Health Services Research and Program Evaluation

Causal Inference and Estimation

MARCELO COCA PERRAILLON AND DONALD HEDEKER

Draft

Contents

Preface

This is a book on quantitative methods in health services research, health economics, and health policy evaluation – more generally referred to as "program evaluation." Health services research is a multidisciplinary field that examines the use, costs, quality, outcomes, and other aspects of health care including the organization of healthcare markets. Evaluating the impact of health policy is central to the field.

Quantitative analyses in health services research apply methods and language developed in econometrics and statistics or biostatistics. In most applications, the goal is to understand the causal impact of policy changes or "treatments," broadly defined, on a set of outcomes. In most circumstances, however, randomized trials are either not feasible or prohibitively expensive, and we must establish causality using observational data; that is, data that were not collected as part of an experiment. The main distinction between experiments and observational studies is that in observational studies treatment assignment is not under the control of the investigator. The consequence is that other factors besides treatment are not held constant so it is more difficult to establish causality.

Most readers have already learned that correlation or association does not imply causation. The goal of causal inference is to understand under which conditions correlation –or any other measure of association– does imply a causal effect. Thus, this book is about the design of observational studies and the estimation of statistical models to answer causal research questions. Or said another way, under which circumstances an approach can *identify* causal effects. However, we also cover the necessary background to understand advanced methods. The background material is focused on understanding the mechanics and properties of parametric and nonparametric statistical models. These models are useful as descriptive and predictive tools, but our ultimate goal is to use them to answer causal research questions.

One feature of our book is that we separate the *design* of an observational study from the *estimation* of statistical models. The separation of design and estimation is one of the most valuable aspects of the potential outcomes framework since causal effects are defined independently of an estimation method. This approach is part of the "new" causal inference field in statistics, although causal inference has always been central to econometrics. In the last two to three decades, these separate but related fields have found plenty of common ground regarding causality. The new part is a clear definition of causal effects and a mathematical notation based on potential outcomes and counterfactuals that continues to expand and clarify our understanding of established methods and facilitates the development of new ones.

Our approach is based on the premise that complex concepts are better understood when first introduced with intuitive examples and graphs, followed by theory, and then practical applications using statistical software. Based on our experience teaching graduate-level classes, we think that students learn best by doing, and "doing" means relating the theory to application using statistical software. Some concepts are difficult to understand in theory but are relatively easy to understand when implemented in practice (and vice versa).

We strive to present theory intuitively but formally to show *how* the theory is applied and *why* methods work, which is essential for understanding *when* specific methods should be used and *what* meaning can be derived from the estimators. It is also the basis for understanding methods that still have not been developed. This is *not* a "cookbook approach" book in the sense that we do not focus on rules for specific situations because more often than not it is not possible to precisely spell out or anticipate the specific situation a rule requires. Instead, we focus on principles, concepts, and assumptions needed – the how, why, when, and what– which can then be evaluated in specific situations.

We do not shy away from presenting complex concepts and mathematical notation because they are essential tools to develop intuition on how and why statistical methods work. Mathematics is a language that makes the job *easier*, not more difficult. Mathematics allows us to represent ideas and concepts using symbols, and we manipulate these symbols to discover new ideas and prove propositions that might not be self-evident. Manipulating complex ideas in our minds without the use of symbols is much more difficult. However, we always provide the intuition behind the mathematics to help students understand how the symbols relate to ideas since not all students are comfortable with mathematics. At the end of the course(s), students should be able to understand the language of mathematics as it applies to statistical analysis. Our recommendation to students is to think of mathematics as a language. Success in understanding a concept expressed with mathematics is to make sure the meaning of the mathematical notation is understood. That is often the first, but often-ignored, first step.

This book is intended for advanced undergraduates, master's students, and doctoral students in health services research, health economics, public policy, public health, and related fields. Students in these disciplines come from diverse backgrounds with different levels of preparation. We assume the same background that is commonly required for admission to these programs: two semesters of calculus and introductory statistics. A class on linear regression would be helpful, but not strictly necessary since we review the essential features of linear models. We keep linear algebra to a minimum. The goal of the mathematical appendix is to review the mathematical background needed to understand the rest of the book. We hope that students go over the introductory material even if it is not assigned by instructors. Each new concept is based on previous concepts; it is a lack of knowledge of the basics, and the corresponding notation, that confuses students the most. Previous knowledge of Stata or R is helpful, although the background chapters also serve as an introduction to Stata, and the supplementary material reproduces the code using R.

Key features of this book include:

- *Semantics Boxes* that clarify how terms are used in different disciplines. Because our field is multidisciplinary, the terms we use can be confusing –sometimes comically so– because the same terms can have different definitions or because the same concept is named differently in other fields.

- *Notation Boxes* that clarify how mathematical symbols are used in different disciplines or by different authors. As we said, mathematics is a language, but it is a language with symbols that are not standard and can be defined in different ways by different authors. We clarify and present alternative mathematical notation because not understanding unfamiliar notation can prevent students and practitioners from grasping the underlying concepts. A variant of this theme is that sometimes the notation is the result of giving statistical models an interpretation tied to an underlying theory, so we also cover different ways of understanding and/or deriving statistical models. We think students will be better equipped to understand theoretical papers and more advanced textbooks if they understand the notation.
- Extensive examples using datasets to illustrate real-life applications. One frustrating aspect of teaching health services research methods is that we usually cannot use the same datasets that are common in the field and our own research because Data Use Agreements (DUA) do not permit the distribution of these data. However, we have created multiple datasets from publicly available sources and include datasets that authors have made publicly available to reproduce published papers. Our goal is to use datasets that reflect how practitioners work in our fields.
- Stata code to reproduce all examples and figures in the book. We use Stata code as a tool for learning. In some cases, like graphs or long output, not all of the code is in the book, but it is available in the online supplemental material.
- Stata version control. We prefer Stata as the main statistics packed for the book because it has the features we need and it has extensive documentation and substantial technical support. Stata is that it is backwards compatible. Regardless of updates, commands will always work provided the code includes a Stata version statement. This ensures that our code will not become obsolete when new versions are released or commands are updated. Most of our code requires Stata 16 to 18. Each program file begins with a version statement.
- Online supplemental material. The online supplemental material includes R code to replicate most of the examples in the book when possible, although some material is specific to Stata. The online supplemental material also covers additional topics that we had to leave out from the text because of space constraints. More graphs are presented in the supplementary material than in the textbook to conserve space.
- End-of-chapter exercises to reinforce key concepts.
- End-of-chapter bibliographical notes with references to books and papers where readers can find additional or complementary material.
- End-of-chapter ”Key points” to highlight the most important message of each chapter.

This book is also intended to be a tool for faculty who teach quantitative methods and a reference for practitioners. We wrote it because we could not find a textbook that fit the needs of our students and our research. In our classes, we ended up assigning book chapters and papers that use different notation and language, which makes both learning and teaching more difficult. We had to complement those materials with extensive lecture notes and “translations” of notation, terms, and subject-matter. Our lecture notes are the basis for this book.

Additional supplemental material for instructors include:

- Solutions to most end-of-chapter exercises.
- Most of the sample datasets contain additional variables that are not part of our analyses. Instructors could use these variables to expand problems sets or create examples focusing on different research questions. In many cases, the variables have missing values. Most textbooks use small sample datasets with non-missing values, but this does not reflect the reality of how research is conducted, so we decided to retain missing values in some of the datasets.
- Lecture notes for most chapters. The lecture notes focus on the most important parts of each chapter. These notes can be used as a starting point for teaching with our book. The lecture notes are in Latex (Beamer) and PowerPoint format. In general, we cover the key parts in class and then assign other section for reading.
- Errata. Despite multiple revisions and editing, the presence of a mistake converges to 1 in probability given the length of our book. We will post a complete list of errors by chapter as we find them, including updates and clarification of some material.

We wrote the book with a two-semester quantitative methods sequence in mind plus additional material for review. We cover topics that should be part of the standard toolkit in health services research and health/public policy doctoral programs as well as applied econometrics courses in economics programs, although most of our examples are about health care.

The book is divided into four parts. Parts I and II introduce the major subjects we cover, including the potential outcomes framework and a review of linear regression. Part III focuses on estimation and inference of statistical models, including interpretation of model parameters (causal or not) and discussion of nonparametric models. In other words, Part III discusses techniques to *estimate* statistical models and the assumptions and properties of these models when applied to a sample, *without necessarily assuming that findings from these models have a causal interpretation*. On the other hand, Part IV covers the most important methods to estimate causal effects using observational data: propensity scores and matching estimators as an alternative and complement to regression adjustment, longitudinal (panel) data, difference-in-differences, regression discontinuity designs, and instrumental variables. Thus, in Part IV, model interpretation implies a causal statement that requires precise definition.

Two chapters are fundamental for students to master: Chapter 3 on the potential outcomes framework and Chapter 6 on marginal effects. Chapter 3 is the foundation to understand the definition of causal effects and the identification of causal effects using a sample, and it presents the potential outcome notation we use in the rest of the book. Chapter 6 on marginal effects is essential for interpreting model parameters and to express model parameters in different scales and metrics regardless of whether the parameters have a causal interpretation. We provide an overview of each chapter and their connections in Chapter 1.

We have tried to make the chapters as self-contained (modular) as possible –particularly in Part IV– so they can be used independently, although this separation is artificial. We refer to other material in the book when we think students would benefit from reading sections in other chapters, but we have tried to keep such references to a minimum.

Each chapter progresses from simple to advanced, from known to unknown, and from concrete to abstract without losing track of practical applications. Instructors could skip the sections that appear towards the end of each chapter if they think the material is too advanced for their students. However, we hope that all of the material can be covered, time permitting. Often, "advanced" really means "unknown." Most concepts are simple once we understand them, and our understanding of "sophisticated" changes with time. What was a sophisticated method a decade ago could be a standard one now.

0.1 Suggested use for classes

A typical two-semester sequence for students starting a sequence of quantitative methods would cover linear regression review, the potential outcomes framework, estimation methods, interpretation and commonly used causal inference methods. Since each chapter is organized from basic to more advance topics, some sections at the end of the chapter could be skipped or only the first part of chapter could be assigned. For example, we frame propensity scores and matching estimators as an alternative to regression adjustment, which can be useful when overlap problems are present (and for uncovering them). Thus, only the first part of Chapter 9 on propensity scores could be assigned.

Our two-semester classes usually cover the following chapters:

First semester:	Chapters 2-6
Second semester:	Chapters 9-12

In some programs, students take a year of mathematical statistics and/or econometric theory before taking applied methods classes. In this case, a two-semester sequence would skip some of the background material on estimation but cover chapters in more detail:

First semester:	Chapters 3,6
Second semester:	Chapters 10-12

Optional: Chapters 7,8,9

Alternatively, the book could be used for a one-semester class on causal methods for the analysis of observational data assuming the statistical/econometrics background material is known (students could review on their own Chapter 3 or other background):

Chapters 9-12
Optional (but suggested): Chapter 6