinations (runs or cells) in a Full Factorial Design as your benchmark. Unless there are only a few attributes, or your company can accommodate a design with a large number of cells, you will often need to create a fractional factorial design.

Step 3: Generating an Orthogonal Fractional Factorial Design – Generate a fractional factorial design in a software such as SAS PROC FACTEX. If the numbers of levels for different attributes are not the same, consider using binary variables for 4- or 8-level attributes, as in Example 7.3 Part II in Subsection 7.2.5.2.2. One can also combine some levels for attributes that have a number of levels not a multiple of two – for example, for a 3-level attribute, one can combine levels 0 and 1 as level 1 from a 4-level attribute that has levels 0, 1, 2, and 3. Examine the alias of the design. If the total number of runs is satisfactory and there is no special business constraint (i.e., no scenario is absolutely wanted or unwanted), jump to Step 6; otherwise, proceed to Step 4.

Step 4: Generating an Optimal Design – Use the orthogonal design from Step 3 as the candidate set to generate an optimal design (alternatively, one may use the full factorial design as the candidate set). Enter all required effects to be estimable in the design. If more than one design is available for the same set of desired effects, statistical measures such as D-optimality and A-optimality can be used to help determine the best design. See Subsection 7.2.5.2.2 for details.

Step 5: Handling Constraints – If there are any business constraints, we can remove the unwanted scenarios and add the desired scenarios to the current design. When necessary, incorporate the constraints in a candidate set and rerun optimal design. See Subsection 7.2.5.3 for an example.

Step 6: Variable Transformation – It is often advisable to center the continuous (quantitative) variables to reduce correlations. Required interaction terms also need to be created explicitly in this step, as mentioned in Subsection 7.2.5.2.2.

Step 7: Checking for Estimability – Compute the correlation matrix of all required variables (main, squared, and interaction terms) and check for any unusually high correlation coefficients. Run a linear model (PROC GLM) to make sure no variable is linked to another variable in the design. One may also combine the treatment group with a control group and conduct a simulation to include some demographic variables to make sure the final model, including all required design variables (and their interactions) as well as their interaction effects with demographic variables, is estimable. If there is any issue from this step, go back to Step 6 to check whether there is any coding error in transforming variables or to Step 4 to see if all desired effects are clearly captured in the design. Report the final design when it is acceptable. See also Subsection 7.2.5.2.2 for details.

7.3 Measurement Metrics for Test and Learn

This section discusses various measurement metrics for Test and Learn. While this section is not very technical, it is an important topic for anyone who is thinking about which metrics to measure and model.

To start with metrics, we will review a concept that the field of Advertising and Marketing has long considered, known as the AIDA model:

$$\textbf{Awareness} \rightarrow \textbf{Interest} \rightarrow \textbf{Desire} \rightarrow \textbf{Action}.$$

AIDA is a four-step process to drive the ultimate action; see Rawal (2013) and Li and Yu (2013).[24] For marketing, while the final desirable action may be a sale (or retention), there are many intermediate metrics we may want to measure in order to track the path to success step-by-step and to understand which stages of the process will require special attention.

In traditional advertising, one measure of the cost of achieving awareness is advertising impression, which is defined as CPM, or Cost per Thousand targets exposed, that is, the cost to deliver an ad to a thousand people (Sissors and Baron 2002). While CPM measures the efficiency of "reach," it is not a measure of customer responses. In this section and this chapter as a whole, we focus on customer responses for Test and Learn measurement and modeling.

We now have many relevant metrics related to the four stages of the AIDA model. From the marketing or general business point of view, there is typically a CTA for each treatment or intervention. The CTA can be visiting a site for education, calling a customer rep for inquiry, buying a product, spending

more, etc. Here is a list of the most important metrics[25] for various stages of the marketing funnel:

1. **Email Open Rate:** The first metric in email marketing is to have the target customers open your email. Most marketers would not consider this much of a success, but at least that is the starting point of the marketing funnel. This corresponds to the "Interest" of the AIDA model. Like all metrics on this list, there is a distinction between baseline measurement (natural response rate) and lift measurement (treatment minus control response rate), where the latter can be referred to as incremental email open rate or lift in email open rate.
2. **Inquiry or Visit Rate:** A similar metric to opening an email is the inquiry or visit rate that can apply to all marketing channels – phone, store visit, web, etc. – and is considered a necessary first milestone of marketing. This metric is related to the "Interest" and "Desire" parts of the AIDA model.
3. **Click-through Rate:** Click-through is one of the most common metrics, whether it is a click on a link in an email, a click on a link on a page served, or a click on a web banner ad. This may be exploring a product, reading a white paper, or signing up for a webinar. It is a metric that marketers desire and is almost second best to the actual sale. This metric can be considered the "Desire" or "Action" of the AIDA model and is a common metric not only for measurement but also for predictive modeling.
4. **Purchase Rate or Conversion Rate:** This is often the ultimate metric business would like to see for measurement and is certainly considered the "Action" stage of the AIDA model. It is also a key metric for predictive modeling. At the segment or entire campaign level, this metric can also be transformed into Cost per Conversion, which is defined as marketing expense divided by the number of sales, as the baseline metric (cost per all kinds of conversion, whether it is due to a treatment or not). With respect to the lift metric, one would measure Cost per Incremental Conversion, where Incremental Conversion (or lift in conversion) is (sample treatment conversion rate – sample control conversion rate) × treatment sample size.[26]
5. **Amount Spent:** While a purchase is good, there are differences between high-amount and low-amount sales. Measuring and modeling the amount spent is another common business activity. Similarly, one can translate sales amount to revenue as a common financial metric or Cost per Revenue as an efficiency measure. As in #4, one can measure Cost per Incremental Revenue using the revenue lift metric as the denominator.

6. **Retention Rate:** While purchase is relevant to acquiring new customers or cross-selling additional products to existing customers, retaining customers is another key marketing metric. Depending on your product or service, retention can be a binary metric (retained or not) or a continuous metric (value retention, e.g., usage of a credit card or balance of a bank account).
7. **Net Revenue:** While having a high sales amount or revenue is good, it is more important to check if it exceeds the marketing expense, whether for the entire marketing campaign, a specific segment, or a particular customer. Subtracting the marketing expense from the revenue is the net revenue metric that is commonly reported and analyzed.
8. **Return on Investment (ROI):** This is simply Net Revenue (revenue minus marketing expense, also known as Return), divided by marketing expense, which can be measured for the entire campaign, a specific segment, or a particular individual. Again, like other metrics, one can compute the baseline ROI, but it is often more useful to compute Incremental ROI (or Lift in ROI), where the latter is the lift in return (return caused by a treatment) divided by the expense associated with the treatment. The Incremental ROI is also known as Marketing ROI; see Lenskold (2003) for some examples.
9. **Lifetime Value (LTV):** While net revenue measures the immediate return of a marketing campaign, LTV (sometimes known as Long Term Value or Customer Lifetime Value, CLV) measures the sum of expected immediate and future values from a customer through the Expected Net Present Value (NPV) formula, typically discounted by estimated future attrition rate (in addition to the usual NPV discount rate). A general LTV formula is, with time t being discrete (say, year):

$$LTV_i = \sum_{t=0}^{\infty} \frac{\text{Return}_i(t)\, S_i(t)}{(1+r)^t} = \sum_{t=0}^{\infty} \frac{\left(\text{Revenue}_i(t) - \text{Expense}_i(t)\right) S_i(t)}{(1+r)^t}, \tag{7.9}$$

where r is the discount rate or hurdle rate in a typical NPV calculation, representing a discount of future values,

$$S_i(t) = \text{Probability of surviving at least to time (year) } t$$

$$= \left(1 - A_i(1)\right)\left(1 - A_i(2)\right)\ldots\left(1 - A_i(t-1)\right) = \prod_{k=1}^{t-1}\left(1 - A_i(k)\right),$$

$A_i(k)$ = Attrition rate of individual i in year k, assuming $\mathfrak{G}_i(\ ) = 1$.

In practice, the upper time limit of Eqn. (7.9) is a finite value $< \infty$, as the value inside the summation approaches zero when t gets very large. Additionally, in industries where some kind of risk is considered, the LTV can be further discounted by risk.[27]

While LTV is quite commonly used as a decision metric to establish how much a company should spend on a customer (or a segment) or to help determine the long-term effects of marketing treatments, it is typically not a metric for predictive modeling (as it is not observable and requires many assumptions). Rather, it takes model scores as an input to LTV calculations, as seen in Eqn. (7.9), where the revenue component $\text{Revenue}_i(t)$ and survival probability $S_i(t)$ are typically predicted by models. See Roberts and Berger (1999) for computations and applications, as well as Rosset et al. (2003), Fader et al. (2005), and Ansari et al. (2008) for alternative methods of computations and applications. While LTV is a good long-term metric for an individual customer or a segment, for treatment prioritization there is a concept called Incremental LTV where the value is driven by a specific treatment, similar to Incremental ROI or Marketing ROI discussed above. The incremental LTV calculation can be accomplished by replacing $\text{Revenue}_i(t)$ by lift in revenue for a revenue-generating campaign or by replacing $S_i(t)$ by lift in survival probability for a retention campaign.

While measurement can include many or all these metrics, as many of them are relatively straightforward to calculate, usually only a couple of them are used for predictive modeling or uplift modeling, given the high time and resource requirement for modeling. Typically, metrics closer to the ultimate goal or action, such as purchase, retention, and amount spent, are used for modeling. When data quantities are insufficient, one would use metrics closer to desire or awareness for modeling, for example, click-through and inquiry. Metrics such as ROI and LTV are typically used for decision-making, such as prioritization of customers for service or prioritization of treatments for customers. As a quick guideline, Table 7.20 summarizes these metrics and their typical usage for measurement, predictive modeling, or as derived metrics for decision-making.

7.4 Opportunities for Continuous Improvement

While some previous sections focus on the mechanics of experimental design, this section discusses the opportunities to keep improving experiments, which ultimately improves response rates.

After the campaign with the chosen design is executed, response data are collected, models are developed, and attribute combinations that appear to have achieved the best outcome (highest response, highest revenue, etc.) are identified,[28] one would naturally want to repeat the best attribute combination

TABLE 7.20

List of Metrics and Their Usage

Baseline Metric	Lift Metric	Standard Measurement	Potential for Predictive Modeling	Derived Metric for Decision-Making
Email Open Rate	Incremental (or Lift in) Email Open Rate	✓	✓	
Inquiry or Visit Rate	Incremental (or Lift in) Inquiry or Visit Rate	✓	✓	
Click-through Rate	Incremental (or Lift in) Click-through Rate	✓	✓	
Purchase (or Conversion) Rate	Incremental (or Lift in) Purchase (or Conversion) Rate	✓	✓	
Amount Spent	Incremental (or Lift in) Amount Spent	✓	✓	
Retention Rate	Incremental (or Lift in) Retention Rate	✓	✓	
Net Revenue	Lift in Net Revenue			✓
Return On Investment (ROI)	Marketing (or Incremental) ROI			✓
Lifetime Value (LTV)	Incremental LTV			✓

in the next campaign in order to maximize return on investment. Should we stop here and simply use this or these attribute combination(s) in the long future without any change? Of course, there is always room for improvement, and below we have some suggestions about improvement opportunities:

1. **Same Set of Attributes (*Response Surface Methodology*):** In situations where the best attribute combination(s) land in a "corner solution" (e.g., lowest APR level, highest rebate, highest credit limit), one may wonder if extending the possible levels further would result in an even better outcome. Likewise, even if the best attribute level so far is somewhere in the middle of the possible levels, one may wonder if it is indeed the "optimal" level (e.g., if the best APR level in terms of generating incremental revenue is 6.9%, one may want to know if 5.9%, 6.5%, 7.5%, or 7.9% would be even better). In these cases, we may want to "keep testing" until a more "globally optimal solution" is obtained. A common method in industrial experiments (industrial product design, quality control, etc.) known as the Response Surface Methodology (RSM) aims at iteratively searching for better, and ultimately optimal, levels with the same set of attributes (by searching in the direction of steepest ascent). While RSM is a common method in industrial experiments, it is less known in

marketing or related business settings but is certainly an opportunity. We do not cover RSM in this chapter (for it would take a whole chapter by itself), but interested readers can refer to Box et al. (1978), Montgomery (1991), Box (2006), Khuri and Mukhopadhyay (2010), or Goos and Jones (2011).

2. **New Attributes:** In the spirit of Kaizen (ongoing improvement) and sustainable competitive advantage, one would try to be creative to dream up other possible attributes that can help improve the response outcome. It is thus natural to test new and potentially useful attributes in an ongoing fashion.
3. **Fundamental Shift:** Changes in Customer Behavior or Other Conditions (Economic, Geographic, Demographic, etc.) can lead to a shift in the response function. For example, in the ever-changing technology world, whether one is selling software or hardware, the competitive landscape, underlying technology, as well as consumer needs and interests can easily and quickly evolve over time. One may need to consider updating the design attributes and predictive models in response to the market change.

Appendix 7.1: Derivation of the Power Calculation and Sample Size Determination for Four Group Comparison in Eqns. (7.6) and (7.7)

Assume that the decision rule is $\Delta\hat{p}_1 - \Delta\hat{p}_2 > a$, for some constant a, where ^ denotes the sample estimate.

$$P(\textit{Type I error}) = \alpha = P(\Delta\hat{p}_1 - \Delta\hat{p}_2 > a \,|\, H_0)$$

$$\cong P\left(\frac{(\Delta\hat{p}_1 - \Delta\hat{p}_2) - 0}{I} > \frac{a}{I} \;\; | H_0 \right), \qquad \text{(A7.1)}$$

where

$$I = \sqrt{\frac{\hat{p}_{1t}(1-\hat{p}_{1t})}{n_{1t}} + \frac{\hat{p}_{1c}(1-\hat{p}_{1c})}{n_{1c}} + \frac{\hat{p}_{2t}(1-\hat{p}_{2t})}{n_{2t}} + \frac{\hat{p}_{2c}(1-\hat{p}_{2c})}{n_{2c}}}, \qquad \text{(A7.2)}$$

which approximates the standard deviation of $\Delta\hat{p}_1 - \Delta\hat{p}_2$ using sample response rates.

Denote $\frac{a}{I}$ by z_α, the critical value associated with α.

Similarly,

$$P\left(\textit{Type II error}\right) = \beta = P\left(\Delta\hat{p}_1 - \Delta\hat{p}_2 \le a \,|\, H_1\right)$$

$$\cong P\left(\frac{(\Delta\hat{p}_1 - \Delta\hat{p}_2) - (\Delta p_1 - \Delta p_2)}{I} \le \frac{a - (\Delta p_1 - \Delta p_2)}{I} \middle| H_1\right). \quad \text{(A7.3)}$$

Now, denote $\frac{a - (\Delta p_1 - \Delta p_{2)}}{I}$ by z_β, the critical value associated with β.

Combining (A7.1) and (A7.3), we have:

$$-z_\beta = z_\alpha - \frac{(\Delta p_1 - \Delta p_2)}{I} \quad \text{(A7.4)}$$

Therefore,

$$\text{Power} = 1 - P\left(\text{Type II Error}\right)$$

$$= 1 - \Phi\left(\frac{a - \left(\Delta p_1 - \Delta p_2\right)}{I}\right)$$

$$= 1 - \Phi\left(z_\alpha - \frac{(\Delta p_1 - \Delta p_{2)}}{I}\right) \quad \text{from } (\text{A7.4}), \text{ which is Eqn. } (7.3),$$

where $(\Delta p_1 - \Delta p_2)$ is estimated by its sample estimate, $(\Delta\hat{p}_1 - \Delta\hat{p}_2)$.

Again, from Eqn. (A7.4),

$$(z_\alpha + z_\beta)^2 I^2 = (\Delta p_1 - \Delta p_2)^2. \quad \text{(A7.5)}$$

For convenience, we define the following ratios, which are the input parameters to the sample design:

$$R_1 = \frac{n_{1t}}{n_{1c}}, \quad R_2 = \frac{n_{2t}}{n_{2c}}, \quad \text{and} \quad R_t = \frac{n_{1t}}{n_{2t}}.$$

Re-expressing Eqn. (A7.2) in terms of the ratios defined above:

$$I^2 = \frac{\hat{p}_{1t}\left(1 - \hat{p}_{1t}\right) + \hat{p}_{1c}\left(1 - \hat{p}_{1c}\right)R_1 + \hat{p}_{2t}\left(1 - \hat{p}_{2t}\right)R_t + \hat{p}_{2c}\left(1 - \hat{p}_{2c}\right)R_tR_2}{n_{1t}} = \frac{J}{n_{1t}}, \quad \text{(A7.6)}$$

where J is defined as the numerator of I^2, which can be easily calculated once the ratios (input parameters) are specified. Substitute Eqn. (A7.6) into (A7.5), we can obtain n_{1t} as follows:

$$n_{1t} = \frac{J\left(z_\alpha + z_\beta\right)^2}{\left(\Delta p_1 - \Delta p_2\right)^2},$$

and we can also obtain the other sample sizes through the input ratio parameters:

$$n_{1c} = \frac{n_{1t}}{R_1}, \quad n_{2t} = \frac{n_{1t}}{R_t}, \quad \text{and} \quad n_{2c} = \frac{n_{2t}}{R_2} = \frac{n_{1t}}{R_t R_2},$$

which are the same as Eqn. (7.6).

Next, we apply the above set of Eqn. (7.6) to determining the control size such that the top decile can be statistically significantly better than that of the overall sample (if such a difference exists). In this case, Δp_1 represents the lift in the top decile and Δp_2 represents the lift in the rest of the deciles (i.e., deciles 2–10). Since the top decile is 10% of the overall sample, from Eqn. (A7.5), we now have:

$$\left(\frac{\Delta p_1 - \Delta p_2}{z_\alpha + z_\beta}\right)^2 = \frac{\hat{p}_{1t}\left(1-\hat{p}_{1t}\right)}{0.1 n_t} + \frac{\hat{p}_{1c}\left(1-\hat{p}_{1c}\right)}{0.1 n_c} + \frac{\hat{p}_{2t}\left(1-\hat{p}_{2t}\right)}{0.9 n_t} + \frac{\hat{p}_{2c}\left(1-\hat{p}_{2c}\right)}{0.9 n_c}.$$

Expressing n_c in terms of n_t and the other components, we have:

$$n_c = \frac{\dfrac{\hat{p}_{1c}\left(1-\hat{p}_{1c}\right)}{0.1} + \dfrac{\hat{p}_{2c}\left(1-\hat{p}_{2c}\right)}{0.9}}{\left(\dfrac{\Delta p_1 - \Delta p_2}{z_\alpha + z_\beta}\right)^2 - \dfrac{1}{n_t}\left[\dfrac{\hat{p}_{1t}\left(1-\hat{p}_{1t}\right)}{0.1} + \dfrac{\hat{p}_{2t}\left(1-\hat{p}_{2t}\right)}{0.9}\right]} \quad \text{which is Eqn. (7.7).}$$

Notes

1. It is quite common that the analyst or marketer has a hypothesis that the proportion (response rate) of one group may be greater than that of another, and thus a one-sided test is more common. A two-sided test will only require a simple adjustment in (7.1) and (7.2): Replacing z_α with $z_{\alpha/2}$.
2. An implementation of the formula in Excel is available from the authors.

3. If one is interested in detecting the lift difference between the top 3 deciles and the rest of the sample, a simple adjustment to (7.7), (7.8a), and (7.8b) is required.
4. In the situation of uplift modeling, to ensure each and every desired parameter is estimated is not straightforward, as it is necessary to have some understanding of what the parameters may be (for main and interaction effects), along with other requirements; see Mathews (2010) and Hsieh et al. (1998) for some situations. A practical method is to use Monte Carlo simulations as described in Subsection 7.2.5.2.2.
5. If explicitly controlling for the confounder in the Randomized Block Design is not feasible, a typical method is to control for the confounder in the analysis phase that includes the confounder as an independent variable in addition to the treatment variable in a regression model, also known as regression adjustment or analysis of covariance (ANCOVA). Another alternative is to apply propensity score matching as described in Chapters 3 and 8. Having confounders eliminated in the design phase is better than doing it in the analysis phase (due to potential model misspecification in analysis).
6. See census regions and divisions of the United States: http://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf.
7. The Latin Square Design is an alternative for more sophisticated situations with multiple blocking factors; see Montgomery (1991) or Box et al. (1978).
8. Another advantage that is more discussed in classical experimental design books is the improvement of significance testing for treatment effects due to smaller variability within block *and* within treatment, compared to variability only within treatment; see, for example, Montgomery (1991) and Ledolter and Swersey (2007).
9. One can imagine having a +1 and –1 for the + and – signs, then the sum of products of any two variables (i.e., time reduction × fee, time reduction × interaction, and fee × interaction) is equal to zero. This corresponds to having zero correlation or complete orthogonality between them.
10. Full factorial designs can be "made" orthogonal with the appropriate way of coding the independent variables (using contrast coding with 1s and –1s). Often we choose some way of coding that is more convenient for interpretation (e.g., dummy coding with 1s and 0s), but the design would not be completely orthogonal (but remains approximately orthogonal).
11. Technically, $\boldsymbol{X}$ here is the design matrix with all the replicates – in other words, it is a stack up of multiple replicates of the original design matrix, or $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_0 \\ \boldsymbol{X}_0 \\ \cdots \\ \boldsymbol{X}_0 \end{pmatrix}$, where $\boldsymbol{X}_0$ is the original design matrix without replicates. Then, it can be shown that $Var\left(\hat{\boldsymbol{\beta}}\right) = \frac{\sigma^2}{m}\left(\boldsymbol{X}_0{}'\boldsymbol{X}_0\right)^{-1}$, where m = number of replicates. This makes intuitive sense as the variance of estimated parameters goes down if number of replicates increases.
12. See also http://www.itl.nist.gov/div898/handbook/pri/section3/pri3347.htm for easy access to several fractional factorial designs.
13. Other software for experimental design include JMP and the R package RcmdrPlugin.DoE.

14. While this book is not meant to be a programming "cookbook," the programs listed in this section are unique enough that the authors feel it would be helpful to include them in the text. Please refer to this book's website for the entire program which includes the step-by-step design process.
15. This program uses two binary attributes to represent a four-level attribute (e.g., apr1 and apr2 together represent APR), which is an efficient way when working with some actual binary attributes (brand and channel). An alternative way which treats the four-level attributes as having four NOMINAL levels in PROC FACTEX (along with the same estimability criterion for main and two-way interaction effects) would end up with the full factorial design of 1,024 runs, i.e., no reduction in size at all, due to more runs required for nominal attributes that are actually continuous.
16. The model statement in PROC FACTEX allows users to directly specify which effects to be estimable, a convenient feature. Alternatively, one can specify the Resolution of the design in the model statement, see Rule of Resolution in Subsection 7.2.5.2.1.
17. Footnote 9 also applies here as there are replicates of runs in practice.
18. An alternative method, maximizing the trace of $\boldsymbol{X'X}$ (i.e., summing the diagonal elements of the matrix, essentially minimizing the sum of variances of all parameter estimates and ignoring the covariances) is called the A-optimal design. Details of alternative designs can be found in the online SAS/QC manual under PROC OPTEX or Kuhfeld (1997, 2010).
19. Alternatively, generating the full factorial design as a candidate set can be easily accomplished using PROC PLAN in SAS.
20. The default optimization method in PROC OPTEX is called the Fedorov algorithm. An alternative method in the same SAS procedure is the sequential algorithm.
21. This centering step is not necessary but is generally a good practice for reducing reduce correlations. For instance, a variable and its squared term are typically highly correlated, and centering the variable would significantly reduce the magnitude of such correlation.
22. The highest correlation coefficient (which involves two-way interactions) in this example is 0.74, and most other correlation coefficients are much lower. Correlation coefficients close to 1.0 or –1.0 would indicate problems.
23. Code that illustrates these steps for the credit card marketing example is available from the authors.
24. See also http://marketing-made-simple.com/articles/purchase-funnel.htm for the purchase funnel concept.
25. The list here is not meant to include all possible metrics; see, for example, Sterne (2002) for other web metrics.
26. This way renormalizes the number of control conversions such that the lift in conversion measures the incremental gain in conversion due to treatment in the treatment group (i.e., Average Treatment effect on the Treated, or ATT).
27. For example, in the mortgage or auto loan business, the risk components include default rate and prepayment, which can be combined with other components in the LTV computation.
28. In addition to identifying the best attribute combination (e.g., color = diamond, APR = 6.9%, etc.) for the entire sample, one can also identify the best attribute combination for each demographic group; see Chapter 7 for treatment optimization discussion.

References

Almquist, Eric, and Gordon Wyner. 2001. Boost Your Marketing ROI with Experimental Design, *Harvard Business Review*.

Ansari, Shahid, Alfred J. Nanni, Dessislava A. Pachamanova, and David P. Kopcso. 2008. "Using Simulation to Model Customer Behavior in the Context of Customer Lifetime Value Estimation". *INFORMS Transactions on Education*, 10(1): 1–9.

Box, George. 2006. *Improving Almost Anything: Ideas and Essays*, Revised edition. Hoboken, NJ: Wiley.

Box, George, William G. Hunter, and J. Stuart Hunter. 1978. *Statistics for Experimenters*. Hoboken NJ: Wiley.

Brunner, Rose, and Tom Kirchoff. 2012. "Experimental Design at Capital One: A Case Study in Auto Finance". *2012 American Statistical Association Conference on Statistical Practice*, American Statistical Association.

Campbell, Andrew, Michael Goold, Marcus Alexander, and Jo Whitehead. 2014. *Strategy for the Corporate Level: Where to Invest, What to Cut Back and How to Grow Organizations with Multiple Divisions*. San Francisco, CA: Jossey-Bass/Wiley.

Clarke, Geoffrey M., and Robert E. Kempson. 1997. *Introduction to the Design & Analysis of Experiments*. London: Arnold.

Cochran, William G. 1977. *Sampling Techniques*, 3rd edition. New York, NY: Wiley.

Davenport, Thomas H. 2009. How to Design Smart Business Experiments. *Harvard Business Review*. https://hbr.org/2009/02/how-to-design-smart-business-experiments

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee. 2005. "Counting Your Customers' the Easy Way: An Alternative to the Pareto/NBD Model". *Marketing Science*, 24(2): 275–284.

Fleiss, Joseph L. 1986. *The Design and Analysis of Clinical Experiments*. New York, NY: Wiley.

Friedman, Lawrence M., Curt D. Furberg, and David L. DeMets. 2010. *Fundamentals of Clinical Trials*, 4th edition. Breinigsville, PA: Springer.

Goos, Peter, and Bradley Jones. 2011. *Optimal Design of Experiments: A Case Study Approach*. Hoboken, JH: Wiley.

Goward, Chris. 2013. *You Should Test That: Conversion Optimization for More Leads, Sales, and Profit OR The Art and Science of Improving Websites*. Indianapolis, IN: Wiley.

Holland, Charles. 2005. *Breakthrough Business Results with MVT*. Hoboken NJ. Wiley.

Hsieh, F. Y., Daniel A. Bloch, and Michael D. Larsen. 1998. "A Simple Method of Sample Size Calculation for Linear and Logistic Regression". *Statistics in Medicine*, 17: 1623–1634.

Khuri, Andre I., and Siuli Mukhopadhyay. 2010. "Response Surface Methodology". *WIREs Computational Statistics*, 2: 128–129.

Kirk, Roger E. 1982. *Experimental Design*, 2nd edition. Pacific Grove, CA: Brooks/Cole Publishing.

Kohavi, Ron, Alex Deng, Brian Frasca, Roger Longbotham, Toby Waker, and Ya Xu. 2012. Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained, *KDD 2012*.

Kuhfeld, Warren F. 1997. Efficient Experimental Designs Using Computerized Searches, Sawtooth Software Research Paper Series. http://homepage.cs.uiowa.edu/~gwoodwor/AdvancedDesign/KuhfeldTobiasGarratt.pdf (downloaded 01/02/2016).

Kuhfeld, Warren F. 2010. *Marketing Research Methods in SAS*, MR-2010. The SAS Institute. https://support.sas.com/techsup/technote/mr2010.pdf (downloaded on 01/03/2016).

Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt. 1994. "Efficient Experimental Design with Marketing Research Applications". *Journal of Marketing Research*, 31: 545–557.

Ledolter, Johannes, and Arthur J. Swersey. 2007. *Testing 1-2-3: Experimental Design with Applications in Marketing and Service Operations*. Redwood City, CA: Stanford University Press.

Lenskold, James D. 2003. *Marketing ROI: The Path to Campaign, Customer, and Corporate Profitability*. New York, NY: McGraw-Hill.

Li, Jiangyu, and Haibo Yu. 2013. "An Innovative Marketing Model Based on AIDA: A Case from E-Bank Campus-Marketing by China Construction Bank". *iBusiness*, 5: 47–51.

Luecke, Richard. 2005. *Strategy: Create and Implement the Best Strategy for Your Business*. Boston, MA: Harvard Business Review Press.

Manzi, Jim. 2012. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. New York, NY: Basic Books.

Mathews, Paul. 2010. *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Ohio, OH: Mathews Malnar and Bailey, Inc.

McFarland, Colin. 2013. *Experiment! Website Conversation Rate Optimization with A/B and Multivariate Testing*. Berkeley, CA: New Riders.

Montgomery, Douglas C. 1991. *Design and Analysis of Experiments*, 3rd edition. Hoboken, NJ: Wiley.

Paige, Christopher H. 2001. "Capital One Financial Corporation". *Harvard Business School Case Number* 9-700-124.

Rao, Vithala R., and Joel H. Steckel. 1998. *Analysis for Strategic Marketing*. Boston, MA: Addison Wesley Longman, Inc.

Rawal, Priyanka. 2013. "AIDA Marketing Communication Model: Stimulating a Purchase Decision in the Minds of the Consumers through a Linear Progression of Steps". *IRC's International Journal of Multidisciplinary Research in Social & Management Sciences*, 1(1): 2320–8236.

Roberts, Mary Lou, and Paul D. Berger. 1999. *Direct Marketing Management*. New Jersey, NJ: Prentice-Hall.

Rosset, Saharon, Einat Neumann, Uri Eick, and Nurit Vatnik. 2003. "Customer Lifetime Value Models for Decision Support". *Data Mining and Knowledge Discovery*, 7: 321–339.

Scheaffer, Richard L., William Mendenhall, and Lyman Ott. 1990. *Elementary Survey Sampling*, 4th edition. Boston, MA: Duxbury/PWS-Kent Publishing.

Siroker, Dan, and Pete Koomen. 2013. *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers*. Hoboken, NJ: Wiley.

Sissors, Jack Z., and Roger B. Baron. 2002. *Advertising Media Planning*, 6th edition. New York, NY: McGraw-Hill.

Sterne, Jim. 2002. *Web Metrics: Proven Methods for Measuring Web Site Success*. Princeton, NJ: Wiley.

8

Uplift Analytics III: Model-Driven Decision-Making and Treatment Optimization Using Prescriptive Analytics

8.1 Introduction

While most of this book discusses methods for measurement and prediction, this chapter answers the question: "if you have some idea about what will happen, whether solid or not, what should you do to make the future better?" in the context of uplift analytics, of course.

The methods discussed in Chapter 6 are useful for handling the most typical or simplest situation when there is only a SINGLE treatment. That is, there is only a single product (with a single feature) available to promote. To choose the right targets with a single treatment is straightforward. This chapter discusses solutions to situations where there is more than one treatment. Additionally, there is often a budget or quantity constraint due to limited resources. Otherwise, one could simply assign the treatment with the highest predicted lift value to each individual as long as the predicted lift value is positive. The constrained optimization, with realistic constraints, while not very hard to write down mathematically, is typically difficult to solve because of its huge size (large number of individuals). Fortunately, heuristics (approximation methods) are available for solving this kind of large optimization problem and are discussed in this chapter.

Another issue discussed in this chapter is how to handle uncertainty. Most common optimization methods, such as linear programming (LP), are designed to solve deterministic problems; that is, all parameters in the optimization model are assumed to be known with complete certainty. In our case, the lift values are all estimated, sometimes with a high degree of uncertainty; directly using the estimated lift values does require someone to take a "leap of faith." However, instead of completely trusting the estimated lift values, can we improve our selection or optimization problem if we have some idea about the uncertainty or range of the lift values? The answer is yes, but it requires a different set of methods, which are discussed later in this chapter.

 DOI: 10.1201/9780429172588-8

The mathematical and computational methods covered in this chapter were mostly founded in the fields of Operations Research, Management Science, and Industrial Engineering (or Prescriptive Analytics as a latest term), and they are quite different from those in the rest of the book where data mining, statistics, and econometrics are utilized. All key algorithms described in this chapter are practical and are ready to use with standard software for practitioners and students.

8.2 Single Treatment and Multiple Treatment Optimization

Before we introduce optimization with multiple treatments, it is helpful to first discuss the simpler single treatment case.

8.2.1 Single Treatment Optimization

This section considers the simplest case where there is only a single treatment (versus control). The goal is to find targets such that overall lift is maximized subject to a quantity constraint. Such optimization is equivalent to a selection problem where the goal is to select the right targets. Formally, it means to select a set of individuals S with the following objective function:

$$\text{Maximize: \# Incremental Responders} = \sum_{i \in S} \Delta p_i \tag{8.1a}$$

subject to the following constraint: No. of elements in $S \le U$, where Δp_i is the lift value for individual i (i.e., treatment response rate minus control response rate), and U is the upper limit for the number of individuals selected with our budget. The actual (future) value of Δp_i is unknown (a random variable), and we may consider using its expected value, $E(\Delta p_i)$, which can be learned from past data and is the mean (average) difference between the sample treatment response rate and sample control response rate.[1] The objective function then becomes:

$$\text{Maximize: Expected \# Incremental Responders} = \sum_{i \in S} E(\Delta p_i) \tag{8.1b}$$

The expectation $E(\Delta p_i)$ is unknown and can be estimated by its individual-level model-based estimate (i.e., model score) denoted by $\Delta \hat{p}_i$.

The objective function in Eqn. (8.1b) is simply the expected (or average) total number of incremental responders from those selected in set S and can be written in a more common mathematical form:

$$\text{Maximize} \sum_{i=1}^{n} E(\Delta p_i) x_i \tag{8.2}$$

subject to:

$$\sum_{i=1}^{n} x_i \le U,$$

$$x_i = 0 \text{ or } 1, \qquad i = 1, \ldots, n.$$

Again, $E(\Delta p_i)$ is replaced by its estimate, $\Delta \hat{p}_i$. In Eqn. (8.2), the decision variable x_i represents the decision of whether individual i should be selected (= 1) or not (= 0), n represents the size of the target population, and U again represents the maximum number of targets we can reach with our budget.

The optimization model (8.2) can be easily solved by selecting the U individuals with the highest values of $\Delta \hat{p}_i$'s, which is exactly the model implementation procedure outlined in Chapter 6. In a slightly more general situation where the contact cost for some individuals is higher than others, we have:

$$\text{Maximize} \sum_{i=1}^{n} \Delta \hat{p}_i x_i \tag{8.3}$$

subject to:

$$\sum_{i=1}^{n} c_i x_i \le \text{Budget}, \quad \text{and } \mathbb{0}_i = \quad \text{or } 1, \quad i = 1, \ldots, n,$$

where c_i is the contact cost for individual i.

The differential cost in the constraint of model (8.3) is needed when the business needs to contact the highest value customers with a more personal touch, such as outbound telemarketing, but the rest with a lower cost touch, such as direct mail or email. Alternatively, higher-value customers may have a different "service package," so the direct mail creative material for them may be different. Another example is simply that individuals in different geographic regions require different costs of contact. So the question is how to solve Eqn. (8.3).

Model (8.3) is set up as a linear integer programming problem, and it can be solved using formal integer programming techniques such as branch and

bound (see Papadimitriou and Steiglitz 1998 or Taha 2010). However, because of the specific simple form of model (8.3), there are reasonably good and simple heuristics available. Model (8.3) is actually called the 0-1 knapsack problem, which has been well studied in the field of Operations Research. As described in Chapter 9 of Williams (2003), the knapsack situation arises when a hiker tries to fill her knapsack to maximize total value (the objective function). Each item i has its own value (represented by Δp_i in our case, or its estimate $\Delta\hat{p}_i$) and a weight (c_i), and there is a weight limit U. The simplest solution is a greedy algorithm proposed by Dantzig (1957), the father of linear programming, and also described in Pisinger (1995). We will rephrase the algorithm with our terminology as follows:

Algorithm 8.1 (Adapted from Dantzig 1957)

1. Calculate the ratio of value to cost for each individual: $\frac{\Delta\hat{p}_i}{c_i}$;
2. Sort the ratios from highest to lowest and list all individuals in that order;
3. Keep selecting those individuals from the top of the list as long as the budget constraint is still met;
4. Stop the selection when the budget constraint is violated.

Notice that the key step is calculating the ratio in step 1. It should be mentioned that the objective function (or its components $\Delta\hat{p}_i$'s) and the cost are *not* on the same scale. If we can translate the sample lift values $\Delta\hat{p}_i$ to an economic value such as revenue (by multiplying $\Delta\hat{p}_i$ by marginal revenue per response), then such revenue and cost are both expressed in a monetary value, so we can sort by the "incremental value over cost," resulting in a slightly different greedy algorithm as follows:

Algorithm 8.2

1. Translate the predicted lift for each individual to economic value: $r_i\Delta\hat{p}_i$, where r_i is the revenue value associated with the response for individual i;
2. Calculate the "incremental value over cost," $r_i\Delta\hat{p}_i - c_i$;
3. Sort the incremental values from highest to lowest and list all individuals in that order;
4. Keep selecting those individuals from the top of the list as long as the incremental value over cost is positive and the budget constraint is still met;
5. Stop the selection when the budget constraint is violated or the incremental value cost turns negative.

Note that Algorithms 8.1 and 8.2 are heuristics and do not guarantee they will find the global optimum. Nevertheless, these heuristics tend to at least find "reasonably" good solutions. These algorithms use model estimates of Δp_i, which are likely far from exact. In practice, a better alternative is to use the holdout sample data to estimate Δp_i at the decile or semi-decile level. Even though individuals within the same decile or semi-decile can no longer be differentiated from each other, the decile/semi-decile level lift estimates are recommended for the following reasons:

1. These group-level lift values "borrow strength" from others in the same group using the sample means at the group level.
2. Using performance metrics from the holdout sample mitigates the well-known "re-substitution" error that may occur when training and evaluation are done with the same sample. Since it is common to use holdout sample performance for assessing and reporting model accuracy, which leads to expected success in the future, it would be natural to use holdout sample performance for decision-making or optimization.

We will come back to this point when discussing multiple treatments in the next section.

Algorithm 8.3

1. Revisit the model evaluation step. Obtain the decile or semi-decile level average lift as the predicted lift values for optimization.
2. Replace $\Delta\hat{p}_i$ with the decile/semi-decile level predicted lift values (i.e., average lift at group level), and proceed to Step 1 of Algorithm 8.1 or 8.2.

There are other more advanced algorithms to solve Eqn. (8.3), such as Dynamic Programming, that are beyond our scope; see Pisinger (1995), Papadimitriou and Steiglitz (1998), or Dasgupta et al. (2006).

8.2.2 Multiple Treatment Optimization

8.2.2.1 Four Targeting Situations and Introduction to Multiple Treatment Optimization

The single treatment optimization situation in Section 8.2.1 served as the simplest case and segues into the more general situation where there is more than one treatment.[2] This situation arises quite frequently in marketing campaigns where multiple treatments are available, and we need to know not

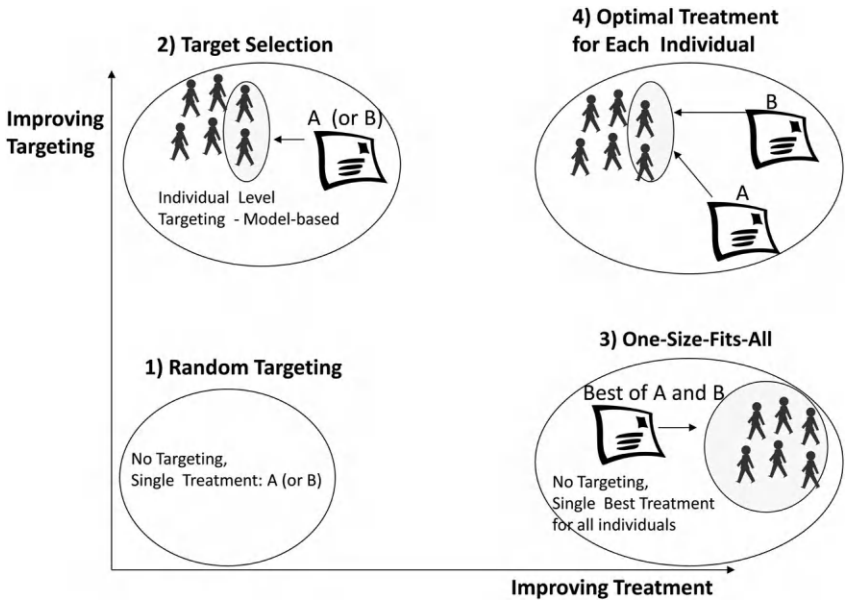


FIGURE 8.1
From targeting to optimization.

only which treatment is the best overall but also the best treatment for each individual. Figure 8.1 describes various situations:

1. **Random Targeting (or No Targeting):** Without any information or any strategy about whom to target, one can resort to "random targeting," that is, no targeting. Selecting a random group of individuals has the advantage of learning about them and doing a better job next time.
2. **Target Selection:** Most marketing campaigns have some targets, for example, focusing on a few customer segments, using data- or rule-based targeting, or predictive models. Uplift modeling is among the most advanced techniques for target selection.
3. **One-Size-Fits-All:** Using A/B testing, one can pick the best of two treatments. One can use multiple iterations of A/B testing to find the champion among multiple treatments. Alternatively, one can conduct an experimental design (also known as multivariate testing) to simultaneously test multiple treatments (see Chapter 7 for details). These techniques allow you to pick the single best treatment for the overall population. Measurement done at a pre-defined segment or group level (e.g., age group, state/region) can allow one to find the single best treatment for the group. Still, it is "One-Size-Fits-All" for a given group or the whole population.

4. **Optimal Treatment for Each Individual:** This is the most granular level of optimization – it aims at finding the best treatment for *each* individual. By treatment, we include "no treatment" as a possible alternative. This is our focus for the rest of the chapter.

8.2.2.2 Optimization Models for Multiple Treatments

We now consider the methodology for Situation 4 in Figure 8.1: Optimal Treatment for Each Individual. We assume that for each of the n individuals, we have m treatments plus the option of not assigning any treatment. Our aim is to optimally determine treatment assignment at the individual level so as to maximize the total number of incremental responders, that is, those who would respond due to a treatment. Note that the treatment assignment optimization can result in individuals who are not assigned any treatment. The optimization model can be formulated as the following linear binary integer program (from Lo and Pachamanova 2015):

$$\text{Maximize} \sum_{i=1}^{n}\sum_{j=1}^{m} \Delta\hat{p}_{ij}x_{ij} \tag{8.4}$$

subject to:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} \le B, \quad \text{Budget Constraint}$$

$$\sum_{j=1}^{m} x_{ij} \le 1, \qquad \text{for } i = 1,\ldots,n,$$

No more than 1 treatment is assigned to each individual, and

$$x_{ij} = 0 \text{ or } 1, \quad i = 1,\ldots,n; \quad j = 1,\ldots,m.$$

where $\Delta\hat{p}_{ij}$ = estimated lift value for individual i and treatment j, x_{ij} (decision variable) = 1 if treatment j is assigned to individual i and 0 otherwise, and c_{ij} = cost of promoting treatment j to individual i.

Quite often, we may also have an operational constraint on the quantity for each treatment:

$$\sum_{i=1}^{n} x_{ij} \le U_j, \qquad \text{for } j = 1,\ldots,m.$$

which gives the upper limit on the number of individuals receiving each treatment. Again, the original unknown expected value $E(\Delta p_{ij})$ is estimated by its estimate $\Delta\hat{p}_{ij}$ in the objective function of model (8.4).

If the group "those who are not assigned any treatment" is also explicitly represented in the decision variables, the optimization model (8.4) has the following equivalent form (with an additional decision variable x_{i0}):

$$\text{Maximize} \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{p}_{ij}x_{ij} + \sum_{i=1}^{n} \hat{p}_{i0}x_{i0} \tag{8.5}$$

subject to:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} \le B,$$

$$\sum_{j=0}^{m} x_{ij} = 1, \qquad \text{for } i = 1,\ldots,n \tag{8.5.1),}$$

and $x_{ij} = 0$ or 1, $i = 1,\ldots,n$; $j = 0, 1,\ldots,m$.

In Eqn. (8.5), $\hat{p}_{ij}$ = estimated probability of individual i responding to treatment j, $\hat{p}_{i0}$ = estimated response probability of individual i when no treatment is used (estimated by the control group response probability), and the decision variable $x_{i0} = 1$ if no treatment is assigned to individual i and 0 otherwise. Note that: $\Delta\hat{p}_{ij} = \hat{p}_{ij} - \hat{p}_{i0}$ for all i and j. In reality, all these response probabilities are estimated.

To see why Eqn. (8.5) is equivalent to Eqn. (8.4), we start with the objective function of model (8.5):

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{p}_{ij}x_{ij} + \sum_{i=1}^{n} \hat{p}_{i0}x_{i0} &= \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\Delta\hat{p}_{ij} + \hat{p}_{i0}\right)x_{ij} + \sum_{i=1}^{n} \hat{p}_{i0}x_{i0} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \Delta\hat{p}_{ij}x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{p}_{i0}x_{ij} + \sum_{i=1}^{n} \hat{p}_{i0}x_{i0} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \Delta\hat{p}_{ij}x_{ij} + \sum_{i=1}^{n} \hat{p}_{i0}, \ \left(\text{using constraint } (8.5.1)\right)
\end{aligned}$$

which is the same as the objective function in Eqn. (8.4) plus a constant that does not depend on the decision variables. Hence, Eqns. (8.4) and (8.5) are equivalent. Additional constraints can also be added to model (8.4) or (8.5), for example, limited quantities available for certain treatments.

Model (8.4) or (8.5) is a binary (or zero-one) integer programming problem and is well-known to be challenging to solve. Unlike LP problems, there is no efficient general algorithm available due to its *NP-completeness*

nature (which is a formal way of saying they are really hard to solve and can take a very long time[3]). Nevertheless, there are common integer programming algorithms such as Brand-and-Bound and Cutting Plane that may take an exponential number of iterations; see Bertsimas and Tsitsiklis (1997), Papadimitriou and Steiglitz (1998), or Taha (2010). General software such as CPLEX ILOG and SAS/OR can handle Eqn. (8.4) or (8.5) for relatively small n and m. However, when n is large, and even for a moderate size m, the problem can become very large. To have some appreciation of the problem size, imagine that we have $n = 10$ million individuals and $m = 5$ treatments only, so we will have 50 million binary decision variables for model (8.4), which translates to $2^{50,000,000}$ possible combinations of decisions to attempt (if one is to use complete enumeration, aka brute force). With the importance of such problems in marketing and related applications, specialized commercial marketing optimization software packages are available to handle this type of large integer programming problem, such as MarketSwitch and SAS Marketing Optimization. Because of their customized nature, ability to integrate with campaign management tools, and their proprietary mathematical algorithms, these software packages tend to be more expensive than general optimization software or data mining software.

An alternative to solving Eqn. (8.4) or (8.5) is to use heuristics such as first grouping individuals into clusters (using cluster analysis, for example, based on certain individual characteristics) and solving the LP problem at the cluster level (as well as individual-level optimization in each cluster); see Storey and Cohen (2002) for general marketing optimization (not specific to uplift problems). General heuristics or approximation algorithms such as simulated annealing, genetic algorithms, and tabu search can also be considered; see Goldberg (1989), Bertsimas and Tsitsiklis (1997), or Michalewicz and Fogel (2002).

In practice, instead of using individual-level model estimates for Δp_{ij}, we recommend using the holdout sample estimates that are less biased, for reasons given in Section 8.2.1 for the single-treatment case. However, unlike the deciling method for the single-treatment situation in Algorithm 8.3, since each treatment has its own set of deciles, we cannot compare the lift estimate for decile 1 of treatment 1 to the lift estimate for decile 1 of treatment 2 in order to pick the better treatment, as the two decile 1s are not the same group of individuals. To address this issue, we employ cluster analysis to group individuals in the algorithm described below (see Chapter 2 for an introduction to cluster analysis).

Algorithm 8.4 (For Multiple Treatment Optimization) (Adapted from Lo and Pachamanova (2015))

1. In the holdout sample, compute the estimates of $E(\Delta p_{ij})$, $\Delta \hat{p}_{ij}$, for all individuals i and treatments j in the data.

2. Perform a cluster analysis of the m model-based lift scores, $(\Delta\hat{p}_{i1}, \ldots, \Delta\hat{p}_{im})$, $i = 1, \ldots, n$, to obtain C clusters of individuals. Some attention should be paid to the tradeoff between sample size and granularity.
3. For each cluster $c = 1, \ldots, C$ in the holdout sample, calculate the cluster-level lift score for each treatment, $(\Delta\hat{p}_{c1}, \ldots, \Delta\hat{p}_{cm})$, using the sample mean difference in response rate between each treatment and the control group within each cluster.
4. If more than one cluster solution is available (because of different clustering algorithms and/or different uplift models), one may choose the cluster solution such that the cluster-level lift scores are as far away from the overall sample lift scores as possible; that is, choose a cluster solution such that the following Euclidean distance is the greatest in order to support optimization:

$$\text{Squared distance to the overall sample mean} = \sum_{c=1}^{C} n_c \sum_{j=1}^{m} \left(\Delta\hat{p}_{cj} - \Delta\hat{p}_j\right)^2, \tag{8.6}$$

   where $\Delta\hat{p}_1, \ldots, \Delta\hat{p}_m$ are the overall sample lift scores for treatments $1, \ldots, m$, respectively (averaging over all individuals in the entire holdout sample), and n_c is the sample size of cluster c.
5. Apply the chosen cluster solution to the new data for the future campaign and report the size of each cluster, N_c.
6. Solve the following LP problem for the future campaign assignment:

$$\text{Maximize} \sum_{c=1}^{C}\sum_{j=1}^{m} \Delta\hat{p}_{cj} x_{cj} \tag{8.7}$$

   Subject to:

$$\sum_{c=1}^{C}\sum_{j=1}^{m} c_j x_{cj} \le Budget, \quad \text{Budget Constraint}$$

$$\sum_{j=1}^{m} x_{cj} \le N_c, \quad \text{for } c = 1, \ldots, C, \quad \text{Cluster Size Constraint, and}$$

$$x_{cj} \ge 0, \quad c = 1, \ldots, C; \quad j = 1, \ldots, m,$$

   where x_{cj} now denotes the number of individuals in cluster c to receive treatment j, and as in model (8.5), c_j = cost of treatment j for each individual (assuming the cost is not individual or cluster-specific).

Again, we may also have an operational constraint on the quantity for each treatment:

$$\sum_{c=1}^{C} x_{cj} \leq U_j, \qquad \text{for } j = 1, \ldots, m.$$

where we have an upper limit on the number of individuals receiving each treatment. Unlike model (8.4), the decision variables x_{cj} in model (8.7) are no longer binary but are simply non-negative real numbers, that is, continuous or quantitative values that can be rounded up to approximate integer values. Model (8.7) can typically be solved by simple LP software, including Excel Solver.

Algorithm 8.4 provides an optimal solution at the cluster level, that is, optimizing the treatment allocation to each cluster (including no allocation to some clusters, which is also part of an optimal solution). So exactly who should receive each treatment at the individual level for a given cluster? If the cluster size for a given cluster in the target population is larger than the quantity recommended by the optimization algorithm, one recommendation is to simply choose the targeted individuals at random. This approach, while simple, has the advantage of having a natural randomized control group (those who are not selected but similar to those selected for the next campaign), a key for measurement and further refinement. After all, the individual-level model-based lift scores should be relatively homogenous within each cluster (through cluster analysis in Step 2 of Algorithm 8.4). Additionally, we may not have good model-based estimates to differentiate between individuals within a given cluster. Alternatively, if we really have to "optimize" at the individual level for a given cluster, we can prioritize the selection by individual-level model-based lift value score, even though those individual-level estimates may not be very accurate (i.e., choose those with the highest individual-level estimates).[4]

8.3 Joint Men's and Women's Merchandise Optimization Example (Using Excel Solver)

Example 8.1 (Extension of Example 6.2), adapted from Lo and Pachamanova (2015)

This data set has two email treatment groups for men's and women's merchandise, respectively, plus a no-mail control group. In this example, we apply the Two Model Approach and Treatment Dummy Approach from

Lo (2002), that is, Methods 1 and 2 from Chapter 6, to maximize overall visits. Using the Two Model Approach (Method 1), we can calculate the lift score for men's merchandise and the lift score for women's merchandise as inputs to cluster analysis (clustering on two variables), and similarly for the Treatment Dummy Approach (Method 2).[5] Following Algorithm 8.4:

1. We first calculate the individual model scores for all individuals using each method (Methods 1 and 2).
2. Perform clustering using k-means with 10 clusters (one may choose other clustering algorithms such as hierarchical or EM-based) under each method. Figure 8.2 illustrates the women's treatment lift score (vertical) and men's treatment lift score (horizontal) using Method 1 (the original clusters 1 and 3 are merged into cluster 1 because of small sizes, a common issue with clustering). It is clear that the two treatment lift scores are related, although not strongly linearly.
3. Cluster-level means and computations for the squared distance to the overall lift means are in Table 8.1.

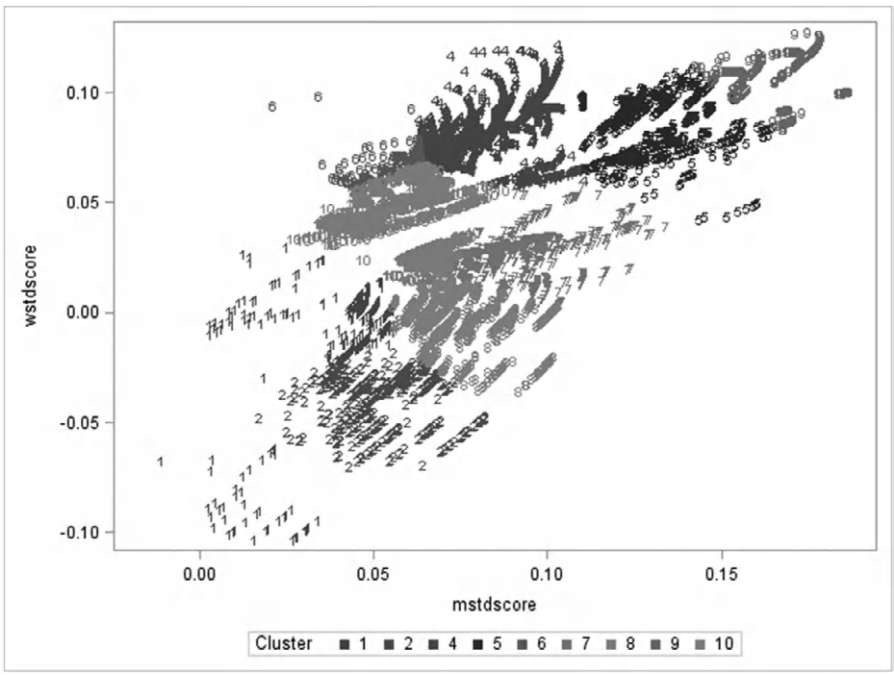


FIGURE 8.2
Lift score for women's merchandise by lift score for men's by cluster.

TABLE 8.1

Lift in Response by Treatment by Cluster and Calculation of the Squared Distance to Overall Means

Obs	Cluster	Size	Obs. Lift in Response: Men's	Obs. Lift in Response: Women's	Sqr dist1	Sqr dist2	Squared Distance to Overall Mean
1	Overall	18985	0.07408	0.043863			
2	1	418	0.15871	0.022388	0.007162	0.000461	
3	2	565	0.06521	–0.005547	7.87E-05	0.002441	
4	4	6022	0.06577	0.062776	6.91E-05	0.000358	
5	5	1237	0.12903	0.06177	0.00302	0.000321	
6	6	894	0.06717	0.076013	4.77E-05	0.001034	
7	7	2924	0.0519	0.021328	0.000492	0.000508	
8	8	2807	0.08684	0.025362	0.000163	0.000342	
9	9	410	0.22491	0.023865	0.02275	0.0004	
10	10	3708	0.05717	0.042621	0.000286	1.54E-06	
	Sum				19.5151	7.662298	27.18

4. Between Methods 1 and 2, Method 1 has the highest squared distance to the overall sample mean and thus is chosen as the cluster solution for the next steps. The bubble chart in Figure 8.3 shows the two-dimensional distance of each cluster (denoted by black bubbles) to the overall sample mean (large gray bubble), where the size of each bubble is proportional to cluster sample size.

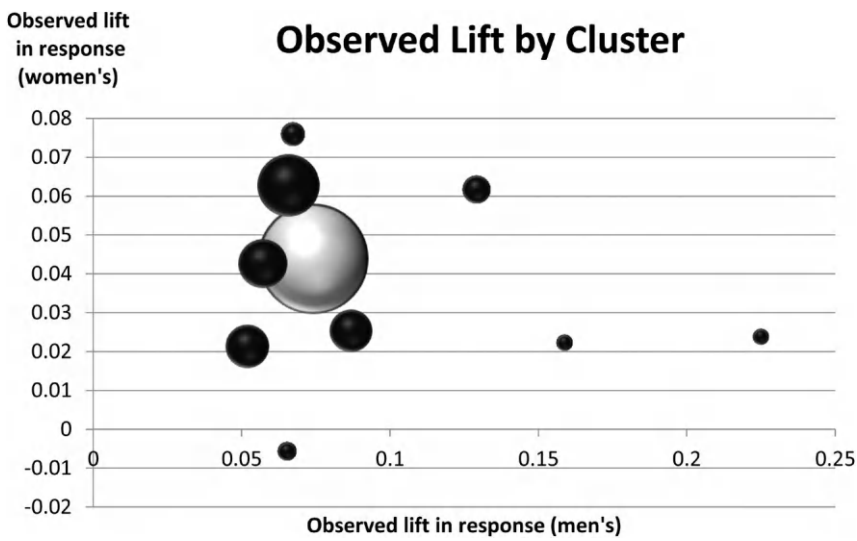


FIGURE 8.3
Bubble chart for observed lift by cluster. (Cluster-level means (black bubbles) versus overall sample mean (large gray bubble); bubble size is proportional to sample size.)

5. We can now apply the chosen cluster solution to *future* data for optimization. For this illustrative example, we assume the future data is 10 times the holdout sample, and the cluster distribution is the same as in the holdout sample.
6. The last step is to solve the LP problem in model (8.7). The results are shown in Table 8.2. Note that cluster sizes (N_c) in the second column are used in the cluster size constraint. The third and fourth columns are simply taken from the holdout sample's observed (sample mean) lift values by cluster and are used in the objective function calculations. The cost per treatment contact is assumed to be \$1.0 each for both men's and women's merchandise, and the total budget is set at \$60,000. The decision variables are set up such that the estimated total number of visits is maximized (shaded cell). The optimal solution, solved by the "Simplex LP" option in Excel Solver, is:
   - 4,180 men's mailings to cluster 1 (again, the original clusters 1 and 3 are merged as cluster 1),
   - 2,340 men's mailings to cluster 4,
   - 12,370 men's mailings to cluster 5,
   - 8,940 to women's mailings to cluster 6,
   - 28,070 men's mailings to cluster 8, and
   - 4,100 men's mailings to cluster 9.

The optimal results here are clearly governed by the observed lift values as well as the cluster sizes. This example will be extended to consider other situations in subsequent parts of this chapter.

TABLE 8.2

Linear Programming Computations Using Excel Solver

Cluster	Cluster Size in New Data	Obs. Lift in Response: Men's	Obs. Lift in Response: Women's	Cost per Treatment ($)	Decision var on Number of Men's	Decision var on Number of Women's	Total Number of Treated by Cluster
1	4,180	0.1587	0.0224	1	4,180	–	4,180
2	5,650	0.0652	–0.0055	1	–	–	–
4	60,220	0.0658	0.0628	1	2,340	–	2,340
5	12,370	0.1290	0.0618	1	12,370	–	12,370
6	8,940	0.0672	0.0760	1	–	8,940	8,940
7	29,240	0.0519	0.0213	1	–	–	–
8	28,070	0.0868	0.0254	1	28,070	–	28,070
9	4,100	0.2249	0.0239	1	4,100	–	4,100
10	37,080	0.0572	0.0426	1	–	–	–
Total	189,850			obj value	5,773	680	6,453
				cost	\$51,060	\$8,940	\$60,000
				Budget			\$60,000

8.4 Random Errors and Bootstrapping

Recall that the optimization methods in Algorithms 8.1, 8.2, and model 8.4 assume that the model scores are *exactly* correct for single treatment or multiple treatment problems. Algorithms 8.3 and 8.4, however, mitigate the inaccuracy of the individual-level model scores by using decile/semi-decile/cluster-level performance in the holdout sample. As explained in Section 8.2.1, this approach of using the grouped data in the holdout sample has the advantages of using group-level mean lift values (lower variance) and taking data from the holdout sample (lower re-substitution bias).

Are the group-level holdout sample lift values "accurate" enough? While they are better than using individual-level model scores, they still have some degree of uncertainty for the following reasons:

1. **Random Errors:** While sample lift values may be good representations of the true mean lift values (when group sizes are sufficiently large), there is no guarantee that a new sample (even following the same distribution) would result in the same set of lift values, due to randomness from sampling.
2. **Population Difference:** If the future target population is changed from the current population (because of demographic or regional differences, for example), one may make some appropriate adjustments – see Section 8.5 for the "data shift" problem.
3. **Systematic Change:** This can be driven by many factors, such as changes in the economy and customer behavioral changes, and is particularly important for products that are highly driven by the economy or market trends (the latter applies to many technology products such as tablet PCs and 3D TVs). Typical ways to handle this problem include updating models frequently enough to address the latest changes and including those factors (such as economic variables) in the model so that a change in the economy can change the model scores.

This section and the next address the first issue – random errors. Following Lo and Pachamanova (2015), a natural choice for mitigating random errors is through resampling, also known as bootstrapping. In other words, it assumes there are very many possible samples (resamples drawn from the original samples with replacement) so that we can calculate the lift values for each sample. Using results from many different samples, we then obtain a distribution of lift values, and, as a result, many useful statistics, such as standard deviations and percentiles, can be calculated (for each group/cluster) to assess the uncertainty. This section discusses how this is done; see Efron and Tibshirani (1993) or Davison and Hinkley (1997) for a full description of bootstrapping.

Algorithm 8.5 (Bootstrapped Holdout Sample Performance) (Adapted from Lo and Pachamanova (2015))

Assume a cluster solution is obtained using Algorithm 8.4.

1. Randomly draw a bootstrapped sample (resample) of the holdout sample with replacement. In SAS, this can be achieved using proc surveyselect.
2. Calculate the lift values $\Delta\hat{p}_{c1(1)}, \ldots, \Delta\hat{p}_{cm(1)}$ by cluster using the bootstrapped sample, where the additional subscript (1) indicates that values are from bootstrap sample 1.
3. Repeat steps 1–2 until B bootstrapped samples and their cluster-level lift values, $(\Delta\hat{p}_{c1(1)}, \ldots, \Delta\hat{p}_{cm(1)}), ,..\ (\Delta\hat{p}_{c1(B)}, \ldots, \Delta\hat{p}_{cm(B)})$, are all obtained.
4. Since we now have B lift values for each treatment and each cluster, we can compute their summary statistics, such as percentiles, medians, and standard deviations, as well as covariances or correlations to assess their uncertainty and dependency. For example, the qth percentile of Δp_{cj} can be determined by the inverse of the empirical distribution of $\Delta\hat{p}_{cj}$ evaluated at q, which can be denoted by $\hat{F}^{-1}_{\Delta\hat{p}_{cj}}(q)$. The bootstrapped standard deviation for each treatment score is estimated by:[6]

$$\widehat{SD}\left(\Delta p_{cj}\right) = \sqrt{\frac{\sum_{b=1}^{B}\left(\Delta\hat{p}_{cj(b)} - \overline{\Delta\hat{p}_{cj}}\right)}{B-1}},$$

where $\overline{\Delta\hat{p}_{cj}}$ is the average of the B bootstrapped estimates $\Delta\hat{p}_{cj(1)}, \ldots, \Delta\hat{p}_{cj(B)}$. Moreover, the covariance between two treatment scores (for treatments k and l) within the same cluster c is estimated by:

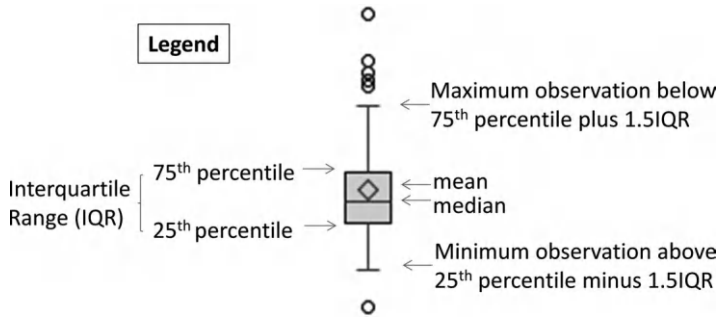
$$\widehat{Cov}\left(\Delta p_{ck},\ \Delta p_{cl}\right) = \frac{\sum_{b=1}^{B}\left(\Delta\hat{p}_{ck(b)} - \overline{\Delta\hat{p}_{ck}}\right)\left(\Delta\hat{p}_{cl(b)} - \overline{\Delta\hat{p}_{cl}}\right)}{B-1}.$$

Similarly, the covariance between two treatment scores (for treatments k and l) from two different clusters c and c' can be estimated by:

$$\widehat{Cov}\left(\Delta p_{ck},\ \Delta p_{c'l}\right) = \frac{\sum_{b=1}^{B}\left(\Delta\hat{p}_{ck(b)} - \overline{\Delta\hat{p}_{ck}}\right)\left(\Delta\hat{p}_{c'l(b)} - \overline{\Delta\hat{p}_{c'l}}\right)}{B-1}.$$

Example 8.2: Continuation of Example 8.1 in Section 8.3

Applying the bootstrapping process in Algorithm 8.5 to the previous example with $B = 60$ bootstrap samples, we have obtained the results in Figures 8.4 and 8.5 and Tables 8.3 and 8.4. First, Figure 8.4 shows the box plot of the distributions of men's and women's lift values by cluster (see the legend of Figure 8.4 for interpretation), indicating that some clusters have a wider dispersion of lift values than others. In Table 8.3, cluster-level lift



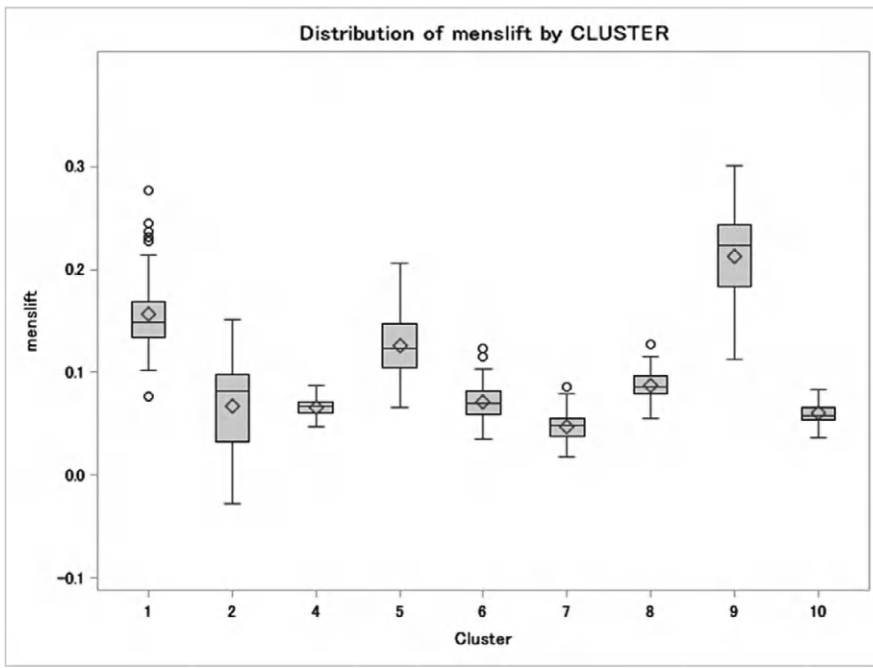


FIGURE 8.4
Distribution of men's and women's lift by cluster using bootstrapping. *(Continued)*

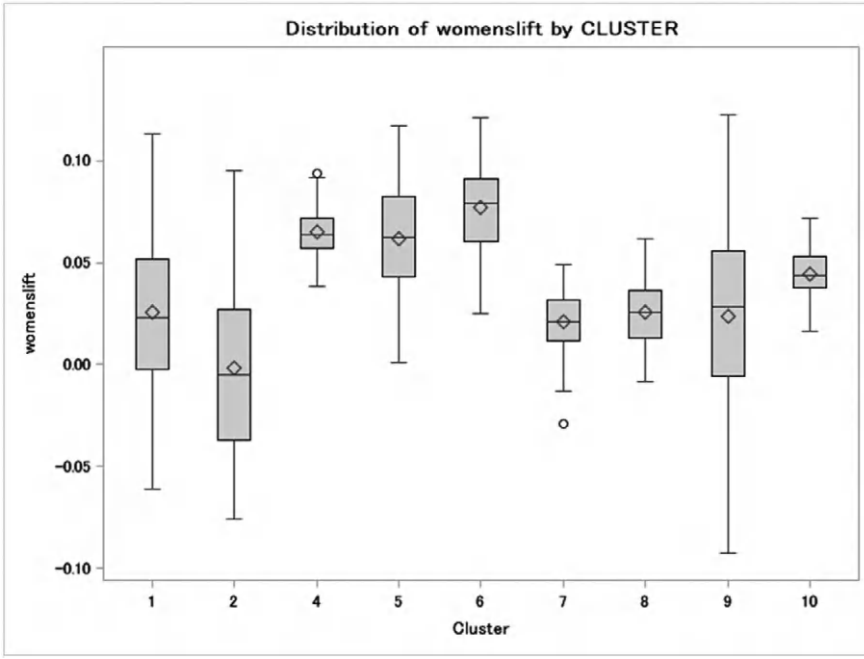


FIGURE 8.4 *(Continued)*

value percentiles are shown along with the bootstrapped standard deviations, which indicate that the uncertainty measured by standard deviation is not uniform across clusters. The correlation matrix of the cluster-level men's lift values and cluster-level women's lift values is shown in Table 8.4 along with a subset of scatter plots shown in Figure 8.5. Notice that the correlations across clusters are generally small for most combinations.

8.5 Optimization under Uncertainty

All algorithms in Sections 8.2 and 8.3 are optimization methods designed to solve *deterministic* problems; that is, all input parameters in the optimization model are assumed to be known with total certainty. In our context, the lift values (even at the cluster level) are estimated and can have a high degree of uncertainty. Rather than taking a "leap of faith" on the estimates, are there ways to handle uncertainty?

The most obvious method from classical *deterministic* optimization textbooks is to conduct a sensitivity analysis (e.g., Ravindran et al. 1987 and Taha 2010), which means the analyst would try different values of the

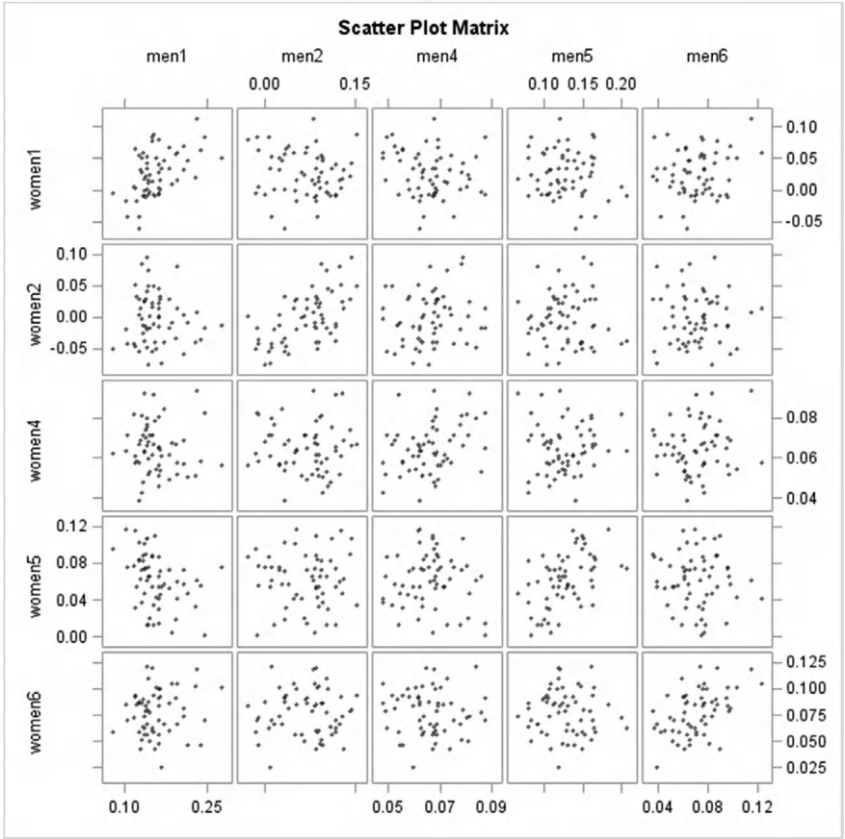


FIGURE 8.5
Scatter plots of men's and women's lift scores by cluster. (Truncated plots: Only clusters 1, 2, 4, 5, and 6 are shown.)

TABLE 8.3

Bootstrapped Results of Lift Values by Cluster

Cluster	Men's 5th Percentile	Women's 5th Percentile	Men's 50th Percentile	Women's 50th Percentile	Men's 95th Percentile	Women's 95th Percentile	Bootstrap SD for Men's	Bootstrap SD for Women's
1	0.1105	–0.0298	0.1491	0.0229	0.2341	0.0840	0.0376	0.0357
2	–0.0111	–0.0575	0.0813	–0.0050	0.1351	0.0774	0.0452	0.0401
4	0.0489	0.0471	0.0671	0.0640	0.0848	0.0879	0.0099	0.0122
5	0.0791	0.0128	0.1233	0.0623	0.1746	0.1083	0.0301	0.0291
6	0.0386	0.0440	0.0703	0.0791	0.1003	0.1140	0.0186	0.0213
7	0.0255	–0.0045	0.0479	0.0209	0.0753	0.0480	0.0149	0.0162
8	0.0611	0.0002	0.0865	0.0256	0.1145	0.0535	0.0159	0.0158
9	0.1231	–0.0543	0.2231	0.0281	0.2912	0.0955	0.0459	0.0457
10	0.0438	0.0281	0.0579	0.0436	0.0787	0.0633	0.0101	0.0111

TABLE 8.4

Correlation Matrix of Men's and Women's Lift Values by Cluster (Men1 = Men's Lift Values for Cluster 1, etc.)

	Men1	Women1	Men2	Women2	Men4	Women4	Men5	Women5	Men6	Women6	Men7	Women7	Men8	Women8	Men9	Women9	Men10	Women10
Men1	1.000																	
Women1	0.498	1.000																
Men2	−0.156	−0.108	1.000															
Women2	−0.058	0.006	0.591	1.000														
Men4	−0.100	−0.187	−0.026	0.133	1.000													
Women4	−0.039	0.158	−0.021	−0.037	0.316	1.000												
Men5	−0.176	−0.133	−0.029	−0.036	−0.037	0.048	1.000											
Women5	−0.373	−0.178	0.011	0.037	−0.068	−0.052	0.499	1.000										
Men6	0.249	0.170	−0.022	0.003	−0.067	0.109	−0.222	−0.069	1.000									
Women6	0.127	0.233	0.014	0.109	−0.044	−0.113	−0.034	−0.139	0.467	1.000								
Men7	0.194	0.073	−0.008	0.058	−0.008	−0.028	0.262	−0.004	0.022	−0.027	1.000							
Women7	0.083	−0.027	0.182	0.132	−0.024	0.106	0.101	0.024	−0.202	−0.226	0.561	1.000						
Men8	0.054	0.004	−0.029	−0.022	−0.226	−0.277	−0.304	−0.131	0.242	0.012	−0.065	0.029	1.000					
Women8	0.148	0.236	−0.030	−0.124	−0.218	−0.296	−0.161	−0.023	0.040	-0.037	−0.157	−0.208	0.538	1.000				
Men9	−0.095	−0.125	0.075	0.125	0.090	−0.044	0.133	−0.089	−0.139	0.049	−0.159	−0.110	−0.156	−0.089	1.000			
Women9	0.008	0.040	−0.009	−0.025	0.239	0.051	−0.158	−0.143	−0.039	−0.058	−0.156	−0.131	−0.099	−0.144	0.264	1.000		
Men10	0.073	−0.002	−0.050	−0.111	−0.197	−0.009	−0.064	−0.277	0.023	0.018	−0.069	−0.050	0.088	0.025	0.120	0.201	1.000	
Women10	0.129	0.187	−0.193	−0.154	0.011	−0.116	−0.066	−0.266	0.130	0.099	−0.062	−0.110	0.170	0.342	0.225	0.110	0.326	1.000

uncertain parameters as an input to optimization and see if the resulting solutions would change. In standard sensitivity analysis in LP, one can also determine the range of input parameters such that the resulting solutions remain the same. As pointed out by relevant literature such as Wallace (2000), Higle and Wallace (2003), and King and Wallace (2012), such sensitivity analysis, or "what-if" analysis, is a post-optimality investigation of how a change in the data may affect the solution and is not a direct way to handle uncertainty.[7] Sensitivity analysis can only determine different solutions that are optimal for different values of uncertain parameters and cannot determine a *single* solution that addresses the variability of parameters. In addition to its theoretical weakness, in our context of treatment optimization in uplift analytics, running a sensitivity analysis will face two practical challenges:

1. What alternative values would the analyst use for sensitivity analysis? One can make some simple assumptions or apply statistical methods such as parametric confidence intervals or bootstrapping (e.g., Section 8.4) to assess the potential range of uncertainty. However, that could mean repeatedly running optimization for *very many* possible scenarios, especially when the number of clusters or the number of treatments is not small. In our relatively simple example in Example 8.1, we have 9 clusters and 2 treatments, resulting in 18 lift values, and each can have a high degree of uncertainty with a correlation structure among them (as investigated in Example 8.2), resulting in a huge optimization task to run all possible scenarios.
2. Once we have many solutions for many possible scenarios, which one should we ultimately select as the final solution? This is not a simple answer if there are many different solutions. One may consider the "worst case" scenario (which will be discussed in Subsection 8.5.2 under Robust Optimization). In this case, the analyst would actually not need to consider so many possible scenarios in the first place. Another way is to look for a common optimization solution (if available) under several scenarios, that is, searching for a solution that is relatively insensitive to many possible scenarios, which would require some effort and "luck" (and what if such insensitive solution does not exist?).

As a result, applying sensitivity analysis may appear simple but is not a very practical way to address many possible scenarios, especially when systematic and scientific methods are actually available to solve stochastic optimization directly.

This section will focus on various optimization methods to handle uncertainty. As discussed before, we will address the uncertainty due to random

errors from the lift value estimates. The bootstrapping or resampling methodology from Section 8.4 will be used as a key input.

We will first discuss the earliest and also the most well-known optimization method for handling uncertainty, Mean-Variance Optimization, followed by more recent techniques including robust optimization (RO) and stochastic programming (SP). All these methods can be found in the optimization literature and advanced textbooks on optimization. We will provide a practical description and tailor our discussion to treatment optimization for uplift analytics.

8.5.1 Mean-Variance Optimization (MVO) – Excel Solver Using Bootstrap Results

Over 60 years have passed since Markowitz (1952) introduced his work on MVO, also known as the Mean-Variance Criterion, as part of the Modern Portfolio Theory that won the Nobel Memorial Prize in Economic Sciences in 1990. MVO is now widely known and is a popular method in finance as well as other fields; see, for example, Cornuejols and Tutuncu (2007), Fabozzi et al. (2007), and Zenios (2007).

The main idea of MVO is to balance return and risk, where risk is measured by standard deviation (aka volatility). The objective is to maximize the overall mean portfolio return (i.e., a weighted average of asset returns, with weights or asset allocations to be determined by optimization) with the risk (measured by standard deviation of the portfolio return) not higher than a specific level. An alternative objective is the opposite: To minimize the overall portfolio variance with the mean portfolio return at least higher than a certain level. Since variance is a quadratic function of portfolio weights, the resulting optimization problem is no longer linear but is a quadratic program, which is still relatively straightforward to solve.

The goal of MVO is to optimally assign weights to assets for asset allocation, and it can be applied to other contexts such as optimizing projects, treatments, products, customers, employees, etc. In our situation, the subject of interest is treatment by individual customer or treatment by cluster (of individuals). Considering m treatments and C clusters of individuals as your "assets" as in model (8.7), we aim at determining the optimal quantity for each treatment and each cluster. Following the framework of MVO, we constrain the variance of the objective function (where the objective function is the estimated number of overall incremental responders):

$$\text{Maximize E}\left(\sum_{c=1}^{C}\sum_{j=1}^{m}\Delta p_{cj}x_{cj}\right)=\sum_{c=1}^{C}\sum_{j=1}^{m}E\left(\Delta p_{cj}\right)x_{cj} \tag{8.8}$$

subject to:

$$\text{Var}\left(\sum_{c=1}^{C}\sum_{j=1}^{m}\Delta p_{cj}x_{cj}\right) \le v, \quad \text{Maximum Uncertainty (or Risk) Constraint,}$$

$$\sum_{c=1}^{C}\sum_{j=1}^{m} c_j x_{cj} \le B, \quad \text{Budget Constraint,}$$

$$\sum_{j=1}^{m} x_{cj} \le N_c, \quad \text{for } c = 1, \ldots, C, \quad \text{Cluster Size Constraint, and}$$

$$x_{cj} \ge 0, \quad c = 1, \ldots, C; \quad j = 1, \ldots, m,$$

where as before x_{cj} is the decision variable for a number of treatments of type j assigned to cluster c, and c_j = cost of treatment j for each individual.

Again, the expected (mean) values inside the objective function of model (8.8) above can be estimated by the holdout sample lift values, $\Delta\hat{p}_{cj}$. The variance component in model (8.8) is slightly more complicated because of all the possible pairwise covariances:

$$\begin{aligned}
&\text{Var}\left(\sum_{c=1}^{C}\sum_{j=1}^{m}\Delta p_{cj}x_{cj}\right) \\
&= \sum_{c=1}^{C}\sum_{j=1}^{m} Var\left(\Delta p_{cj}\right)x_{cj}^{\,2} + 2\sum_{c}\sum_{c'>c}\sum_{j}\sum_{j'>j} x_{cj}x_{c'j'}\, Cov\left(\Delta p_{cj}, \Delta p_{c'j'}\right) \\
&= \boldsymbol{x}'\boldsymbol{\Omega}\boldsymbol{x}, \qquad (8.9)
\end{aligned}$$

where $\boldsymbol{x} = (x_{11}, \ldots, x_{1m}, \ldots\ldots, x_{C1}, \ldots, x_{Cm})'$ and Ω = variance-covariance matrix of $(\Delta p_{11}, \ldots, \Delta p_{1m}, \ldots\ldots, \Delta p_{C1}, \ldots, \Delta p_{Cm})'$. One advantage of expressing Eqn. (8.9) in a matrix form is to facilitate numerical computations using matrices (see Example 8.3a below where Excel is used). The variances and covariances inside Eqn. (8.9) can be estimated by bootstrapping in Algorithm 8.5 of Section 8.4.

By changing the value of v in model (8.8), we can trace out the "efficient frontier" by obtaining different optimal solutions based on tradeoffs between the mean and variance, where the "efficient frontier" is defined as the set of solutions that have the highest possible mean return given a risk level or the lowest possible standard deviation given a mean return, an idea that will be made clearer in the numerical example below. The bootstrapping procedure in Algorithm (8.5) can be employed to estimate the variances and covariances in Eqn. (8.9).

According to the theory of MVO, an alternative formulation[8] to model (8.8) is minimizing the overall variance subject to a minimum constraint for the expected value:

$$\text{Minimize Var}\left(\sum_{c=1}^{C}\sum_{j=1}^{m}\Delta p_{cj}x_{cj}\right) \tag{8.10}$$

subject to:

$$\sum_{c=1}^{C}\sum_{j=1}^{m}E\left(\Delta p_{cj}\right)x_{cj}\geq e \quad \text{Minimum Mean Value Constraint,}$$

$$\sum_{c=1}^{C}\sum_{j=1}^{m}c_j x_{cj}\leq B, \quad \text{Budget Constraint,}$$

$$\sum_{j=1}^{m}x_{cj}\leq N_c, \quad \text{for } c=1,\ldots,C, \quad \text{Cluster Size Constraint, and}$$

$$x_{cj}\geq 0, \quad c=1,\ldots,C; \quad j=1,\ldots,m.$$

The efficient frontier using model (8.10) can be obtained by changing the value of e. Note that, strictly speaking, MVO is generally not regarded a "true" stochastic optimization method as the "asset expected returns," which are equivalent to $E(\Delta p_{cj})$ are assumed known with certainty (estimated by holdout sample lift values $\Delta\hat{p}_{cj}$ in practice). Nevertheless, it does handle uncertainty by explicitly incorporating the variance in optimization.

Example 8.3a (MVO version of Example 8.1 using results from Example 8.2), adapted from Lo and Pachamanova (2015)

Recall that in Example 8.1, we solved a joint men's and women's merchandise optimization using the lift values estimated from the holdout sample. Since we simply plugged the estimated values into a deterministic LP model, we ignored uncertainty. From Example 8.2, we learned that the standard deviations of the men's and women's lift values are quite different across clusters (Table 8.3), and we also calculated the correlation matrix in Table 8.4. We now apply these numerical bootstrapped results from Tables 8.3 and 8.4 to model (8.10).

We first calculate the total variance in Excel as a function of variances and covariances using Eqn. (8.9).[9] We then again employ Excel Solver to minimize the total variance by adding a constraint for the minimum mean value (first constraint of model (8.10)). Since this is a nonlinear (quadratic) programming model as opposed to an LP model, the "GRG Nonlinear" option in Solver

TABLE 8.5

MVO Illustration – Minimizing Variance with the Minimum Mean Value Set at 4,000

Cluster	Cluster Size in Holdout Sample	Obs. Lift in Response: Men's	Obs. Lift in Response: Women's	Cost per Treatment ($)	Decision var on Number of Men's	Decision var on Number of Women's	Total Number of Treated by Cluster
Overall	18985	0.07408	0.043863				
1	418	0.1587	0.0224	1	3,146	1,034	4,180
2	565	0.0652	–0.0055	1	–	–	–
4	6,022	0.0658	0.0628	1	713	3,254	3,967
5	1,237	0.1290	0.0618	1	1,232	7,991	9,223
6	894	0.0672	0.0760	1	7,477	1,463	8,940
7	2,924	0.0519	0.0213	1	9,453	278	9,731
8	2,807	0.0868	0.0254	1	2,069	559	2,628
9	410	0.2249	0.0239	1	4,100	–	4,100
10	3,708	0.0572	0.0426	1	190	7,907	8,097
Total	18,985			obj value	2,811	1,189	4,000
				cost	$28,380	$22,487	$50,867
				Budget			$60,000

is selected instead of "Simplex LP." Table 8.5 shows the results for a special case where the minimum mean value is set at 4,000 (recall that the optimal solution to the corresponding deterministic optimization model from Table 8.2 in Example 8.1 has a mean value of 6,453). If we focus on the two decision variable columns in Table 8.5 and compare them to those in Table 8.2, the assignment appears to be more diversified with generally smaller quantities in each cell.

Figure 8.6 summarizes several possible optimal solutions – obtained by minimizing the overall variance with a changing minimum mean value, resulting in a mean-standard deviation "efficient frontier." For example, at the upper right of Figure 8.6, the squared point at (standard deviation, mean) = (1018, 6453) is the original solution from Example 8.1. We then set the minimum mean incremental responders at various values, solve the associated quadratic programming model to minimize the variance, and plot these different solutions[10] in Figure 8.6. The one closest to the original solution near the upper right has the minimum mean value set at 6,400 with a resulting standard deviation of 939. At the lower left, we set the minimum mean value at 1,000 and obtained an optimal solution with a standard deviation of 67. Which solution to ultimately select for an actual application depends on the user's balancing art of risk (standard deviation) and expected return of the objective function.

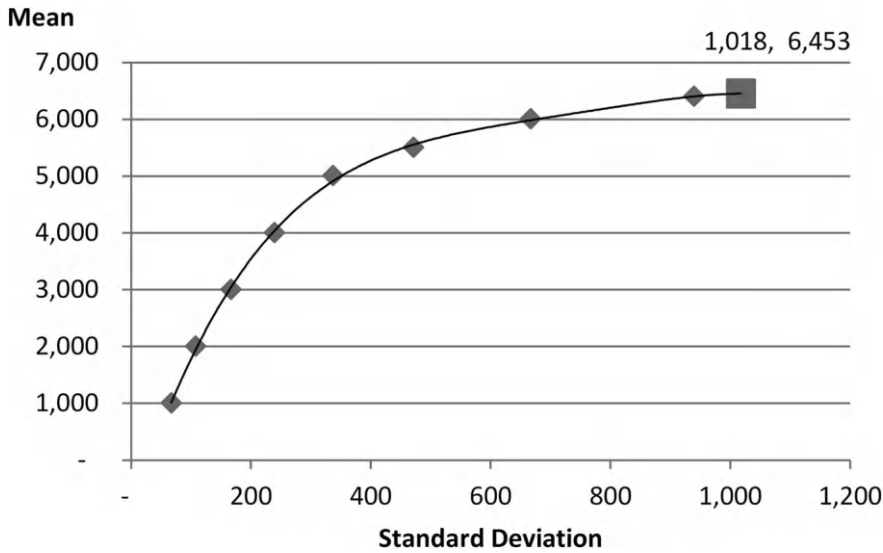


FIGURE 8.6
Mean and standard deviation solution tradeoff – "Efficient Frontier".

8.5.2 Robust Optimization (RO)

RO aims at finding the best solution when the parameters in an optimization problem are not fixed but are allowed to vary in pre-specified uncertainty sets. In practice, the RO approach often reduces to solving the optimization problem when the uncertainties take on "worst-case values." In a maximization problem, it will be maximizing the objective function under some kind of worst-case scenario for the coefficients in the problem.

RO itself is a vast area of research, and there are multiple ways to formulate the problem as well as different choices for uncertainty sets; see Tutuncu and Koenig (2004), Fabozzi et al. (2007), or Bertsimas et al. (2011). We will only illustrate with a simple and practical method here.

The simplest example is when the input parameters in the optimization problem are allowed to take values within interval uncertainty sets, that is, when we specify an upper and lower bound for each parameter. Recall model (8.7) from Algorithm (8.4), if we can establish a lower bound for each lift value, $\Delta p_{cj(l)}$ where the subscript (l) denotes the lower bound, we can then simply replace the lift values with their associated lower bounds ("worst case") while keeping everything else the same, resulting in a standard deterministic LP model:

$$\text{Maximize} \sum_{c=1}^{C} \sum_{j=1}^{m} \Delta p_{cj(l)} x_{cj} \tag{8.11}$$

subject to:

$$\sum_{c=1}^{C}\sum_{j=1}^{m} c_j x_{cj} \le B, \quad \text{Budget Constraint,}$$

$$\sum_{j=1}^{m} x_{cj} \le N_c, \quad \text{for } c = 1, \ldots, C, \quad \text{Cluster Size Constraint, and}$$

$$x_{cj} \ge 0, \quad c = 1, \ldots, C; \quad j = 1, \ldots, m.$$

How do we find the lower bounds $\Delta p_{cj(l)}$? The simplest method is to use a parametric lower bound based on the usual one-sided confidence interval, which assumes asymptotic normality for the sample estimator of the cluster-level lift value. Specifically, we first estimate the standard deviation of the sample lift value for cluster c and treatment j from the holdout sample:

$$\widehat{SD}\left(\Delta\hat{p}_{cj}\right) = \sqrt{\frac{\hat{p}_{cj}\left(1-\hat{p}_{cj}\right)}{n_{cj}} + \frac{\hat{p}_{c0}\left(1-\hat{p}_{c0}\right)}{n_{c0}}}, \tag{8.12a}$$

where $\hat{p}_{cj}$ is the sample response rate for cluster c and treatment j, and $\hat{p}_{c0}$ is the sample response rate for cluster c in the control group, and n_{cj} and n_{c0} are their associated sample sizes.

Then, the parametric lower bound is given by:

$$\Delta p_{cj(l)} = \Delta\hat{p}_{cj} - \lambda\widehat{SD}\left(\Delta\hat{p}_{cj}\right). \tag{8.12b}$$

For example, if we are to approximate the 5th percentile using this parametric lower bound, $-\lambda$ can be set to $\Phi^{-1}(0.05) = -1.645$, or $\lambda = 1.645$, assuming $\Delta\hat{p}_{cj}$ is (asymptotically) normally distributed. Similarly, if we would like to be more conservative, using the 1st percentile as the lower bound to present the "worst case," λ can be set to $-\Phi^{-1}(0.01) = 2.326$.

A generally better approach is to employ the bootstrapping procedure in Algorithm 8.5 of Section 8.4, as it is nonparametric or distribution-free; that is, no normality assumption is required. We will illustrate both the parametric and nonparametric lower bounds below with an example.

Example 8.3b (Continuation of Example 8.3a)

Replacing the estimated lift values in columns 3 and 4 of Table 8.2 by the 5th percentiles from columns 2 and 3 in Table 8.3, we have a lower bound, or the "worst case," for each lift value (for each treatment at the cluster level). Repeating the process outlined in Example 8.1 with the 5th percentiles, we solve a new problem using Excel Solver ("Simplex LP" is selected as it becomes a standard LP model).

Tables 8.6a and 8.6b report the new LP solutions using the parametric and nonparametric (bootstrapped) lower bounds, respectively. The standard

TABLE 8.6a

Linear Programming Computations Using Cluster-level "Worst" Lift Values: Using Parametric Lower Bounds

Cluster	Cluster Size in New Data	SD of Men's Lift	SD of Women's Lift	5th Percentile of Men's	5th Percentile of Women's	Cost per Treatment ($)	Decision var on Number of Men's	Decision var on Number of Women's	Total Number of Treated by Cluster
1	4,180	0.041	0.034	0.09073	−0.03386	1	4,180	–	4,180
2	5,650	0.041	0.038	−0.00144	-0.06767	1	–	–	–
4	60,220	0.011	0.011	0.04790	0.04521	1	11,280	–	11,280
5	12,370	0.029	0.027	0.08154	0.01754	1	12,370	–	12,370
6	8,940	0.022	0.023	0.03109	0.03887	1	–	–	–
7	29,240	0.016	0.015	0.02632	−0.00296	1	–	–	–
8	28,070	0.017	0.016	0.05868	−0.00057	1	28,070	–	28,070
9	4,100	0.055	0.048	0.13404	−0.05485	1	4,100	–	4,100
10	37,080	0.012	0.011	0.03800	0.02419	1	–	–	–
Total	189,850					obj value	4,125	–	4,125
						Cost	$60,000	$	$60,000
						Budget			$60,000

TABLE 8.6b

Linear Programming Computations Using Cluster-level "Worst" Lift Values: Using Bootstrapping

Cluster	Cluster Size in New Data	5th Percentile of Men's	5th Percentile of Women's	Cost per Treatment ($)	Decision var on Number of Men's	Decision var on Number of Women's	Total Number of Treated by Cluster
1	4,180	0.11048	−0.02982	1	4,180	–	4,180
2	5,650	−0.01111	−0.05751	1	–	–	–
4	60,220	0.04894	0.047102	1	11,280	–	11,280
5	12,370	0.07908	0.012833	1	12,370	–	12,370
6	8,940	0.03861	0.044034	1	–	–	–
7	29,240	0.02547	−0.00449	1	–	–	–
8	28,070	0.0611	0.000176	1	28,070	–	28,070
9	4,100	0.12311	−0.05429	1	4,100	–	4,100
10	37,080	0.04383	0.028138	1	–	–	–
Total	189,850			obj value	4,212		4,212
				cost	$60,000	$	$60,000
				Budget			$60,000

deviations in Table 8.6a are computed using the parametric formula in Eqn. (8.12a). Note that in this example the 5th percentiles (of both men's and women's lift values) are relatively close in Tables 8.6a and 8.6b, which means the parametric lower bounds are a good approximation, at least in this example. As a result, Tables 8.6a and 8.6b result in the same solution for the decision variables (the optimal objective values are slightly different because they are based on different [but similar] estimates of lower bounds). Comparing this new solution to the original deterministic solution in Table 8.2, cluster 6 is no longer assigned any treatment, and cluster 4 now has a higher quantity. This should not be a surprise because cluster 6's 5th percentiles (both men's and women's) are relatively small compared to cluster 4's, while previously in Table 8.2 cluster 6's lift values are slightly higher than the corresponding values in cluster 4 in Table 8.2. Another observation is that the objective function value (mean value of total incremental responders) is now down to 4,212 (in Table 8.6b, or 4,215 in Table 8.6a) but it is based on the 5th percentiles of lift values. If we keep this solution and replace the 5th percentiles with the original lift values at the cluster level to compute the objective function value, we obtain 6,361 as the mean value, which is only marginally (1.1%) lower than the original number (6,453) reported in Table 8.2. As a result, if uncertainty is taken into account in this way, the new solution seems a reasonable way to diversify.

Comparing the MVO solution in Table 8.5 to the RO solutions in Tables 8.6a and 8.6b, we see that the MVO solution is more diverse in the sense of assigning quantities to more clusters, which is not a surprise given the goal of minimum variance in MVO versus the goal of maximizing a lower bound (a conservative solution) in RO.

8.5.3 Stochastic Programming (SP)

Stochastic Programming (SP) directly addresses any uncertainty through probability distributions as opposed to employing a pessimistic scenario as in RO or managing variances as in MVO. Because of its combination of a wide range of applications and strong and established theories, SP itself is a key subfield of optimization theory and Operations Research and has been around for many decades – one of the earliest contributions is from Dantzig (1955). For further study, see Wallace and Ziemba (2005), Cornuejols and Tutuncu (2007), Zenios (2007), or King and Wallace (2012) for practical introductions and applications, and Birge and Louveaux (1997) or Shapiro et al. (2014) for the more mathematically inclined. We will discuss techniques from SP that are most relevant and practical to uplift optimization.[11]

Considering again model (8.7) in Section 8.2.2, one may impose an additional probability constraint to "guarantee" that a given result is achieved with a certain probability (known as chance constraint optimization, a branch of SP):

$$\text{Maximize} \sum_{c=1}^{C} \sum_{j=1}^{m} \Delta \hat{p}_{cj} x_{cj} \tag{8.13}$$

subject to:

$$P\left(\sum_{c=1}^{C}\sum_{j=1}^{m}\Delta p_{cj}x_{cj} \geq a\right) \geq b \text{ Probability Constraint } (0 \leq b \leq 1),$$

$$\sum_{c=1}^{C}\sum_{j=1}^{m}c_j x_{cj} \leq Budget, \text{ Budget Constraint,}$$

$$\sum_{j=1}^{m}x_{cj} \leq N_c, \text{ for } c = 1, \ldots, C, \text{ Cluster Size Constraint, and}$$

$$x_{cj} \geq 0,\ c = 1, \ldots, C;\ j = 1, \ldots, m.$$

Alternatively, rather than maximizing the estimated (mean) number of incremental responders, one may be interested in maximizing the probability of achieving a given number of incremental responders:

$$\text{Maximize } P\left(\sum_{c=1}^{C}\sum_{j=1}^{m}\Delta p_{cj}x_{cj} \geq z_0\right) \tag{8.14}$$

subject to (as before):

$$\sum_{c=1}^{C}\sum_{j=1}^{m}c_j x_{cj} \leq Budget, \text{ Budget Constraint,}$$

$$\sum_{j=1}^{m}x_{cj} \leq N_c, \text{ for } c = 1, \ldots, C, \text{ Cluster Size Constraint, and}$$

$$x_{cj} \geq 0,\ c = 1, \ldots, C;\ j = 1, \ldots, m.$$

Furthermore, one may even change the objective function to another probability-related function, such as an αth percentile (also known as Value-at-Risk or VaR[12]), for a small value of α:

$$\text{Maximize the } \alpha \text{ th percentile of } \sum_{c=1}^{C}\sum_{j=1}^{m}\Delta p_{cj}x_{cj}, \tag{8.15}$$

while keeping the same set of constraints in Eqn. (8.14).

On the surface, model (8.15) looks similar to the Robust Optimization (RO) model (8.11). In fact, they are different because model (8.11) ignores the dependency structure (correlations) among the lift values Δp_{cj}. For example, if two of the lift values are very negatively correlated, the chance that *both* of them will achieve their respective lower bound (say, 5th percentile) is very low. As a result, model (8.11) is more conservative than model (8.15) as the latter takes into account the dependency structure of the lift values.

Note that, unlike LP problems for deterministic optimization or even quadratic programming problems for MVO, models, Eqns. (8.13)–(8.15) may not satisfy the convexity property (see Appendix 8.1) and thus cannot guarantee global optima. Nevertheless, good heuristics are available to solve them, attempting to achieve some "good enough" solutions even though global optima are not guaranteed. See Appendix 8.2 for A Brief Introduction to Simulation Optimization, a relatively simple technique employed in Example 8.3c.

Example 8.3c (Continuation of Example 8.3b)

We now include a probability constraint in the original model in Table 8.1. Using Oracle's Crystal Ball (CB),[13] we first input the distributions for all men's and women's lift values, each following the bootstrap sample's marginal distribution (called "Custom Distribution" in the CB software). We then enter the entire correlation matrix of the bootstrapped lift values from Table 8.4 into the "Defined Correlations" area.

Knowing that the optimal solution in Example 8.1 results in 6,453 expected incremental responders (using the sample mean lift values), we now would want to have some guarantee that the actual number of incremental responders would be at least a given reasonably high quantity. Applying model (8.13), the following probability constraint is included to have an 80% probability that at least 5,500 incremental responders are achieved:

$$P\left(\text{no. of incremental responders} \geq 5,500\right) \geq 0.80,$$

$$\text{i.e., } P\left(\sum_{c=1}^{C}\sum_{j=1}^{2} \Delta p_{cj} x_{cj} \geq 5,500\right) \geq 0.80,$$

where C = number of clusters, while keeping the same objective and other constraints in Example 8.1.

See EPM Information Development Team (2009) for details on how to use the CB software. The solution[14] with the above additional probability constraint is shown in Table 8.7,[15] which is only slightly different from the original deterministic solution in Table 8.2 (highlighted columns indicate that those values have gone through random number generations in CB). The top panel of Figure 8.7 shows the histogram of the simulated optimal objective

TABLE 8.7

Probability Constrained Optimization – Summary of Results

Cluster	Cluster Size in New Data	Obs. Lift in Response: Men's	Obs. Lift in Response: Women's	Cost per Treatment ($)	Decision var on Number of Men's	Decision var on Number of Women's	Total Number of Treated by Cluster
1	4,180	0.1587	0.0224	1	4,180	–	4,180
2	5,650	0.0652	–0.0055	1	2,340	–	2,340
4	60,220	0.0658	0.0628	1	–	–	–
5	12,370	0.1290	0.0618	1	12,370	–	12,370
6	8,940	0.0672	0.0760	1	–	8,940	8,940
7	29,240	0.0519	0.0213	1	–	–	–
8	28,070	0.0868	0.0254	1	28,070	–	28,070
9	4,100	0.2249	0.0239	1	4,100	–	4,100
10	37,080	0.0572	0.0426	1	–	–	–
Total	189,850			obj value	5,772	680	6,451
				cost	$51,060	$8,940	$60,000
				Budget			$60,000

function value (no. of incremental responders) and its fitted curve using a lognormal distribution (best fit).[16] The estimated probability (based on sample proportions) of achieving 5,500 is 96.81% at the bottom below the histogram and is reflected by the darker portion of the histogram. The lower panel of Figure 8.7 shows the simulation history of the solution, showing that after about 300 simulations, the algorithm converges to the final solution (there is no infeasible solution shown in the graph even though the legend includes an infeasible solution line).

Instead of including a probability constraint, we now use a probability in the objective function in the form of model (8.14):

$$\text{Maximize } P\left(\text{no. of incremental responders} \geq 6,000\right),$$

$$\text{i.e., Maximize } P\left(\sum_{c=1}^{C}\sum_{j=1}^{2} \Delta p_{cj} x_{cj} \geq 6,000\right),$$

while again keeping all constraints in Example 8.1. Note that the value 6,000 inside the probability objective is relatively large given that the optimal objective function value of the original deterministic model is 6,453 from Table 8.2. The new solution with this probability objective function is summarized in

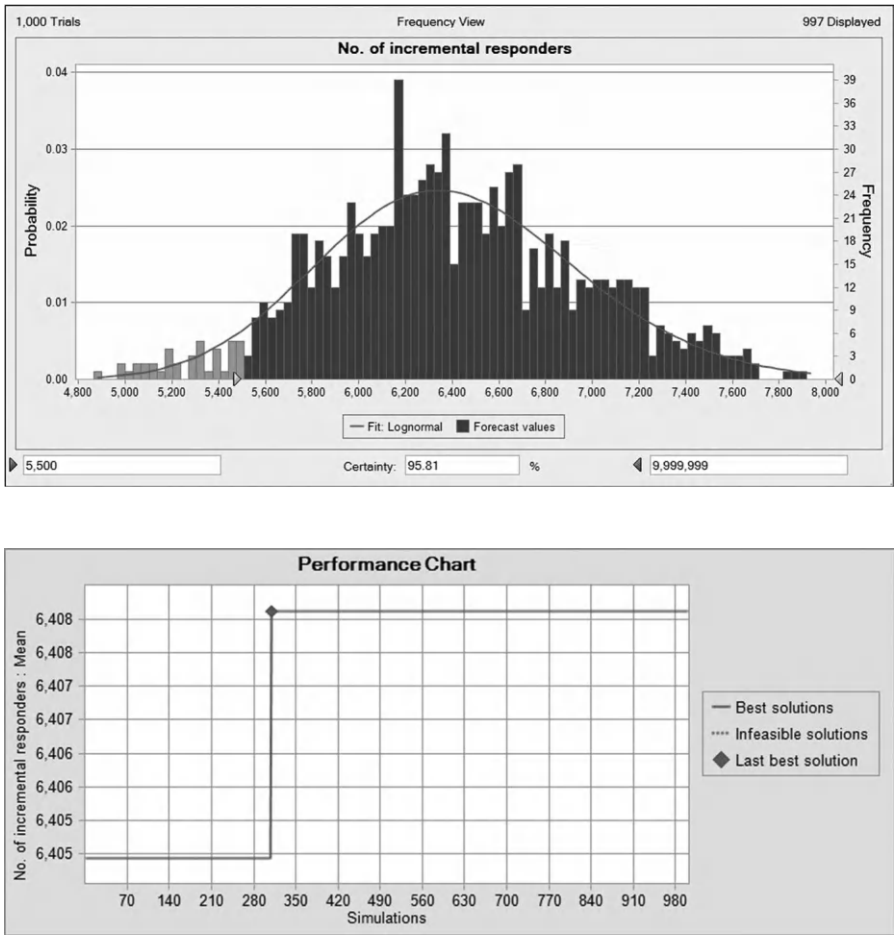


FIGURE 8.7
Probability constrained optimization – Distribution of no. of incremental responders.

Table 8.8, which is not very different from the solution in Table 8.2, other than cluster 2 is also assigned with a small quantity and thus is slightly more diverse. The top panel of Figure 8.8 shows the estimated probability distribution of the number of incremental responders obtained from this optimization. Since 6,000 is used in the probability objective function, one would be interested in the probability of achieving at least 6,000, which is 76.82% as listed at the bottom of the histogram and also reflected by the darker portion of the histogram in Figure 8.8. The bottom panel of Figure 8.8 shows the simulation history, which shows that it slowly and gradually converges to the final solution (76.82% probability).

TABLE 8.8

Probability as Objective Function – Summary of Results

Cluster	Cluster Size in New Data	Obs. Lift in Response: Men's	Obs. Lift in Response: Women's	Cost per Treatment ($)	Decision var on Number of Men's	Decision var on Number of Women's	Total Number of Treated by Cluster
1	4,180	0.1587	0.0224	1	4,180	–	4,180
2	5,650	0.0652	–0.0055	1	305	–	305
4	60,220	0.0658	0.0628	1	1,573	462	2,035
5	12,370	0.1290	0.0618	1	12,370	–	12,370
6	8,940	0.0672	0.0760	1	–	8,940	8,940
7	29,240	0.0519	0.0213	1	–	–	–
8	28,070	0.0868	0.0254	1	28,070	–	28,070
9	4,100	0.2249	0.0239	1	4,100	–	4,100
10	37,080	0.0572	0.0426	1	–	–	–
Total	189,850			obj value	5,743	709	6,451
				cost	$50,598	$9,402	$60,000
				Budget			$60,000

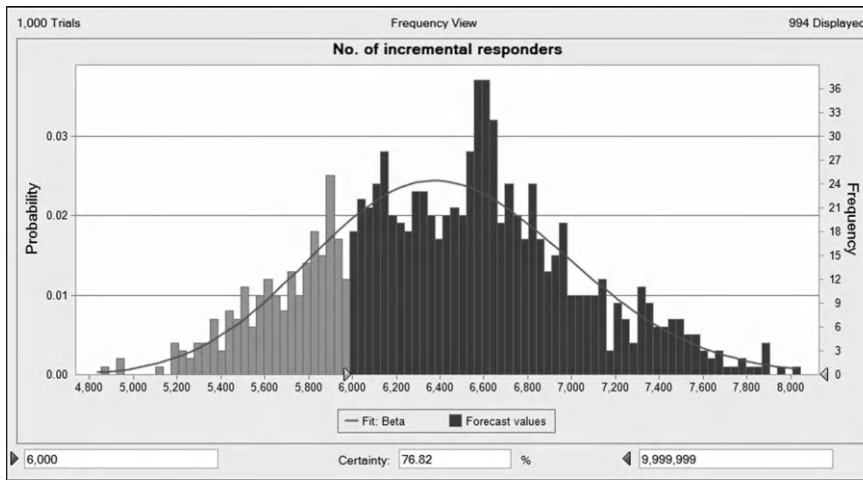


FIGURE 8.8

Probability as objective function – Distribution of no. of incremental responders. *(Continued)*

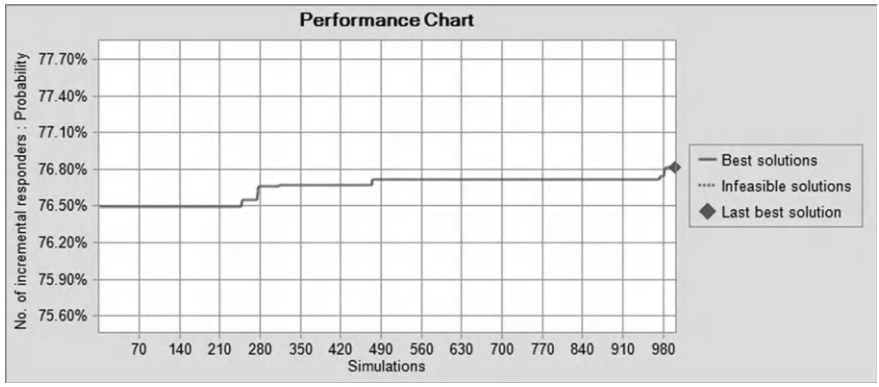


FIGURE 8.8 *(Continued)*

Last but not least, we follow model (8.15) to maximize the "worst" case scenario, represented by the 5th percentile:

$$\text{Maximize the 5th percentile of } \sum_{c=1}^{C}\sum_{j=1}^{2} \Delta p_{cj}\, x_{cj},$$

while again keeping the same constraints in Example 8.1. The resulting solution in Table 8.9 shows that it appears slightly more diverse than the original

TABLE 8.9

Maximizing 5th Percentile – Summary of Results

Cluster	Cluster Size in New Data	Obs. Lift in Response: Men's	Obs. Lift in Response: Women's	Cost per Treatment ($)	Decision var on Number of Men's	Decision var on Number of Women's	Total of Treated Women's by Cluster
1	4,180	0.1587	0.0224	1	4,180	–	4,180
2	5,650	0.0652	−0.0055	1	200	–	200
4	60,220	0.0658	0.0628	1	4,274	–	4,274
5	12,370	0.1290	0.0618	1	12,370	–	12,370
6	8,940	0.0672	0.0760	1	–	6,806	6,806
7	29,240	0.0519	0.0213	1	–	–	–
8	28,070	0.0868	0.0254	1	28,070	–	28,070
9	4,100	0.2249	0.0239	1	4,100	–	4,100
10	37,080	0.0572	0.0426	1	–	–	–
Total	189,850			obj value	5,913	517	6,431
				cost	$53,194	$6,806	$60,000
				Budget			$60,000

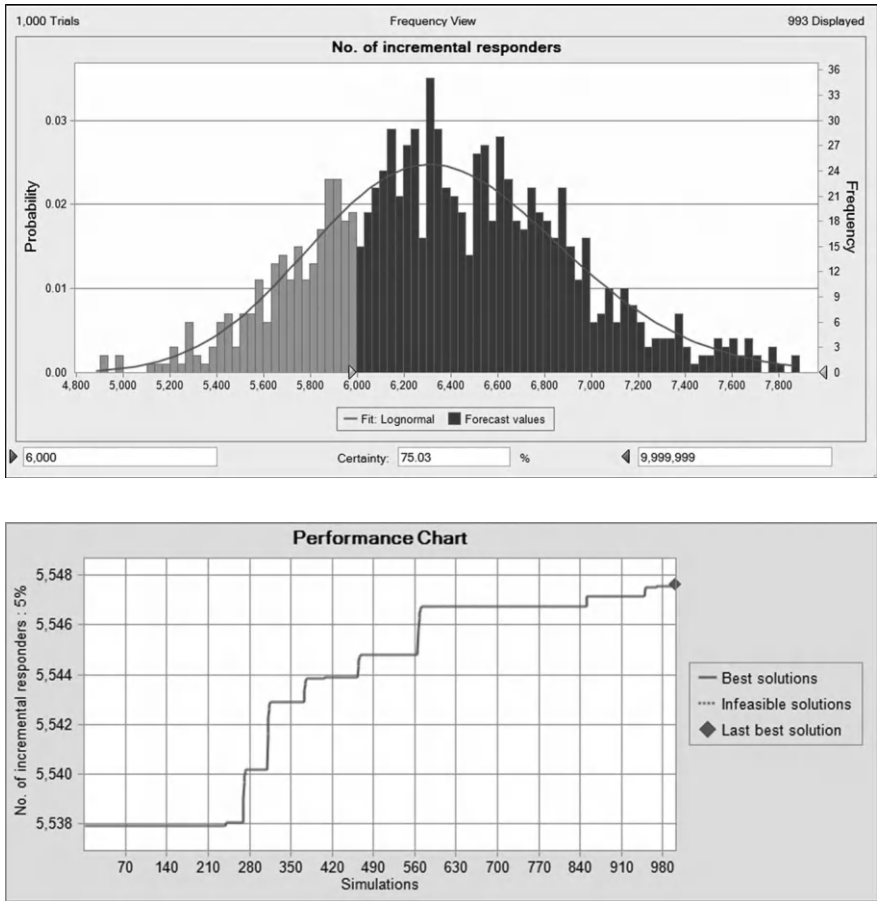


FIGURE 8.9
Maximizing 5th percentile – Distribution of no. of incremental responders.

deterministic solution in Table 8.2 (because cluster 2 is also assigned). The top panel of Figure 8.9 shows the estimated probability distribution of the objective function value, with the simulation history shown in the lower panel of Figure 8.9.

8.6 Concluding Remarks

Chapter 6 discusses the basics of uplift modeling, where the objective is to construct a useful predictive model for predicting lift values. The goal is to select the right individual targets for a future campaign using an uplift model.

This chapter discusses mathematical and computational algorithms for treatment optimization in an uplift analytics setting. These algorithms serve different purposes in various situations. When there is only a single treatment used in the previous marketing campaign and also the same single treatment will be considered for a future campaign, model (8.3), Algorithm 8.1, or Algorithm 8.2 can be applied to select the right targets for treatment. Under a more general situation where multiple treatments are involved, a binary integer programming model (model (8.4)) can be employed to select the right individual targets for the right treatment. Due to its tremendous computational complexity and the difficulty in accurately estimating *individual-level* lift values, Algorithm 8.4 is proposed as a practical way to solve a much simpler optimization problem. The key to Algorithm 8.4 is using holdout sample lift estimates at the cluster level as an input to optimization.

Even though holdout sample data are used, Algorithm 8.4 is still based on statistical estimation, and as a result, the estimates are not known with complete certainty. While it is quite common to use statistical estimates as an input to *deterministic* optimization with the assumption that the estimates are completely correct, there are mathematically advanced yet practical ways to handle uncertainty in optimization. The degree of uncertainty can be assessed using the bootstrapping method described in Algorithm 8.5. Section 8.5 describes multiple methods to address optimization under uncertainty, including MVO, a simple version of RO, and SP – the latter is achieved by addressing uncertainty directly through probabilities using specific algorithms such as Simulation Optimization. These methods for optimization under uncertainty are practical to employ, especially with relatively small problems. Whether or not the analyst chooses to use them depends on how important he/she thinks it is to handle uncertainty due to statistical estimation. More sophisticated methods are introduced in Lo et al. (2017).

Last but not least, the methods introduced in this chapter may serve only as starting points to consider, and there can be other issues that arise in practice to make things a little different (and more interesting), such as additional constraints and sensitivity analysis of controllable parameters (e.g., budget).

Appendix 8.1: A Few Words on Convexity

The optimization problems in this chapter cannot guarantee global optima unless they belong to a class of "convex minimization" problems (or, equivalently, "concave maximization" problems). The convexity concept applies to problems with continuous decision variables only and thus does not apply to problems with binary decision variables. A convex minimization problem requires that the objective function is a convex function (e.g., U-shaped) and the constraints are convex inequalities. For linear constraints, they are always convex (and also concave). We now focus on the objective function.

The simplest case of optimization is from single-variable calculus, where we want to minimize a (smooth) nonlinear function of a single variable, $f(x)$, without any constraints. To solve this problem, we take the first derivative and set it to zero: $f'(x)=0$ and then determine the root, x^*. Then we determine if x^* is a local minimum by checking whether the second derivative evaluated at x^*, $f''(x^*)\geq 0$ (one may imagine a U-shaped curve). If the latter inequality is satisfied, we know this x^* is a local minimum. And if $f''(x)\geq 0$ *for all* x, which means $f(x)$ is a *convex function*, then x^* is also a *global* minimum of $f(x)$.

The conditions for multiple variables are similar. Again, if our goal is to find the values of a vector $(x_1,\ldots,x_m)'$ such that $f(x_1,\ldots,x_m)$ is minimized (assuming again it is a smooth function), from standard calculus or operations research textbooks (e.g., Ravindran et al. 1987), we equate the first partial derivative vector to zero, $\nabla f(\boldsymbol{x})=\left(\frac{\partial f}{\partial x_1},\ldots,\frac{\partial f}{\partial x_m}\right)'=0$, in order to determine the root, $\boldsymbol{x}^*=(x_1^*,\ldots,x_m^*)'$. To check whether $\boldsymbol{x}^*$ is a local minimum, we would check whether the second partial derivative (also known as the Hessian matrix) evaluated at $\boldsymbol{x}^*$ is positive definite (matrix version of "greater than zero"): $\nabla^2 f(\boldsymbol{x}^*)\geq 0$,where the i,jth element of $\nabla^2 f(\boldsymbol{x}^*)=\left.\frac{\partial^2 f}{\partial \boldsymbol{x}_i \boldsymbol{x}_j}\right|_{x=x^*}$. To guarantee it is also the global minimum, one would check whether $f(\boldsymbol{x})$ is a convex function, that is, $\nabla^2 f(\boldsymbol{x})\geq 0$ (positive semi-definite) *for all* values of $\boldsymbol{x}$, that is, $\boldsymbol{u}'\,\nabla^2 f(\boldsymbol{x})\,\boldsymbol{u}\geq 0$ for any values of $\boldsymbol{u}$ (one may imagine a bowl-shaped surface in a three-dimensional space with $m = 2$).

Let's get back to our context of uplift treatment optimization; suppose we would like to minimize the overall variance as in model (8.10). It can be proved that the objective function, which is a quadratic function of decision variables, is positive semi-definite, and hence it is guaranteed that model (8.10) has a global minimum.[17] However, for the probability objective functions in Example 8.3c, there is no guarantee that the local optima are global. The general practical advice is to start with various initial solutions to arrive at (potentially) different final solutions and then pick the one that has the best objective function value.

Appendix 8.2: A Brief Introduction to Simulation Optimization

Stochastic Programming (SP) has a wide range of applications in any field that requires optimization under uncertainty. One of the key existing algorithms available is Simulation Optimization, a computationally intensive method that combines Monte Carlo simulation and optimization, as the name suggests.

Simulation Optimization iterates between two steps:

1. **The Optimization Step:** It suggests a possibly improved solution given the results of previous iterations. It is typically based on a heuristic search algorithm as the optimizer, taking outputs from the Simulation step. In CB, the user has to indicate the number of simulations for this step under "Optimization Control."
2. **The Simulation Step:** It evaluates a given solution using Monte Carlo Simulation. The strength of simulation allows this algorithm to evaluate almost any situation with a given solution suggested by the optimization step (in Example 8.3c, bootstrapping is employed for simulation). In CB, the user has to specify the number of random trials for this step.

The above two steps feed each other until a certain termination criterion is satisfied. For further information, see Better et al. (2008) or EPM Information Development Team (2009).

Notes

1. Formally, $\Delta p_i = E\left(\Delta p_i\right) +$ uncertainty (noise) with mean zero.
2. Technically, the single treatment case includes a control group, so some literature refers to this simplest situation as a "two-treatment" case.
3. NP-completeness and its related terms are key concepts in Computer Science and Operations Research for measuring the complexity of algorithms. They are beyond the scope of this book; see Cook (2012), Cormen (2013), or Fortnow (2013) for an introduction or Bertsimas and Tsitsiklis (1997), Papadimitriou and Steiglitz (1998), or Dasgupta et al. (2006) for a formal discussion.
4. Specifically, in this approach, if the cluster-level solution from Algorithm 8.4 recommends only a *single* treatment for a given cluster, we can prioritize targets using Algorithm 8.1 or 8.2 at the cluster level. However, if the cluster-level solution involves more than one treatment for a given cluster, one may solve a binary integer program similar to model (8.4), except that the problem size is much smaller as we are solving for each cluster as opposed to the entire population.
5. Additionally, the Four Quadrant Method described in Section 9.2 of Chapter 9 can also be used for multiple treatments, as outlined in Section 8.2.1.
6. While standard deviations can be estimated using the usual parametric formula (see Eqn. (8.12a)), estimating covariances will require methods such as bootstrapping.
7. A common appropriate situation where sensitivity analysis makes sense is on *controllable* parameters. For example, one may be interested in assessing the impact of an additional budget or a lower budget on the optimal solution.

8. Lo and Pachamanova (2015) outlined another alternative for the MVO formulation using both the mean and variance in the objective function.
9. The multifunction in Excel is used here for the matrix computation of Eqn. (8.9); see Jackson and Staunton (2003) for an illustration.
10. The deterministic solution obtained from Example 8.1 is used as the starting solution for each preset minimum mean value.
11. We only discuss a subset of Stochastic Programming (SP) that can be practically applied to uplift optimization. For example, one branch of SP called scenario optimization requires finding a common solution to all possible simulated scenarios (see Wallace and Ziemba (2005), Zenios (2007), or King and Wallace (2012)) and is particularly suitable for multistage optimization problems. Scenario optimization is not discussed here due to its complexity for the uplift situation and the fact that most uplift problems are of a single stage.
12. See, for example, Jorion (2006) for details of VaR and other risk measures such as Conditional Value-at-Risk, CVaR.
13. Stochastic optimization in CB is based on a heuristic method called Simulation Optimization; see Appendix 8.2 for a brief introduction, or EPM Information Development Team (2009) or Better et al. (2008) for further details. Other Excel add-on software such as @Risk and Frontline Systems also have a Simulation Optimization capability. In all solutions in Example 8.3c, the following setting is used: No. of simulations = 1,000 and no. of trials = 1,000.
14. All solutions to stochastic optimization in Example 8.3c use the optimal solution from Table 8.2 in Example 8.1 as the initial solution.
15. The bootstrapped results from Example 8.2 are used in "Custom" Distribution (i.e., discrete uniform distribution with equal probability for each bootstrapped observation) in Crystal Ball for each uncertain (random) variable. Additionally, the correlations from Table 8.4 are taken as an input so the random number generation mechanism will simulate values in a multivariate fashion – by default in Crystal Ball, a common dependence structure called Gaussian copula is used for multivariate simulation; see, for example, Trivedi and Zimmer (2005) or Meissner (2014).
16. Crystal Ball selects the lognormal distribution as the best-fit distribution in Figure 8.7. Similarly, Crystal Ball chooses the best-fit distributions for other solutions in Figures 8.8 and 8.9.
17. Strictly speaking, only *population* variance-covariance matrices are guaranteed to be semi-positive definite, but the same is not necessarily true for all *sample* variance-covariance matrices.

References

Bertsimas, Dimitris, and John N. Tsitsiklis. 1997. *Introduction to Linear Programming*. Belmont, MA: Athena Scientific.

Bertsimas, Dimitris, David B. Brown, and Constantine Caramanis. 2011. "Theory and Applications of Robust Optimization". *SIAM Review*, 53(3): 464–501.

Better, Marco, Fred Glover, Gary Kochenberger, and Haibo Wang. 2008. "Simulation Optimization: Applications in Risk Management". *International Journal of Information Technology & Decision Making*, 7(4): 571–587.

Birge, John R., and Francois Louveaux. 1997. *Introduction to Stochastic Programming*. New York, NY: Springer.

Cook, William J. 2012. *In Pursuit of the Traveling Salesman: Mathematics of the Limits of Computation*. Princeton, NJ. Princeton University Press.

Cormen, Thomas H. 2013. *Algorithms Unlocked*. Cambridge, MA: MIT Press.

Cornuejols, Gerard, and Reha Tutuncu. 2007. *Optimization Methods in Finance*, Cambridge, England: Cambridge University Press.

Dantzig, George B. 1955. "Linear Programming under Uncertainty". *Management Science*, 1, 3 & 4: 197–206.

Dantzig, George B. 1957. "Discrete Variable Extremum Problems". *Operations Research*, 5: 266–277.

Dasgupta, Sanjoy, Christos H. Papadimitriou, and Umesh V. Vazirani. 2006. *Algorithms*. New York, NY: McGraw-Hill.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge, England: Cambridge University Press.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

EPM Information Development Team. 2009. Oracle Crystal Ball Decision Optimization, Fusion Edition: Release 11.1.1.3.00. http://docs.oracle.com/cd/E12825_01/epm.111/cb_oq_user.pdf

Fabozzi, Frank J., Petter N. Kolm, Dessislava A. Pachamanova, and Sergio M. Focardi. 2007. *Robust Portfolio Optimization and Management*. Hoboken, NJ: Willey.

Fortnow, Lance. 2013. *The Golden Ticket: P, NP, and the Search for the Impossible*. Princeton, NJ: Princeton University Press.

Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization & Machine Learning*. Boston, MA: Addison-Wesley.

Higle, Julia L., and Stein W. Wallace. 2003. "Sensitivity Analysis and Uncertainty in Linear Programming". *Interfaces*, 33(4): 53–60.

Jackson, Mary, and Mike Staunton. 2003. *Advanced Modeling in Finance Using Excel and VBA*. Hoboken, NJ: Wiley.

Jorion, Philippe. 2006. *Value At Risk: The New Benchmark for Managing Financial Risk*, 3rd edition. New York, NY: McGraw-Hill.

King, Alan J., and Stein W. Wallace. 2012. *Modeling with Stochastic Programming*. New York NY: Springer.

Lo, Victor S. Y. 2002. "The True-Lift Model – A Novel Data Mining Approach to Response Modeling in Database Marketing". *ACM SIGKDD Explorations*, 4(2): 78–86.

Lo, Victor S. Y., and Dessislava Pachamanova. 2015. "Prescriptive Uplift Analytics: A Practical Approach to Solving the Marketing Treatment Optimization Problem and Accounting for Estimation Error Risk". *Journal of Marketing Analytics*, 3(2): 79–95.

Lo, Victor S. Y., Dessislava Pachamanova, and Nalan Gulpinar. 2017. "Uncertainty Representation and Data-Driven Heuristics for Prescriptive Uplift Analytics in Segmented Marketing". Technical Paper.

Markowqitz, Harry. 1952. "Portfolio Selection". *Journal of Finance*, 7(1): 77–91.

Meissner, Gunter. 2014. *Correlation Risk Modeling and Management*. Hoboken, NJ: Wiley.

Michalewicz, Z., and D. B. Fogel. 2002. *How to Solve It: Modern Heuristics*. New York NY: Springer.

Papadimitriou, Christos H., and Kenneth Steiglitz. 1998. *Combinational Optimization: Algorithms and Complexity*. Garden City, NY: Dover.

Pisinger, David. 1995. "Algorithms for Knapsack Problems," *PhD Thesis*, Department of Computer Science, University of Copenhagen.

Ravindran, A., Don T. Phillips, and James J. Solberg. 1987. *Operations Research: Principles and Practice*, 2nd edition. Hoboken, NJ: Wiley.

Shapiro, Alexander, Darinka Dentcheva, and Andrzej Ruszczynski. 2014. *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia, PA: SIAM.

Storey, A., and M. Cohen. 2002. "Exploiting Response Models: Optimizing Cross-sell and Up-sell Opportunities in Banking," *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM's SIGKDD 325–331.

Taha, Hamdy A. 2010. *Operations Research*, 9th edition. London, England: Prentice Hall.

Trivedi, Pravin K., and David M. Zimmer. 2005. "sn Introduction for Practitioners". *Foundations and Trends in Econometrics*, 1(1): 1–111.

Tutuncu, R. H., and M. Koenig. 2004. "Robust Asset Allocation". *Annals of Operations Research*, 132: 157–187.

Wallace, Stein W. 2000. "Decision Making under Uncertainty: Is Sensitivity Analysis of Any Use?" *Operations Research*, 48(1): 20–25.

Wallace, Stein W., and William T. Ziemba (eds.). 2005. *Applications of Stochastic Programming*. Philadelphia, PA: MPS-SIAM Series on Optimization.

Williams, H. Paul. 2003. *Model Building in Mathematical Programming*, 4th edition. Hoboken, NJ: Wiley.

Zenios, Stavros A. 2007. *Practical Financial Optimization: Decision Making for Financial Engineers*. Oxford, England: Blackwell Publishing.

9

Uplift Analytics IV: Advanced Modeling Techniques for Randomized and Non-Randomized Experiments

9.1 Introduction

This chapter discusses advanced topics on Uplift/True-lift modeling. Chapter 6 introduced two fundamental modeling methods that are straightforward to apply. Section 9.2 of this chapter considers more advanced modeling techniques that may seem less straightforward but are still very practical to use.

Additionally, the causal inference discussion in Chapters 3–5 introduced methodologies for measuring causality in *observational* or *non-randomized/non-experimental* data. Chapters 6–8 are concerned with measuring or estimating causal effect (or targeting) at the individual level in *randomized experiments* and optimizing treatment assignment. By *randomized* experiments, we mean there is a random split between treatment and control groups so that any difference between them can be attributable to the treatment. Section 9.4 combines the methodologies of these previous chapters to address the intersection of the two issues: (1) Measuring causal effect or targeting at the individual level and (2) in *non-randomized experiments or observational* data.

In some business situations, randomized experiments are available and easy to execute, such as direct marketing (paper mail, email, or online). However, there are situations where randomized experiments are not feasible. As previously mentioned, uplift analytics (in randomized experiments) is still an emerging area, and uplift analytics in non-randomized experiments has received limited attention in the literature. Section 9.4 provides some methodological details to address this issue.

Section 9.5 discusses the presence of "direct response," which means we know directly whether individual customers actually respond or not (through click-through or coupon scans, for example) and how to integrate uplift modeling with direct response modeling through a state-of-the-art approach.

DOI: 10.1201/9780429172588-9

9.2 Advanced Uplift/True-lift Modeling Techniques

Chapter 6 describes two simple methods for uplift modeling. This section introduces a more advanced method, proposed by Kane et al. (2014). Consider Table 9.1, where all individuals in the sample data are classified into a 2×2 table. Response or not is an action taken by the individuals while the treatment/control split is determined by the campaign designer.

Define our estimation objective (i.e., lift) as a function of covariates x, giving:

$$Z(x) = P(R|T,x) - P(R|C,x), \text{ where } R = \text{event of response}$$

Here $Z(x)$ is the response probability difference (i.e., lift) between the treatment and control groups, given a set of characteristics x. As in Chapter 6, our objective is to find a set of individuals such that their sum of lift values is maximized. Using Bayes' rule from probability theory, Appendix 9.1 shows that the above lift function $Z(x)$ can be re-expressed as:

$$Z(x) = \frac{1}{2}\left[\frac{P(TR|x)}{P(T)} + \frac{P(CN|x)}{P(C)} - \frac{P(TN|x)}{P(T)} - \frac{P(CR|x)}{P(C)}\right]. \tag{9.1}$$

In Eqn. (9.1), the numerators $P(TR|x)$, $P(CN\ x)$, $P(TN|x)$, and $P(CR|x)$ are probabilities that an individual is a treatment responder, a control non-responder, a treatment non-responder, or a control responder, respectively. Hence, we call this the *Four Quadrant Method* or *KLZ* (named after Kane et al. 2014). These probabilities can be predicted by a model with four categorical outcomes using statistical and data mining techniques such as a multinomial logit model, CART/CHAID/C4.5 (decision tree algorithms), random forest, boosted tree (MART also known as TreeNet), or neural network. The key requirement is to be able to estimate probabilities associated with multiple categorical outcomes. The denominators in Eqn. (9.1), $P(T)$ and $P(C)$ are simply the treatment and control probabilities or proportions, respectively, that is, percentages of individuals assigned to treatment and control, which are known quantities as they are determined by the campaign designer.

Equation (9.1) is essentially a modified (or corrected) version of Lai (2006) who proposed the following scoring equation:

$$P(TR|x) + P(CN|x) - P(TN|x) - P(CR|x),$$

which can be simply expressed as:

$$P(TR \text{ or } CN|x) - P(TN \text{ or } CR|x) \tag{9.2}.$$

TABLE 9.1

Treatment by Response 2 × 2 Table

	Response	**No Response**
Treatment	**Treatment Responders (TR)**	**Treatment Non-responders (TN)**
Control	**Control Responders (CR)**	**Control Non-responders (CN)**

Note that Eqn. (9.2) contrasts the two sets of diagonal elements in Table 9.1. At first glance, this may look a little strange. The logic can be explained *intuitively* as follows, with reference to Table 9.1:

- In the treatment group, Event TR represents those who respond in the treatment group, so they are who we like.
- Event TN indicates those who do not respond in the treatment group, certainly not who we like.
- In the control group, Event CR indicates those who respond in the control group, so this group responds *anyway* regardless of the treatment and can be considered who we do not like.
- Event CN has those who do not respond in the control group – whether or not they would respond if they received a treatment is not known but at least there is a chance to persuade them to respond with a treatment, so they can be considered who we like.

Another way to explain this is to note that, as described in Section 6.3 of Chapter 6, Persuadables (positive lift) are hidden within the Treatment Responders and Control Non-responders (P(TR)+P(CN)), while Do-Not-Disturbs (negative lift) are hidden within the Treatment non-Responders and Control Responders (P(TN)+P(CR)). This method tries to maximize the probability of being a Persuadable and minimize the probability of being a Do-Not-Disturb within the highest-scoring observations. Nevertheless, Kane et al. (2014) and Appendix 9.1 show that Eqn. (9.2) is only mathematically correct if $P(T) = P(C) = 0.5$, that is, treatment and control groups are of the same size. In many practical cases, the treatment group is larger or much larger than the control group, as marketers tend to maximize the overall campaign value whenever possible, with their assumption (or hope) that the campaign will really deliver a positive lift. Kane et al. (2014) show that Eqns. (9.1) and (9.2) can result in highly correlated score values for a relatively wide range of $P(T)$. Similar methods have also been independently proposed by Tian et al. (2014) and Weisberg and Pontes (2015) for the case of $P(T) = P(C) = 0.5$. An example of the KLZ technique is set out in the Example 9.1 box.

Example 9.1 (Continuation of Example 6.2)

Recall that Example 6.2 from Chapter 6 describes a case from online retail data where the goal in the example is to develop an uplift model for web visits. To apply the KLZ discussed in this section to Example 6.2, we first fit a multinomial logit model for the four outcomes in Table 9.2 using Control Non-responders (CN) as the base case (reference category). Then the predicted lift is calculated by Eqn. (9.1) with the predicted probabilities for the four outcomes. The lift chart is displayed in Figure 9.1a, which shows that the KLZ method performs the best in the top semi-decile, with a general downward slope similar to other lines. Given the zig-zag pattern, it is sometimes easier to use the gains chart in Figure 9.1b, which shows the cumulative percentage of incremental responders captured. Figure 9.1b shows that the KLZ appears to be leading in the top 20%+ targets. Table 9.2 summarizes the numerical performance using the three performance metrics introduced in Chapter 6, again showing the clear leading performance of the KLZ in the top 15% but cannot beat the Treatment Dummy Approach with the overall Gini curve.

TABLE 9.2

Performance Evaluation of Alternative Uplift Models

	Gini	Gini 15%	Gini Repeatability (R^2)
Baseline	1.8556	–0.0240	0.2071
Two Model Approach	2.0074	0.0786	0.2941
Treatment Dummy (Lo (2002))	2.4392	0.0431	0.2945
Four Quadrant Method (KLZ (2014))	2.3703	0.2288	0.3290

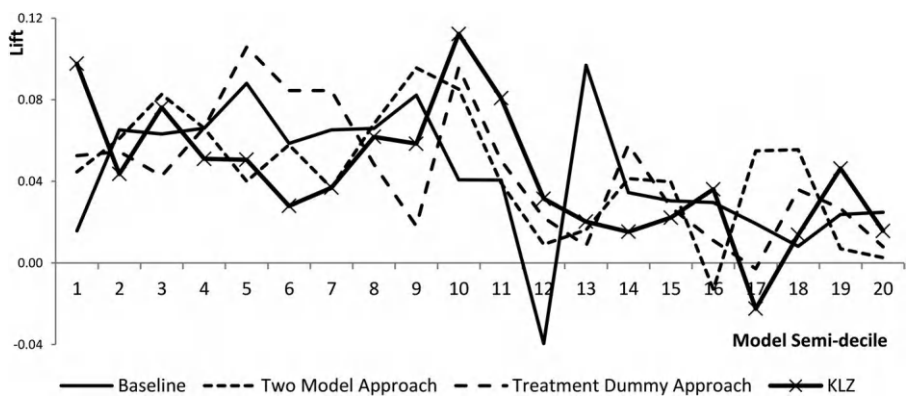


FIGURE 9.1a
Lift chart for various uplift models.

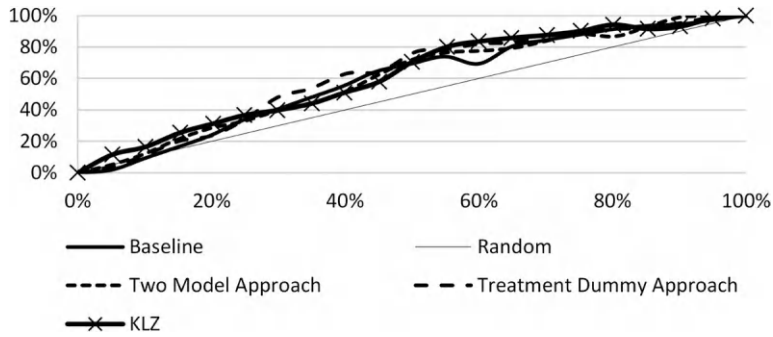


FIGURE 9.1b
Gains chart for various uplift models.

9.2.1 Extension of the Four Quadrant Methods to Multiple Treatment and Continuous Response Variable Cases

To extend from the binary treatment case (one treatment, one control) to the case of multiple treatments plus a control group is straightforward for the Two Model Approach (simply fit one model for each treatment group; see the Separate Model Approach in Lo and Pachamanova 2015) and the Treatment Dummy method (one dummy variable for each treatment along with their interaction effects). For the KLZ, it is also relatively straightforward. Consider comparing each treatment group to the control; one can arrive at a similar 2×2 table as in Table 9.1, except that the treatment in this table would represent only one of the treatments. Then Eqn. (9.1) follows for EACH treatment group compared to the control group. As a result, Eqn. (9.1) will result in a model score for each treatment group and for each individual. Such treatment-specific scores can be computed for individual-level treatment optimization as discussed in Chapter 7.

What if the response variable is continuous instead? The two methods outlined in Chapter 6 can be readily used by using an OLS-type regression (for continuous response variables) instead of a logistic regression (for binary response variables). However, the KLZ outlined in this section would not be applicable. A transformation method for continuous response variables R has been proposed by Weisberg and Pontes (2015) for the case of $P(T) = P(C) = 0.5$ for all individuals:

$$R_i^* = \begin{cases} 2\left(R_i - \bar{R}\right) \text{ if } i \text{ is in treatment,} \\ -2\left(R_i - \bar{R}\right) \text{ if } i \text{ is in control,} \end{cases} \tag{9.3}$$

where $\bar{R} = \dfrac{\dfrac{\sum_{i \text{ in treatment}} R_i}{n_T} + \dfrac{\sum_{i \text{ in control}} R_i}{n_C}}{2}$, and n_T and n_C are sample sizes of the treatment and control groups, respectively. The transformation in Eqn. (9.3)

works because the expected value of the transformed response variable is the average treatment effect (ATE), conditional on covariates:

$$E\left(R_i^*|x_i\right) = 0.5*2\left(E\left(R_i|T_i=1,x_i\right)-E\left(\bar{R}\right)\right)-0.5*2\left(E\left(R_i|T_i=0,x_i\right)-E\left(\bar{R}\right)\right)$$
$$= E\left(R_i|T_i=1,x_i\right)-E\left(R_i|T_i=0,x_i\right).$$

Another transformation for continuous response variables is proposed by Athey and Imbens (2015), for the general situation where $P(T) \neq P(C)$ (but still is a constant for all individuals):

$$R_i^* = R_i \frac{T_i - P(T)}{P(T)\left(1-P(T)\right)}. \tag{9.4}$$

Similar to Eqn. (9.3), the expected value of Eqn. (9.4) can be shown to be the same as the ATE, given covariates (also known as Conditional Average Treatment Effect, or CATE). With the transformation Eqn. (9.4), one can simply fit regular OLS-type regression models or any supervised learning models for continuous response variables to *directly* predict the lift as a function of covariates, similar to the KLZ for binary outcomes in Eqn. (9.1).

9.3 Situations Where Randomized Experiments Are Not Available

Several years ago, one of the authors was glad to accept a project for selecting targets for outbound telemarketing. This was considered a great opportunity as the impact or lift from outbound telemarketing, where a well-trained professional customer rep makes a phone call to a customer or potential customer, is usually much higher than that from direct mail, email, or an online message. However, when we learned that the customer reps previously "cherry-picked" customers so those who were contacted were different from those who were not contacted, the team hesitated to do that work as all the Uplift/True-lift modeling techniques (as introduced in Chapters 5 and 6) require a random split between treatment and control – in this situation, can the team still learn something from the existing data? In the business world, we do not say NO when there are opportunities to learn from "imperfect" data. Let us consider more examples below.

1. **Telemarketing:** To expand on the above example, outbound telemarketing or in-person customer visits, which theoretically can be done in a randomized experimental fashion, are often "cherry-picked" due to the much higher cost of contact. This means those with a

higher potential (greater monetary value or likelihood of responding) are often chosen as the targets, so there may not be a proper control group set up (although this is still desirable).

2. **Product Educational Workshop:** You would like to conduct an educational workshop or seminar on your product (say, a new kitchen product in a department store or a new car model in a car dealer), and in such a case you may send invitations to certain potential customers (random or targeted) and hope some of them will show up. However, those who see the workshop sign on the street or receive a referral invitation from their friends and family may also show up, and we cannot block them out by telling them, "You are in the control group."
3. **Retail Chain Design:** The unit of analysis is not necessarily an individual. For example, a retail chain is interested in measuring the effectiveness of a certain treatment (e.g., change of uniforms, store colors, and local marketing strategies) on sales. We would like to learn whether the treatment is effective overall (across the country) at the store group level and at the individual store level. Incorporating store-level variables in uplift modeling, we can understand what characteristics are associated with a higher sales level, and thus we can apply appropriate treatment to more stores. While some control over the similarity of treatment and control stores is possible, such analysis design may not guarantee a completely random split between treatment and control.
4. **Car Safety Program:** A car insurance firm or a government agency may be interested in testing the effectiveness of a car safety program on the behavior of young drivers. However, randomization is not feasible because of the "opt-in" option. Not only it is of interest to measure the overall effectiveness but also it is key to see whether the program is more effective for certain types of drivers, for example, by age, geography, education, and profession.
5. **Preclinical Analysis:** In biomedicine, a drug (treatment) needs to be proven to be successful in a randomized clinical trial. However, in a preclinical stage, especially in epidemiology (see Rothman et al. 2008) and genomics, observational data are commonly collected to understand the predictors of certain diseases, which may infer causes and appropriate treatments (e.g., a healthy diet, regular exercise, or cleaner air, which is mostly self-selected and cannot be easily randomly assigned).
6. **Talent Development:** For an application in human resources, observational data are available to understand treatments for talent development, such as the impact of an extensive training program on employee performance, where other available variables may include education, tenure, years of relevant experience, salary, and

geography. Due to the high cost of the program per employee, not everyone has received the training, so there is no randomized experiment. And also because of the high cost, it would be important to understand who would benefit most from such a program.

7. **Pre-experimental Marketing and Sales Programs:** In industries where randomized experimental data are not yet common but historical observational data are available, including various marketing and sales treatments for different customers at different points of time, it would be natural to learn as much as we could through causal inference and uplift modeling based on big data (which sometimes include time series data for each individual, leading to panel data) or small data before a randomized experiment is available to further confirm the results. For example, a pharmaceutical company is interested in physician targeting and they have not done any experimental tests yet. Nevertheless, they have observational data that list what marketing and sales efforts were used for each physician in their database as well as the response outcome. One can utilize such observational data without waiting to collect experimental data. In fact, it is quite common that one can use such data to narrow down the list of potentially effective treatments and then run them in an experimental treatment/control setting later for confirmation and refinement.

Measuring the overall effect (or segment-level effect) in these situations can possibly be achieved using the methodologies in Chapters 3 and 4. This chapter is about targeting them at the individual level, that is, uplift modeling, and we describe how to handle this in a non-randomized experiment or observational data setting below.

9.4 Uplift Modeling for Observational Data

We assume the terms "observational data" and "non-randomized experiments" are interchangeable and both indicate that it was not possible to have a random split of treatment and control. As a result, the treatment and control groups may not look alike. We also assume that we have some known factors that determine the difference between the treatment and control groups. For example, in the outbound telemarketing example discussed above, the professional customer reps "cherry-picked" the best targets for contacts. One can ask the reps what criteria they used – did they use recency (how recently the customers purchased a product), frequency (how many times they bought from or interacted with us), monetary inflows (how much money they spent), or some other knowledge to indicate their potential value

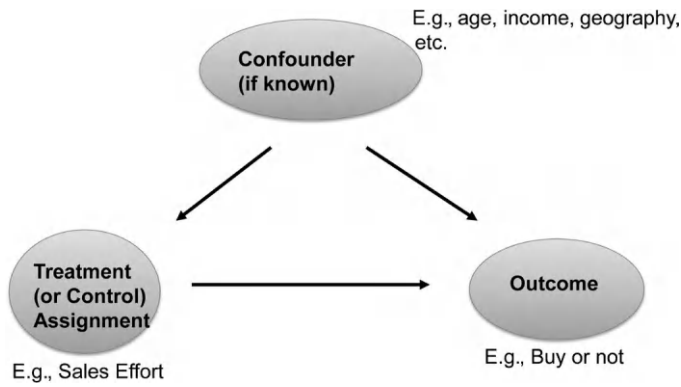


FIGURE 9.2
Conceptual causal diagram with confounders.

(e.g., income, where they live). While every rep may have used slightly different criteria, we should still be able to gather some ideas from them about which variables they generally used. Such knowledge can be very valuable for us in the techniques to be described below.

Consider Figure 9.1, where we would like to build an uplift model for improving sales efforts. However, previous sales data are observational, and past sales efforts may have depended on confounders such as age, income, and geography. For instance, past sales efforts may have focused more on older customers with higher incomes. As a result, the effect of sales effort can pass indirectly to the outcome variable through the confounders, as shown on the upper path in Figure 9.2 from treatment to confounder to outcome, a process known as a "back-door" in the causal inference literature (see Chapter 3 for further discussion).

First let us consider how we generally handle this in measuring the *population* (or overall) effect of treatment (causal inference as described in Chapter 3). There are two general ways to block (or control for) the confounders in order to isolate the treatment effect: (1) By blocking the link between the confounders and the outcome variable (through an outcome regression model), or (2) by blocking the link between the confounders and the treatment/control assignment (through propensity score matching), or a combination of the two by blocking both links, which is a doubly-robust estimation method. In the uplift context, we are interested in the individual-level impact, which requires us to focus on measuring not only the main effect of treatment but also the heterogeneous treatment effects (also known as effect modifiers in epidemiology literature) on outcome.

Figure 9.3 reveals the confounder blocking process in the uplift context. The interaction between treatment and confounder (or a set of confounders) allows us to model the individual-level treatment effect. In order to isolate the direct main effect of treatment on the outcome and the

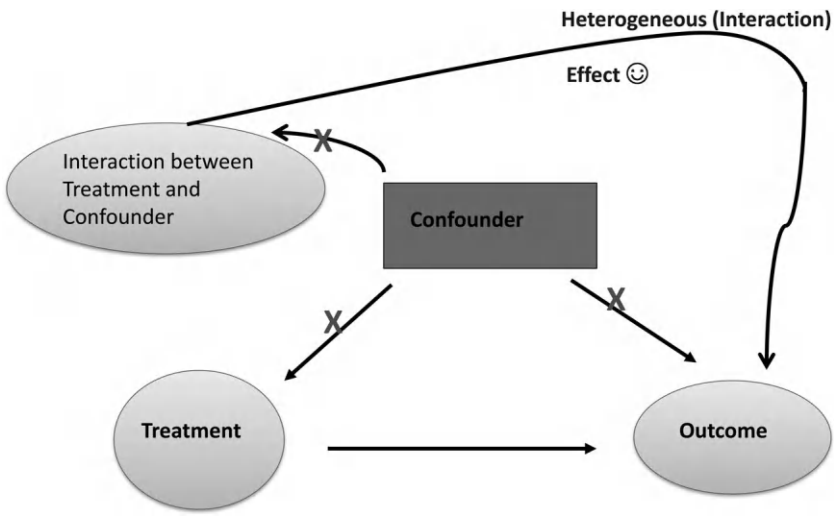


FIGURE 9.3
Conceptual causal diagram to block the confounder set.

heterogeneous (interaction) effect on the outcome, our goal is to block the following links:

1. Between the confounder and the interaction;
2. Between the confounder and the treatment variable; and
3. Between the confounder and the outcome.

Blocking the first two links can be achieved by propensity score (PS) matching, while the last link can be blocked by making the confounder present in an outcome regression model, essentially controlling for the confounder. This is similar to the doubly-robust estimation method or the Marginal Structural Model with *effect modification* (see Section 12.5 of Hernan and Robins 2016 for reference). We set out the detailed methodologies in the following subsections.

9.4.1 Modified Two Model Approach and Treatment Dummy Approach Using Propensity Score

Recall that Chapter 6 describes two approaches to uplift modeling: Two Model Approach (Method 1) and Treatment Dummy Approach (Method 2). For non-randomized experiments, treatment and control groups are not necessarily comparable, but if selection to treatment or control depends on a set of observable variables, a PS approach can be employed to make the treatment and control groups more comparable (see Chapter 3). In the uplift context

here, the Inverse Probability Weighting (IPW) method, a popular method under the PS approach, can be easily used for adjustment (e.g., Chapter 7 in Morgan and Winship 2014[1]).

Using the IPW method, the weights[2] are constructed as follows:

$$W_i = \begin{cases} \dfrac{P(T)}{P(T_i = 1|x_i)} & \text{if } i \text{ is in treatment,} \\ \dfrac{1-P(T)}{1-P(T_i = 1|x_i)} & \text{if } i \text{ is in control.} \end{cases} \tag{9.5}$$

The denominators of the weight in Eqn. (9.5) are simply the PS (that an individual belongs to the treatment or control group) given covariates, and can be estimated through logistic regression (or any binary classification model that can generate probability estimates). The numerators[3] denote the unconditional (overall) sample proportions of treatment and control. IPW creates a pseudo-population in which the arrow from confounders to the treatment variable is removed; see Figure 9.3. Equation (9.5) shows the ratio of the unconditional treatment (or control) proportion to the conditional treatment (or control) proportion. Intuitively, treated individuals who are highly likely to be in the treatment group, say, $P(T_i = 1|x_i) = 0.9$ while $P(T) = 0.5$, would be overrepresented in the treatment group, and thus we would want to have a lower weight. On the other hand, those treated individuals with a low likelihood to be in the treatment group, say, $P(T_i = 1|x_i) = 0.1$, would be underrepresented in the treatment group, and we would like their representation to be higher, leading to a higher weight.

The weight formula in Eqn. (9.5) is used to measure the ATE of the *whole* population. It is not uncommon that we may be interested in the treatment effect on the treated group or control group only. For example, in an observational study that has treatment and control groups that are not randomly split, we might like to infer from the treatment data what would happen to the control group (those not yet treated) if they receive a treatment. In this case, we would make the treatment group look like the control group in terms of their composition (individual characteristics). This is the case for measuring the Average Treatment effect on the Controls (ATC), with the following weight formula:

$$W_i(ATC) = \begin{cases} \dfrac{\left(1-P\left(T_i = 1 \mid x_i\right)\right)/\left(1-P(T)\right)}{P\left(T_i = 1|x_i\right)/P(T)} & \text{if } i \text{ is in treatment,} \\ 1 \text{ if } i \text{ is in control.} \end{cases} \tag{9.5a}$$

Similarly, if one is interested in what would happen to the treated group if they were NOT treated, one would make the control group look like the

treatment group in terms of its composition. This is the case for measuring the Average Treatment effect on the Treated (ATT), with another weight formula:

$$W_i(ATT) = \begin{cases} 1 \text{ if } i \text{ is in treatment,} \\ \dfrac{P(T_i = 1|x_i)/P(T)}{\left(1 - P(T_i = 1|x_i)\right)/\left(1 - P(T)\right)} \text{ if } i \text{ is in control.} \end{cases} \tag{9.5b}$$

Applied to the uplift context where the aim is to estimate treatment effect given certain individual characteristics,[4] the IPW method is similar to measuring "Effect Modification" (i.e., heterogeneous treatment effect) in the Epidemiology literature.[5] While IPW is straightforward to apply, it has a potential issue with too large or too small weights (due to the reciprocal of PS or one minus PS).[6] As a remedy to avoid over- or under-stated weights, two steps are commonly recommended in the literature:

1. **Overlap Analysis:** Check the max/min of the PS for treatment and control groups, respectively, and also plot their distributions (using boxplot or histogram/kernel density function) to examine the overlap of PSs for treatment and control groups. Discard those data points that do not fall in the overlap (or "region of common support") of the two groups; see, for example, Sturmer et al. (2014).
2. **Windsorization of Weight:** This is also known as Trimming in the PS literature. Essentially, extreme weight outliers are replaced by more reasonable values. For instance, we may cap ("windsorize") the weights at 95th and 5th percentiles. While this is quite commonly mentioned in the literature and has been shown to be beneficial to logistic regression-based PS models (Lee et al. 2011), there is no consistent recommendation on the exact cutoff percentiles; see Crump et al. (2009) and Morgan and Winship (2014). Sensitivity analysis using various cutoffs can be used.

The weighting scheme in Eqn. (9.5) can be easily applied to the Treatment Dummy Method and the Two Model Approach for balancing treatment and control compositions. What about the KLZ? It is not that much harder, but it requires some explanation, as in the following subsection.

9.4.2 Modified Four Quadrant Method (Modified KLZ)

Recall from Eqn. (9.1) that the lift function for each individual can be expressed as a linear combination of four probabilities:

$$Z(x) = \frac{1}{2}\left[\frac{P(TR|x)}{P(T)} + \frac{P(CN|x)}{P(C)} - \frac{P(TN|x)}{P(T)} - \frac{P(CR|x)}{P(C)}\right]. \tag{9.1}$$

Equation (9.1) is valid only for randomized experiments where $P(T)$, or $P(C) = 1 - P(T)$, is a known constant driven by campaign design. For example, by design, we may have a randomized experiment with 80% treatment and 20% control. This implies that, given a target group, the probability of assigning treatment is 0.8 for each and every individual. Such a randomization process has a known and identical probability of treatment for everyone. In non-randomized experiments (or observational data), however, $P(T)$ or $P(C)$ is no longer a constant as some individuals have received treatment with a higher or lower probability depending on certain characteristics (e.g., through self-selection bias). Our methodology introduced here is simply to replace $P(T)$ or $P(C)$ with $P(T|x)$ or $P(C|x)$, which means the treatment assignment probability is now a function of some predictors (or covariates):

$$Z(x) = \frac{1}{2}\left[\frac{P(TR|x)}{P(T|x)} + \frac{P(CN|x)}{P(C|x)} - \frac{P(TN|x)}{P(T|x)} - \frac{P(CR|x)}{P(C|x)}\right] \quad (9.6).$$

Equation (9.6) is a simple extension of Eqn. (9.1). Hence, Eqn. (9.6) can be recognized as the Modified KLZ or the Modified KLZ (modified for non-experimental data). Here, $P(T \mid x)$ is simply the PS that can be estimated using a logistic regression or any type of method that can predict a binary outcome.[7] However, in our case, having a separate model for $P(T \mid x)$ is optional, because we can simply sum up the appropriate components in the numerators, as follows:

$$P(T|x) = P(TR|x) + P(TN|x) \quad (9.7a)$$

and

$$P(C|x) = 1 - P(T|x). \quad (9.7b)$$

Then one can easily compute the predicted lift values using Eqn. (9.6) along with Eqns. (9.7a) and (9.7b).

The above description only applies to binary response variables. What about continuous response variables for observational data? A transformation for continuous response variables for observational data is proposed by Athey and Imbens (2015):

$$R_i^* = R_i \frac{T_i - P(T|x_i)}{P(T|x_i)\left(1 - P(T|x_i)\right)}, \quad (9.8)$$

where $P(T|x_i)$ is simply the PS that can be estimated. Note that Eqn. (9.8) is a natural extension of Eqn. (9.4) for observational data, where predictors (or covariates) are included.[8]

9.4.3 Example for Uplift Modeling for Observational Data

Example 9.2 (Continuation of Example 6.1)

Recall from Example 6.1 in Section 6.7.1 that the response variable, Donated, is driven by a set of variables. Assume again that we have the same set of covariates as in Example 6.1. Further, we assume the treatment assignment is a logistic function of spent (average amount donated in the past) and frequency (number of times a donation was made in the past) in the simulation.

Following the methodologies outlined earlier in Section 9.4:

1. We first fit a logistic regression to recover the relationship between treatment assignment and confounders (spent and frequency in this case), using the whole data set.
2. The training data is scored to determine the PS for each observation.
3. We next apply Eqn. (9.5) to construct a set of weights using the IPW method based on the calculated PS from the previous step.
4. We then follow the overlap analysis and windsorization steps outlined in Section 9.4.1 to process the weights.
5. Next, this set of weights is used in conjunction with various uplift modeling methods to fit the training data.
6. Finally, the holdout sample is used to evaluate the various uplift modeling methods mentioned in Sections 9.4.1 and 9.4.2 (Two Model Approach, Treatment Dummy Method, and Modified KLZ), where the holdout sample, like the training sample, also has to be weighted using the IPW method to compute unbiased lift estimates.

Following Eqn. (9.5), the following SAS statement is applied to construct the weights:

$$\text{if trt} = \mathbf{1} \text{ then wei} = \mathbf{0.8033}/\text{ps; else wei} = \mathbf{0.1967}/(\mathbf{1} - \text{ps});$$

We summarize the distribution of PS for the treatment and control groups in Table 9.3, Figures 9.4a and 9.4b. Since the two distributions almost completely overlap (with similar minimums and maximums), no data points are discarded. As mentioned in Section 9.4.1, high variability of PS has been a

TABLE 9.3

Summary Statistics of Propensity Score

Analysis Variable: ps					
trt	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
0	59008	0.707418	0.146591	0.268396	0.99723
1	240992	0.826781	0.124055	0.268396	0.999999

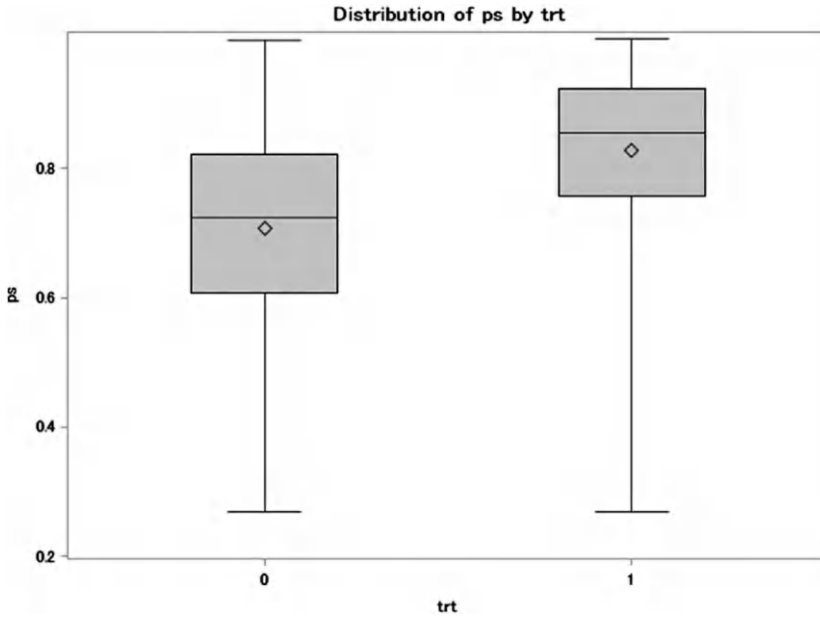


FIGURE 9.4a
Box plot of propensity score for treatment (trt = 1) and control (trt = 0).

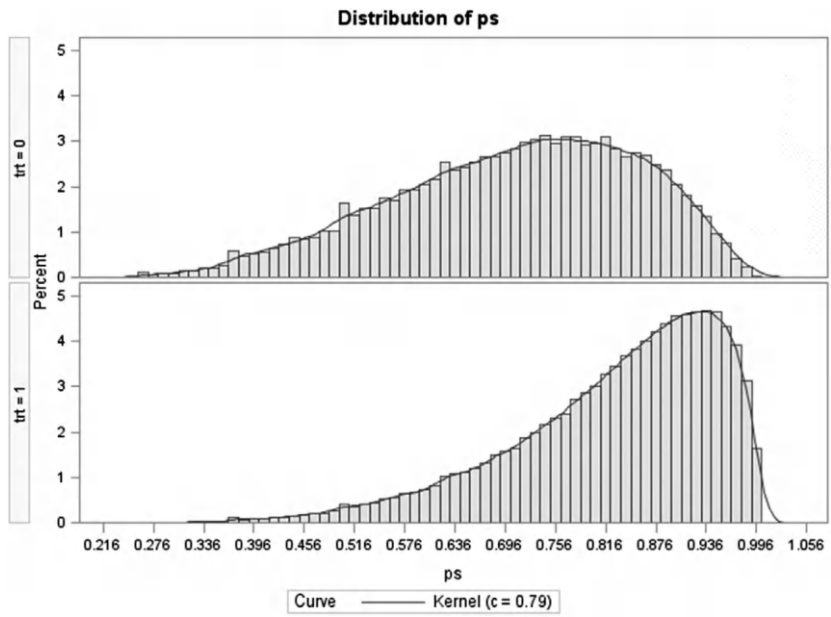


FIGURE 9.4b
Frequency distribution of propensity score for treatment (trt = 1) and control (trt = 0).

discussion point in the literature, so a windsorization step is performed to cap ("windsorize") the data at 95th and 5th percentiles, which are 1.51 and 0.5 in this example, essentially limiting the weights to be no more than 50% higher or lower for each observation.

With the weights constructed, we can now apply various uplift modeling methods using the weights.[9] The results are summarized in Figures 9.5a,b and Table 9.4, showing that, in this illustrative example, the three uplift models

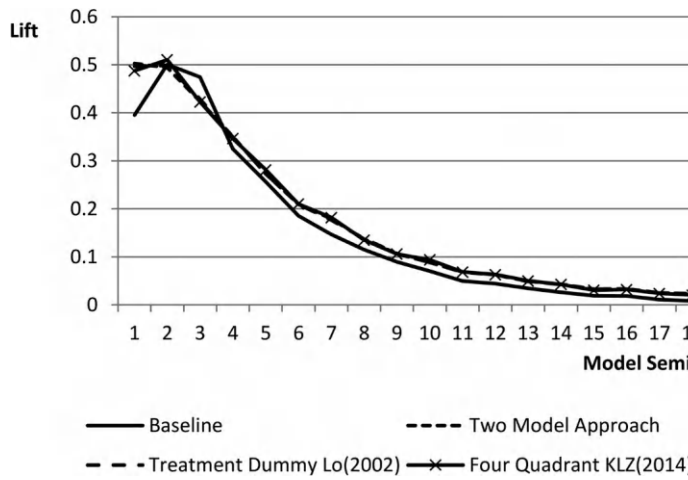


FIGURE 9.5a
Lift chart for uplift modeling for observational data example.

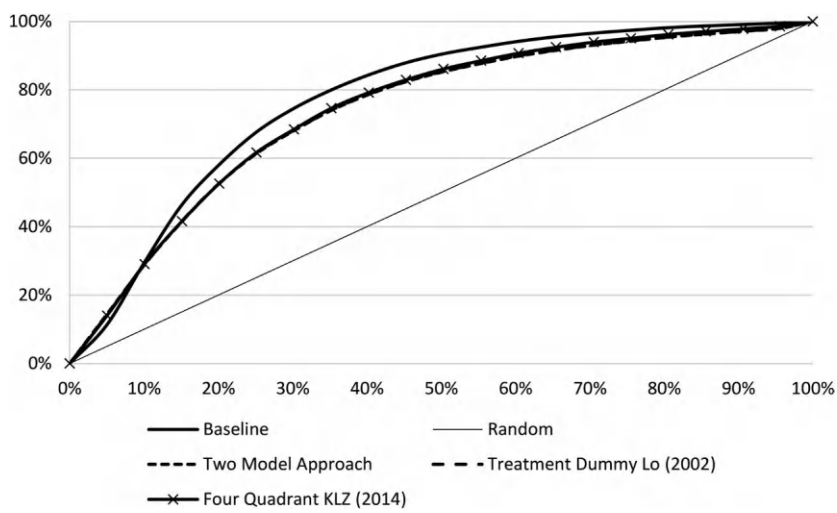


FIGURE 9.5b
Gains chart for uplift modeling for observational data example.

TABLE 9.4

Summary of Model Performance

	Gini	Gini 15%	Gini 5%	Gini Repeatability (R^2) (%)
Baseline	5.4611	0.5714	0.0635	78.21
Two Model approach	4.7428	0.5464	0.0943	78.63
Treatment Dummy (Lo (2002))	4.7556	0.5459	0.0924	78.66
Modified Four-Quadrant Method (KLZ (2014))	4.8431	0.5445	0.0897	80.49

are very close to each other in performance and are all better than the baseline (treatment only) model for the first 5% but are not better when the whole sample is considered. It should be mentioned that in Table 9.4, because we are dealing with non-randomized treatment and control data, the Gini coefficient and the Top 15% Gini coefficient need to incorporate the set of weights derived from the PSs; see Appendix 9.2 for the computational formulas.

9.5 Direct Response Modeling and Integration of Direct Response and Uplift Modeling

We have been discussing uplift modeling since Chapter 6. This set of techniques for uplift handles the situation where we do NOT know *exactly* who responds to the treatment, that is, we can only infer by collecting lots of data and using the difference between the treatment and control response rates. In a nutshell, this requires us to analyze treatment and control data at the aggregate level, group level, or sub-sub-subgroup level (through predictive modeling), so granular that it is *almost* at the individual level. However, what if we actually know who EXACTLY responded to the treatment? That would be a much simpler situation, and would only need us to develop a regular supervised learning mod to predict direct response (or sales) as a function of covariates without worrying about the control group. Is it really that simple? Let us consider two cases where direct response data are available.

1. **Retailer Couponing:** Imagine a common situation for retailers (or restaurants) that often use coupons or some form of promotional codes to attract customers. A typical way is to include a coupon on a postcard. If the customer decides to use it for a purchase, the postcard will be scanned and the "direct response" will actually be captured. Hence, the direct response data can tell EXACTLY who has responded to the coupon. However, those without a coupon (in

the control group) may also purchase something from the retailer. Additionally, some in the treatment group (those who received the postcard with the coupon) may purchase WITHOUT using the coupon – they may have forgotten to use it despite receiving the treatment, or they may not have seen the postcard at all.[10] So it does sound like there is still an uplift modeling opportunity. Should we simply build a direct response model or an uplift model? Why not both?

2. **Email Click-through:** Another example is an email marketing campaign to sell a retail product online. Those who receive the email with a specific URL link (or a QR code) are in the treatment group. Customers in the treatment group may click the link ("click-through") to purchase a product. But some customers who received the email may also purchase the product from the general website instead of clicking the link in the email. Similarly, those who did not receive the email (in the control group) can purchase the product from the general website. This is similar to the retailer example with coupons. That is, one can develop a direct response model for click-through as a function of covariates, and one may also develop an uplift model using the treatment and control data. Which one should we do? Why not both?

9.5.1 Uplift on Response Probability

Consider only the blue boxes in Figure 9.6. We first split the customer data into treatment and control groups. Within treatment (T), some directly responded (D) and some did not (D^c). For those who directly responded (D), they already responded (R). For those who did not directly respond (D^c), for example, those who did not use the coupon, some may have made a purchase anyway (R) and others may not (N). Similarly, in the control group (C), some may have made a purchase (a response, R) and others may not (N).

We are now expressing the tree diagram in Figure 9.5 mathematically. Using our notation for the lift function of covariates:

$$Z(x) = P\left(R|T,x\right) - P\left(R|C,x\right), \text{ where } R = \text{event of response.}$$

$$= P\left(R|T,D,x\right)P\left(D|T,x\right) + P\left(R|T,D^c,x\right)\left(1 - P\left(D|T,x\right)\right) - P\left(R|C,x\right),$$

following the direct and no direct branches under Treatment (T) in Figure 9.6, where D = direct response and D^c = no direct response, and applying the standard rules of conditional probability

$$= P\left(D|T,x\right) + P(R|T,D^c,x)\left(1 - P\left(D|T,x\right)\right) - P\left(R|C,x\right), \tag{9.9}$$

since $P\left(R|T,D,x\right) = 1$, that is, response (purchase) probability is 100% if the coupon is actually used.[11]

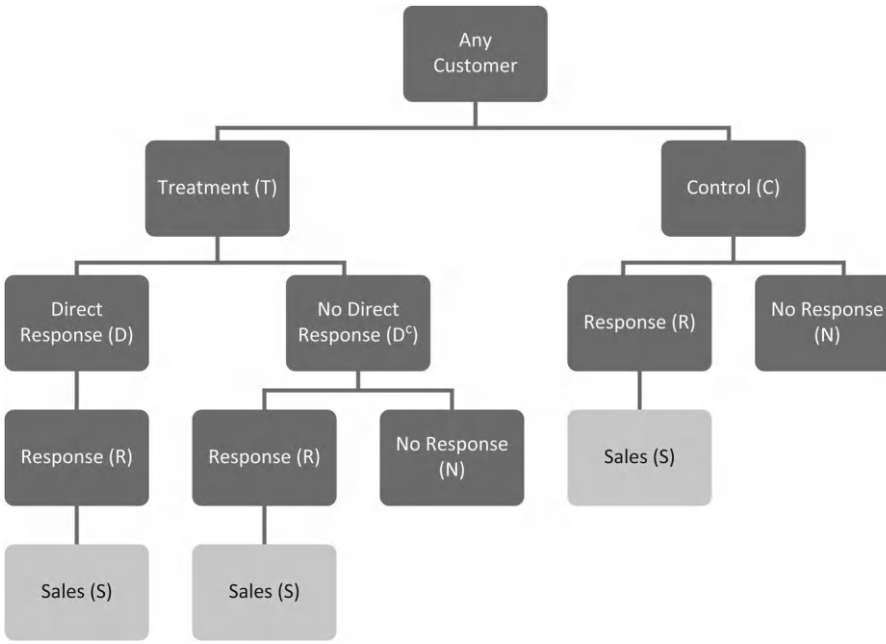


FIGURE 9.6
Tree diagram of direct response and response data.

Equation (9.8) has the following probabilities to be modeled:

1. $P(D|T,x)$ = Direct response probability as a function of covariates for the treatment group – this can be developed using regular supervised learning for binary outcomes (note that direct response is only available in the treatment group in this model),
2. $P(R|T,D^c,x)$ = Treatment response probability among those who did not use the coupon, as a function of covariates (i.e., the directly responded individuals would need to be *excluded* from modeling for this probability), and
3. $P(R|C,x)$ = Control response probability as another function of covariates.

Note that the probabilities in (2) and (3) above can be handled by any uplift modeling technique that provides *separate* estimates from the treatment and control response models, respectively, that is, the Two Model Approach or the Treatment Dummy Approach but not the Four Quadrant Model[12] (as the decomposition in Eqn. (9.1) does not capture direct response).

Equation (9.9) can also be rearranged as:

$$= P(D|T,x)\left[1-P(R|T,D^c,x)\right]+\left[P(R|T,D^c,x)-P(R|C,x)\right]. \qquad (9.10)$$

In Eqn. (9.10), the component in the first pair of parentheses is the difference in response probability between a direct responder (which is always 1.0) and someone in the treatment group who did not use a coupon. The first component, $P(D|T,x)\left[1-P(R|T,D^c,x)\right]$, is always ≥ 0. The second component, $\left[P(R|T,D^c,x)-P(R|C,x)\right]$, measures the difference in response probability between the treatment group (without using the coupon to directly respond) and the control group. The second component may not necessarily be ≥ 0. Note that Eqns. (9.9) and (9.10) are derived from the fundamental theory of conditional probability and do not require any statistical model assumptions. Modeling the various probability components in those equations as functions of covariates, however, does require empirical data and statistical/data mining models.

Example 9.3: Integration of Direct Response Model and Uplift Model

In this simulated example, a clothing retailer is interested in maximizing the response (buying) rate of their products through mailing coupons to the right customers. The retailer stores historical data of each customer as follows:

- **Recency:** number of months ago for its most recent purchase (0, 1, 2, …, 12)
- **Frequency:** number of times purchases were made in the past year (0, 1, …)
- **Spent:** average spent amount on past purchases
- **Demographics:** age, income

In our simulations, the direct response using coupons is a function of recency, while the response rate (if coupon is not used) is a function of age and frequency. Using Eqn. (9.8), we need to estimate:

- **Direct Response Model:** Model 1: $P(D|T,x)$, Direct response probability as a function of covariates (where recency is the only covariate used in the actual theoretical model), developed using regular logistic regression;
- **Uplift Model:** Model 2: $P(R|T,D^c,x)$ and Model 3: $P(R|C,x)$, response probabilities in the treatment group (when direct response is not used) and control groups, respectively, both as a function of age, amount spent, and frequency in the actual models; estimated using a Treatment Dummy variable approach (from Chapter 6 or Lo 2002).

The estimated results[13] are reported in Table 9.5 (only significant coefficients at 5% level are kept, and Models 2 and 3 are estimated together using

TABLE 9.5

Estimated Direct Response and Uplift Models

	Baseline	Model 1 $P(D\|T,x)$	Model 2 $P(R\|T,D^c,x)$	Model 3 $P(R\|C,x)$	Uplift $P(R\|T,x)$	Uplift $P(K\|C,x)$
	Treatment Only Baseline	Direct Response	Treatment Response	Control Response	Treatment Response	Control Response
Intercept	–1.9697	–1.6077	–15.9304	–15.9304	–1.9707	–15.7035
age	0.0468		0.1494	0.0897	0.0468	0.0882
income						
frequency	0.0322		0.1003	0.1003	0.0326	0.0326
spent	0.000352	0.000994	0.000994	0.000349	0.000999	
wealth	0.000823				0.000823	
recency	–0.239	–0.3946		–0.239		
log(age)						
log(income)	–0.9036			–0.9036		
log(spent)						
log(wealth)						

the Treatment Dummy variable approach). Note that in this example, the significant drivers in the direct response model (Model 1) and the uplift model (Models 2 and 3) are quite different, which is a sign that the integrated model (uplift + direct response) should improve over either one alone. It is expected that the more different the drivers are in the two models, the more powerful the integrated model would become compared to having only either the direct response model or the uplift model.

For comparison, we fit a baseline model using treatment data only (whether they had a direct response or not), that is, predicting $P\left(R|T,x\right)$. Additionally, we fit a "standard" uplift model using the Treatment Dummy Approach (again from Chapter 6 or Lo 2002), without using direct response data, to predict treatment and control response rates, respectively, $P\left(R|T,x\right)$ and $P\left(R|C,x\right)$. And we also evaluate the performance of a direct response-only model (i.e., Model 1 only, $P\left(D|T,x\right)$). The lift chart in Figure 9.7 shows that the direct response model does not differentiate well in this example (as it is driven by only one variable by simulation design). All the other models are better, and the integrated model (direct response + uplift) clearly outperforms all, at least in the top semi-decile. The Gini metrics in Table 9.6 confirm the performance comparison.[14]

9.5.2 Uplift on Sales Revenue

Let's revisit Figure 9.6, this time including the green boxes (Sales, S). Whenever a purchase (response) is made, we can collect the sales revenue, S.

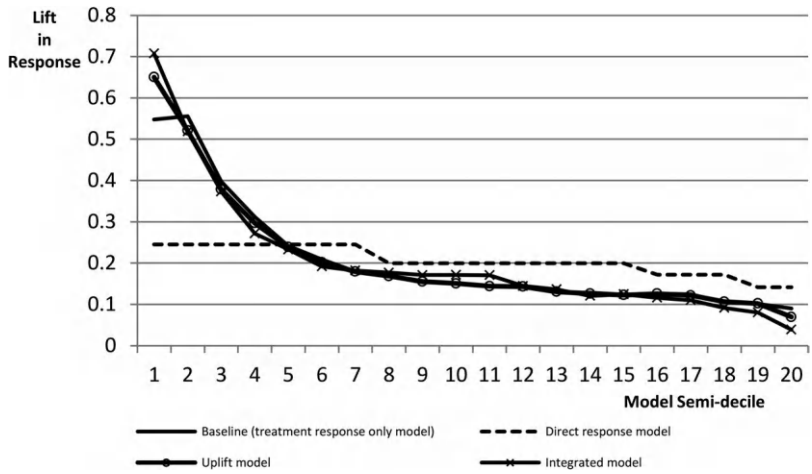


FIGURE 9.7
Lift charts for direct response only, uplift, and integrated direct response and uplift methods.

TABLE 9.6

Performance Metrics for Comparison between Direct Response Model, Uplift Model, and Integrated Model for Response Lift

	Gini	Gini 15%	Gini Repeatability (R^2) (%)
Baseline (Treatment Only)	3.1199	0.4612	68.0
Direct Response Model	0.8909	0.0593	88.2
Uplift Model	3.2694	0.5151	65.2
Integrated Model	3.5165	0.5553	65.7

In addition to modeling the purchase response probabilities (*R*, a binary Y/N variable), we are now also interested in modeling sales revenue (*S*, a continuous variable in $value). Equations (9.9) and (9.10) can be generalized to model sales revenue, *S*:

$$Z(x) = E(S|T,x) - E(S|C,x), \text{ where S = sales revenue.}$$

$$= E(S|T,D,x)P(D|T,x) + E(S|T,D^c,x)(1 - P(D|T,x)) - E(S|C,x) \quad (9.11a)$$

using the standard conditional expectation formula

$$= E(S|T,D,x)P(D|T,x) + E(S|T,D^c,R,x)P(R|T,D^c,x)(1 - P(D|T,x)) - E(S|C,R,x)P(R|C,x) \quad (9.11b)$$

where D = direct response, and D^c = no direct response, and R = response (in the control group); or

$$= P(D|T,x)\left[E(S|T,D,x) - E(S|T,D^c,x)\right] + \left[E(S|T,D^c,x) - E(S|C,x)\right]. \quad (9.12a)$$

$$= P(D|T,x)\left[E(S|T,D,x) - E(S|T,D^c,R,x)P(R|T,D^c,x)\right]$$
$$+\left[E(S|T,D^c,R,x)P(R|T,D^c,x) - E(S|C,R,x)P(R|C,x)\right]. \quad (9.12b)$$

Equation (9.12a) is the sum of two components:

1. The difference in expected sales due to direct response (over non-direct response), discounted by the direct response probability, and
2. The difference in expected sales between treatment (among the non-direct responders) and control.

In Eqns. (9.11b) and (9.12b), one will need to estimate the following models as functions of covariates:

1. $P(D|T,x)$ = Direct response probability in the treatment group,
2. $E(S|T,D,x)$ = Expected sales of the direct responders in the treatment group,
3. $P(R|T,D^c,x)$ = Response probability of those who did not respond directly in the treatment group,
4. $E(S|T,D^c,R,x)$ = Expected sales of the indirect responders (those who did not respond directly) in the treatment group,
5. $P(R|C,x)$ = Response probability of the control group (i.e., probability of natural response), and
6. $E(S|C,R,x)$ = Expected sales of the control responders.

Even though the above process involves six models, it is a relatively straightforward extension of those models from Eqns. (9.9) and (9.10). It is also quite possible that the two expected sales models in (4) and (6) are similar or identical, that is, for those who did not respond directly, it is possible that their expected sales are the same if they decide to purchase. We can, of course, let the data speak, that is, testing whether some of the sales models are the same.

Example 9.3 (continued)

Continuing with our last clothing retailer example in Section 9.5.1, let's say we are now interested in maximizing sales revenue. In our simulation, we assume log(sales) is a function of log(income) and log(spent), where spent = average spent amount of past purchases.

For model estimation, in addition to the probability models, Models 1 for direct response and Models 2 and 3 for uplift, as outlined in the example of the previous section, we have the following log(sales) models:

- **Model 4:** Expected log(sales) of the direct responders, $E(S|T,D,x)$, where S = log(sales), and log(age), log(income), and log(spent) are its predictors in the actual theoretical model for simulation,
- **Model 5:** Expected log(sales) of the indirect responders (those who did not respond directly) in the treatment group, $E(S|T,D^c,R,x)$, which is a function of log(income) and log(spent) in the actual model, and
- **Model 6:** Expected log(sales) of the control responders, $E(S|C,R,x)$, also a function of log(income) and log(spent) in the actual model.

The estimated models[15] are summarized in Table 9.7. Note that Model 4 has a higher intercept than that of Models 5 and 6 because the direct responders purchased more in this data. As in the previous section, the baseline model, which predicts $E(S|T,R,x)$ using the treatment responders only, and the "pure" uplift model, which requires predictions of $E(S|T,R,x)$ and $E(S|C,R,x)$ are included for comparison.

The results are summarized in Figure 9.8 and Table 9.7. Lift in sales is defined as Average Sales in Treatment minus Average Sales in Control. As Eqn. (8.11b) or (8.12b) indicates, the model system includes a combination of response probability models and sales models (conditional on response)

TABLE 9.7

Estimated Log(Sales) Models

	Baseline $E(S\|T, R,x)$	**Model 4** $E(S\|T, D,x)$	**Model 5** $E(S\|T, D^c,R,x)$	**Model 6** $E(S\|C,R,x)$	**Uplift** $E(S\|T,R,x)$	**Uplift** $E(S\|C,R,x)$
	Expected lnsale of Treatment	**Expected lnsale of Direct**	**Expected lnsale of the Indirect**	**Expected lnsale of the Control**	**Expected lnsale of Treatment**	**Expected lnsale of Control**
Intercept	0.312	4.76349	2.89786	2.89786	4.10314	–1.2974
age	0.02507		–0.00189			
income					–0.00368	0.00161
frequency	0.01775				–0.01727	
spent	0.00024534				–0.00023731	
wealth	0.00032151				0.00040808	
recency	–0.19996	–0.01327	–0.00955		–0.12989	
log(age)					1.04325	1.04325
log(income)	–0.37153					
log(spent)		0.11467	0.12614	0.12614	0.25961	0.25961
log(wealth)		0.02859			–0.21083	–0.21083

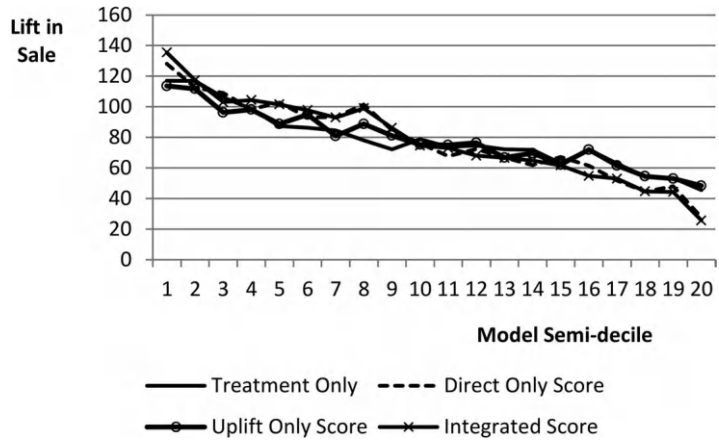


FIGURE 9.8
Lift chart for various models.

TABLE 9.8

Performance Metrics of Various Log(Sales) Models

	Gini	Gini 15%	Gini Repeatability (R^2) (%)
Baseline (Treatment Only)	1.3274	0.1395	89.9
Direct Response Model	1.8164	0.1583	94.8
Uplift Model	1.2829	0.1215	93.5
Integrated Model	1.9491	0.1741	96.0

to estimate the overall sales. Similar to other uplift modeling examples, the baseline model is the estimated probability from the treatment-only response model multiplied by the exponential of the log(sales) model using all treatment responders.[16] Similar to the example in the previous section that was focused on response rate, the integrated model appears to be better than others (baseline, direct response, and "pure" uplift model), as also confirmed by the Gini metrics in Table 9.8.

9.6 Concluding Remarks and Opportunities for Improvement

This chapter introduced a more recently developed uplift modeling technique, the KLZ, in Section 9.2, which has been empirically shown to be very useful, at least in limited data sets. Sections 9.3 and 9.4 discuss a more general situation when randomized experiments are not available, and

methodologies for uplift modeling on observational data, utilizing a combination of PS matching[17] and uplift modeling, can be applied. Section 9.5 introduces an integrated modeling method in the situation where direct response data are available, along with randomized treatment and control response data. The methodologies described in this chapter should be considered mostly state-of-the-art.

Even though this chapter concludes the uplift modeling methodologies, there are many opportunities to further improve the techniques introduced in this chapter and Chapter 6:

1. **Advanced Statistical and Machine Learning Methods for Supervised Learning:** Although uplift modeling is a little different from regular supervised learning as it aims at modeling the difference between treatment and control response rates, it can still benefit from advanced techniques developed for *regular* supervised learning. For example, lasso and elastic net can possibly improve variable selection in the KLZ (Section 9.2), Modified KLZ (Section 9.4.2), or the Two Model Approach and the Treatment Dummy Method (Sections 6.5), whether we are working with experimental data (Chapter 6 and Section 8.2) or observational data (Sections 9.3 and 9.4). Likewise, these methods (Two Model Approach, Treatment Dummy, or KLZ) could possibly be improved with advanced techniques that can better capture nonlinear relationships and interaction effects such as random forests (e.g., Guelman et al. 2014), boosted tree (MART), and neural network (see, e.g., Hastie et al. 2013). Similarly, for uplift modeling on observational data, the PS model (Section 9.4) has a binary dependent variable (treatment or control) that is commonly handled by logistic regression but could be improved with more advanced modeling techniques for detecting and capturing nonlinearities and interaction effects (Westreich et al. 2010).
2. **Model Ensemble:** Similar to applying more advanced modeling techniques, the performance of all the methods described in this chapter could be empirically improved with model ensemble methods by averaging out available models. A commonly used method for improving any supervised learning is Bootstrap Aggregate or Bagging, using bootstrapping and averaging to arrive at a lower expected mean squared error (through lower variability with mostly unchanged bias; see, for example, Hastie et al. 2013). Other model ensemble methods for uplift modeling have also been considered in the literature, for example, by Grimmer et al. (2016).
3. There are software tools available for uplifting modeling which include a variety of techniques for comparison analysis, as summarized in Table 9.9. Some software are applicable to experimental (RCT) data only, and others can be employed for both experimental

TABLE 9.9

Summary of Available Software for Uplift Modeling

Open Source or Commercial	Uplift on RCT Data	Uplift on RCT or Observational Data
Open Source: R	• Uplift (Gruelman 2015) – tree, RF • Tools4uplift (2019) – two-model, interaction • quint (2020) – tree, based on effect size • mr_uplift (2020) – multiple treatments, neural net	• grf (Athey et al. 2019; Tibshirani et al. 2023) – transformed outcome, tree, RF, PSM-IPW • rlearner (2020) – meta learner, PSM-IPW
Open Source: Python	• Pylift (Wayfair 2019) – transformed outcome • scikit uplift (2022) – single model, two-model, transformed outcome	• CausalLift (2019) – two-model, PSM-IPW • CausalML (Uber 2025) – tree, RF, meta learners, PSM-IPW • EconML (Microsoft 2025) – RF, meta learners, PSM-IPW
Commercial Software	• JMP Pro: Uplift Model – tree • SAS Enterprise Miner: Incremental Response Model – two-model, interaction	

or observational data, where propensity score matching (PSM)-inversely proportional weighting (IPW) is often applied to the latter. In particular, grf (generalized random forest) in R provides a random forest equivalent for uplift modeling and may be favored by academic researchers for publications given the strong theoretical support (see Athey et al. 2019; Tibshirani et al. 2023). In commercial applications, CausalML developed by Uber and EconML developed by Microsoft Research (both in Python) appear to be attractive choices by business practitioners, especially given its coverage of techniques including tree-based methods such as random forest version for uplift as well as various "meta-learners" briefly described as follows (see Kunzel et al. 2019; CausalML 2023; EconML 2023; Facure 2023; Molak 2023, for further description):

a. S-learner: A single (S) model capturing treatment effect as a dummy variable using machine learning, essentially the same as the Treatment Dummy variable approach in Lo (2002) with flexibility on the choice of predictive modeling techniques.

b. T-learner: Developing two (T) separate models for treatment and control groups, respectively, which is the same as the Two Model approach previously discussed (e.g., Lo and Pachamanova 2015), and machine learning techniques can be used.

c. X-learner: Similar to T-learner, X (Cross) learner estimates two separate machine learning models for treatment and control

groups and uses them to predict the "counterfactual" estimates (i.e., estimated control response rates for the treatment group and estimated treatment response rates for the control group). Next, this approach proceeds to take the difference between the actual response and its counterfactual (estimated) rate for each individual and then estimate two separate models using the difference as the outcome variable for the treatment and control groups, respectively. Finally, the two estimates are combined in a weighted average for the X-learner model score.

d. R-learner: This method employs cross validation with a specific loss function to be minimized after adjusting for an outcome model and a propensity score model, see Nie and Wagner (2021) for details.

e. Doubly Robust (DR) learner: This methodology estimates both the propensity score and the outcome model and is effective if at least one of them is correct. This approach is an extension of the DR method for estimating average treatment effect and is particularly useful when the data are observational (i.e., non-experimental), see Kennedy (2023) for details.

Appendix 9.1: Proof of the Four Quadrant Method – Modifying the Lai Method with Addition of Probability Weights

Consider the 2×2 table in Figure 9.1. Define our estimation objective (i.e., lift) as a function of covariates x:

$Z(x) \equiv P(R|T,x) - P(R|C,x)$, where R = event of response. $Z(x)$ is the response probability difference (lift) between the treatment and control groups given a set of characteristics, x. It can be re-expressed as follows.

$$\begin{aligned} Z(x) &= P(R|T,x) - \left(1 - P(N|C,x)\right), \quad \text{where } N = \text{no response} \\ &= P(R|T,x) + P(N|C,x) - 1 \\ &= \frac{P(TR|x)}{P(T|x)} + \frac{P(CN|x)}{P(C|x)} - 1, \quad \text{by Bayes' rule} \\ &= \frac{P(TR|x)}{P(T)} + \frac{P(CN|x)}{P(C)} - 1, \end{aligned} \tag{A9.1}$$

due to randomization of treatment and control, that is, assignment of treatment/control is random and does not depend on x. Note that Eqn. (A9.1) indicates that $P(TR|x)$ and $P(CN|x)$ have positive contributions to the lift.

Similarly, if we look at $Z(x)$ in another way,

$$\begin{aligned} Z(x) &= \big(1 - P\big(N|T,x\big)\big) - P\big(R|C,x\big) \\ &= 1 - \frac{P\big(TN|x\big)}{P(T)} - \frac{P\big(CR|x\big)}{P(C)}, \end{aligned} \tag{A9.2}$$

where Eqn. (A9.2) indicates that $P(TN\,|\,x)$ and $P(CR\,|\,x)$ have negative contributions to the lift.

Note that Eqns. (A9.1) and (A9.2) are the same equations except that we are expressing them differently. Adding Eqns. (A9.1) and (A9.2) together, we have:

$$2Z(x) = \frac{P\big(TR|x\big)}{P(T)} + \frac{P\big(CN|x\big)}{P(C)} - \frac{P\big(TN|x\big)}{P(T)} - \frac{P\big(CR|x\big)}{P(C)}, \quad \text{or}$$

$$\frac{P(C)\big(P\big(TR|x\big) - P\big(TN|x\big)\big) + P(T)\big(P\big(CN|x\big) - P\big(CR|x\big)\big)}{P(T)P(C)} \tag{A9.3}$$

Quite often, marketing programs have a larger sample in the treatment group than the control group; that is, $P(C) < P(T)$, as marketers often aim at gaining more revenue by contacting more individuals. It can be easily shown that Lai (2006)'s method is a special situation where $P(T) = P(C) = 0.5$, which is not mathematically correct in general cases. However, Lai (2006) also proposed using a weight based on empirical findings, but, in fact, there is a simple mathematical equation as shown in Eqn. (A9.3). Note that although this method has the same mathematical objective as Methods A1 and A2, that is, maximizing $P(R|T) - P(R|C)$, empirically, because of the different estimation methods, it results in different estimates.

Appendix 9.2: Computations of Weighted Gini Coefficient and Weighted Top 15% Gini for Non-Randomized Data

This appendix is a generalization of Appendix 6.1, that is, extending the randomized data to the non-randomized data situation, where weights are available for adjustment. As in Appendix 6.1, assume we rank the *holdout* sample

by semi-decile, that is, 20 groups with 5% in each group. Define the average lift at group j as:

$$\text{lift}(j) = P(R|T, j) - P(R|C, j),$$

where $P(R|T, j)$ and $P(R|C, j)$ represent the response probabilities in semi-decile subgroup j in the treatment and control groups, respectively, and can be estimated by the relative frequencies of response in the *holdout* sample. Then,

$$\text{Weighted Gini coefficient} = \sum_{g=1}^{20} (\text{cum\%lift}(g) - \text{cum\%sam}(g)),$$

where

$$\text{cum\%lift}(g) = \text{cumulative \% lift up to semi} - \text{decile group } g$$

$$= \frac{\overline{lift}(1,\ldots,g)\sum_{j=1}^{g}\sum_{i} w_{tji}}{\overline{lift}\sum_{j=1}^{20}\sum_{i} w_{tji}} = \frac{\overline{lift}(1,\ldots,g)\sum_{j=1}^{g} w_{tj}}{\overline{lift}\sum_{j=1}^{20} w_{tj}},$$

where

$$\overline{lift} = \text{overall weighted lift (of all 20 semi} - \text{deciles)}$$

$$= \frac{\sum_{j=1}^{20}\left(\sum_{i} w_{tji}\right)\overline{y}_{tj}}{\sum_{j=1}^{20}\sum_{i} w_{tji}} - \frac{\sum_{j=1}^{20}\left(\sum_{i} w_{cji}\right)\overline{y}_{cj}}{\sum_{j=1}^{20}\sum_{i} w_{cji}},$$

and

$\overline{lift}(1,\ldots,g)$ = weighted lift from semi-deciles 1, …, g, (where $g = 1,\ldots, 20$)

$$= \frac{\sum_{j=1}^{g}\left(\sum_{i} w_{tji}\right)\overline{y}_{tj}}{\sum_{j=1}^{g}\sum_{i} w_{tji}} - \frac{\sum_{j=1}^{g}\left(\sum_{i} w_{cji}\right)\overline{y}_{cj}}{\sum_{j=1}^{g}\sum_{i} w_{cji}}$$

$$= \frac{\sum_{j=1}^{g} w_{tj}\ \overline{y}_{tj}}{\sum_{j=1}^{g} w_{tj}} - \frac{\sum_{j=1}^{g} w_{cj}\ \overline{y}_{cj}}{\sum_{j=1}^{g} w_{cj}},$$

And w_{tji} = weight associated with the treatment group, semi-decile j, and individual i in the *holdout* sample; and similarly, w_{cji} = weight associated

with the control group, semi-decile j, and individual i in the *holdout* sample. Additionally, $w_{tj} = \sum_i w_{tji}$ and $w_{cj} = \sum_i w_{cji}$, representing the total weight at semi-decile j for the treatment group and control group, respectively. Further,

$$\text{cum\%sam}(g) = \text{cumulative \% sample up to semi} - \text{decile group } g$$

$$= \frac{\sum_{j=1}^{g} \sum_i w_{tji}}{\sum_{j=1}^{20} \sum_i w_{tji}} = \frac{\sum_{j=1}^{g} w_{tj}}{\sum_{j=1}^{20} w_{tj}} .$$

Similarly, the Weighted Top 15% Gini is simply focused on the top 15%, or the top 3 semi-deciles, of the Weighted Gini coefficient formula:

$$\text{Weighted Top 15\% Gini} = \sum_{g=1}^{3} (cum\%lift(g) - cum\%pop(g)).$$

Appendix 9.3: Proof of Propensity Score Weighting for Non-Randomized Data

Define Y_1 and Y_0 as the potential outcomes for treatment and control (untreated), respectively. In reality, only one of these two is observed for an individual. One can use the IPW method to compute ATE for the whole population, Average Treatment Effect on Treated (ATT) or Average Treatment Effect on Control (ATC). We will use ATC below for illustration.

The ATC is defined as:

$$ATC = E\left(Y_1 | T = 0\right) - E\left(Y_0 | T = 0\right)$$

where the first component on the right is defined as the mean potential outcome of treated in the control group and the second component is the mean potential outcome of untreated (control) in the control group. The second component can be estimated from the observed control group data using the sample mean, but the first component will require some adjustment.

To compute the first component, we prove the IPW (Inversely Probability Weighting) method below:

$$E\left(Y_1 | T = 0\right)$$

$= E_X E_Y\left(Y_1 | T=0, x\right)$, where the first expectation is w.r.t. covariates X and the second is w.r.t. outcome Y

$$= \iint y\, p\left(y_1 | T=0,\, x\right) p\left(x | T=0\right)\, dy\, dx$$

$= \iint y p\left(y_1 | T=1,\, x\right)\, p\left(x | T=0\right)\, dy dx$, because $T \perp (Y_1,\, Y_0) | x$, that is, by the conditional ignorability (or exchangeability) condition

$$= \iint y p\left(y_1 | T=1,\, x\right) p\left(x | T=1\right) \frac{p\left(x | T=0\right)}{p\left(x | T=1\right)}\, dy\, dx$$

the beginning step to change the domain[18] from $T=0$ to $T=1$

$$= \iint y p\left(y_1 | T=1,\, x\right) p\left(x | T=1\right) \frac{\frac{P\left(T=0|x\right) p(x)}{P(T=0)}}{\frac{P\left(T=1|x\right) p(x)}{P(T=1)}}\, dy dx \text{ by Bayes' Theorem}$$

$$= \iint y p\left(y | T=1, x\right) p\left(x | T=1\right) \frac{\frac{P\left(T=0|x\right)}{P(T=0)}}{\frac{P\left(T=1|x\right)}{P(T=1)}}\, dy\, dx$$

by the consistency assumption and canceling a common term

$$= \iint y p\left(x, y | T=1\right) w_{1\to 0}(x)\, dy\, dx$$

where $w_{1\to 0}(x) = \frac{\frac{P\left(T=0|x\right)}{P(T=0)}}{\frac{P\left(T=1|x\right)}{P(T=1)}} = \frac{1-PS(x)}{PS(x)} \frac{P(T=1)}{P(T=0)}$, representing the weight function to map from treatment to control, and $PS(x) =$ propensity score as a function of covariates, x.

Note that the last integral form of $E\left(Y_1 | T=0\right)$ is simply the weighted expected value of Y in the treatment group and can be estimated by the following sample average in the treatment group:

$$\hat{E}\left(Y_1 | T=0\right) = \frac{\sum_{k\in T} w_k y_k}{\sum_{k\in T} w_k},$$

where $w_k = \frac{1-\widehat{PS_k}(x)}{\widehat{PS_k}(x)} \frac{P(T=1)}{P(T=0)}$, $k \in T$, and $\widehat{PS_k}(x) =$ estimated propensity score for individual k. Note that $\sum_{k \in T} w_k$ can be shown to approximate (or asymptotically converge to) n_T, the original sample size of the treatment group, and the weight w_k can be numerically normalized to satisfy $\sum_{k \in T} w'_k = n_T$, by defining $w'_k = w_k \frac{n_T}{\sum_{l \in T} w_l}$. This weight is the same as the formula in Eqn. (9.5a) for measuring ATC.

The above weighting scheme can be easily extended to multiple treatments with a control group (i.e., moving each treatment group to the control group using a weight).

In the context of uplift modeling for non-randomized experiments, we aim at estimating the individual-level treatment effect, and the above proof can be extended to condition on covariates, x, for this purpose. In other words, the outer expectation (integral) E_X would be taken off for the uplift situation. The key is that the weight would be the same as estimating the population-level *ATC*.

Appendix 9.4: Different Training Sample and Usage Population

All the examples described in Section 9.3 can be considered to have selection bias (in social sciences, economics, or statistics), which implies that the treatment and control are not homogeneous with each other. In a similar but different situation, where the training sample (including treatment and control) is different from the usage population, it is called a *data shift* problem in the machine learning literature and is a relatively new subfield (e.g., Shimodaira 2000, Bickel and Scheffer 2007, Bickel 2009, Bickel et al. 2009, and Sugiyama and Kawanabe 2012).

Let's suppose a training sample is available for Uplift modeling. If treatment and control are similar, methodologies in Chapter 6 or Section 9.2 can be applied; otherwise, methods introduced in Section 9.4 can be employed. What if we know the training sample is not identical to the usage population? For example, the training sample has more of the younger age group or more of the higher income prospects, but we are interested in applying the model to a broader population. Obviously, if there is very little overlap between the training sample and the broader population, it will not be a good idea to "extrapolate" the model out too far. However, in many cases, there are still quite a lot of overlaps, and the question is how to adjust our methodologies, assuming we have some information about the usage population.

The solution from the data shift literature states that all we need is to apply a weight in the model estimation of the training sample you have, even though the training sample is known to be different from the future *applied* sample. The weight has the following form:

$$W_i = \frac{P(applied)}{P(training)}\left(\frac{1}{P(applied|x_i)} - 1\right), \tag{A9.4a}$$

which can be arranged as:

$$W_i = \frac{P(applied)/(1-P(applied))}{P(applied|x)/(1-P(applied|x_i))}, \tag{A9.4b}$$

where $P(applied)$ and $P(training)$ are the probabilities that a data point is drawn from applied or training set (imagine that the applied and training data are mixed together randomly), respectively, and $\frac{P(applied)}{P(training)}$ can be estimated by $\frac{\text{Sample size of applied data}}{\text{Sample size of training data}}$. Likewise, $P(applied|\, x_i)$ is the probability that a data point is drawn from the applied set given covariates or predictors x_i, and can be estimated by supervised learning techniques for binary dependent variables such as logistic regression.

Note that Eqn. (A9.4b) resembles Eqn. (9.5). Both the data shift problem and the selection bias problem in Section 9.4 can be solved by some adjustment using a set of weights, as the two problems are methodologically quite similar. For the data shift problem, once the weight from Eqn. (A9.4b) is estimated, the uplift model can be estimated along with the estimated weight, similar to the PS matching adjustment method in Section 9.4.

Notes

1. The reason for choosing the IPW method as opposed to other common methods of propensity score matching, such as principal stratification (also known as sub-classification), is that the weight can be readily used in an outcome regression for uplift modeling. However, principal stratification is essentially weighting at the stratum (group) level as opposed to the individual level. Therefore, to use principal stratification, one simply needs to use stratum-specific proportions of treatment and control as the denominators of (8.5); see Section 17.8 of Imbens and Rubin (2015) for details of the stratum-level weighting description and Lunceford and Davidian (2004) for comparisons between the two methods. Abrevaya et al. (2015) and Athey and Imbens (2015) also propose to use the IPW method for computing Conditional Average Treatment Effect (CATE), which is equivalent to uplift modeling on observational data.

2. The weight in Eqn. (9.5) is essentially measuring the Average Treatment Effect (ATE) on the whole population rather than Average Treatment Effect on Treated (ATT) or Average Treatment Effect on Control (ATC). Alternatively, unlike measuring effect on the population, uplift modeling may search for individual (or subgroup) causal effects for untreated individuals; that is, what would be the effect if an untreated individual were treated? As a result, applying weight for ATT or ATC is more appropriate in some situations; see Appendix 9.3 for a mathematical proof for the case of calculating ATC.
3. This is called the stabilized weights as opposed to unstabilized weights where numerators equal 1. The stabilized weights have a potential advantage of reduced variability; see Section 12.4 of Hernan and Robins (2016). They also make the pseudo-sample the same size as the original sample; see Hernan and Robins (2005) for details. The $P(T)$ and $1-P(T)$ constant components in (9.5), (9.5a), and (9.5b) all serve a stabilizing purpose.
4. Since we are interested in the treatment effect conditional on some characteristics rather than the overall effect, instead of Average Treatment Effect (ATE), Average Treatment effect on Control (ATC), and Average Treatment effect on Treated (ATT), we can technically call them Conditional ATE (or CATE), Conditional ATC, and Conditional ATT, respectively.
5. See, for example, Section 12.5 of Hernan and Robins (2016), where Marginal Structural Model along with effect modification is estimated. See also VanderWeele (2009) for the distinction between "interaction" (between two treatments) and effect modification (interaction between treatment and covariates) in Epidemiology.
6. See, for example, Freedman and Berk (2008) and Posner and Ash (2012) for potential issues with this approach.
7. Although logistic regression is the most commonly used method for propensity score estimation, one can use other methods such as decision trees, random forests, neural networks, etc.
8. While this Section (9.4) is about applying propensity score matching for observational data, there is a parallel method discussed in the machine learning literature where the training data and the data where the developed model is applied to are not the same. Such method is briefly discussed in Appendix 9.4.
9. Proc logistic in SAS is used for all the uplift modeling methods in this example, using the weights constructed. The academic literature stated that the standard errors of coefficients can be biased when the weights are used, resulting in potentially biased p-values. To incorporate the weights properly for variance estimation, one may use proc surveylogistic which has a practical disadvantage of not being able to perform stepwise procedure automatically, at least in the current version of SAS (9.4). We tried proc surveylogistic for some models, and the results are only minimally different from proc logistic.
10. Some readers may recall such experience of forgetting to use a coupon while making purchases at a store.
11. In the click-through case, if someone clicked but did not proceed with a purchase, this component would not be 1 and become another probability to be determined.
12. Equation (8.1) estimates the lift directly and does not provide separate estimates for treatment and control response rates, although one could possibly extract the treatment and control response rate components from (8.1).

13. Since variables including age, income, spent, and wealth are estimated along with their respective log variables, we monitor the output of the stepwise regression to reduce collinearity using this rule: When the raw variable and its log counterpart both show up as main effects, only the one with a lower p-value is selected. Similarly for the interaction effects with the treatment dummy.
14. In Table 9.6, while the direct response model has the lowest performance in terms of the Gini metrics, it has a higher R^2, due to a more linear relationship between lift and semi-decile numbers, as seen in Figure 9.6. As explained in Chapter 6, the Gini metrics are more important metrics for comparison.
15. Similar to the example in the previous section, variables including age, income, spent, and wealth are estimated along with their respective log variables, and collinearity is reduced with this rule: When the raw variable and its log counterpart both show up as main effects, only the one with a lower p-value is selected. Similarly for the interaction effects with the treatment dummy.
16. Since S represents log(sales), the estimated value of sales is simply an exponential function of the model estimate of $E(S|\ldots)$.
17. A similar method for adjusting training sample when the usage population is different is briefly discussed in Appendix 9.4.
18. Such a change of probability measure is known as the *Radon-Nikodym derivative* in probability theory literature; see Billingsley (1995).

References

Abrevaya, Jason, Yu-Chin Hsu, and Robert P. Lieli. 2015. "Estimating Conditional Average Treatment Effects". *Journal of Business & Economic Statistics*, 33: 485–505.

Athey, Susan, and Guido W. Imbens. 2015. "Machine Learning Methods for Estimating Heterogeneous Causal Effects". Working Paper, Stanford Graduate School of Business.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests". *The Annals of Statistics*, 47(2): 1148–1178.

Bickel, Steffen. 2009. "Learning under Differing Training and Test Distributions". *PhD Dissertation*, University of Potsdam.

Bickel, Steffen, Michael Bruckner, and Tobias Scheffer. 2009. "Discriminative Learning Under Covariate Shift". *Journal of Machine Learning Research* 10: 2137–2155.

Bickel, Steffen, and Tobias Scheffer. 2007. "Dirichlet-Enhanced Spam Filtering Based on Biased Samples". In Bernhard Schölkopf, John Platt, and Thomas Hofmann (eds.), *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*. Cambridge, MA: MIT Press, 161–168.

Bickel, Steffen, Michael Bruckner, and Tobias Scheffer. 2009. "Discriminative Learning under Covariate Shift". *Journal of Machine Learning Research*, 10: 2137–2155.

Billingsley, Patrick. 1995. *Probability and Measure*, 3rd edition. Hoboken, NJ: Wiley.

Uber (2025) CausalML. https://github.com/uber/causalml

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects". *Biometrika*, 96: 187–199.

Facure, Matheus. 2023. *Causal Inference in Python: Applying Causal inference in the Tech Industry*. Sebastopol, CA: O'Reilly.

Freedman, David A., and Richard A. Berk. 2008. "Weighting Regressions by Propensity Scores". *Evaluation Review*, 32: 392–409.

Grimmer, Justin, Solomon Messing, and Sean Westwood. 2016. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods". Working paper, Department of Political Science, Stanford University.

Gubela, Robin, Stefan Lessmann, Johannes Haupt, Annika Baumann, Tillmann Radmer, and Fabian Gebert. 2017. *Revenue Uplift Modeling*. Conference Paper, Thirty Eighth International Conference on Information Systems (ICIS), Seoul, Korea: ICIS.

Guelman, Leo, Montserrat Guillen, and Ana M. Perez-Marin. 2014. "A Survey of Personalized Treatment Methods for Pricing Strategies in Insurance". *Insurance: Mathematics and Economics*, 58: 68–76.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2013. *The Elements of Statistical Learning*, 7th printing edition. New York, NY: Springer-Verlag.

Hernan, Miguel A., and James M. Robins. 2005. "Estimating Causal Effects from Epidemiological Data". *Journal of Epidemiology Community Health*, 60: 578–586.

Hernán, Miguel, and James Robins. 2025. *Causal Inference: What If*. Boca Raton, FL: CRC Press.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, England: Cambridge University Press.

Kane, Kathleen, Victor S. Y. Lo, and Jane Zheng. 2014. "Mining for the Truly Responsive Customers and Prospects Using True-Lift Modeling: Comparison of New and Existing Methods". *Journal of Marketing Analytics*, 2(4): 218–238.

Kennedy, Edward H. 2023. "Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects". Retrieved on January 1, 2024, https://arxiv.org/abs/2004.14497

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning". *Proceedings of the National Academy of Sciences*, 116(10): 4156–4165.

Lai, Lilly Y.-T. 2006. Influential Marketing: A New Direct Marketing Strategy Addressing the Existence of Voluntary Buyers. *Master of Science thesis*, Simon Fraser University School of Computing Science.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. 2011. "Weight Trimming and Propensity Score Weighting". *PLOS ONE*, 6(3): e18174.

Lo, Victor S. Y. 2002. "The True-Lift Model – A Novel Data Mining Approach to Response Modeling in Database Marketing". *ACM SIGKDD Explorations*, 4(2): 78–86.

Lo, Victor S. Y., and Dessislava Pachamanova. 2015. "Prescriptive Uplift Analytics: A Practical Approach to Solving the Marketing Treatment Optimization Problem and Accounting for Estimation Error Risk". *Journal of Marketing Analytics*, 3(2): 79–95.

Lunceford, Jared, and Marie Davidian. "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study". *Statistics in Medicine*, 23(19): 2937–2960.

Molak, Aleksander. 2023. *Causal Inference and Discovery in Python*. Birmingham, England: Packt Publishing.

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference*, 2nd edition. New York, NY: Cambridge University Press.

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-oracle Estimation of Heterogeneous Treatment Effects," *Biometrika*, 108(2): 299–319.

Posner, Michael A., and Arlene S. Ash. 2012. "Comparing Weighting Methods in Propensity Score Analysis". Unpublished working paper, Columbia University.

Quinonero-Candela Joaquin, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2009. *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika*, 70(1), 41–55.

Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash. 2008. *Modern Epidemiology*, 3rd edition. Riverwoods, IL: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Shimodaira, Hidetoshi. 2000. "Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function". *Journal of Statistical Planning and Inference*, 90: 227–244.

Sturmer, T., R. Wyss, R. J. Glynn, and M. A. Brookhart. 2014. "Propensity Scores for Confounder Adjustment When Assessing the Effects of Medical Interventions Using Nonexperimental Study Designs". *Journal of Internal Medicine*, 275(6): 570–580.

Sugiyama, Masashi, and Motoaki Kawanabe. 2012. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaption*. Cambridge, MA: MIT Press.

Tian, Lu, Ash A. Alizadeh, Andrew J. Gentles, and Robert Tibshirani. 2014. "A Simple Method for Estimating Interactions between a Treatment and a Large Number of Covariates". *Journal of the American Statistical Association*, 109(508): 1517–1532.

Tibshirani, Julie, Susan Athey, et al. 2023. "Package 'grf' version 2.3.1". Retrieved on January 1, 2024. https://cran.r-project.org/web/packages/grf/grf.pdf

VanderWeele, Tyler J. 2009. "On the Distinction between Interaction and Effect Modification". *Epidemiology*, 20(6): 863–871.

Weisberg, Herbert I., and Victor P. Pontes. 2015. "Post Hoc Subgroups in Clinical Trials: Anathema or Analytics?" *Clinical Trials*, 12(4): 357–364.

Westreich, Daniel, Justin Lessler, and Michele Jonsson Funk. 2010. "Propensity Score Estimation: Machine Learning and Classification Methods as Alternatives to Logistic Regression". *Journal of Clinical Epidemiology*, 63(8): 826–833.

10

Causality in Times Series Data

10.1 Introduction

Time series data, which track variables hourly, daily, weekly, monthly, quarterly, or annually, are ubiquitous. Firms track sales, purchases, profits, payroll, employment, and numerous other series, often displaying the results in stylish dashboards and graphs.

Managers want to extract usable lessons from these data. For instance, data on advertising in different channels (TV, print, social media, etc.) may be correlated with sales of the company's services, but the firm would like to know what mix of media will maximize the return to their advertising budget: Media Mix Modeling (MMM) models like this are widely used. Another company may be considering raising the price of one of its products but first wants to estimate the effect that such a change would have on sales. This can only be done if there is a plausible causal story behind the data.

In this chapter, we first work through a straightforward example where a firm wants to measure the effect of changes in prices. We then introduce the idea of "Granger causality," and evaluate the extent to which such an approach is, or is not, causal. There are good recent academic reviews of causality in time series in Runge et al. (2023) and Palshikar et al. (2023).

10.2 Raise the Price?

Your company, Federal Oil, produces and sells gasoline, currently charging $3.51 per gallon, and would like to estimate the effects of raising this price. The company's analyst has data going back to 2005 on its sales, regional Gross Domestic Product and population, a price index, and the prices of gasoline and diesel fuel (which latter is produced by a rival firm). The numbers are shown in Table 10.1 (and are from a real, but somewhat different, setting).

The basic idea is to relate consumption to price, but some adjustments to the dataset are called for, and this is quite typical. It is likely that an

 DOI: 10.1201/9780429172588-10

TABLE 10.1

Data for Gasoline Demand Modeling

Year	Qgas	GDP	CPI	Pop	Pgas	Pdiesel	Qgaspc	RealPgas	RealPdiesel	RealGDP/cap
2005	111.3	30.1	14.6	8.2	0.18	0.10	13.6	4.10	2.35	83.3
2006	94.3	36.8	16.7	8.5	0.23	0.13	11.1	4.61	2.51	85.9
2007	108.2	42.7	19.7	8.8	0.31	0.16	12.3	5.15	2.74	81.6
2008	93.0	48.8	25.7	9.0	0.50	0.24	10.3	6.39	3.06	69.9
2009	77.3	61.7	33.9	9.3	0.61	0.28	8.3	5.92	2.74	64.8
2010	76.5	75.6	40.5	9.4	0.65	0.30	8.1	5.28	2.42	65.7
2011	70.9	84.8	44.4	9.7	0.72	0.34	7.3	5.38	2.57	65.1
2012	72.7	94.7	49.1	10.0	0.76	0.37	7.3	5.15	2.51	64.0
2013	75.3	110.2	56.3	10.1	0.76	0.37	7.5	4.49	2.19	64.2
2014	73.4	137.2	64.7	10.4	0.95	0.47	7.1	4.86	2.39	67.7
2015	69.1	171.8	82.1	10.6	1.35	0.68	6.5	5.43	2.73	65.2
2016	72.5	200.3	89.5	10.9	1.34	0.67	6.6	4.97	2.49	67.9
2017	75.8	230.2	100.0	11.2	1.44	0.73	6.8	4.76	2.41	68.1
2018	67.4	245.7	108.5	11.5	1.65	0.89	5.9	5.04	2.73	65.2
2019	74.0	279.7	124.4	12.1	1.68	0.99	6.1	4.46	2.63	61.6
2020	82.7	322.5	136.8	12.4	1.69	1.14	6.7	4.09	2.76	62.9
2021	87.3	456.6	190.1	12.8	2.16	1.32	6.8	3.76	2.31	62.3
2022	95.9	682.0	276.0	13.1	3.07	2.40	7.3	3.69	2.88	62.3
2023	103.5	820.2	331.1	13.5	3.51	2.89	7.7	3.51	2.89	60.9

Note: Qgas: quantity of gasoline sold; GDP: Gross Domestic Product; CPI: Consumer price index; Pop: Population; Pgas: Price of gasoline; Pdiesel: Price of diesel fuel; Qgaspc: Quantity of gasoline per capita; RealPgas: Real price of gasoline (i.e., price of gasoline in prices of 2023); RealPdiesel: Real price of diesel; RealGDP/cap: Real GDP per capita.

appropriate outcome variable would be sales of gasoline per capita. More importantly, prices and GDP should all be adjusted for inflation – we bring them to the prices of 2023 – to give "real" prices and GDP per capita. When time series stretch over at least some months, deflation such as this is likely to be necessary.

It is often helpful to graph the key series before proceeding to further modeling. In our example, the relevant series are shown in Figure 10.1. The top panels show the real prices of gasoline (on the left) and diesel fuel (on the right); the bottom left panel shows real GDP per capita (which dropped sharply early in the period in question), and the bottom right panel shows the sales of gasoline per capita. We want to know what would happen to the quantity (and value) of sales were we to change the real price of gasoline.

A priori, we believe that the quantity of gasoline bought will be influenced by the level of real GDP per capita, the price of gasoline itself – higher price, lower quantity sold, as per the "law" of demand – and also the price of diesel fuel, which represents the cost of a close substitute for gasoline (because drivers could switch to diesel-powered cars, given enough time). Most practical

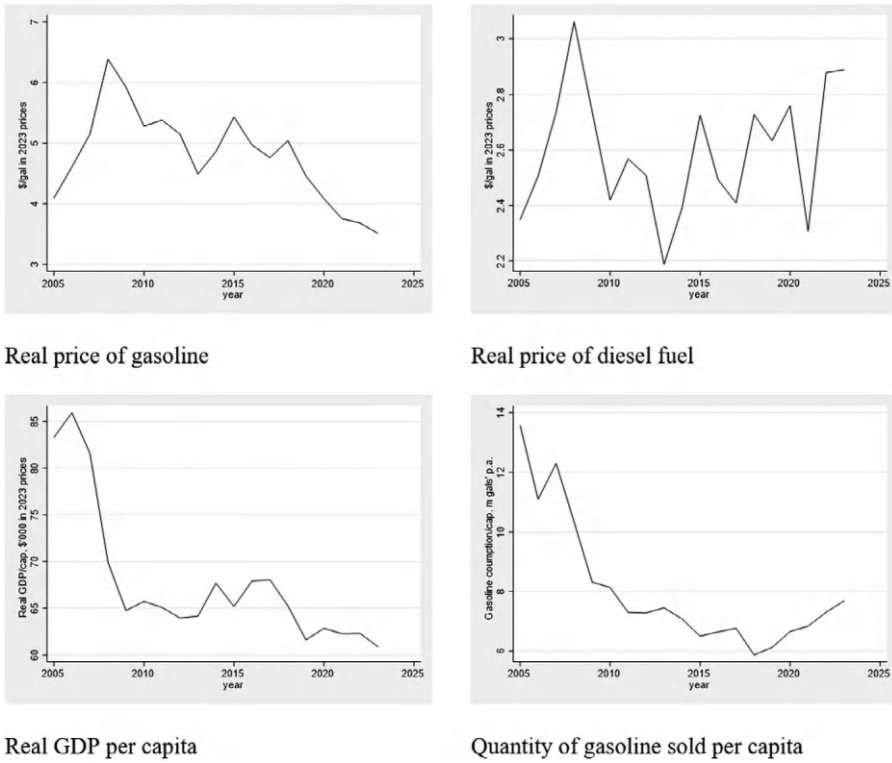


FIGURE 10.1
Visual representation of key time series in gasoline demand example.

models are more elaborate than this one, but all require some thought about the direction in which we believe causality moves.

The results of several regression models are shown in Table 10.2. Column (2) shows the estimated coefficients of a simple linear model ("Ordinary least squares"). It may be written as follows:

$$\text{qgaspercap} = -13.095 - 0.119\,\text{rpgas} + 1.556\,\text{rpdie} + 0.261\,\text{rGDPpercap} \qquad (10.1)$$

The table shows an adjusted R^2 of 0.74, which is quite a close fit. The coefficients have the expected signs: If this is indeed measuring a demand curve, we expect a higher price of gasoline to be associated with a lower quantity sold, while a higher price of diesel fuel will have a positive effect, as consumers switch to buying gasoline. Column (3) shows the p-values, where a low value – typically taken to be 0.1 or lower – shows that the estimated coefficient is statistically significantly different from zero, and so is measuring something real.

A lower row of Table 10.2 shows the Durbin-Watson statistic. This is a test to determine whether there might be autocorrelation in the residuals of the

TABLE 10.2

Estimation Results, Gasoline Demand Model

		Linear		Linear with Time		Logs	
	Mean	**Coeff.**	**p-Value**	**Coeff.**	**p-Value**	**Coeff.**	**p-Value**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Dependent variable							
Quantity of gasoline per capita	8.07						
Independent variables							
Real price of gasoline ($/gal)	2.90	–0.119	0.75	–1.790	<0.01	–0.097	0.64
Real price of diesel ($/gal)	1.57	1.556	0.22	3.480	<0.01	0.446	0.25
Real GDP per capita ($, '000)	68.80	0.261	<0.01	0.062	0.18	2.029	<0.01
Year				–0.399	<0.01		
Intercept		–13.095	0.01	806.458	<0.01	–6.761	<0.01
Adjusted R-squared		0.74		0.90		0.68	
Durbin Watson/Durbin		1.28		2.44		0.92	
Dickey-Fuller (approx p-value)		0.01		<0.01		0.14	

	Autocorrelation adjustment		**Partial adjustment**		**Error Correction**	
	Prais-Winsten		**Logs, lagged y**		**Δlogs, lagged y**	
	Coeff.	**p-Value**	**Coeff.**	**p-Value**	**Coeff.**	**p-Value**
	(8)	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**
Dependent variable						
Quantity of gasoline per capita						
Independent variables						
Lagged qty of gasoline per capita			0.718	<0.01	–0.033	0.87
Real price of gasoline ($/gal)	–0.348	0.27	–0.263	0.07	–0.792	0.03
Real price of diesel ($/gal)	0.152	0.62	0.081	0.74	0.390	0.12
Real GDP per capita ($, '000)	1.086	0.03	0.416	0.24	–0.094	0.83
Year						
AR(1) term (rho)	0.808					
Error correction term (e from Step 1)					–0.684	0.05
Intercept	–2.080	0.31	–-0.864	0.52	–0.012	0.09
Adjusted R-squared	0.88		0.86		0.43	
Durbin Watson/Durbin	2.18		1.78	0.18		
Dickey-Fuller (approx p-value)	0.34		<0.01			

series. In the simplest case of an independent variable X and an outcome variable Y, subscripted for time, we have

$$Y_t = \alpha + \beta X_t + \varepsilon_t \tag{10.2}$$

Here, α and β are the true parameters (which can only be estimated), and ε is the residual that is not "explained" by the model, usually assumed to be randomly distributed with zero mean. In time series data, the residuals are often serially correlated, so that a large positive random shock last period is more likely to be followed by another positive shock this year (if we have positive autocorrelation), in which case shocks may persist over time. Formally, we may represent first-order autocorrelation as

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t \tag{10.3}$$

where ρ is the autocorrelation coefficient, and v_t is random noise. If there is autocorrelation in the residuals ε_t in Eqn. (10.1), then tests of the estimated coefficients will be too confident, or (equivalently) the p-values will be too low. A Durbin-Watson statistic well below 2 (as we have in column (2)) indicates positive autocorrelation, while a value well above 2 reflects negative autocorrelation (Gujarati and Porter 2009).

The presence of autocorrelation is evidence that we have not specified our model correctly: We may have omitted important variables, or be using the wrong functional form. One common fix is to add a time trend, with the results shown in columns (4) and (5) of Table 10.2. In our example, the equation fits better, but now we appear to have negative autocorrelation. The time trend may also be picking up some of the effects on gasoline sales that would more rightfully be attributable to movements in real GDP per capita, yielding a model that risks being less useful for policy purposes.

Economists often estimate equations such as this one in log-log form. It is still a linear regression, but now the variables are in log form. One reason for doing this is that the estimated coefficients now give elasticities, which are unit-free measures of how a percentage change in an "explanatory" variable is associated with a percentage change in the outcome variable. For instance, in Column (6) in Table 10.2, the own-price elasticity of demand for gasoline is given by −0.097, which means that if the price were to rise by (say) 10%, the quantity demanded would fall by approximately 0.97%. This is convenient because now the firm has an easy way to estimate the effects of changes in the price of gasoline on the quantity it sells.

There is strong positive autocorrelation in the equation whose estimates are shown in Column (6), as shown in the Durbin-Watson statistic of 0.92. Some manipulation gives

$$Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta\left(X_t - \rho X_{t-1}\right) + v_t \tag{10.4}$$

Once we have a value for ρ, we can calculate $Y_t - \rho Y_{t-1}$ and $X_t - \rho X_{t-1}$ and apply ordinary least squares to get our estimate of β. The results, based on the Prais-Winston technique (Johnston 1972), are shown in Columns 8 and 9 of Table 10.2. The fit is good, and the estimated coefficients are reasonable, and there is no autocorrelation in the residuals. This is a viable model.

A recurrent issue in time-series analyses is that shocks, even if random, may affect the outcome variable over a longer period than the one used in the model. For instance, a hurricane might reduce the demand for gasoline this year, but also next year, especially if consumers take time to adjust. There may also be a slow response to changes in, for instance, the price of gasoline: In the short run, a higher price will prompt consumers to drive less, but in the long run they may replace large cars with small ones or move closer to work.

To reflect such lags, it may be useful to estimate a partial adjustment model. Formally, we may write

$$y_t = y_{t-1} + k\left(y_t^* - y_{t-1}\right). \tag{10.5}$$

This says that the observed level of the variable (y_t) is equal to the level of the variable in the previous period (y_{t-1}) and an adjustment factor that is k times the gap between the desired level of the variable (y_t^*) and its previous value. If $k = 1$, the adjustment is complete in the time period, and the observed value of y is equal to its desired level. A low value for k implies that the consumer moves slowly from their previous level of y to their desired level. The need to allow for partial adjustments is likely to be greater when the time period is shorter (e.g., weeks rather than months).

We assume that the independent variables drive the desired outcome, which is the outcome we would like to achieve if we could do so without any adjustment costs. In the simple case, this gives

$$y_t^* = \alpha + \beta X_t + \varepsilon_t \tag{10.6}$$

which, with some rearrangement, gives

$$y_t = k\alpha + (1-k)y_{t-1} + k\beta X_t + \varepsilon_t. \tag{10.7}$$

This can be estimated directly, with the key change being the inclusion of the lag of the outcome variable on the right-hand side. The results are shown in columns (10) and (11) of Table 10.2. The equation fits well enough, and Durbin's h-statistic is 1.78, close enough to 2 to suggest that autocorrelation is not a problem. The estimated value of k is 0.282 (= 1 – 0.718), which implies that about 28% of the adjustment of quantity demand to changes in prices or income takes place within a year – a relatively slow reaction. The coefficients in column (10) give short-run elasticities – for instance, –0.263 for the own-price elasticity of demand – but the long-run elasticities are found by dividing these by k. Thus, the long-term elasticity of demand for gasoline in this case would be –0.93 (= –0.263/0.282).

10.3 Stationarity

Up to this point, we have implicitly assumed that the time series with which we are working are stationary, meaning that they do not show any clear trend over time, or more precisely, their mean and variance do not change over time. If two or more series are not stationary, they may seem to be closely related because they are drifting together (or apart) over time, even though there may not be a relationship between them. This can easily lead to spurious correlations and to incorrect inferences about possible causal relationships.

In passing, we might note that while spurious correlations are not confined to time series, this is where they are most common. For instance, over the period 1960 to 2019, the correlation between the number of beaver colonies in Ohio and real GDP per capita in Ireland was 0.88. It is impossible to think of a causal link between these two measures, or even a common cause ("confounder"), and neither follows a consistent trajectory over time. Clearly, correlation is not causality, and in this case, it is not even a hint!

A common solution to nonstationarity is to purge the series of the time trend, typically by taking differences ("integrating of degree 1") (Enders 2010). Now, instead of regressing y_t on x_t, we would regress $y_t - y_{t-1}$ on $x_t - x_{t-1}$. When this is done for Ohio beavers and Irish real GDP per capita, any correlation disappears!

In this context, it is helpful to test whether time series are stationary. This is often done using the augmented Dickey-Fuller test, although other tests are sometimes applied (Enders 2010). The simplest version of the test consists of estimating an equation of the form

$$\Delta X_t = \varphi_0 + \varphi_1 t + \varphi_2 X_{t-1} + \varepsilon_t. \tag{10.8}$$

If the estimated value of φ_2 is significantly negative, then one rejects the hypothesis of a "unit root," meaning that it is reasonable to suppose that the series X_t is stationary. Testing for stationarity for the four variables in our example, plus the consumer price index, gives the results shown in Table 10.3. It appears that the price of gasoline and the real GDP per capita are not stationary

TABLE 10.3

Results of Dickey-Fuller Tests for Stationarity

Variable (in Logs)	Test Statistic (φ_2)	Approx p-Value	Stationary?
Quantity of gasoline sold	–2.914	0.04	Yes
Real price of gasoline	–0.828	0.81	No
Real price of diesel	–3.267	0.02	Yes
Real GDP per capita	–2.090	0.25	No
Consumer price index	5.570	1.00	No

but instead show some systematic movement over time, as does (of course!) the consumer price index.

A subset of our variables are nonstationary, which suggests that it may be useful to estimate an Engle-Granger error correction model. There are two steps:

Step 1: Estimate a linear equation in levels (as in Eqn. (10.1)). If the equation is statistically significant and the residuals are stationary, we have a cointegrating vector, which gives the long-run relationship between the variables (and in our case, the long-run elasticities).

Step 2: Estimate the relationship in its differenced form, and include the lagged residuals from Step 1 (and typically the lagged value of the dependent variable). In its simplest form, it will look like:

$$\Delta \ln(Y_t) = a + b\,\Delta \ln(Y_{t-1}) + c\,\Delta \ln(X_{t-1}) + d\,e_{t.} \tag{10.9}$$

The equation should also include the variables that are stationary, since they may have a short-run effect on the outcome variable, even if they cannot have a permanent effect on it. The estimate of the coefficient d measures how quickly the outcome variable adjusts toward its long-run level.

The results of this estimation, for our example, are shown in column (12) in Table 10.2. The error correction coefficient is -0.68, so if we are off the long-run demand curve, gasoline consumption will adjust back to its long-run level, with about two-thirds of the adjustment taking place in the first year.

In following these procedures, we are assuming a direction of causality, and we consider that the model is complete. This is where judgment comes in, and a model, or at least a clear structure, is needed. In Figure 10.2, we have created a path diagram with nodes and edges in an effort to capture the relevant relationships. We have focused on the relationship between the price of gasoline and the quantity of gasoline sold. It is likely that the price of crude oil influences both the price of gasoline and the price of diesel fuel since crude oil constitutes the great bulk of the cost of producing gasoline

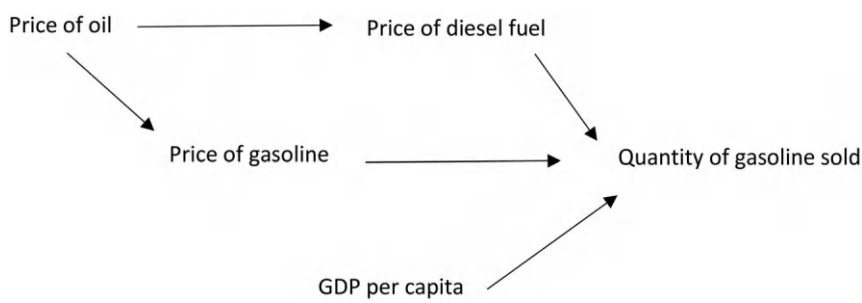


FIGURE 10.2
Path diagram for Gasoline Sales Model. (T is the treatment, Y the outcome of interest, and C, X, S, and Z are other variables.)

and diesel fuel. To block this backdoor effect, we needed to include the price of diesel fuel in the regression; otherwise, the measured effect of the price of gasoline on sales of gasoline would pick up some of the effects of the price of diesel fuel as well, and we want to change one without the other – that is, to "do" a change in the price of gasoline.

Of course, our model is likely to be too simple. Other variables may affect the quantity of gasoline purchased, such as the price of cars, technical changes (like the development of electric vehicles), or other shocks such as a hurricane. If these are completely unrelated to the price of gasoline, then we may still be able to identify the effects of gasoline prices on sales, but complete independence is rare. We discussed Bayesian Networks more fully in Chapter 5.

There is also the serious problem of endogeneity. While it is logical that a higher price of gasoline can be expected to reduce the quantity demanded, it may also be true that an exogenous increase in demand for gasoline would provide an opportunity or need for suppliers to raise the price – in effect reversing the assumed direction of causality and moving us up the supply curve. There is a large literature on the conditions needed in order to successfully identify demand and supply curves, especially in broad or partly noncompetitive markets (MacKay and Miller 2024).

10.4 Granger Causality

Consider a variable, such as the price of gold, for which we have data over time. We would like to know whether the price of oil helps drive ("cause") the price of gold. This might happen, for instance, if a higher oil price is a signal of uncertainty, which in turn pushes investors to hold gold. Nimble asset managers would be interested in knowing about such a link because they might then respond to a higher price of oil by buying gold.

Following the seminal paper by Clive Granger (1969), we start by modeling the price of gold (y_t) as being largely driven by its prior values. A linear autoregressive form would give

$$y_t = a_0 + \sum_{i=1}^{n} a_i y_{t-i} + \varepsilon_t \tag{10.10}$$

where we allow n lagged values of the price to influence the current price, and ε_t is a random error with zero mean.

Now expand the model in Eqn. (10.10) to include lagged values of the price of oil (x_t), as follows:

$$y_t = a_0 + \sum_{i=1}^{n} a_i y_{t-i} + \sum_{i=1}^{n} b_i x_{t-1} + \varepsilon_t \tag{10.11}$$

If at least some of the x_{t-1} terms are statistically significantly different from zero, then this equation will fit the data better than Eqn. (10.10), and we reject the null hypothesis that the price of oil does not "Granger-cause" the price of gold.

The intuition here is that if past values of x contain information that improves the prediction of y, then we may think of x as in some sense "causing" y. The use of lagged values is important here because of the assumption that "cause cannot come after an effect."

This model picks up the effect of precedence, but this is not necessarily the same as causality in a meaningful sense. Some writers prefer to think of it as "predictive causality" when lagged values of x help predict y; more commonly, we say that in this case, x "Granger causes" y.

We are typically interested not only in whether x Granger causes y but also whether y Granger causes x. If both are present, we have bidirectional Granger causality. In the case of two variables, we have a two-dimensional vector autoregressive (VAR) model (Enders 2010), which would look like this:

$$y_t = a_0 + \sum_{i=1}^{n} a_i y_{t-i} + \sum_{i=1}^{n} b_i x_{t-i} + \varepsilon_t \tag{10.12a}$$

$$x_t = \tilde{a}_0 + \sum_{i=1}^{n} \tilde{a}_i y_{t-i} + \sum_{i=1}^{n} \tilde{b}_i x_{t-i} + \varepsilon_t \tag{10.12b}$$

It is possible to add additional variables, but the models and testing quickly become more complicated. There are various types of VAR: A recursive VAR allows for some concurrent variables on the right-hand side, which can pick up the effect of shocks, for instance. A structural VAR includes restrictions on the parameters of the model – for instance, we allow rainfall to affect the wheat crop, but not the reverse – which can help sharpen our estimates of Granger causality and may sometimes be essential for identifiability.

Several steps are needed in order to test for Granger causality, even once we have identified the variables of interest and assembled the time series. We illustrate the process using daily data on the US dollar prices of gold and oil, from 13 January 2015 through 1 November 2024. Days for which we have data on both prices are counted consecutively, giving us 2,441 observations.

Step 1: Graph. It is usually helpful to graph time series before working with them. Figure 10.3 shows the prices of gold (top panel) and oil (bottom panel). The price of gold appears to be trending upward over time – perhaps not surprisingly, because no adjustment has been made for inflation here – while any trend in the price of oil is harder to discern.

FIGURE 10.3
Price of gold (USD/oz) and oil (USD/barrel), daily, 13 January 2015 through 1 November 2024.

TABLE 10.4

Results of Dickey-Fuller Tests for Stationarity

Variable	Test Statistic (φ_2)	Approx p-value	Stationary?
Price of gold	1.243	0.996	No
Change in price of gold	–38.261	<0.001	Yes
Ln(price of gold)	0.620	0.998	No
Change in Ln(price of gold)	–38.052	<0.001	Yes
Price of oil	–2.486	0.119	No
Change in price of oil	–40.116	<0.001	Yes
Ln(price of oil)	–4.860	<0.001	Yes
Change in Ln(price of oil)	–45.826	<0.001	Yes

Step 2: Stationarity. We now need to check that the series are stationary because otherwise we again have the problem of spurious correlation. This may be done using a test such as the Augmented Dickey-Fuller test (as discussed above) or the Andrews-Ploberger test (if we think there are structural breaks in the time series). The results in Table 10.4 show that the price series are not stationary, but their differences are. The differences in logs, which may be interpreted as growth rates, are also stationary.

Step 3: Lag Length. In our example, we used lags of one and two periods, but this is somewhat arbitrary. Frequently, researchers first estimate the VAR using different lags and choose the model (i.e., the number of lags) that strikes a balance between having a close fit and having a parsimonious number of variables. In our case, the Akaike Information Criterion (AIC) suggests four lags, while the Schwartz/Bayesian Information Criterion (BIC) indicates no lags. For now, we stay with two lags, in part to illustrate the procedures that follow. Ozcicek and McMillin (1999) have an extended discussion of the issue, including allowing for asymmetric lags.

Step 4: Estimate the VAR Model. The results, for our example, are set out in Table 10.5, using differences in logs (i.e., percentage changes). In the top panel, we see that the price of oil Granger causes the price of gold: A higher price of oil is associated with a higher price of gold in the next period and a lower price in the period (i.e., day) after that: In both of these cases, the p-value is less than 0.05. It seems that the price of gold Granger causes the price of oil, with a lag of two working days.

Step 5: Test for Granger Causality. For this, we want to know whether the inclusion of lagged values of the other variable (e.g., the price of oil, in the equation where the price of gold is the outcome variable) adds to the fit of the equation. The results are summarized in

TABLE 10.5

Vector Autoregression Model

		Coefficient	p-Value
D_lpgold equation	lpgold(–1)	–0.00009	0.996
	lpgold(–2)	–0.02219	0.272
	lpoil(–1)	**0.01571**	**0.004**
	lpoil(–2)	**–0.01276**	**0.020**
	Intercept	0.00033	0.068
D_lpoil equation	lpgold(–1)	0.10254	0.171
	lpgold(–2)	**–0.17885**	**0.017**
	lpoil(–1)	**–0.03605**	**0.075**
	lpoil(–2)	0.00537	0.791
	Intercept	0.00020	0.770

TABLE 10.6

Tests of Granger Causality

Equation	Potential Granger Cause	p-value, Chi-Square Test
D.lpgold equation	D.lpoil	0.001
D.lpoil equation	D.lpgold	0.022

Table 10.6. In both cases, we reject the null hypothesis of no causality (because the p-values of the Chi-Square statistics are small), which is evidence of bidirectional Granger causality.

Step 6: Graph Impulse Response Functions. This is an optional, but useful, last step. Given the dynamic nature of a VAR model, a change imposed on one variable will have effects over time, both on its own values, and on the values of the other variables in the model, and these in turn may feed back to affect future values of the original variable. These effects may be shown by graphing an impulse response function (IRF). The top panels of Figure 10.4 show the effect of a unit change in the price of gold on the price of oil, with the left panel showing the response in each period and the right panel showing the cumulative response. The gray area shows the 95% confidence zone. A higher price of oil Granger causes a rise in the price of gold one period (day) later, followed by an offsetting drop the next day. The cumulative effect is small. There is a similar effect from a change in the price of gold on the price of oil, with only short-run effects, and no long-run cumulative change.

It might have been more appropriate to start rather than finish with Figure 10.5! It is a time-series path diagram, which includes a time dimension. The black lines show the directions in which Granger causality flowed

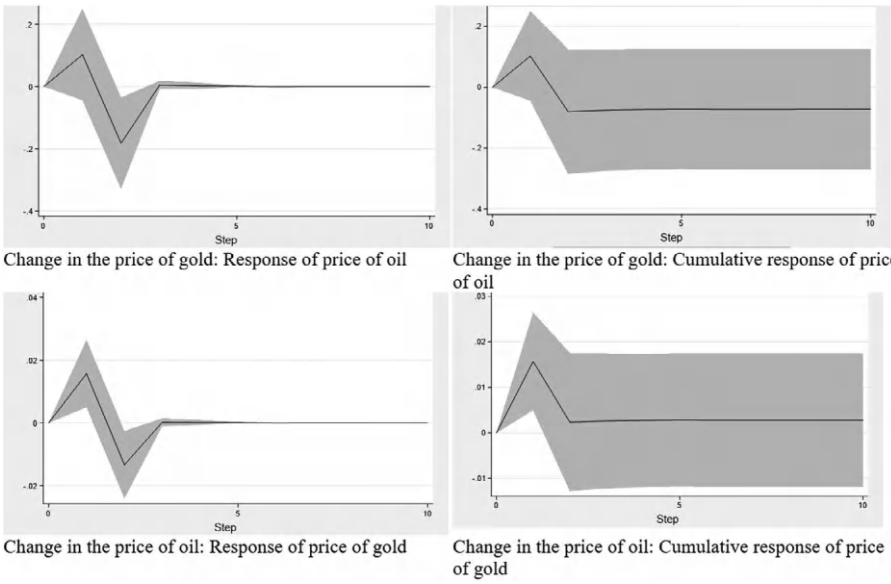


FIGURE 10.4
Impulse response functions for VAR model of prices of gold and oil. (*Note:* Line shows impulse-response function; shaded area show 95% confidence interval.)

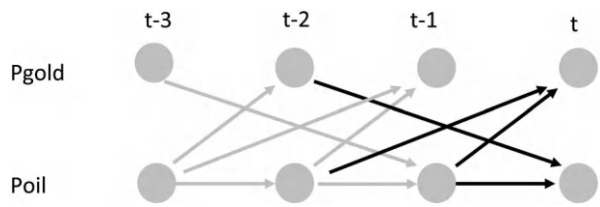


FIGURE 10.5
Time Series directed acyclic graph. (The most recent relationships between nodes and their parents are shown in black arrows; the gray arrows mirror these relationships in earlier periods.)

in our model, and the gray lines repeat the pattern for earlier periods. There are more possible edges (i.e., arrows), but some have been ruled out in the process of estimating the VAR.

10.5 Problems with Granger Causality

The tests for Granger causality are only compelling if several underlying conditions hold. The series under consideration needs to be continuous, so the technique does not apply to, for instance, binary variables. The relationship

between a variable and its lagged values must be linear. The time periods need to be discrete; the number of lags should ideally be known; the series must be stationary; there should be no measurement errors; and the VAR should be complete, in the sense of including all relevant variables.

These conditions are stringent, and so it is no surprise that they rarely, if ever, hold. This explains the assertion by Shojaie and Fox (2022), in their review of Granger causality, that "while limited and not generally informative about causal effects, the notion of Granger causality can lead to useful insights about interactions among random variables observed over time" (p. 5).

10.6 Business and Finance Applications

While there have been many applications of such techniques in macroeconomics, for instance, to try to determine whether consumption Granger causes GDP, or the reverse (e.g., Wen 2007), applications to business are less common, but by no means unheard of. For instance, Banerjee and Siddhanta (2015) try to tease out the "causal" relations between spending on marketing and profit for firms in the personal care industry in India, using quarterly data from 2001:Q2 through 2011:Q1. In plain English, they want to know whether spending on marketing works to raise profit. While it might seem self-evident that causality runs from marketing expenditure to profit, it is also possible that higher profit enables more spending on marketing, in which case (Granger) causality might flow in the other direction.

Banerjee and Siddhanta first check for stationarity and find that the two series – ln(profit) and ln(spending on marketing) – are nonstationary, but the differenced series are stationary. They then use the AIC criterion to determine the best lag length for the VAR, which they find to be nine quarters. This is relevant because marketing spending may have a lagged effect on profits – indeed, spending more on marketing now may reduce profits in the short-run, but boost them in a year or two.

Their next step is to check whether the two series are co-integrated, which will be the case if the residuals of a levels regression that includes the two (and lags) are stationary. It turns out that they are, so the errors from the cointegrating equation need to be included in the VAR, which is now, technically, a Vector Error-Correction (VEC) model. Their final step is to test for Granger causality. They find that they cannot reject the null hypothesis that profits do not drive marketing spending, but they do reject the null hypothesis that marketing spending does not drive profit. In other words, marketing spending Granger causes profit (with a lag of one-and-a-half to two years), but not vice versa.

Another example is to understand the advertising patterns of competitive firms. For example, in a consulting engagement for the telecom industry

using weekly advertising spend data, we have studied whether a corporation's advertising spend is followed by its major competitor's or vice versa. If it is the former, for example, such insight enables the corporation to adjust or "optimize" advertising spend more holistically by taking into account its competitor's reaction.

10.7 Media Mix Modeling

More generally, time-series data are widely used in MMM. The idea is to quantify the impact of measures such as marketing spend for different media on performance indicators such as sales or profits. The standard technique is regression, which is "used to infer causation from observational data" (Sun et al. 2017), although as we have seen, regression can only measure association, not causality.

Media mix models typically use weekly or monthly data over a period that rarely exceeds five years; older data may reflect a different market structure. This makes for small datasets unless there is regional information, in which case the analysis can be pursued at a subnational level, and benefits from a larger sample, which in turn may give greater precision. Even so, such models are often overfit, as analysts are under pressure to find strong fits ("hunting for R") and high statistical significance ("p-hacking").

A common weakness of media mix models is that they lack information on campaign spending by their competitors. Haughton et al. (2014) show how it is sometimes possible to infer information about whether competitors are spending on a marketing campaign by applying a Hidden Markov Chain. In their study of marketing a drug to physicians, Haughton et al. (2014) first use a regression approach and then formulate a directed acyclic graph (DAG), as discussed in Chapter 5. But they also note that their data do not incorporate autocorrelation effects, which are at the heart of trying to identify Granger causality.

References

Banerjee, Neelotpaul, and Somroop Siddhanta. 2015. "An Empirical Investigation on the Impact of Marketing Communication Expenditure on Firms' Profitability: Evidence from India". *Global Business Review*, 16(4): 609–622.

Enders, W. 2010. *Applied Econometric Time Series*, 3rd edition. Hoboken, NJ: John Wiley & Sons.

Gujarati, Damodar N., and Dawn Porter. 2009. *Basic Econometrics*, 5th edition. Boston, MA: McGraw-Hill Irwin.

Haughton, Dominique, Guangying Hua, Danny Jin, John Lin, Qizhi Wei, and Changan Zhang. 2014. "Imputing Unknown Competitor Marketing Activity with a Hidden Markov Chain". *Journal of Direct, Data and Digital Marketing Practice*, 15: 276–287.

Johnston, John. 1972. *Econometric Methods*, 2nd edition. New York, NY: McGraw-Hill.

MacKay, Alexander, and Nathan Miller. 2024. "Estimating Models of Supply and Demand: Instruments and Covariance Restrictions". Working Paper 19-051, Harvard Business School.

Ozcicek, Omeer, and W. Douglas McMillin. 1999. "Lag Length Selection in Vector Autoregressive Models: Symmetric and Asymmetric Lags". *Applied Economics*, 31: 517–524.

Palshikar, Girish Keshav, Manoj Ape, Sushodhan Vaishampayan, and Akshada Shinde. 2023. Causality in Time-Series: A Short Review. In Satyasai Jagannath Nanda, and Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, India: SCRS, 723–748.

Runge, Jakob, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. 2023. "Causal Inference for Time Series". *Nature Reviews Earth & Environment*, 4: 487–505.

Shojaie, Ali, and Emily B. Fox. 2022. "Granger Causality: A Review and Recent Advances". *Annu Rev Stat Appl*, 9(1): 289–319.

Sun, Yunting, Yueqing Wang, Yuzue Jin, David Chan, and Jim Koehler. 2017. *Geo-Level Bayesian Hierarchical Media Mix Modeling*. Google Inc, Mountain View, CA.

Wen, Yi. 2007. "Granger Causality and Equilibrium Business Cycle Theory". *Review, Federal Reserve Bank of St. Louis*, 89(3): 195–205.

11

Structural Equation Models

11.1 Introduction

Your marketing staff tell you that most of the customers who buy upscale handbags are "fashion conscious" and "materialistic." This certainly sounds plausible, but you would like to test whether they are right and quantify the strength of such effects. But before you can do that, you need to define what is meant by "fashion conscious" and "materialistic," because these are latent variables that are not directly measurable.

Structural Equation Modeling (SEM) provides a way to address problems such as this. It typically consists of two parts – a measurement ("outer") model that allows one to construct the latent variables, and a causal structural ("inner") model that defines and quantifies the links among the latent variables, and between the latent variables and an outcome of interest (such as sales of handbags).

At its most fundamental, SEM is a way of thinking that forces us to be rigorous about how variables relate to one another and the direction of causality. It also relies on a system of notation in the form of path diagrams, which we lay out in this chapter. And it brings structure to our models in ways that we explain more completely below. Sometimes it is useful to think of SEM models as a subset of Structural Causal Models (SCMs), which use Directed Acyclic Graphs (DAGs) to help think through the paths of causality, as discussed more fully in Chapter 5.

The SEM approach is widely used in academic studies of marketing – there is a good review of practice by Hair et al. (2012) – where it has become "quasi-standard" (p. 414). Hair et al. (2021) refer to it as a second-generation technique, in contrast to first-generation techniques such as multiple or logistic regression.

SEM is also popular in operations research (see Sarstedt et al. 2014) and in the study of issues related to managing human resources (such as job satisfaction) and is one of the key methodological tools employed by social scientists, especially in sociology and psychology. Its roots are to be found mainly in the work of Wold (1982) and his student Jöreskog (1978). Until recently, it has been less widely used in the business world, where descriptive and

DOI: 10.1201/9780429172588-11

exploratory methods are far more common and where the rigor required in constructing a good SEM can be intimidating.

A few years ago, this prompted someone to ask, in a Web Chat Room: "Does anyone ever use structural equation modeling in the business world?" In response, Joseph Lurchman of Fors Marsh Group wrote:

> The company I work for has many clients that want to understand the problems they're having as much as they're interested in getting an answer to them. Data mining algorithms … provide good answers, but are borderline useless for presenting to clients who aren't interested in the technical aspects of their problems. By contrast, we have used SEM for many clients interested in getting more invested in understanding their customers/clients/markets.
>
> **(Lurchman 2013).**

11.2 Basic Concepts

It is helpful to think of Structural Equation Models (SEMs) as coming in two varieties: Confirmatory and exploratory. Confirmatory SEM, often referred to as correlation-based SEM (CB-SEM), sets up a strong model structure and tests it using data. This is the more traditional form of SEM, and we consider it first. More recently, a form of exploratory SEM, partial least squares SEM (PLS-SEM), has become popular; the model here has less structure, the assumptions are less restrictive, and while it does not allow for model testing in the same way, it is relatively easier to apply. There has been an explosion of interest in SEMs over the past two decades, helped by the development of software that makes the techniques accessible to most researchers – short comments on the most popular programs are given in Appendix 11.1 – and by persuasive books (e.g., Bollen 1989) and hundreds of academic articles (e.g., Buckler 2009, Bollen and Pearl 2013).

We now develop the key ideas of SEMs. In Chapter 2, we discussed linear regression, where the variable y is associated with variables X_1 and X_2 as follows:

$$y_i = \alpha_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{11.1}$$

With observations on the y and X variables we can estimate the coefficients α_0, β_1, and β_2 (giving the estimates $\hat{\alpha}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$). Here ε_i is an unobserved error term for observation i, which stands in for measurement error and any variables that are not included – perhaps because they cannot be measured – but may matter.

We usually assume that y is the outcome ("dependent" or "target") variable, and the X_k are "independent" variables (also known as covariates or features) that influence y. When we do this, we are assuming a direction of causality, and if our estimated coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are statistically significantly different from zero, then our causal assumption is not contradicted. On the other hand, if, say, $\hat{\beta}_1$ were statistically indistinguishable from zero, this would be evidence against (part of) our causal assumptions. The estimate of this regression model cannot establish causality, but it can help contradict our initially imposed structure, and this brings an element of falsifiability into the analysis (Popper 1959).

A nice way to show this is with a *path diagram*, which in the regression case looks like the one in Figure 11.1.

The convention here is that observed ("manifest") variables are placed in boxes, while unobserved ("latent") variables, including in this case the error term, are shown in circles or ovals. The arrows show the assumed direction of causality and represent much of the structure that we impose on the data. The numbers shown next to the arrows (called "path coefficients") measure the linear association between the variables. Thus, the model in Figure 11.1 may be written as

$$y_i = \alpha_0 + 0.04X_{1i} - 0.61X_{2i} + \varepsilon_i.$$

Path diagrams impose discipline and force us to think through the pattern of causality and the structure of relationships between variables. A *strong causal assumption* would, for instance, consist of setting a path coefficient to zero, which is tantamount to saying that the variable has no place in the model. A *weak causal assumption* might, for instance, force a coefficient to be positive.

To see the importance of structure, imagine a model where we have information on three variables (call them Y_1, Y_2, and Y_3), but no prior idea of how they are related. Our path diagram is then the one shown in Figure 11.2.

Here we have "error" terms associated with each variable. We have also added curved lines; when they are present, they indicate that we are allowing

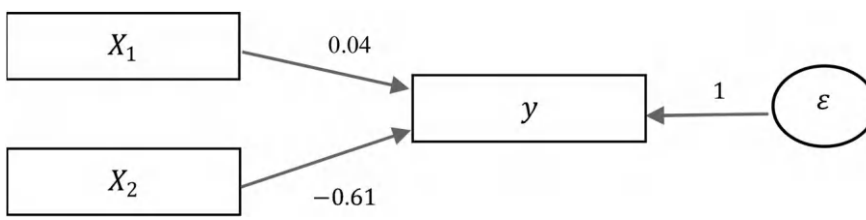


FIGURE 11.1
Path diagram representation of a regression model.

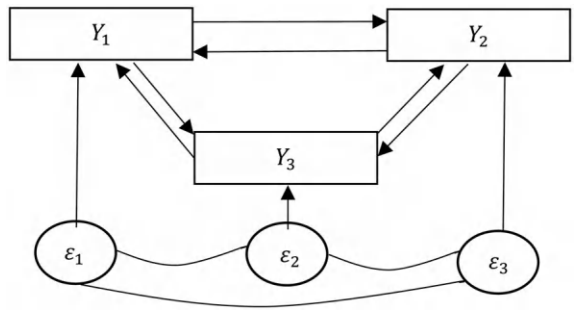


FIGURE 11.2
Path diagram for three variables and no structure.

correlations among the variables. The model in Figure 11.2 has no structure and may be written as:

$$Y_1 = \alpha_1 + \beta_{12}Y_2 + \beta_{13}Y_3 + \varepsilon_1$$
$$Y_2 = \alpha_2 + \beta_{21}Y_1 + \beta_{23}Y_3 + \varepsilon_2$$
$$Y_3 = \alpha_3 + \beta_{31}Y_1 + \beta_{32}Y_2 + \varepsilon_3$$
$$\text{cov}\left(\varepsilon_i\varepsilon_j\right) \neq 0, \forall\ i, j.$$

This model is not identifiable – we cannot estimate the coefficients uniquely – and so we need to impose some structure in order to progress. Often, we do this without thinking it through, but SEM forces us to be more rigorous.

Figure 11.3 shows a path diagram for a more-structured model, representing a causal chain.

Perhaps Y_1 measures the quality of a product, Y_2 measures customer satisfaction with the product, and Y_3 gives the value of repeat sales. Instead of using the path diagram, we might equally well write the model as:

$$Y_1 = \alpha_1 + \varepsilon_1$$
$$Y_2 = \alpha_2 + \beta_{21}Y_1 + \varepsilon_2$$

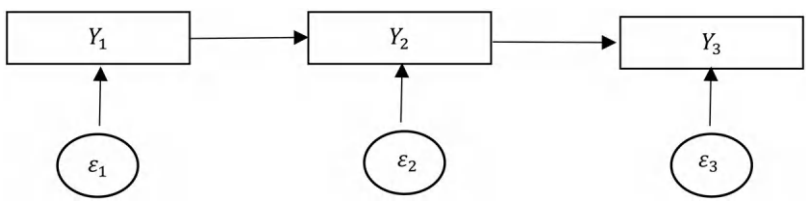


FIGURE 11.3
Path diagram of a causal chain.

$$Y_3 = \alpha_3 + \beta_{32} Y_2 + \varepsilon_3$$
$$\text{cov}\left(\varepsilon_i \varepsilon_j\right) = 0, \forall\ i, j.$$

In this case we are making strong causal assumptions that $\beta_{12} = \beta_{13} = \beta_{23} = \beta_{31} = 0$, and the error covariances are zero. Our weak causal assumptions are that $\beta_{21} \neq 0$ and $\beta_{32} \neq 0$, and these are testable.

11.2.1 Latent Variables

SEM becomes more interesting when we begin to model latent variables. Suppose we believe that our employees will be more productive if they have greater job satisfaction (denoted by Z_1) and stronger work incentives (Z_2). The problem here is that none of these broad variables – job satisfaction, work incentives, or job performance – can be measured simply and directly. They are thus latent variables. But they are, in turn, related to underlying measurable ("manifest") variables.

For instance, to measure job satisfaction, imagine that we have surveyed our staff and asked a set of questions like these:

On a scale of 1 (not at all) to 5 (very much):

Is your work fulfilling?	1	2	3	4	5
Do you enjoy your work?	1	2	3	4	5
Is it a pleasure to come to work every morning?	1	2	3	4	5

These three questions, with responses on a five-point Likert scale, may be thought of as measuring different dimensions of job satisfaction. Denote the responses to these questions by X_1, X_2, and X_3. Then we may graph this as:

Note the direction of the causal paths here: Job satisfaction leads to the responses to the questions. It is not the answers to the questions that cause job satisfaction; instead, they measure it. That is why the model in Figure 11.4

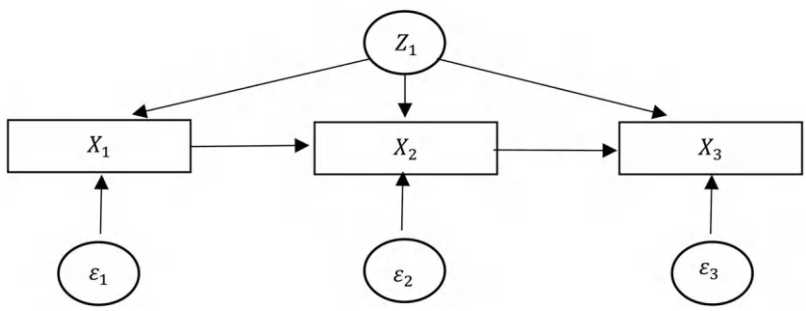


FIGURE 11.4
Path diagram with latent variables.

is a *measurement model*, and it is an integral part of almost every SEM. In the terminology of SEMs, Z_1 is a *reflective* latent variable. Formally,

$$X_i = \alpha_{0i} + \alpha_{1i}\ Z_1 + \varepsilon_i,\ i = 1,2,3,\ E(\varepsilon_i) = 0,\ cov(Z_1\varepsilon_i) = 0,$$

and the challenge is to find the α's that link the latent Z_1 to the observable $X_i's$. This is discussed in more detail below.

Schematically, our SEM may be represented as in Figure 11.5, simplified only inasmuch as it excludes error terms and correlations among variables. The structural (inner) model is shown in the middle and consists of hypothesized causal relations among the latent variables. Here Z_1 and Z_2 are exogenous latent variables, but Z_3 is an endogenous latent variable because it is determined within the structural model. Note too that the model is recursive, flowing in a single direction; there must be a (hypothesized) causal chain, without any feedback loops. Broadly, we have for the structural model:

$$Z_i = \beta_{i0} + \sum_j \beta_{ij} Z_j + v_i,\ i = 1,2,3,\ E(v_i) = 0.$$

The outer model in Figure 11.5 is concerned with the determinants of the latent variables. Here Z_1 (job satisfaction) is a reflective measurement (or Mode A) model. For purposes of illustration, we have defined Z_2 (work incentives)

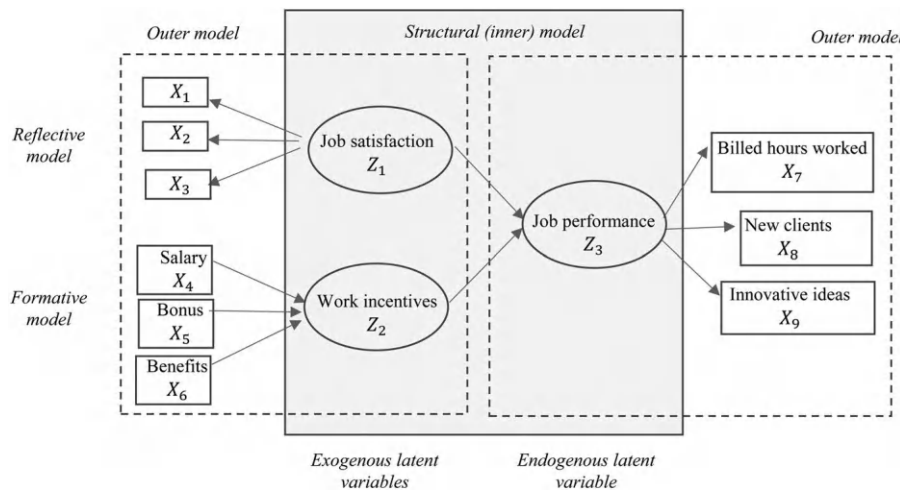


FIGURE 11.5
Typical basic structural equation model for job performance. (*Note:* The model for "job satisfaction" is the same as in Figure 11.4. For simplicity, the error terms have been omitted here. Created by authors, inspired by Figure 1 in Sarstedt et al. 2014.)

to be a *formative* latent variable, giving a Mode B model. Here the arrows go from the manifest variables (X_4, X_5, X_6) to Z_2. The idea here is that a higher salary is not a reflection of work incentives; it *is* a work incentive, and indeed the work incentives variable (Z_2) is formed by its components. Formally:

$$Z_2 = \beta_{20} + \beta_{24} X_4 + \beta_{25} X_5 + \beta_{26} X_6 + \delta,\ E(\delta) = 0,\ cov(\delta\ X_i) = 0.$$

Here too there is the challenge of finding the appropriate coefficients (the β'*s*), but this is not so much a measurement model as a definitional model. Usually, CB-SEM models cannot easily accommodate formative latent variables, but PLS-SEM can.

To give a better feel for how the measurement models in an SEM may be estimated – for instance, to create the weights that we need to put on the paths from X_1, X_2, and X_3 to Z_1 – it is helpful to take a short detour to discuss factor analysis. While the algorithms that estimate SEMs will generate these path coefficients, the procedure is very close to that used by factor analysis, and an understanding of how factor analysis works helps us to appreciate the mechanisms behind SEMs.

11.3 A Short Detour: Factor Analysis

Consider a teacher who, at the end of every semester, asks every student to complete a course evaluation questionnaire covering perhaps 20 questions related to their experience with the course and teacher. Typical questions include the following.

On a scale of 1 (not at all) to 5 (very much):

The teacher:					
Was well organized	1	2	3	4	5
Arrived in class on time	1	2	3	4	5
Explained things clearly	1	2	3	4	5
Showed enthusiasm for the subject	1	2	3	4	5
Was knowledgeable about the subject	1	2	3	4	5

Some of the questions are largely redundant – the answers are essentially the same as the answers to other questions – while a few of the questions are informative on their own. Therefore, it might be helpful to reduce the many answers to just a few *factors*.

A popular data reduction technique is *exploratory factor analysis*, which seeks to create a small number of factors – that is, unobserved latent

TABLE 11.1

Factor Loadings for Student Course Evaluation Example

Question	Factor 1 Loadings	Factor 2 Loadings
Organized	**0.82**	0.22
Arrives on time	**0.76**	0.03
Explained things clearly	0.13	**0.92**
Enthusiastic	0.09	**0.68**
Knowledgeable	0.14	**0.97**
Possible label	"Organized"	"Competence"

variables – that still contain almost all the information but are perhaps easier to comprehend. There are a number of techniques for creating factors – any good textbook will give the details – but the essential idea is that highly correlated variables should be grouped together. The result is a set of factors that are most commonly (in the principal component approach to factor extraction) defined as weighted averages of the underlying manifest (i.e., observed) variables; the weights of each variable on the factors are the *factor loadings*, equivalent to path coefficients in the path diagram in a reflexive model.

For example, exploration of the course evaluation data may give two factors, with the loadings shown in Table 11.1. Factor 1 mainly reflects the two variables that show that the teacher is well organized, so we might label it "organized." And the weights on factor 2 are related to the teacher's mastery of the material, so we might label it "competence."

Exploratory factor analysis is a data compression exercise, but it lacks all structure, so there are no *a priori* restrictions about which factors may represent which indicators. The method of principal components is by far the most common extraction method for exploratory factor analysis, and a rotation of factors is often performed, most commonly a Varimax rotation that ensures as much as possible that indicators "load" onto (have a high correlation with) at most one latent factor, to facilitate the interpretation and naming of the factors.

Exploratory factor analysis often has its uses, but it does not allow one to test anything, only to describe and summarize. A related technique, *confirmatory factor analysis*, is closer in spirit to CB-SEM. Instead of letting the data determine the number of factors and the variables that load on the factors, the analyst specifies these and then uses the data to "confirm" whether these are appropriate choices.

For instance, in our example, we might believe that enthusiasm should load on factor 1 but not factor 2. We can try this and test whether this improves or worsens the ability of the factors to summarize the data. CB-SEM models are confirmatory. The measurement model components may be thought of as reflecting a confirmatory factor analysis.

11.4 Back to SEMs

As noted above, most SEM models have a measurement model and a structural model. The latter sets out, and measures, the causal relations among the latent variables. To illustrate, let us go back to our first example, where we postulate that "fashion-conscious" and "materialistic" consumers will have a greater propensity to buy high-end handbags. Following MacLean and Gray (1998), there are three latent variables here, which are related as shown in Figure 11.6.

The "structural" part of SEM largely resides in the causal relations we attach to these latent variables. But each latent variable needs its measurement model: Perhaps "fashion-conscious" is measured based on the results of an attitude survey that includes questions such as:

	Indicate how much you agree with these statements (1 = strongly disagree, 5 = strongly agree)					
X_1	Fashion is an important means of self-expression	1	2	3	4	5
X_2	I like high-class items	1	2	3	4	5
X_3	I'm usually the first among my friends to learn about a new brand or product	1	2	3	4	5

Meanwhile, "materialistic" could be gauged by questions such as these:

	Indicate how much you agree with these statements (1 = strongly disagree, 5 = strongly agree)					
X_4	I am extravagant about my clothes and food	1	2	3	4	5
X_5	I'm the type to buy something I want immediately even if I have to borrow money	1	2	3	4	5

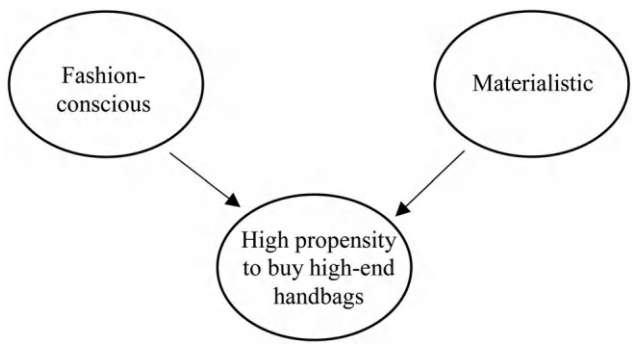


FIGURE 11.6
Relationships between latent variables in the fashion consumer example. (Graph by authors, based on study by MacLean and Gray 1998.)

The propensity to buy a handbag might be measured by a composite of questions like these:

	Indicate how much you agree with these statements (1 = strongly disagree, 5 = strongly agree)					
Y_1	The chances of me buying something today are high	1	2	3	4	5
Y_2	I'm not really interested in buying anything today	1	2	3	4	5

Then our model becomes as shown in Figure 11.7.

By the standards of SEMs, this is still a very straightforward model – MacLean and Gray develop it somewhat further (see their Figure 7) – but it captures the spirit of a typical CB-SEM exercise.

Given the model structure, the links between the variables can be quantified (if the model has sufficient structure to be identified), and one usually can test whether the *model* performs well, relative to a model with more or fewer causal relations. The traditional CB-SEM allows for hypothesis testing – that is, it is confirmatory – because it assumes that the variables are multivariate normal (Gaussian), and the relations between the variables are taken to be linear. The CB-SEM can usually be constructed using maximum likelihood methods just with information on the covariance matrix of the indicator variables, which is why this method is often referred to as the "covariance approach" to SEM. We set out a fuller example in the next

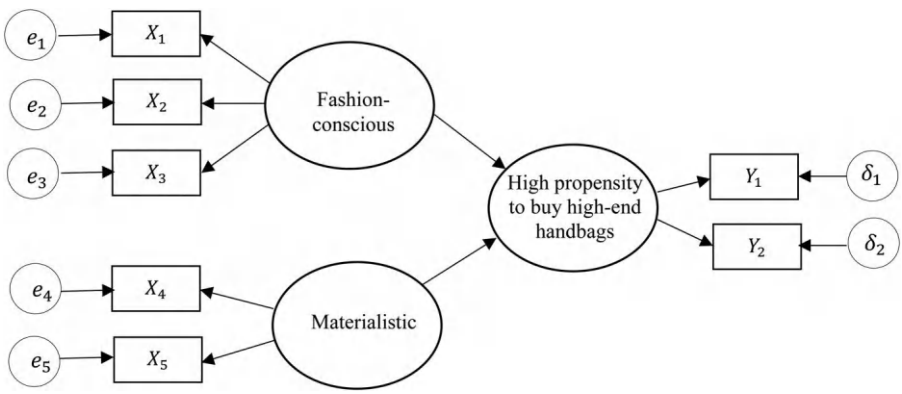


FIGURE 11.7
Basic CB-SEM model of fashion consumer example. (*Note:* X_i and Y_i are manifest (measured, observable) variables; "Fashion-conscious" and "Materialistic" are exogenous latent variables, and "High propensity to buy high-end handbags" is an endogenous latent variable. The e_i and δ_i are normally distributed errors.)

section. Jannoo et al. (2014) argue, based on Monte Carlo simulations, that CB-SEM can perform well using non-normal data, provided the sample is large enough.

We would like to emphasize that for all CB-SEMs, a model is needed. It could be specified *a priori*, based on theory, or developed based on the preliminary construction of a DAG from data – see Chapter 5 for details – or by using a combination of these two approaches. But in a process that parallels factor analysis, it is also possible to estimate a more "exploratory" SEM, which is generally done using the partial least squares algorithm (see, e.g., Tenenhaus et al. 2005). We return to this issue below, after first giving another illustration of a CB-SEM, also developed in a marketing context.

11.5 The Covariance Approach: An Example

In order to illustrate more completely the covariance approach to the estimation of SEMs, we draw on an example by Vij and Farooq (2014) in which the authors investigate the impact of Knowledge Orientation (KSO) on performance for different manufacturing and service organizations in the National Capital Region, India. This example is typical of a very large number of such models used in several areas of business studies, notably marketing and information systems.

The data were obtained from 240 questionnaires returned by managerial-level employees. Variables (on a scale of 1–5) are listed in Table 11.2. The authors identified and validated via a confirmatory factor analysis four factors to represent KSO: Idea Sharing Propensity (ISP), Good Organizational Climate (GOC), Top Management Support (TMS), and Knowledge Sharing Culture (KSC). In the same fashion, three factors found to represent performance are Satisfaction relative to Major Competitor (PER_SAT), Profitability relative to Major Competitor (PER_PRO), and Innovativeness Relative to Major Competitor (PER_INN). The authors propose and then validate an SEM, shown for large organizations (with more than 250 employees) in Figure 11.8; as usual, the manifest variables are shown in boxes and the latent variables (including the e_i error terms) in ovals. The numbers show the path coefficients, which measure the linear association between the variables. The model extracts a single constructed KSO from its four factors and a single constructed Performance Relative to Major Competitor (PER_COM) from its three factors and then examines the relationship between KSO and PER_COM.

This study finds that KSO has a positive impact on performance and that this impact depends on the size of the organization. Indeed, the significant

TABLE 11.2

Variables for the KSO Example

Code	Questionnaire Statement (1. Strongly Disagree – 5. Strongly Agree)
S1	A climate of openness and trust permeates my organization.
S2	In our organization, everyone speaks up if they have an opinion or idea to offer.
S3	We do not share ideas with other people of similar interest, especially when they are based in different departments.
S4	Knowledge-sharing behavior is built into the performance appraisal system in my organization.
S5	Our company culture welcomes debates and stimulates discussions.
S6	In our organization, we are rewarded for sharing knowledge with our colleagues.
S7	There is no restriction on employees if they want to talk to anyone in the organization, including top management.
S8	In my organization, relatively more-committed employees are more willing to share their learning and experiences with others.
S9	Top managers provide most of the necessary help and resources to enable employees to share knowledge.
S10	My organization's culture encourages and facilitates knowledge sharing.
S11	Top managers do not support and encourage employees to share their knowledge with colleagues.
Code	Compared to the major competitor in your industry, in the last three years, how has your business performed on the following parameters?
CC1	Sales Growth
CC2	Return on Investment
CC3	Market Share
CC4	Service Quality
CC5	Customer Satisfaction
CC6	Employee Satisfaction
CC7	Employee Turnover
CC8	Product Innovation
CC9	Process Innovation
CC10	Product Quality

Source: Vij and Farooq 2014.

standardized coefficient for KSO to PER_COM is 0.742 for large organizations (more than 250 employees). For smaller organizations, this coefficient is 0.678. In such cases, the size of the organization is said to be a moderator for the relationship between KSO and PER_COM. In statistical terms, we might say that size interacts with KSO in its relationship with PER_COM. On the other hand, the relationship between KSO and PER_COM does not depend on industry (manufacturing versus services).

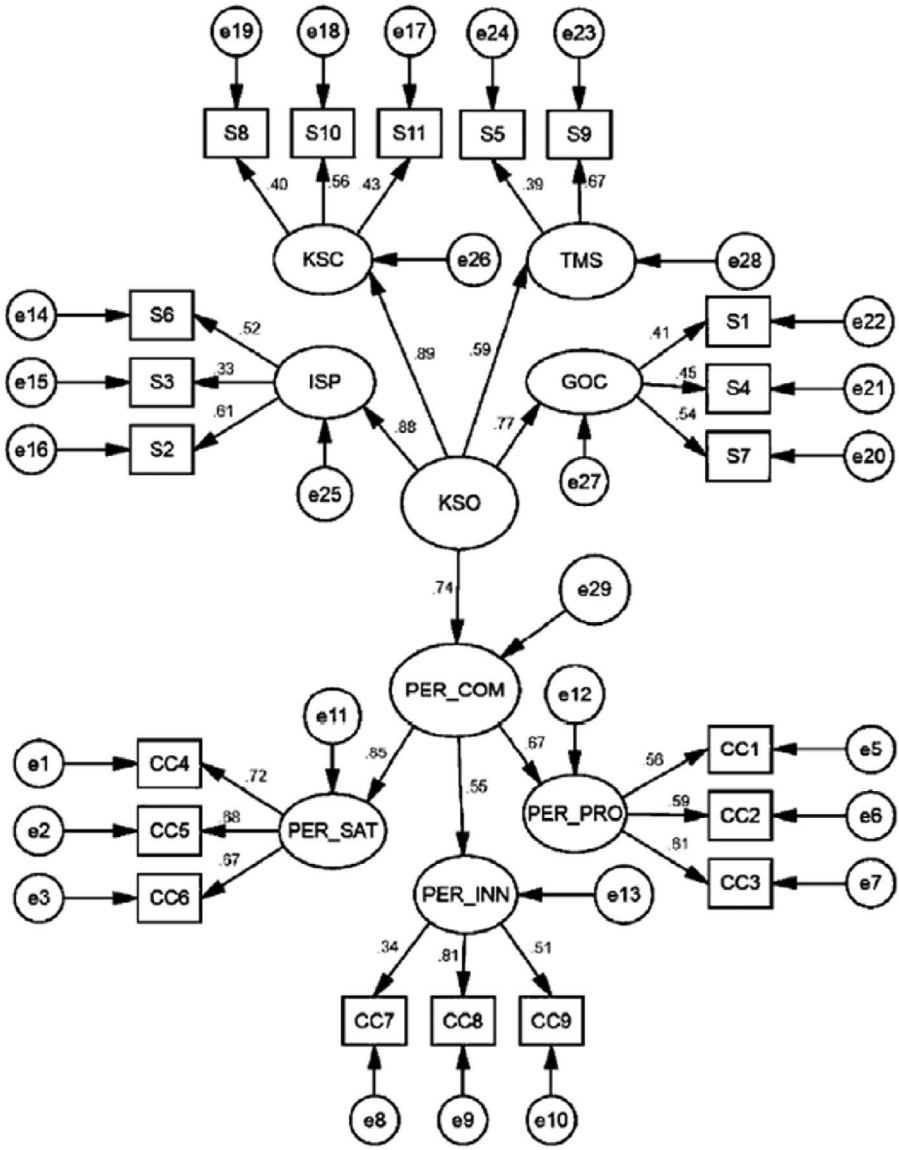


FIGURE 11.8
Covariance-based structural equations model of KSO (knowledge sharing orientation and performance). (From Vij and Farooq 2014, Figure 2.)

11.6 The Partial Least Squares Approach

The traditional approach to SEM is often referred to as covariance-based SEM (CB-SEM) because the model can be optimized and solved based just on the covariances among the manifest variables, provided one is willing to assume that the underlying variables are normally distributed. This in turn allows one to test theories and to use standard tools of statistical inference, since "the objective in using [CB-]-SEM is to determine whether the a priori model is valid, rather than to 'find' a suitable model" (Shah and Goldstein 2005, p. 149). The downside is that normality may be an unrealistic assumption, the system may not solve satisfactorily, and relatively large samples are usually needed.

An alternative approach is to relax the assumption of normality. On the surface, the model looks just the same, but underneath it is quite different. A new solution algorithm is needed because the optimization associated with CB-SEM is no longer possible. In most cases, the relationships in the structural (inner) model are now estimated using ordinary least squares, which is why this approach is called partial least squares SEM (PLS-SEM). This flexibility allows one to estimate more complex models, but it comes at a cost: PLS-SEM models may be used for prediction and exploration, but not for testing whether a theory holds or not, or for judging causal effects.

Hair et al. (2021) argue that the strength of PL-SEM is that its "statistical properties provide very robust model estimations with data that have normal as well as extremely non-normal distributional properties" (p. 14).

PLS-SEM models have become popular in management, especially since about 2000 (see Sarstedt et al. 2014, Figure 2). Table 11.3 shows that PLS-SEM models have, on average, more latent variables and almost twice as many

TABLE 11.3

Comparison of Model Size, CB-SEM versus PLS-SEM, in Operations Research and Market Research

	CB-SEM	PLS-SEM
	Per Model	
Number of indicators (manifest variables)	16.3	29.6
Number of latent variables	4.7	7.9
Number of observations	246	211
Number of parameters	37.5	
Number of relationships in the inner model		10.6

Note: CB-SEM is covariance-based SEM. Results are from Shah and Goldstein (2005) and refer to 75 SEMs in operations research articles published in four top disciplinary journals between 1984 and 2003. PLS-SEM is partial least squares SEM. Results are from Hair et al. (2012), and refer to 311 models in 204 articles appearing in 24 top journals in market research between 1981 and 2010.

measured (manifest) variables as conventional CB-SEM models. For instance, Eberl (2010) uses a PLS-SEM model to untangle the effect of corporate reputation (separated into "competence" and "likeability") on customer satisfaction and hence loyalty.

When would a PLS-SEM be preferred over a CB-SEM? Sarstedt et al. (2014, Table 1) provide a useful summary; briefly, a PLS-SEM may be more useful if:

- The data are non-normal;
- The goal is exploration or prediction rather than model testing;
- There are more indicators (i.e., manifest variables), more latent variables, and greater model complexity;
- Some of the latent variables are formative rather than reflective;
- There are fewer observations.

Where CB-SEM only needs information on covariances to be able to find an optimal solution, PLS-SEM uses a more elaborate algorithm, which can require substantial computing power. Garson (2016) provides a clear description of the typical solution algorithm for PLS-SEM: Standardize all indicator (manifest) variables to mean zero and standard deviation one, note that latent variables are linear combinations of the indicator variables (the outer weights), and constrain the (inner) path coefficients among the latent variables to vary from –1 to +1. Now iterate through the following four steps until not much change occurs in the outer weights:

1. Define latent variables with initially equal weights on indicator variables;
2. Assign initial weights to inner paths so as to maximize the R-squared of the regression determining each endogenous latent variable;
3. Use these inner path weights to compute latent variable scores;
4. Adjust the outer weights connecting the latent variable scores to their indicator variables as follows:
   a. Reflective latent variables: Use the covariance between the estimated latent variable and each indicator;
   b. Formative latent variables: Regress the estimated construct on its indicators and use the coefficients.

While this algorithm will typically yield a model, it is a relatively ad hoc procedure, there is no global optimum, and little is known about the sampling distribution of the weights (i.e., path coefficients) generated (Wikipedia).

Hair et al. (2021) argue that PLS-SEM models are "particularly appealing for research in fields that aim to derive recommendations for practice … [including] recommendations in managerial implication sections in business

research journals." However, PLS-SEM models cannot handle non-recursive models – that is, models that have causal loops or circular relationships – and do not yield a generally accepted measure of model goodness of fit, making them difficult to evaluate.

11.7 Conclusion

As with many other techniques that we address in this book, it is not possible to prove the existence of causality in SEMs, but the data can often indicate that causality is not present. This is true of covariance-based SEMs too, but not of partial-least-squares SEMs, which can be used for exploration and prediction but not for model testing.

SEM is not easy. Hair et al. (2012, Table 5) set out a list of 36 best practices that they recommend researchers use when conducting and reporting SEMs; Shah and Goldstein have a similar list, as do Sarstedt et al. But when one wants insight rather than "black box" numbers, and when the relations between variables are somewhat complex, then the trouble required to develop an SEM may be well worth the effort.

Appendix 11.1: SEM Software

Here is a list of the main software packages that are used for SEM. No one package is best for all applications, and many researchers work with multiple packages.

Package	Comments	Link
LISREL	"Linear structural relations." The original CB-SEM software. Requires facility with matrix manipulation.	https://ssilive.com/
MPlus	Praised for its good support from its developers, Linda and Bengt Menthén.	https://www.statmodel.com/glance.shtml
R	The lavaan ("latent variable analysis") and OpenMx packages are popular. SEMinR is a recent well-documented package.	http://lavaan.ugent.be/ http://www.openmx-square.org/ https://cran.r-project.org/web/packages/seminr/vignettes/SEMinR.html

(Continued)

Package	Comments	Link
SAS: CALIS	The SAS implementation of structural equation modeling. SAS is widely used by business, but it is also expensive.	https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#calis_toc.htm
SmartPLS	Specializes in partial least squares estimation and exploratory SEM.	https://www.smartpls.com/
SPSS: AMOS	Part of IBM's SPSS package. Widely used, and relatively easy to learn.	https://www.ibm.com/us-en/marketplace/structural-equation-modeling-sem
Stata: sem	The sem command is powerful and flexible, and is part of a statistical package widely used by researchers.	https://www.stata.com/features/structural-equation-modeling/ (see Stata 2013).
ΩNYX	A free stand-alone SEM package.	http://onyx.brandmaier.de/

References

Bollen, Kenneth. 1989. *Structural Equations with Latent Variables*. Hoboken, NJ: Wiley.

Bollen, Kenneth, and Judea Pearl. 2013. Eight Myths about Causality and Structural Equation Models. In Stephen Morgan (ed.), *Handbook of Causal Analysis for Social Research*, Dordrecht: Springer, 301–328.

Buckler, Frank. 2009. "Causal Analysis to the Rescue: How to Find Success Factors from Survey Data". *Marketing Research*, Fall: 6–11.

Eberl, Markus. 2010. An Application of PLS in Multi-Group Analysis: The Need for Differentiated Corporate-Level Marketing in the Mobile Communications Industry. In Vincenzo Esposito Vinzi, Wynne Chin, and Jörg Henseler (eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications*, New York: Springer. 487–514.

Garson, G. D. 2016. *Partial Least Squares: Regression and Structural Equation Models*. Asheboro, NC: Statistical Associates Publishers.

Hair, Joseph, Marko Sarstedt, Christian Ringle, and Jeannette Mena. 2012. "An Assessment of the Use of Partial Least Squares Structural Equation Modeling in Marketing Research". *Journal of the Academy of Marketing Science*, 40: 414–433.

Hair, Joseph, Tomas Hult, Christian Ringle, Marko Sarstedt, Nicholas Danks, and Soumya Ray. 2021. *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook*. Cham, Switzerland: Springer.

Jannoo, Z., B. W. Yap, N. Auchoybur, and M. A. Lazun. 2014. "The Effect of Nonnormality on CB-SEM and PLS-SEM Path Estimates". *International Journal of Mathematical and Computational Sciences*, 8(2): 286–291.

Jöreskog, K. G. 1978. "Structural Analysis of Covariance and Correlation Matrices". *Psychometrika*, 43(4): 443–477.

Lurchman, Joseph. 2013. Response on blog. https://www.researchgate.net/profile/Joseph_Luchman [Accessed June 5, 2019.]

MacLean, Scott, and Kevin Gray. 1998. Structural Equation Modelling in Market Research. *Journal of the Australian Market Research Society*. http://www.smallwaters.com/whitepapers/marketing/

Popper, Karl. 1959. *The Logic of Scientific Discovery*. Abingdon-on-Thames, Oxfordshire, England: Routledge.

Sarstedt, Marko, Christian Ringle, Donna Smith, Russell Reams, and Joseph Hair Jr. 2014. "Partial Least Squares Structural Equation Modeling (PLS-SEM): A Useful Tool for Family Business Researchers". *Journal of Family Business Strategy*, 5(1): 105–115.

Shah, Rachna, and Susan Goldstein. 2005. "Use of Structural Equation Modeling in Operations Management Research: Looking Back and Forward". *Journal of Operations Management*, 24: 148–169.

Stata. 2013. *Structural Equation Modeling Reference Manual, Release 13*. College Station, TX: Stata Press.

Tenenhaus, Michel, Vincenzo Esposito Vinzi, Yves-Marie Chatelin, and Carlo Lauro. 2005. "PLS Path Modeling". *Computational Statistics & Data Analysis*, 48: 159–205.

Vij, S., and R. Farooq. 2014. Knowledge Sharing Orientation and Its Relationship with Business Performance: A Structural Equation Modeling Approach. *IUP Journal of Knowledge Management*, 12(3): 17–41.

Wikipedia. Structural Equation Modeling. https://en.wikipedia.org/wiki/Structural_equation_modeling

Wold, H. 1982. Soft Modeling: The Basic Design and Some Extensions. In K. G. Jöreskog and H. Wold (eds.), *Systems under Indirect Observations, Part II*. Amsterdam: North Holland, 1–54.

12

Discussion and Summary

12.1 General Discussion

This final chapter provides some discussion and a summary of our book. This book covers a variety of techniques for analyzing cause-and-effect relationships at the population or group level. We also provide coverage for uplift analytics or individual/heterogeneous treatment effects from experimental design to model development, treatment optimization, and handling model uncertainty (see Kane et al. 2014). Compared to uplift modeling, causal inference is a more mature field with different techniques originating from statistics, economics, and computer science/artificial intelligence (see, for instance, Hernan and Robins 2005, 2016, Morgan and Winship 2014, Imbens and Rubin 2015). We offer some guidelines for the choice of causal inference techniques in Section 12.2. Section 12.3 provides an overview of business and other applications of the techniques covered in this book, followed by emerging research opportunities in Section 12.4.

12.2 Guidelines of Causal Inference Methodologies

This book has described several causal inference techniques, including when randomized experiments (i.e., randomized control trials or A/B testing) are not available. These techniques are appropriate for different kinds of data with different assumptions. To put them together, we summarize the techniques in Figure 12.1.

12.2.1 Description of the Causal Inference Flowchart

At the top node, we ask whether we have a randomized experiment. If we do, things will be much easier as we can do experimental design on the left branch. Going down the top left rectangular box, if we only have a single treatment variable, for example, price, then from (box 1) we see that it will be an A/B test (if there are only two levels or treatment versus control) or

DOI: 10.1201/9780429172588-12

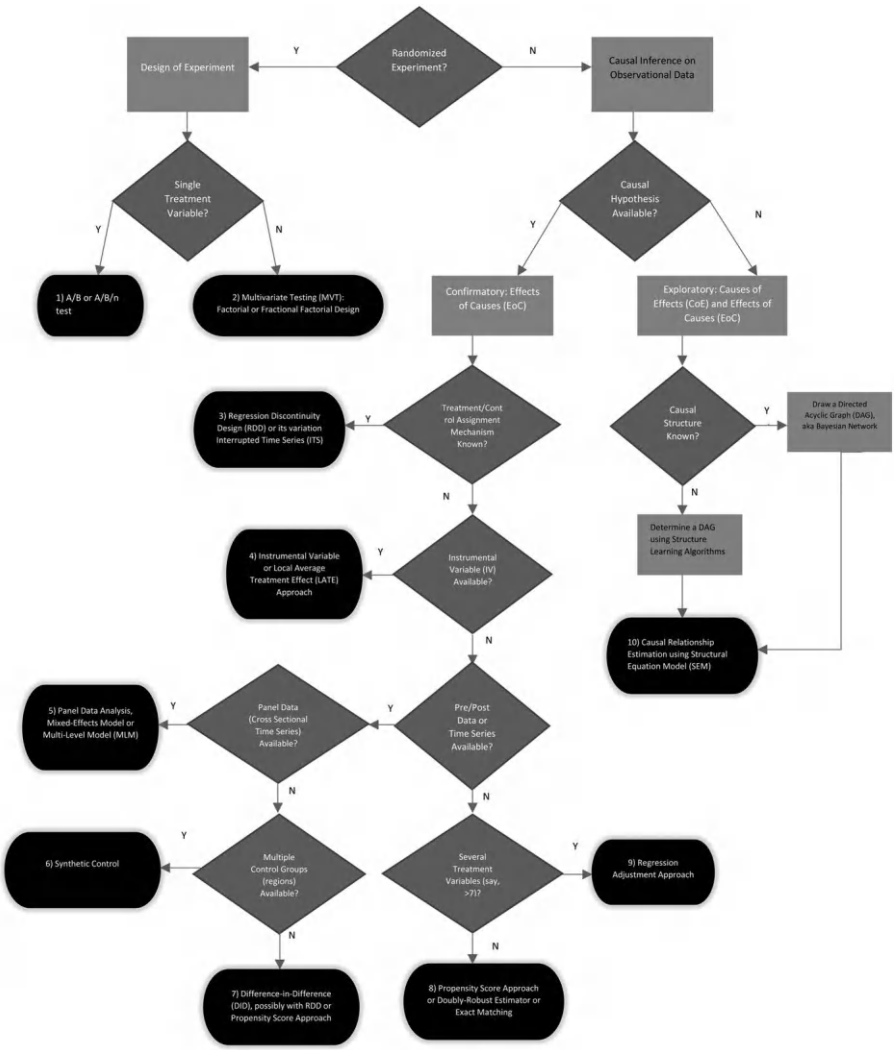


FIGURE 12.1
Flowchart of causal inference techniques.

A/B/n test if there are multiple levels (e.g., low, medium, high; placebo, low dose, high dose; or treatment 1, treatment 2, control). If we have multiple treatment variables (e.g., interest rate, annual fee, benefit features for credit cards; each treatment variable can take on multiple levels), we have (in box 2) Multivariate Testing (MVT) design that can be supported by full factorial or fractional factorial design.

The branch starting at the upper right is where we start with causal inference techniques on observational data. The first question underneath

is whether we have a strong causal hypothesis, that is, whether we know some treatment may be causing some outcome. For instance, we might like to evaluate the medical treatment effect on health or the effect of a pricing change on demand. If this hypothesis is clearly known, we go left to answer the "effects of causes" (EoC) question. Otherwise, we go right to answer both "causes of effects" (CoE) and "effects of causes" (EoC).

First, let us go down the left path to answer the EoC question. We ask whether or not we know the treatment/control assignment mechanism, that is, whether or not we know how the treatment and control groups were assigned (we already know they were not randomly split). For example, does a higher academic score lead to a clear school acceptance? If so, the small neighborhoods around the academic score cutoff may serve as treatment (if higher than the cutoff) and control (if lower) groups, as they should be otherwise similar, which allows us to use (as per box 3) the Regression Discontinuity Design (RDD). If the treatment/control assignment is determined by time only, it is reduced to the Interrupted Time Series (ITS). If the treatment/control assignment mechanism is not clearly known, then we next ask whether an instrumental variable (IV) is known and available. Recall that a satisfactory IV only impacts the outcome indirectly through the treatment variable but does not directly affect the outcome. If such variable is available, we will use (see box 4) the IV technique to obtain the Local Average Treatment Effect (LATE). If an IV is not available, we check if pre/post (before/after treatment happened) data series are available, that is, sequential data over time for the outcome and treatment variables, sometimes along with covariates as well. If yes, we check if panel data are available (i.e., both cross-sectional and time series data).

For example, in an advertising campaign, some geographics received higher ad spend while others did not, and we have the outcome (sales), treatment (ad spend), and other variables (say, socioeconomic conditions and demographics) over time, resulting in a panel data scenario, which can be handled by econometric or statistical techniques such as mixed effects modeling (see box 5). If we do not have panel data, that is, we only have pre and post (say, before and after a policy treatment) as well as treatment and control (say, some states received the treatment and others did not), then we ask if there are multiple control groups available (e.g., multiple states used as a potential control group compared to the treated state), in which case we can consider Synthetic Control (box 6), which takes a weighted average of multiple states to mimic the treated state. Otherwise, from box 7, we can use Difference-in-Difference (DID), or DID along with a propensity score approach (if confounders are available). Note that DID relies on the assumption of parallel outcome trends in treatment and control groups over time.

If time series data are not available (and neither treatment/control assignment nor IVs are available), which is a common situation, we can apply the general propensity score approach (box 8) or regression adjustment (box 9). Regression adjustment is a traditional method that is subject to model

specification error, while the propensity score approach, or its variant, doubly-robust estimation (which combines regression adjustment and propensity score), is highly recommended in general. That said, the propensity score approach needs to focus on one treatment variable at a time, while regression adjustment can handle multiple treatment variables easily. Note that exact 1:1 matching can sometimes be feasible when the dataset is large enough to locate a control subject matchable to each treated subject in each subgroup defined by multiple variables.

We now return to the decision box in the upper right. If a causal hypothesis is not available, we ask if the causal structure is known; that is, even though we are not sure which one is the treatment (or exposure) variable and which is the outcome, do we at least have some ideas how various variables are connected to each other? For example, in the business setting, we may want to know which (non-randomly assigned) customer interactions might lead to cross-sale of any products, but we do not have a hypothesis on which ones may cause which, and there can be many intermediate (mediator) variables such as opening an email, clicking the website, or researching a product, as well as some potential behavioral and demographic variables that may impact all other variables. If the causal structure is known, we then go right and draw the causal diagram, directed acyclic graph (DAG), or Bayesian Network. If the causal structure is unknown, we may apply a structure learning algorithm to estimate the DAG, with or without prior knowledge. Either way, it will be followed by the final step of estimating the relationships in (see box 10) using a structural equation model (SEM).

12.3 Overview of Application Areas

While this book focuses on business analytics applications of causal inference and uplift modeling, the techniques discussed can be applied in other fields. We will discuss some of them as follows.

12.3.1 Economic and Social Science Analysis

Economics has a long history of empirical work, most of it geared toward measuring causal relations. By 2010, 72% of the articles in the top scholarly journals in economics were empirical, up from 38% in 1983 (Hamermesh 2013).

While the workhorse of applied economics continues to be regression in one form or another, economists have been early adopters, or even developers, of many of the techniques discussed in this book. Athey and Imbens (2017) have written an excellent summary of the ways in which economists approach causality in policy evaluation. They emphasize the importance

of justifying the identifying assumptions that help one to measure causality in applied work in economics and argue that machine learning may help by reducing the need to defend the choice of potentially restrictive parametric estimators.

Here we summarize a few examples to give a flavor of how economists apply the statistical techniques that are relevant for measuring causality. The first example comes from labor economics. In 1980, over 125,000 people left Cuba for the United States as part of the Mariel Boatlift. About half of them settled in Miami, and an important question is whether this influx depressed wages in that area. David Card compared the evolution of wages in Miami, prior to and after the boatlift, with the evolution of wages in other apparently comparable cities, including Atlanta and Houston, where few Cubans settled (Card 1990). This was an early informal version of synthetic control, and Peri and Yasenov (2019) re-ran Card's analysis with a formal synthetic control design. Both studies found that the inflow of Cuban migrants had little to no effect on wages (in Miami).

The importance of a good education to economic mobility has long been recognized. But do charter schools, which are supported by public money, but have greater operating flexibility than traditional public schools, lead to improved educational outcomes? A simple comparison of charter and non-charter schools will not resolve the issue, in part because most charter schools are designed to serve children from low-income backgrounds, whose educational performance tends to be lower than that of their more-affluent peers. Abdulkadiroğlu et al. (2011) took advantage of the random assignment of children to charter schools in Boston to determine that charter schools there had a measurable positive effect on student outcomes such as grades and graduation rates.

Randomized field experiments have been widely used in labor economics. In a classic study, whose methods have been widely copied, Bertrand and Mullainathan (2004) sent job applications to employers in the United States that were identical in every respect except that one version had a black-sounding name while the other had a white-sounding name. The latter were significantly more likely to receive a call for an interview, which provides clear evidence of racially biased labor market bias. Similar techniques have been used to measure the extent of racial discrimination in the housing market in the United States, providing clear evidence of its persistence (Langowski et al. 2020).

Randomized controlled trials (RCTs) have been widely used to measure the impact of focused interventions in less-developed countries. Using RCTs, Michael Kremer and his colleagues found that cheap and simple deworming interventions in Kenya had large positive effects on educational outcomes, effects that appear to be robust (Hamory et al. 2021). Banerjee et al. (2015) found that microfinance in an urban setting in India had very limited impacts, again using an RCT. Banerjee et al. (2010) applied a similar approach to evaluating educational programs in India.

Discontinuity designs have been helpful in a number of contexts (Imbens and Lemieux 2008). For example, Bleemer and Mehta (2022) compared students who just qualified to major in economics at the University of California Santa Cruz with those who just failed to make the cut (a GPA of 2.8 in basic economics courses). These economics majors earned about $22,000 more annually in early-career wages than those that did not quite make it, which provides clear evidence of the monetary value of majoring in economics.

University graduates earn more than non-graduates, but it is not clear a priori to what extent this is because those who go to university are more capable or because the university education adds value. Fan et al. (2018) used an RDD to estimate the return on a four-year degree in a Florida public university, finding a substantial net gain.

In the educational context, Fredriksson et al. (2013) took advantage of discontinuities to measure the effect of class size on educational and economic performance. There is a clear correlation between small class size and subsequent performance, but this might be due to a common cause such as wealthy communities affording small classes while separately boosting performance. In Sweden, the law requires many classes to have no more than 25 pupils. Some schools, with cohorts greater than 25, are then required to create additional classes. Cohort size is essentially random, so this creates a natural experiment based on observational data. The study found that smaller classes do have statistically significant positive lasting effects.

Matching techniques are widely used in economic studies. To take just one example, Habimana et al. (2021) match households in Rwanda that receive unconditional cash transfers with households that have the same profile but do not receive transfers. They use an inverse probability-weighted regression adjustment estimator and find that transfers have a small positive effect on household consumption and a large effect on the amount of food derived from home production (which falls).

Geographic variation can be a useful source of exogenous differences. Raj Chetty and his collaborators find sharp variations in mobility across census tracts in the United States and are able to use this information to hone in on the drivers of mobility, including (most recently) variations in "social capital" (defined as an individual's Facebook connections with people in other socioeconomic groups), for which they have 21 billion observations on friendships (Chetty et al. 2022). Chay and Greenstone (2003) use geographic variations in the effects of recession on pollution to identify the causal impact of pollution on infant mortality, applying double differences to panel data.

Discontinuities across US states are commonly used to identify causal effects. Li et al. (2014) use changes in gasoline taxes in some midwestern states to measure the responsiveness of gasoline demand to prices. Given the important economic role of gasoline prices, this subject has received a lot of attention. For instance, Haughton and Sarkar (1996) arrived at similar results using a panel of data from US states for 1970–1991, with state-level fixed effects.

A widely cited and controversial study by Card and Krueger (1994) found that an increase in the minimum wage in New Jersey in April 1992 – the discontinuity – was associated with no reduction in, and perhaps higher levels of, employment, compared to neighboring parts of Pennsylvania (where the minimum wage did not change).

There is a huge literature on IVs, well summarized by Imbens (2014). A classic problem is identifying how price affects the quantity demanded. Observed market prices are driven by both demand and supply, and exogenous variation in supply is needed in order to trace out points on the demand curve. Hammarlund et al. (2022) use information on wind speed to instrument the supply of Norway lobster and estimate own-price elasticities of demand that are substantially larger (absolutely) than more-naïve regression methods.

12.3.2 Healthcare

Healthcare and biomedical applications have been dependent on causality for a very long time, as there is a strong need to scientifically understand the causes of diseases and the effectiveness of medical treatments and health recommendations. Experimental techniques such as RCTs and adaptive clinical trials are widely used; see Akobeng (2005) and Chow and Chang (2008) for reviews of these approaches in biomedical applications. For instance, the clinical trials of COVID-19 vaccine candidates are a highly critical application of adaptive designs using interim analyses and early stopping criteria, as reported in Pfizer (2020), Polack et al. (2020), and Moderna (2020). While randomized experiments are the key methodology for measuring treatment effectiveness before a medicine or vaccine is made available for the public, observational studies may be used to understand cause-and-effect relationships using "real-world data" after the medicine or vaccine becomes available in the market. For example, Barda et al. (2021) analyzed a large-scale observational dataset in Israel to conclude that a COVID-19 vaccine booster shot is effective, and Dickerman et al. (2021) applied observational data analysis to compare the effectiveness of the Pfizer and Moderna vaccines among U.S. veterans.

In addition to analyzing treatment effectiveness for the overall population (or groups of patients), uplift modeling can also be employed to "optimize" treatment at the individual level, which is under the field of Personalized (or Precision) Medicine, with the overall objective of maximizing population health through individual treatment effectiveness; see Hamburg and Collins (2010) and Yong (2015). For example, when the number of vaccines is limited in a nation, health officials may decide who should receive vaccination first in order to achieve the maximum protection for the entire population.

Another growing area of analytics application in healthcare is Digital Health. For example, wearable devices not only report patient vitals but can

also be used as a platform for recommendations to patients, where decisions include which messages to be displayed (and when) for each patient in order to achieve positive outcomes such as minimizing emergency department visits and medical costs. Measuring and optimizing these messages can be supported by RCT or causal inference on observational data. More advanced methods include optimizing sequential recommendations in response to past patient responses and outcomes via reinforcement learning (RL); see Menictas et al. (2019) and Carpenter et al. (2020).

12.3.3 Political Elections

Analytics have been commonly used for political elections for some time, including the usage of randomized experiments for finding the best messages and images in presidential elections; see, for example, Siroker (2010). Such an approach can be helpful in finding the best treatment for the population. Another example is the groundbreaking application of uplift modeling (also known as persuasion modeling in this context) for the 2012 U.S. presidential campaign, which targeted voters who were most persuadable. Instead of finding voters who had already made up their minds, the model was built on a pre-election survey to predict who would be most likely to switch their vote from one candidate to another, followed by outreach efforts to focus on those voters; see Porter (2013) and Siegel (2013) for a description. Many political campaigns have followed similar kinds of methodologies since then.

12.4 Emerging Research and Research Opportunities

The general field of causal inference for estimating average treatment effect for a population has been around for decades, with the potential outcomes notation begun in the 1970s, though it could even go back to the 1920s (see Pearl 2010), but formal and practical techniques such as Propensity Score Matching started only in the early 1980s. The Structural Causal Model (SCM) approach based on DAGs and Bayesian Networks started in the 1990s (see Pearl 2010). And methods such as RDD and Synthetic Control are even more recent, as are machine learning techniques (Athey and Imbens 2015). Researchers continue to develop new methods in this area, and we will discuss some of the latest methodologies below.

Additionally, methodologies for uplift modeling, which aim to estimate the heterogeneous treatment effect or conditional average treatment effect (CATE), mostly started in the 2000s, and many were developed much later (Lo 2002).

12.4.1 Survey Research Methods-Based Cause-and-Effects Analysis

While this book is focused on analyzing observational and experimental data, there is a series of techniques based on quantitative survey research. When observational or experimental data are available, it is generally preferable to use those for measurement and modeling as they represent *actual* behavior. However, there are situations when those data are unavailable or not easily available. For instance, when learning about the causal effects of complex physical items such as the size of an airplane or high-speed train on consumer choice or market share, it would be much easier to gather data from potential customers in a survey than to vary the size or shape of transportation just for measurement. These techniques are based on statistical experimental designs for varying scenarios (e.g., in transportation, these can be different levels of price, sizes of seats, and travel schedules) to assess the preferences of survey respondents. Statistical (or machine learning) models can then be developed using the survey respondent preference data to infer cause-and-effect relationships. The techniques include the classical conjoint analysis to discrete choice analysis[1] and the more recent Best-Worst Scaling (also known as maximum difference scaling or maxdiff); see Ben-Akiva and Lerman (1985), Louviere et al. (2015), and Paczkowski (2019), as well as contingent valuation (Mitchell and Carson 1989). Similar to the techniques introduced in this book, these methodologies can be deployed to measure overall treatment effects and conditional treatment effects (specific to demographics, for example). Such a survey-based approach can also be used to narrow down the treatment options before running in-market experiments, as discussed in Section 6.2.1 of Chapter 6.

12.4.2 Advanced Experimentation (and Optimization) Methods

Chapters 2 and 6 introduced more traditional randomized experiments where the target population is randomly split between treatment and control.[2] MVT methods such as factorial design and fractional factorial design are covered in Chapter 6 for measuring the effects of multiple treatments and their combinations. In addition to these techniques, there are other advanced experimental methods that are beyond the scope of this book, such as adaptive clinical trial, multi-arm bandit (MAB), and RL. Adaptive clinical trial design allows the analyst to check for statistical significance of treatment effect multiple times (known as interim analyses) before the end of the entire trial, potentially shortening the study time and reducing efforts, which is especially important for clinical studies and some expensive business experiments such as business-to-business programs. However, having multiple lookups would inflate the type I error, so statistical methodologies are required to control the overall type I error (known as family-wise error rate or FWER) so we will not have an unusually high false positive result. Such methodology is commonly used for testing medical treatments, including

many COVID-19 vaccines (e.g., Polack et al. 2020). While it is mainly used in biomedicine, it has been proposed by Legare et al. (2023) for business applications. Similarly, MAB is another sequential design method that can lead to a shorter study period. MAB is designed to find the "best" treatment as fast as possible and is not based on statistical significance (Villar et al. 2015). MAB is also used to enable RL (now an increasingly popular method in machine learning and artificial intelligence) for optimizing the sequence of treatments or interventions so as to maximize a longer-term outcome and is based on approximate dynamic programming from the field of Operations Research. See Theocharous et al. (2015), Arulkumaran et al. (2017), and Carpenter et al. (2020) for applications of RL in biomedicine, video games, and online marketing, respectively.

12.4.3 Incorporation of Behavioral Economics and Psychology – for Behavioral Intervention

This book is dedicated to utilizing experimental and observational data for testing treatments or interventions mostly for influencing human behavior. Psychologists and behavioral economists[3] have long been applying empirical analyses to draw conclusions on human behavior, with tremendous knowledge accumulated over the past few decades, rather than assuming traditional rational choice theories for prediction from classical economics. For example, applying "choice architecture" can enable organizations to frame the choice environment in order to influence human decisions such as product or service selection. Another example is that, according to Prospect Theory, humans generally react much more strongly to losses than to gains of equal magnitude. These empirical findings can help significantly reduce the number of possible treatments to save time and efforts. See Cartwright (2018), Hallsworth and Kirkman (2020), and Thaler and Sunstein (2021) for applications.

12.4.4 Multi-Criterion Decision-Making or Advanced Decision-Making

Chapters 6–9 of this book introduce various techniques for finding individuals with the best treatment effects, also known as lift at the individual level. This is mainly driven by a business objective, which can be revenue, cost, customer satisfaction, customer needs, and so on. In practice, organizations often have multiple business goals that can conflict with each other. For instance, business may look for a balance between customer preference and business value or employee needs and corporate needs. In the context of marketing analytics, as an example, one may balance the expected number of incremental responders with a measure of estimation risk such as the probability of achieving (or variability of) a number of incremental responders (Lo and Pachamanova 2015, Pachamanova et al. 2020). This multi-objective problem can be handled by treating one as an objective and

another as a constraint, or through techniques such as goal programming, TOPSIS, and two-sided matching algorithms.[4] See Roth and Sotomayor (1990), Antunes et al. (2016), Kaliszewski et al. (2016), and Rahim et al. (2018) for further reviews.

12.4.5 Model Explainability and Fairness

As predictive models through AI, machine learning, and statistics grow, there has been increasing attention from academics, governments, industry, and consulting firms to model explainability and fairness. There are a variety of techniques to help explain models (e.g., SHAP, LIME, partial dependency plot, and explainable boosted tree) and to evaluate model fairness through different metrics (e.g., Equal Opportunity by comparing the true positive rates of two groups, or Statistical Parity by checking the independence of a sensitive attribute and the predicted value); see Rothman (2020), Barocas et al. (2021, Ch. 3), and Mehrabi et al. (2022). Researchers have also considered employing causal inference to help explain the causal effect of an attribute on the predicted value (e.g., is someone declined a mortgage because of race?) or to understand the counterfactual fairness through learning the effect of a hypothetical change to a personal attribute (e.g., gender, race) on the predicted value (e.g., would the insurance premium be higher or lower if the race were different?). This is an emerging area, as there can be challenges or debates around how a hypothetical attribute change could propagate through downstream attributes on the causal pathway (e.g., for car insurance premiums, would a hypothetical gender change also lead to potential changes in size and color of their vehicle? And for university applications, would a hypothetical change in race also alter other factors such as family income and availability of resources?). There are discussions of these issues in Kusner et al. (2017), Loftus et al. (2018), Zhang and Bareinboim (2018), and Barocas et al. (2021, Ch. 5).

12.4.6 Integration of Experimental and Observational Studies and Multiple Studies

To assess cause-and-effect relationships at the population or individual level, the data sets assumed for all techniques in this book are either experimental (through RCT) or observational. What if both experimental and observational data are available? And even for experimental data that may sound perfect in assessing causality, the training data may not be fully representative of the target population (because of different geographies, different time periods, or simply different groups of people), in the sense that the distributions of variables in the training data and the target population may not be the same. Pearl and Mackenzie (2018, Ch. 10) outline the concept of data fusion for handling scenarios like these, with more details described in Bareinboim and Pearl (2016) and Hunermund and Bareinboim (2021).

Unlike the more typical methodologies from transfer learning, domain adaptation, or data shift, which are mostly based on weighting the training data to look like the target population, their methodology is based on applying a DAG to carefully codify the data dependencies and all sources of bias, such as selection bias, before making appropriate data adjustment and integration.

Notes

1. Daniel McFadden was awarded the Nobel Prize in Economics in 2000 for his development of discrete choice analysis.
2. Abhijit Banerjee, Esther Duflo, and Michael Kremer were recipients of the Nobel Prize in Economics in 2019 for their applications of randomized experiments to alleviate global poverty.
3. Daniel Kahneman and Richard Thaler were awarded the Nobel Prize in Economics in 2002 and 2017, respectively, for their contributions to behavioral economics.
4. Alvin Roth received the Nobel Prize in Economics in 2012 for his contributions to and applications of two-sided matching.

References

Abdulkadiroğlu, Atila, et al. 2011. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots". *The Quarterly Journal of Economics*, 126(2): 699–748.

Akobeng, A. K. 2005. "Understanding Randomised Controlled Trials". *BMJ: Disease in Childhood*, 90(8): 840–852.

Antunes, Carlos H., Maria J. Alves, and Joao Climaco. 2016. *Multiobjective Linear and Integer Programming*. Switzerland: Springer.

Arulkumaran, Kai, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath 2017. "A Brief Survey of Deep Reinforcement Learning". *IEEE Signal Processing Magazine: Special Issue on Deep Learning for Image Understanding*. https://arxiv.org/abs/1708.05866

Athey, Susan, and Guido W. Imbens. 2015. "Machine Learning Methods for Estimating Heterogeneous Causal Effects". Working Paper, Stanford Graduate School of Business.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation". *Journal of Economic Perspectives*, 31(2): 3–32.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2015. "The Miracle of Microfinance? Evidence from a Randomized Evaluation". *American Economic Journal: Applied Economics*, 7(1): 22–53.

Banerjee, Abhijit V., et al. 2010. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India". *American Economic Journal: Economic Policy*, 2(1): 1–30.

Barda, Noam, Noa Dagan, Cyrille Cohen, Miguel A. Hernan, Marc Lipsitch, Isaac S. Kohane, Ben Y. Reis, and Ran D. Balicer. 2021. "Effectiveness of a Third Dose of the BNT162b2 mRNA COVID-19 Vaccine for Preventing Severe Outcomes in Israel: An Observational Study". 398: 2093–2100.

Bareinboim, Elias, and Judea Pearl. 2016. "Causal Inference and the Data-Fusion Problem". *PNAS*, 113(27): 7345–7352.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2021. "Fairness and Machine Learning: Limitations and Opportunities," pre-print at https://fairmlbook.org/pdf/fairmlbook.pdf.

Ben-Akiva, Moshe, and Steven R. Lerman. 1985. *Discrete Choice Analysis: Theory and Analysis to Travel Demand*. Cambridge MA: MIT Press.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". *American Economic Review*, 94(4): 991–1013.

Bleemer, Zachary, and Aashish Mehta. 2022. "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major". *American Economic Journal: Applied Economics*, 14(2): 1–22.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market". *International Labor Review*, 43(2): 245–257.

Card, David, and Alan Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". *American Economic Review*, 84(4): 772–793.

Carpenter, Stephanie M., Marianne Menictas, Inbal Nahum-Shani, David W. Wetter, and Susan A. Murphy 2020, "Developments in Mobile Health Just-in-Time Adaptive Interventions for Addiction Science". *Current Addiction Reports*, 7, 280–290. https://static1.squarespace.com/static/5599a76ce4b0af241ed78134/t/5f702db2641b936671fd55ab/1601187251089/Carpenter2020_Article_DevelopmentsInMobileHealthJust+%281%29.pdf

Cartwright, Edward. 2018. *Behavioral Economics*, 3rd edition. London: Routledge.

Chay, Kenneth Y., and Michael Greenstone. 2003. "The Impact of Air Pollution on Infant Mortality: Evidence from Geographic Variation in Pollution Shocks Induced by a Recession". *The Quarterly Journal of Economics*, 118(3): 1121–1167.

Chetty, Raj, et al. 2022. "Social Capital I: Measurement and Associations with Economic Mobility". *Nature*, 608(7921): 108–121.

Chow, Shein-Chung, and Mark Chang. 2008. "Adaptive Design Methods in Clinical Trials – a Review". *Orphanet Journal of Rare Diseases*, 3: 11.

Dickerman, Barba A., Hanna Gerlovin, Arin L. Madenci, Juan P. Cases, and Miguel A. Hernan 2021, "Comparative Effectiveness of BNT162b2 and mRNA-1273 Vaccines in U.S. Veterans". *The New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa2115463

Fan, Elliott, Xin Meng, Zhichao Wei, and Guochang Zhao. 2018. "Rates of Return to Four-Year University Education: An Application of Regression Discontinuity Design". *The Scandinavian Journal of Economics*, 120(4): 1011–1042.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek. 2013. "Long-Term Effects of Class Size". *The Quarterly Journal of Economics*, 128(1): 249–285.

Habimana, Dominique, Jonathan Haughton, Joseph Nkurunziza, and Dominique Marie-Annick Haughton. 2021. "Measuring the Impact of Unconditional Cash Transfers on Consumption and Poverty in Rwanda". *World Development Perspectives*, 23: 100341.

Hallsworth, Michael, and Elspeth Kirkman. 2020. *Behavioral Insights*. Cambridge, MA: MIT Press.

Hamburg, M. A., and F. S. Collins. 2010. "The Path to Personalized Medicine". *The New England Journal of Medicine*, 363(4): 301–304.

Hamermesh, Daniel. 2013. "Six Decades of Top Economics Publishing: Who and How?". *Journal of Economic Literature*, 51(1): 162–172.

Hammarlund, Cecilia, Johan Blomquist, and Staffan Waldo. 2022. "The Way the Wind Blows: Tracing Out the Demand for Norwegian Lobster Using Instrumental Variables". *Marine Resource Economics*, 37(3): 263–282.

Hamory, Joan, et al. 2021. "Twenty-Year Economic Impacts of Deworming". *Proceedings of the National Academy of Sciences* 118(14): e2023185118.

Haughton, Jonathan, and Soumodip Sarkar. 1996. "Gasoline Tax as a Corrective Tax: Estimates for the United States, 1970–1991". *The Energy Journal*, 17(2): 103–126.

Hernan, Miguel A., and James M. Robins. 2005. "Estimating Causal Effects from Epidemiological Data". *Journal of Epidemiology Community Health*, 60: 578–586.

Hernan, Miguel A., and James M. Robins. 2016. *Causal Inference*. Boca Raton, FL: CRC Press.

Hunermund, Paul, and Elias Bareinboim. 2021. "Causal Inference and Data Fusion in Econometrics," *Technical Report* R-51. https://arxiv.org/pdf/1912.09104.pdf.

Imbens, Guido. 2014. "Instrumental Variables: An Econometrician's Perspective". *Statistical Science* 29(3) 323–358. https://doi.org/10.1214/14-STS480

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, England: Cambridge University Press.

Imbens, Guido W., and Thomas Lemieux. 2008. "The Regression Discontinuity Design – Theory and Applications. Special Issue". *Journal of Econometrics*, 142(2): 611–614.

Kaliszewski, Ignancy, Janusz Miroforidis, and Dmitry Podkopaev. 2016. *Multiple Criteria Decision Making by Multiobjective Optimization: A Toolbox*. Switzerland: Springer.

Kane, Kathleen, Victor S. Y. Lo, and Jane Zheng. 2014. "Mining for the Truly Responsive Customers and Prospects Using True-Lift Modeling: Comparison of New and Existing Methods". *Journal of Marketing Analytics*, 2(4): 218–238.

Kusner, M. J., C. Russell, J. Loftus, and R. Silva. 2017. Counterfactual Fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.), *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. NIPS Proceedings, Long Beach, CA

Langowski, Jamie, William Berman, Grace Brittan, Catherine LaRaia, Jee-Yeon Lehmann, and Judson Woods. 2020. "Qualified Renters Need Not Apply: Race and Voucher Discrimination in the Metro Boston Rental Housing Market". *Georgetown Journal on Poverty Law and Policy*, 28(1): 35–74.

Legare, Jonathan, Ping Yao, and Victor S. Y. Lo. 2023. "A Case for Conducting Business-to-Business Experiments with Multi-arm Multi-stage Adaptive Designs". *Journal of Marketing Analytics*, 11: 490–502. https://doi.org/10.1057/s41270-022-00177-4

Li, Shanjun, Joshua Linn, and Erich Muehlegger. 2014. "Gasoline Taxes and Consumer Behavior". *American Economic Journal: Economic Policy*, 6(4): 302–342.

Lo, Victor S. Y. 2002. "The True-Lift Model – A Novel Data Mining Approach to Response Modeling in Database Marketing". *ACM SIGKDD Explorations*, 4(2): 78–86.

Lo, Victor S. Y., and Dessislava Pachamanova. 2015. "Prescriptive Uplift Analytics: A Practical Approach to Solving the Marketing Treatment Optimization Problem and Accounting for Estimation Error Risk". *Journal of Marketing Analytics*, 3(2): 79–95.

Loftus, J. R., C. Russell, M. J. Kusner, and R. Silva. 2018. "Causal Reasoning for Algorithmic Fairness". Working paper. https://arxiv.org/abs/1805.05859

Louviere, Jordan J., Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge, England: Cambridge University Press.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. "A Survey on Bias and Fairness in Machine Learning". *ACM Computing Surveys*, 54(6): 1–35.

Menictas, Marianne, Mashfiqui Rabbi, Predrag Klasnja, and Susan Murphy. 2019. "Artificial Intelligence Decision-Making in Mobile Health". *The Biochemist*, 41: 20–24.

Mitchell, Robert, and Richard Carson. 1989. *Using Surveys to Value Public Goods*. New York, NY: RFF Press.

Moderna. 2020, "Vaccines and Related Biological Products Advisory Committee Meeting". *FDA Briefing Document: Moderna COVID-19 Vaccine*.

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference*, 2nd edition. New York, NY: Cambridge University Press.

Pachamanova, Dessislava, Victor S. Y. Lo, and Nalan Gulpinar. 2020 "Uncertainty Representation and Risk Management in Direct Segmented Market". *Journal of Marketing Management*, 36(1–2): 1–27.

Paczkowski, Walter R. 2019. *Pricing Analytics: Models and Advanced Quantitative Techniques for Product Pricing*. London: Routledge.

Pearl, Judea. 2010, "An Introduction to Causal Inference". *The International Journal of Biostatistics*, 6(2): 1–59.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Peri, Giovanni, and Vasil Yasenov. 2019. "The Labor Market Effects of a Refugee Wave". *Journal of Human Resources*, 54(2): 267.

Pfizer. 2020. "A Phase 1/2/3 Placebo-Controlled, Randomized, Observer-Blind, Dose-Finding Study to Evaluate the Safety, Tolerability, Immunogenicity, and Efficacy of SARS-COV-2 RNA Vaccine Candidates Against COVID-19 In Healthy Individuals". Retrieved November 05, 2021, from https://cdn.clinicaltrials.gov/large-docs/28/NCT04368728/Prot_000.pdf

Polack, Fernando P., et al. 2020, "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine". *The New England Journal of Medicine*, 383(27): 2603–2615.

Porter, Daniel. 2013. "Pinpointing the Persuadables: Convincing the Right Voters to Support Barack Obama". Predictive Analytics World; Oct, Boston. MA; http://www.predictiveanalyticsworld.com/patimes/pinpointing-the-persuadables-convincing-the-right-voters-to-support-barack-obama/ (available with free subscription).

Rahim, Robbi, et al. 2018. "TOPSIS Method Application for Decision Support System in Internal Control for Selecting Best Employees". *Journal of Physics: Conference Series,* 1028: 012052.

Roth, Alvin E., and Marilda A. Qliveira Sotomayor. 1990. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monographs No.18, Cambridge, England: Cambridge University Press.

Rothman, Denis. 2020. *Hands-On Explainable AI (XAI) With Python*. Birmingham, England: Packt Publishing.

Siegel, Eric. 2013. "The Real Story Behind Obama's Election Victory". *The Fiscal Times*. https://www.thefiscaltimes.com/2013/01/21/Real-Story-Behind-Obamas-Election-Victory

Siroker, Dan. 2010. "Obama's $60 Million Dollar Experiment," Optimizely. https://www.optimizely.com/insights/blog/how-obama-raised-60-million-by-running-a-simple-experiment/

Thaler, Richard H., and Cass R. Sunstein. 2021. *Nudge: The Final Edition*. London: Penguin Books.

Theocharous, Georgios, Philip S. Thomas, and Mohammad Ghavamzadeh 2015, "Ad Recommendation Systems for Life-Time Value Optimization". *WWW'15 Companion: Proceedings of the 24th International Conference on World Wide Web.* IW3C2 1305–1310.

Villar, Sofia S., Jack Bowden, and James Wason. 2015. "Multi-Armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges". *Statistical Science*, 30(2): 199–215.

Yong, Florence H. 2015, "Quantitative Methods for Stratified Medicine". *PhD Dissertation*, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University.

Zhang, Junzhe, and Elias Bareinboim. 2018. "Fairness in Decision-Making – The Causal Explanation Formula". *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). https://ojs.aaai.org/index.php/AAAI/article/view/11564

Index

Note: Page numerals in **bold** refer to tables, and those in *italics* to figures.