Distributional and Nonparametric Econometrics

Second edition

David M. Kaplan

First nice edition, May 2020; second edition, January 2021
Updated October 9, 2023

To the tails.
—DMK

An economist was standing with one foot in a bucket of boiling water and the other foot in a bucket of ice. When asked how he felt, he replied, "On average I feel just fine."

Variation of quote attributed to Mark Twain
As retold by Hansen (2020a, p. 29)

Brief Contents

Contents

Preface

This text was prepared for a 15-week semester Advanced Econometrics course for 2nd-year economics PhD students at the University of Missouri. The class focuses on two general themes: 1) learning about aspects of distributions besides the mean, and 2) nonparametric methods. Other topics naturally arise.

The assumed background is the first-year core PhD econometrics at the University of Missouri, which uses (roughly) the first nine chapters of Hansen (2020a) and related material from Hansen (2020b) in the first semester and a subset of Wooldridge (2010) covering basics like IV, GMM, and potential outcomes in the second semester.

As with my *Introductory Econometrics* text (Kaplan, 2022b), this text's source files are freely available. Instructors may modify them as desired, or copy and paste LaTeX code into their own lecture notes, with usage subject to the Creative Commons license linked on the copyright page. I wrote the text in Overleaf, an online (free) LaTeX environment that includes knitr support. You may see, copy, and download the entire project from Overleaf[1] or from my website.[2]

Another unusual feature is the prevalence of in-class discussion questions. I find these very helpful (for more actively engaging students, for gauging how students are tracking, and for breaking up my lecturing), and students seem to appreciate them, too.

Thanks to everyone for their help and support: my past econometrics instructors, my colleagues and collaborators, my students, and my family.

David M. Kaplan
Spring, 2020
Columbia, Missouri, USA

[1] https://www.overleaf.com/read/bbmwhsvfwgfc
[2] https://kaplandm.github.io/teach.html

Textbook Learning Objectives

For good reason, it has become standard practice to list learning objectives for a course as well as each unit within the course. Below are the learning objectives corresponding to this text overall. In the future, each chapter will additionally list more specific learning objectives that map to one or more of these overall objectives. I hope you find these helpful guidance, whether you are a solo learner, a class instructor, or a class student.

The textbook learning objectives (TLOs) are the following.

1. For a variety of econometric methods, describe their critical assumptions, output, and interpretation (economic and statistical), with some understanding of how these relate.
2. Develop intuition for fundamental concepts to enable you to understand econometrics papers/books that you need to read later for your own research.
3. Judge which of two methods is "better" in a given situation, including in others' research.
4. Gain familiarity with LaTeX and R.
5. Produce new empirical (or methodological) econometric research, aware of its potential flaws (accepting that it won't be perfect), able to articulate (defend) how you've successfully extracted new knowledge about the world from the raw data.

Notation

Variables

Usually, uppercase denotes random variables, whereas lowercase denotes fixed values. The primary exception is for certain counting variables, where uppercase indicates the maximum value and lowercase indicates a general value; e.g., time period t can be $1, 2, 3, \ldots, T$, or regressor k out of K total regressors. Scalar, (column) vector, and matrix variables are typset differently. For example, an n-by-k random matrix with scalar (random variable) entries X_{ij} (row i, column j) is

$$\underline{\boldsymbol{X}} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}$$

and a k-dimensional non-random vector is

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{pmatrix}$$

Unless otherwise specified, vectors are column vectors. The transpose of a column vector is a row vector. For example, using the $\boldsymbol{z}$ defined above,

$$\boldsymbol{z}' = (z_1, z_2, \ldots, z_k)$$

Note: displayed math like above should always have appropriate punction (comma, period) at the end! . . . unless you are defining notation and worry about confusing people.

Greek letters like $\boldsymbol{\beta}$ and θ generally denote fixed population parameters.

I sometimes make exceptions to match convention. For example, ϵ is a Greek letter but is conventionally used for a regression error term or white noise.

Estimators usually have a "hat" on them. Since estimators are computed from data, they are random from the frequentist perspective. Thus, even if θ is a non-random population parameter, $\hat{\theta}$ is a random variable.

I try to put "hats" on other quantities computed from the sample, too. For example, a t-statistic would be $\hat{t}$ (a random variable computed from the sample) instead of just t (which looks like a non-random scalar). Or, a J-statistic would be $\hat{J}$, even though J is already uppercase, to emphasize that it is computed from data (rather than data itself).

Besides hats, tildes and bars may indicate estimators of parameters, and bars indicate sample averages. For example, there may be multiple alternatives for estimating θ: $\hat{\theta}$, $\tilde{\theta}$, and $\bar{\theta}$. The sample average of $Y_1, \ldots, Y_n$ is $\bar{Y}$.

Estimators and other **statistics** (i.e., things computed from data) may sometimes have a subscript with the sample size n to remind us of the asymptotic perspective of a sequence (indexed by n) of random variables. For example, with n denoting sample size, $\hat{\theta}_n$, $\hat{t}_n$, and $\bar{Y}_n$.

The following is a summary.

y	scalar fixed (non-random) value
Y	scalar random variable
θ	scalar non-random value
$\hat{\theta}$	scalar random variable
$\boldsymbol{x}$	non-random column vector
$\boldsymbol{x}'$	transpose of $\boldsymbol{x}$
$\boldsymbol{X}$	random column vector
$\boldsymbol{\beta}$	non-random column vector
$\hat{\boldsymbol{\beta}}$	random column vector
$\underline{\boldsymbol{w}}$	non-random matrix
$\underline{\boldsymbol{w}}'$	transpose of $\underline{\boldsymbol{w}}$
$\underline{\boldsymbol{W}}$	random matrix
$\underline{\boldsymbol{\Omega}}$	non-random matrix
$\hat{\underline{\boldsymbol{\Omega}}}$	random matrix

Symbols

In addition to the following symbols, vocabulary words and abbreviations (like "quantile" or "IVQR") can be looked up in the Index in the very back of the text.

$\implies$	implies; see Chapter 3
$\impliedby$	is implied by; see Chapter 3
$\iff$	if and only if; see Chapter 3
$\lim_{n\to\infty}$	limit
$\operatorname{plim}_{n\to\infty}$	probability limit
$\to$	converges to (deterministic)

$\xrightarrow{p}$	converges in probability to; see Hansen (2020b, §7.3)
$\xrightarrow{\text{a.s.}}$	converges almost surely to; see Hansen (2020b, §7.14)
$\xrightarrow{d}$	converges in distribution to; see Hansen (2020b, §8.2)
$\rightsquigarrow$	converges weakly to
$\equiv$	is defined as
$\approx$	approximately equals
$\doteq$	equals when ignoring smaller-order terms
$\sim$	is distributed as
$\overset{a}{\sim}$	is distributed approximately (or asymptotically) as
$X \perp\!\!\!\perp Y$	X and Y are statistically independent
$\mathrm{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathrm{N}(0, 1)$	standard normal distribution
$\Phi(\cdot)$	cumulative distribution function (CDF) of $\mathrm{N}(0, 1)$
$\phi(\cdot)$	probability density function (PDF) of $\mathrm{N}(0, 1)$
$F_Y(\cdot)$	cumulative distribution function (CDF) of Y
$Q_Y(\cdot)$	quantile function of Y
$f_Y(\cdot)$	probability density function (PDF) of Y (or PMF if discrete)
$\mathbb{1}\{\cdot\}$	indicator function: $\mathbb{1}\{A\} = 1$ if event A occurs, else $\mathbb{1}\{A\} = 0$
$\mathrm{P}(A)$	probability of event A
$\mathrm{P}(A \mid B)$	conditional probability of A given B
$\mathrm{E}(Y)$	expected value of Y
$\widehat{\mathrm{E}}(Y)$	expectation for sample distribution; same as $\frac{1}{n}\sum_{i=1}^{n} Y_i$
$\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$	CEF (function of $\boldsymbol{x}$); see Hansen (2020a, §2.5)
$\mathrm{E}(Y \mid \boldsymbol{X})$	expected value of Y given $\boldsymbol{X}$; this is a random variable
$\mathrm{Q}_\tau(Y)$	τ-quantile of Y; see Section 4.2
$\mathrm{Q}_\tau(Y \mid \boldsymbol{X} = \boldsymbol{x})$	conditional τ-quantile function (τ-CQF); see Section 5.1
$\mathrm{Var}(Y)$	variance of Y
$\mathrm{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x})$	conditional variance (a non-random value)
$\mathrm{Var}(Y \mid \boldsymbol{X})$	conditional variance (a random variable)
$\mathrm{Cov}(Y, X)$	covariance
$\mathrm{Corr}(Y, X)$	correlation
$b \in \{a, b, c\}$	b is in the set containing a, b, and c
$\mathcal{S}_1 \cup \mathcal{S}_2$	the union of sets $\mathcal{S}_1$ and $\mathcal{S}_2$
$\bigcup_{j=1}^{J} \mathcal{S}_j$	the union of $\mathcal{S}_1, \ldots, \mathcal{S}_J$
$\mathcal{S}_1 \cap \mathcal{S}_2$	the intersection of sets $\mathcal{S}_1$ and $\mathcal{S}_2$
$\bigcap_{j=1}^{J} \mathcal{S}_j$	the intersection of $\mathcal{S}_1, \ldots, \mathcal{S}_J$
$\mathbb{N}$	the set of natural numbers, $\{1, 2, 3, \ldots\}$
$\mathbb{R}$	the set of real numbers (which excludes $\pm\infty$)
$\mathbb{R}_{\geq 0}$	the non-negative real numbers

$\mathbb{R}_{>0}$	the strictly positive real numbers
$\bar{\mathbb{R}}$	the extended real numbers, $\mathbb{R} \cup \{-\infty, \infty\}$
$\mathbb{R}^k$	k-dimensional Euclidean space
$\mathbb{Z}$	the set of integers, $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{Z}_{\geq 0}$, $\mathbb{Z}_{>0}$	analogous to $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$
$\mathrm{SE}(\hat{\theta})$	standard error of estimator $\hat{\theta}$
$\arg\min_g f(g)$	the value of g that minimizes $f(g)$
$\boldsymbol{I}_k$	$k \times k$ identity matrix (ones on main diagonal, zeros elsewhere)
$\|\cdot\|$	norm (Euclidean unless otherwise defined)
$\mathrm{tr}(\underline{\boldsymbol{v}})$	trace of matrix $\underline{\boldsymbol{v}}$
$\underline{\boldsymbol{v}}'$	transpose of matrix $\underline{\boldsymbol{v}}$
$\underline{\boldsymbol{v}}^{-1}$	inverse of matrix $\underline{\boldsymbol{v}}$
$\underline{\boldsymbol{v}} > 0$	matrix $\underline{\boldsymbol{v}}$ is positive definite
$\underline{\boldsymbol{v}} \geq 0$	matrix $\underline{\boldsymbol{v}}$ is positive semi-definite

Statistical Software Overview

Note #1: if links don't work, try Google. (That's how I found them, after all.) Google is often able to track down helpful pages.

Note #2: in general, I would trust a random (Googled) page's tips for R much more than its econometric advice. It's easy to try the code they provide and see if it does what you need it to do. It's difficult or impossible to quickly see whether their econometric suggestion is appropriate for your data (or if what they are saying is even correct at all).

As a student at Mizzou, you can use Software Anywhere for free.[3] Even if you are off-campus, that webpage gives instructions for connecting first with VPN.

The on-campus computing sites also provide a variety of statistical software. You can check which computing sites/labs have your favorite software on the Computing Sites Software web page.[4]

If you ever need help beyond what you can find on Google, please feel free to come to my office hours—that's what they are for.

R

Nice things about R:

1. It's free. (As are RStudio and other related products.)
2. It's open-source. (Is that nice? I'm not sure I care.)
3. It's popular:
    - Companies use it.
    - Academics use it across many fields.
    - Statisticians/econometricians often contribute code/packages for new methods in R. (My guess is new econometric methods are provided most commonly in R, then Stata, then Matlab.)
    - There are many online resources for learning R.
4. The syntax is relatively straightforward (i.e., it's not SAS; similar to Matlab, S-plus, etc.).

[3] https://doit.missouri.edu/services/software/software-anywhere
[4] https://doit.missouri.edu/services/computing-sites/sites-software

5. The graphics look the nicest to me.
6. It's flexible (easy to create new functions, etc.).
7. Can do parallel processing to speed up computations (even on your personal computer).

Drawbacks:

1. It can be more complicated to do common econometric tasks in R than in Stata; e.g., cluster-robust standard errors, 2SLS.
2. It's slower than FORTRAN and such.

Sometimes I have used a combination of Stata and R to analyze data. Often datasets are available online in .dta format, and it is easy to do simple manipulations in Stata (filtering, reshaping, merging, etc.). Then, you can load the prepared ("prepped") dataset into R to run whatever special function you want to use in R. (Or if you're just running OLS or something basic, just stick with Stata.) There is actually an R package (`haven`) that loads Stata .dta files (up to version 15 as of March 2020); or you can just export from Stata into .csv format, which is easy to read into R.

More detailed help getting started in in Chapter 2.

Stata

Nice things about Stata:

1. Very intuitive and simple; easy to do most common tasks.
2. Popular among applied economists $\implies$ lots of support, data often available in Stata format, used in jobs, etc.
3. I think the help files within Stata are very helpful (once you know the basic structure and syntax).

Drawbacks:

1. Not as many fancy functions as R, although econometricians are getting better about providing code in Stata (e.g., lots of the new RD methods).
2. Not as easy to code your own functions (vs. R, based on my experiences doing both).
3. Can only have one dataset in memory at a time.
4. Slower? Most expensive version does support parallel processing now, and Mata is compiled (I think).

Suggestion: if you get a job (or research project) where you'll be using Stata for a while, it is definitely worth the investment to learn the commands (rather than using the menus/buttons) and to write DO-files that can be saved and replicated.

UCLA has some respected Stata resources.[5]

The first Google hit is currently a Princeton professor's tutorial; I haven't looked through it, but it's probably pretty good, right?[6]

[5] https://stats.idre.ucla.edu/stata
[6] http://data.princeton.edu/stata

Matlab

Good: ok syntax and speed (including parallel processing).

Bad: much (most?) of the Matlab functionality is in the "toolboxes" that must be separately purchased, so if you don't have access to all the toolboxes (and can't buy them when necessary), functionality is restricted. Also, not as much econometrics-specific code available since R and Stata are more popular for most but not all fields within econ.

See http://people.duke.edu/~hpgavin/matlab.html for a (curated?) list of tutorials, or try Google.

SAS

I primarily used SAS when working at the economic consulting firm NERA for two years. I didn't like it as much as other statistical software options, but (at the time) Stata couldn't handle the big files we had, and my boss got his start as a dedicated SAS programmer. I hope you aren't ever forced into using SAS, but I'm happy to try to help if you'd like to learn it.

Others

Julia: supposed to be great, but less widely used, so maybe have to write more of your own code from scratch. (But I don't think it's like the Esperanto of programming languages or anything.)

Python, Fortran, C, GAUSS, Eviews....

Part I

Writing, Coding, and Logic

Chapter 1
Writing and Typesetting

Unit learning objectives for this chapter

1.1. Get started with an online LaTeX editor and modify some templates [TLO 4]

1.2. Learn best practices for effective writing

Optional resources for this chapter

- Overleaf registration: https://www.overleaf.com?r=63e2691f&rm=d&rs=b
- My LaTeX templates, including job market candidate (JMC) templates: https://www.overleaf.com/read/gtfzpkwrzhhw
- Article on scientific writing: https://pdfs.semanticscholar.org/73e3/171fc0ef4aa6d1d92cff07085f41e94907a6.pdf

1.1 LaTeX

In class, we'll spend one day on LaTeX. I know some of you may not ever use it, which is fine. (Unless you're doing econometric theory: then it's a negative signal if your papers/slides are not in LaTeX, and all the math will [eventually] be much easier in LaTeX.)

Overleaf is an online LaTeX editor that offers free accounts.[1] I've used Overleaf (and its predecessor ShareLaTeX) for many years now, and I like it because: 1) it's free, 2) it's online (so I can easily work from any computer), 3) they do all the work of updating packages and compilers, 4) you can collaborate easily (concurrent editing, etc.), 5) the

[1] Register at https://www.overleaf.com?r=63e2691f&rm=d&rs=b to get me a referral bonus!

compiler is fast. Especially for a beginner, I'd strongly recommend Overleaf over maintaining your own LaTeX system locally. Overleaf also has a WYSIWYG option that may be helpful for beginners, although I've never tried it.

The quickest way to get started is by looking at examples. Here are a few of my projects (read-only access, so don't worry about deleting something by mistake; but, you can copy code or download files):

- https://www.overleaf.com/read/sxqrmqymbktz (job market stuff)
- https://www.overleaf.com/read/snjcmmshybtk (paper and talk/slides)
- https://www.overleaf.com/read/jzkzmyvqgcqx (older paper with older slides template)
- https://www.overleaf.com/read/xzwpqpnpcmdv (other paper)

If you're starting to write a paper, I'd suggest doing what I do: take the `_paper.tex` file from one of the links above, then delete all the content but keep the **preamble** (loading packages, etc.) and structure. Just, remember to delete my name. I'd suggest doing the same for making a CV (in the "job market stuff" project), or slides, etc.

Other than examples, you can largely learn from Google. You can also browse Overleaf's learning materials,[2] which have sometimes popped up on Google for me and seem to be helpful. The StackExchange site[3] is also helpful (and is usually the first Google result).

To get you oriented within the templates/files, the main structure/elements of a paper are the following.

- Preamble: loading packages and defining macros and such; doesn't "do" anything that gets displayed directly, but gets prepared to do things later.
- Title/author/etc.
- Text: a paper is divided into different sections (Introduction, etc.); each section is started with a `\section{Section Title}` command. You can use the command `\subsection{Subsection title}` to start a subsection; subsubsections are possible but should be used rarely. In the line after such a command, add something like `\label{sec:intro}` to let you later use `\cref{sec:intro}` instead of needing to type out Section 1.
- Tables and figures ("floats"): almost all econ papers have at least one table or figure. These are sometimes called "floats" because LaTeX helps decide where to place them to look best; they are not forced to be in a particular place. Sometimes you can find a better place manually, but people (who have studied econ but not typesetting)

[2] https://www.overleaf.com/learn
[3] https://tex.stackexchange.com

are wrong more often than LaTeX, and it saves you time to just let LaTeX handle it. Within a "table" environment there is a caption, a label, and a "tabular" (the part that actually looks like a table). Within a "figure" environment, there is also a caption and label, in addition to the picture itself. Table captions go above the tabular; figure captions go below the picture. For an environment named "env" you put the code between `\begin{env}` and `\end{env}` commands. If your table has `\label{tab:OLS}` then later use `\cref{tab:OLS}` to refer to it; the numbering updates automatically.

- Math: can be either "inline" like $e^{i\pi} + 1 = 0$ or "display" (on its own line) like

$$e^{i\pi} + 1 = 0.$$

To figure out if/where to put punctuation, pretend you just replaced all the math with words (like "e to the power i times pi plus one equals zero"). There are different environments for display math, like **`equation*`** for a single line with no label, or **`equation`** for a single line with a label (and you put a `\label{eqn:Euler}` just inside the equation environment so you can refer to it later like `\cref{eqn:Euler}`). With multiple lines, you have to decide if you also want multiple labels, if you want them aligned in a particular way (like by the = in each line), etc., so there are more options like **`aligned`** and **`aligned*`** as well as nesting a **`split`** environment inside an **`equation`** environment; try to just find an example in one of my files.

- Bibliography: the actual bibliographic information is in a separate .bib file, so you just need `\bibliographystyle{jpe}` to set the style and `\bibliography{_bib}` if your other file is named **`_bib.bib`**.

1.2 Writing Advice

These are my current opinions on effective academic writing. I think they're good opinions, but they're still just opinions, not absolute truths.

Below I refer repeatedly to the well known Gopen and Swan (1990) article, "The Science of Scientific Writing."[4]

Another great resource (i.e., whose advice overlaps with mine) is from the finance professor John Cochrane.[5]

Try to remember the following five S's when you write (and revise, and revise again, and revise again...): Striving, Suppositions, Structure, Simplicity, and Segues. (Ok, some of those are not the best words, but I enjoyed starting all with S.) The first two are more high-level perspectives; the others mix in more concrete suggestions.

One over-arching theme is that the reader has a fixed time/effort budget, and you want to maximize how much they learn (about your research) subject to the budget constraint.

[4] http://stat.wharton.upenn.edu/~buja/sci.pdf
[5] https://drive.google.com/file/d/19S5BJFUY0JIMW4SQloKRVy6iMiIN7xXS/view

You don't want to waste their mental budget on tasks like parsing complex grammar or staring at a results table with 200 numbers in it.

1.2.1 Striving

What is your goal when writing? That is, what are you **striving** for? When I was an undergrad, my goal was to convince my professors that I was smart and deserved a good grade. Although I indeed got good grades, the writing was not something I'd want to read: it was too long and complex.

Instead of trying to convince the reader that *you're* smart, I suggest trying to make *the reader* feel smart. This is partly a goal in itself (people like feeling smart), but also a proxy for the goal of effectively communicating your research to the reader. Think about when you've tried to read academic papers. As a reader, which type of writing do you prefer: long, complicated, unconnected, unintuitive details, or a concise, intuitive narrative?

This emphasis on the reader (instead of the writer) and on communication (instead of presentation) appears in the aforementioned article by Gopen and Swan. In their second paragraph, they put it this way:

> The fundamental purpose of scientific discourse is not the mere presentation of information and thought, but rather its actual communication. It does not matter how pleased an author might be to have converted all the right data into sentences and paragraphs; it matters only whether a large majority of the reading audience accurately perceives what the author had in mind.

I hope the specific suggestions in subsequent sections help you achieve the writing goals that you (should) strive for.

Summary: think about the reader, and how to help them learn something and feel smart.

Discussion Question 1.1 (writing: striving)**.** In light of Section 1.2.1, discuss the following sentences that could be included in an academic paper. Suggest improvements.

a) As the reader can easily surmise from the Monte Carlo simulation study results exhibited in Tables 102–119 in the Appendix of this manuscript, an idiosyncratic pattern is manifest amongst the panoply of DGPs, wherein some computation times reflect superlative celerity yet others demonstrate inordinately pronounced durations.

b) Although this infinite-dimensional result requires additional technical considerations, the intuition follows from the following finite-dimensional example.

c) Subsequent to considerable deliberation and excessive pontification, random forest has been designated as the ML (which stands for Machine Learning) algorithm of choice in our modeling efforts to appropriately discern the complex, sophisticated relationship between the raw textual document data and the corresponding HMV (our novel acronym for "h-metric values").

1.2.2 Suppositions

Don't **suppose** your reader knows anything you didn't know before you started working on your topic. Of course, readers all have different background knowledge, but you only have one paper, so it will not be perfectly tailored to each reader. If you structure your paper well (see Section 1.2.3), it should make it easy for more familiar readers to quickly skim the parts they already know, while less familiar readers can still have enough to learn about your topic and results.

Imagining yourself as the reader helps, but it is difficult. The biggest difficulty is that you think you are a dumb grad student and the reader is a really smart professor. I don't necessarily disagree, but it is more relevant that you have spent a year (or more) working on the same, specific topic, whereas the reader may not be very familiar with your topic, let alone your specific results. Try to remember before you started working on your research: what was difficult for you to understand, what was most helpful, etc. Through your writing, you are trying to condense your year+ of learning into just minutes for your reader.

Summary: imagine going back in time to talk to yourself before you started working on your current research topic; how would you quickly and intuitively explain the background and results?

1.2.3 Structure

The most important S in writing is **structure**. Even if your diction, grammar, and spelling are bad, if your ideas are presented in an effective structure, then your reader will understand what you mean. (Even if you write "casual" when you mean "causal"!)

The structure should make it easy for the reader to skim your paper to find the content they care most about. It should be easy for them to find the "low-hanging fruit," the parts with the highest marginal benefit, which may be different parts for different readers. How often do you (yes, you) read an entire paper carefully from beginning to end? More often, you are trying to find something specific: an empirical result related to yours, a model description, a lit review, etc.

Part of making your paper easy to skim is using a conventional structure. For example, theoretical econometrics papers usually have an introduction (with lit review toward the end, along with paper structure and sometimes notation), a section with the model and assumptions, a section with theoretical results, a section with an empirical application, and a section with simulation results, before a short conclusion that includes possible extensions; and an appendix with proofs, and (now more commonly) a supplemental appendix with more proof details and more simulation or empirical results. I'm less familiar with conventional structures in other fields, but I presume they exist (ask your advisor). Besides section order, other conventional structures include putting the main theoretical results in theorems, labeling assumptions clearly, putting standard errors in parentheses below point estimates (instead of t-statistics or p-values), etc. Imagine how difficult it would be to find the main results if they were buried in the text instead of set

out in a theorem or table.

$\Longrightarrow$ The next paragraph is the single most important one! Read it twice.

Another helpful, conventional structure allocates a single idea ("topic") to each paragraph, whose first sentence ("**topic sentence**") states this idea. Of course, "topic" is ambiguous: you could argue that an entire paper is about only one "topic." Thus, length is also a factor: paragraphs should not be too long (maybe a half-page at most, and usually much shorter). If you notice yourself writing a sentence unrelated to the topic in the topic sentence, congratulations: it's time to start a new paragraph. Like other structures, the topic sentence helps readers skim: they can just read the first sentence of each paragraph, only reading further into the paragraph if the topic interests them. The topic sentence should also help connect the new paragraph to the preceding paragraph, as discussed in Section 1.2.5.

For paragraphs about figures or tables, I suggest putting the name of the figure or table first in the topic sentence. (This is less important for theorems or other "text" since you know where it will be; for "floats" like figures and tables, they may not even be on the same page as the discussion in the text.) If somebody is trying to find the discussion of Figure 1, it is very easy to find if "Figure 1" is the beginning of a paragraph. It also makes clear that the entire paragraph is about Figure 1.

Conventional structures help not only because they are usually pretty good, but also because they are what the reader expects. Gopen and Swan explicate the importance of reader expectations in great detail, drawing from cognitive psychology and linguistics. They apply the framework of reader expectations to structure at all levels: sections within a paper, paragraphs within a section, sentences within a paragraph, and clauses and phrases within a sentence. Discussion of lower-level structure is in Section 1.2.5 below.

Summary: use conventional structures, including topic sentences.

Discussion Question 1.2 (writing: structure)**.** In light of Section 1.2.3, discuss the following whole paragraphs that could be included in an academic paper. (That is, the first sentence of each example is the topic sentence.) Suggest improvements related to structure (not grammar, diction, etc.).

a) These estimates could be explained by statistical discrimination. That is, the estimates may reflect optimal decisions under uncertainty when conditioning on the observable variables. The estimates decrease with experience in the first column of Table 3. Each row in Table 3 is a cross-sectional regression for a different experience level. Following convention, (potential) experience is age minus years of schooling minus five. However, although it does not directly refute statistical discrimination, the pattern in Table 4 makes taste-based discrimination seem more plausible.

b) To compare distributions, the most common statistical tests answer one of two questions: (1) Are the distributions identical or different? (2) Do the distributions differ at the median (or another pre-specified quantile)? Often, the more interesting question is: (3) Across the entire distribution, at which quantiles do the distributions differ? Figure 1 illustrates the difference among these questions. Alternatively,

instead of quantile functions, one could compare CDFs. However, in two-sample settings, the theory is simpler for CDFs, due to Donsker's Theorem. This provides an asymptotic Gaussian process approximation for the centered and scaled empirical CDF.

c) Increasingly, economic datasets are too large to fit on a single computer. For example, text data is often converted into many regressors using individual word or multi-word phrase frequencies. A particular version of this known as term frequency–inverse document frequency (TF–IDF) has been around for decades but proved very successful. For example, TF–IDF has high accuracy for author classification, at least for prose. For poetry, this so-called "bag-of-words" aproach is less appropriate. Instead, stylistic features like sound devices (rhyme, alliteration, etc.) and part-of-speech frequencies take prominence.

1.2.4 Simplicity

Readers have (very) limited attention budgets. There may be outliers, but most will spend only maybe 10 minutes with your paper. It requires mental energy to think about your research results, and to read your paper. If you can write **simply** and minimize the energy required for reading, then the reader has more energy to think about your actual research, which is what you want.

Remember that while you (will) have spent 1–2 years working on your research topic, most of your readers may not even be familiar with the topic at all. (This is especially true when you are trying to get a job.) As suggested in Section 1.2.2, try to remember when you were first learning about your research topic. Always think about ways to simplify, while retaining the core implication or intuition of your results. Write mostly about a special case that captures the intuition. Write mostly about an empirical specification that's a little too simple (in your opinion) but gets similar results. Then present your general results or all your sensitivity checks and alternative specifications, maybe partly in the appendix (or supplemental appendix).

Here are a few specific ideas for keeping things simple and easy to read.

- Put the subject and verb together. Gopen and Swan lament that having widely separated subject and verb is an "all-too-common structural defect," also noting, "Readers expect a grammatical subject to be followed immediately by the verb."

- Try to use as few commas as possible, and (almost) never use dashes. (The journal *Biometrika* forbids dashes.) Obviously, you should use commas wherever grammatically required, but sometimes you can move around phrases to eliminate the need for commas.

- Use short words. English has many words, some with nearly identical meanings. For example, write "use" instead of "utilize," and "titled" instead of "entitled" (if referring to a paper). Short words save the reader time and energy. Long words (with identical meaning) make the reader think you're trying to sound sophisticated at

the expense of communicating your ideas efficiently. (Ok, probably not all readers, but that's how I feel.) Avoid sesquipedalian writing! (That was a joke.)

- Similarly: shorten excessively long phrases. Sometimes it's not a single long word, but a phrase of many short words, that can be replaced by something shorter.
- Paragraphs can be short. If you only need two short sentences to say what you want about a topic, that's fine. Don't waste the reader's time/attention with unnecessary detail.
- Don't write everything you know. Think about the most important 2–3 points you want to communicate, and ask whether each sentence you write contributes to those points. If not, maybe delete it, or at least relegate it to a footnote or appendix (that only very motivated readers will look at).

Summary: simplify.

Discussion Question 1.3 (writing: simplicity)**.** In light of Section 1.2.4, discuss the following sentences that could be included in an academic paper. First, identify the problem(s), using the above list. Second, suggest improvements.

a) Computational quagmires notwithstanding, I venture to put forth the suggestion proposing that estimation utilize GMM.
b) The shortest of the unemployment durations, defined as the number of business days without any reported earnings (regardless of "actively seeking" employment or not), in the Current Population Survey 2010–2011 dataset are associated with the lowest education levels, even controlling for wage and experience.
c) Such rates would, intuitively, be pro-cyclical, going down in recessions—defined, e.g., per the NBER dates, which, though not perfect, are widely used—and in expansions, going up, usually.
d) The important feature is the non-zero skewness, which can actually be derived analytically by using Skorohod's representation $X = F_X^{-1}(U)$ for $U \sim \text{Unif}(0,1)$ and the implication for order statistics $X_{n:k} = F_X^{-1}(U_{n:k})$ along with the skewness $2(\beta-\alpha)\sqrt{\alpha+\beta+1}/[(\alpha+\beta+2)\sqrt{\alpha\beta}]$ of the underlying $U_{n:k} \sim \text{Beta}(\alpha,\beta)$ with $(\alpha,\beta) = (k, n+1-k)$, although the "central" order statistic asymptotics specifies $k/n \to \lambda \in (0,1)$ rather than allowing $k/n \to 0$ or $k/n \to 1$ as in the "intermediate" or "extreme" order statistic asymptotics (the latter of which even allows fixed k or fixed $n-k$ as $n \to \infty$).

1.2.5 Segues (and Sentence Structure)

Each sentence should have **segues** (transitions) to link the previous and current ideas. This is especially true for "topic sentences" that link the prior paragraph to the current paragraph.

These segues should appear in the "topic position" (Gopen and Swan's term), i.e., the beginning of a unit of structure (like a paragraph or sentence). The first part of the

sentence provides the context for the new information that you provide in the second part of the sentence. Part of that context is the relationship with the prior sentence (or paragraph): maybe you are adding supporting evidence, or moving to a different property of the same estimator, or providing evidence that in fact contradicts what you just said, etc.

Here are some examples of transition words or phrases, along with the relationship they imply.

- "However": something providing the opposite argument. Example: "The KS test has all these great properties. However, it has low power in the tails."
- "In contrast" or "Alternatively": something different. Example: "The KS test only tests 'if' two distributions differ. In contrast, `distcomp` tests 'where' two distributions differ."
- "That is": explaining the same idea another way. That is, offering a different perspective on the same substantive content.
- "For example": providing an example to support the prior idea. For example, this sentence.
- "More specifically" or "Specifically" or "Further" or "Moreover" or "Additionally": adding details to the prior idea. Example: "This model allows for observable heterogeneity through the interaction terms. Additionally, it allows for unobservable heterogeneity through the random coefficients."
- "More generally": generalizing the prior idea. Example: "The first sentence of a paragraph is called the topic sentence. More generally, the beginning of any structural unit is called the topic position."
- "The corresponding [something]": like, "The KS MTP implicitly weights the tails much less than the middle of the distribution. The corresponding uneven allocation of pointwise power. . . ."
- "This [something]": like, ". . . the beta CDF evaluated at τ. This CDF can be computed by. . . ."
- "Therefore" or "Consequently": a logical implication. Example: "Readers automatically put emphasis on the last part of a sentence. Therefore, the last part of your sentences should contain important information."

The topic position ideally also provides a perspective from which to see the subsequent information. That is, whose story is this? Gopen and Swan write:

> Readers expect a unit of discourse to be a story about whoever shows up first. "Bees disperse pollen" and "Pollen is dispersed by bees" are two different but

> equally respectable sentences about the same facts. The first tells us something about bees; the second tells us something about pollen. The passivity of the second sentence does not by itself impair its quality; in fact, "Pollen is dispersed by bees" is the superior sentence if it appears in a paragraph that intends to tell us a continuing story about pollen. Pollen's story at that moment is a passive one.

(Note: "passivity" refers to the grammatical term "passive voice," referring to verb constructions like "is dispersed" or "was increased," contrasting the "active voice" like "dispersed" or "increased.")

After establishing the context in the "topic position," you can put the new information in the "stress position" (Gopen and Swan's term) at the end of the sentence. Gopen and Swan write, "It is a linguistic commonplace that readers naturally emphasize the material that arrives at the end of a sentence." So, if readers emphasize the end of the sentence simply due to its location, you need to make sure that the content you put there is worthy of emphasis. More informally, Gopen and Swan describe this by the aphorism, "Save the best for last."

Summary (Gopen and Swan): "Put in the topic position the old information that links backward; put in the stress position the new information you want the reader to emphasize."

Discussion Question 1.4 (writing: segues)**.** In light of Section 1.2.5, consider the following sentence pairs. Suggest a transition word or phrase (e.g., "However," or "That is") to add to the beginning of the second sentence.

a) Table 1 shows an increasing pattern. Table 2 shows a decreasing pattern.
b) Unemployment means zero hours worked. Earnings are zero.
c) Generally, returns to education depends on unobserved "ability." College might increase human capital more for individuals with high ability, which would result in higher future earnings.
d) The problem is misspecification. The implicit assumption of constant partial effects $\frac{\partial m(\boldsymbol{x})}{\partial x_k} = \beta_k$ is incorrect.
e) Latent stochastic dominance implies ordinal stochastic dominance. Ordinal dominance is necessary (but not sufficient) for latent dominance.
f) An individual's ordinal health status can be modeled in terms of a latent, continuously distributed health value. Any ordinal variable can be modeled in terms of a latent variable.
g) Asymptotically, OVB equals $\text{plim}\,\hat{\beta} - \beta = \rho\delta$. OVB is zero if (but not only if) $\rho = 0$.

Discussion Question 1.5 (writing: sentence structure)**.** In light of Section 1.2.5, discuss the following sentence pairs that could be included in an academic paper. Suggest improvements.

a) Intuition may suggest a positive coefficient. This is wrong since such intuition ignores the substitution effect, accounting only for the income effect, like how people spend more on housing when their income increases.

b) Sales tax receipts often form the majority of state government budget revenue. In this paper, I examine annual revenue for 38 states, to see how they were affected by the Great Recession.
c) In turn, this results in a phenomenon called "budget compression," where salaries of new and very senior employees are actually very similar. This actually accurately reflects productivity, despite it seeming counterintuitive to have very similar salaries for more and less experienced employees, who may also have different job titles.

1.2.6 Summary

These are the seven guiding principles from Gopen and Swan, plus a few more.

1. Follow a grammatical subject as soon as possible with its verb.
2. Place in the stress position the "new information" you want the reader to emphasize.
3. Place the person or thing whose "story" a sentence is telling at the beginning of the sentence, in the topic position.
4. Place appropriate "old information" (material already stated in the discourse) in the topic position for linkage backward and contextualization forward.
5. Articulate the action of every clause or sentence in its verb.
6. In general, provide context for your reader before asking that reader to consider anything new.
7. In general, try to ensure that the relative emphases of the substance coincide with the relative expectations for emphasis raised by the structure.
8. Use conventional structures for sections, tables, paragraphs (topic sentence), etc.
9. Consider the reader's perspective (and limited attention), and try to make them feel smart.
10. Simplify.

Try to follow all of these. If it seems too hard for a specific sentence you're writing, try again. If it still seems too hard, think about the overall goal (communicating with the reader) and which principles best help achieve that in your specific sentence, and don't worry about ignoring the rest. You will certainly come back to that sentence again, and you will probably understand your own research better when you do, which will make the writing easier to revise, too.

Discussion Question 1.6 (writing: summary)**.** In light of all you've now learned, discuss the following whole paragraphs that could be included in an academic paper. Suggest improvements.

a) This identification strategy relies critically on the program's staggered rollout schedule being "as good as random." For example, it is problematic if regions with the largest treatment effects were treated first. In fact, any association (even if, say, rollout depended on observables, which then, in turn, are correlated with treatment effects) between rollout schedule and potential outcomes, whether directly causal or a purely "statistical" association, precludes causal identification.

b) Despite this literature, the causal effect of stay-at-home orders during COVID-19 remains an open question. Part of the difficulty is simply articulating the appropriate counterfactual. For example, the stay-at-home order in Columbia, MO was initially issued in late March. It essentially closed all businesses deemed non-essential, although some residents disagreed with the stated classifications. Further, it closed certain amenities like playgrounds and even tennis courts. Even the revised, more lenient stay-at-home order a month later failed to re-open tennis courts. As far as I can tell from the medical and epidemiological literature, there do not seem to have been any documented cases of transmission due to tennis, or any other outdoor, net-based sports in which the ball is primarily (though not exclusively) only contacted by implements such as racquets (i.e., pickleball).

c) However, this non-rejection of $H_0 \colon \beta = 0$ does not mean the articulation agreement has zero effect. Type II error could explain the non-rejection result. The standard error is large. The 95% confidence interval includes negative values. The interval includes values as large as 73 additional nursing bachelor of science degrees per year. 73 and negative together suggest bifurcated beliefs on burgeoning bachelors bereft of bombast; *quod erat demonstrandum*.

1.2.7 Shawn's Suggestions (bonus!)

These suggestions are courtesy of Shawn Ni. He has advised countless PhD students at Mizzou and led the PhD Research Workshop for many years. I added some details, so any errors are probably mine.

Outline: start with an outline. Write the names of each section, usually something like:

- Introduction: what you do, why you do it, roughly how you do it (no math, very few details), and what you discover
- Literature Review: for journal articles this would be shortened and moved inside the Introduction, but for PhD research it helps to have a separate, longer lit review
  - Purposes: make sure you're doing something new/different; make sure it's important
  - Organization: discuss papers in groups based on different issues related to your own research (e.g., the data, the methodology, the relevant policy, etc.)
  - If you can find a recent paper on generally the same topic, you can read their literature review, as well as the (relevant) papers they cite
  - Be complete: find *all* the existing papers with the same research question as you

  - Conversely: if something isn't related to your research question, don't cite it (even if written by someone famous)
  - Use Google Scholar to find related papers (published and unpublished)
  - For good higher-level reviews with many references, check the *Handbook of ...* from Elsevier.[6]
- Model; Data (if applicable); Results; Conclusion

possibly with other sections inserted as needed (Simulations, Identification Strategy, etc.). Then within each section, write an outline. For example, following the above suggestions for the Introduction, your outline might be:

1. WHAT I DO: propose new statistical inference for "stochastic dominance" based on expected utility comparisons
2. WHY: compared to existing CDF-based SD tests, has more statistical power/precision and more economic interpretation
3. HOW: establish Donsker property of utility function classes, apply empirical process theory
4. DISCOVER: new, valid bootstrap confidence sets and multiple testing to compare two distributions

I use UPPERCASE (sometimes for all the text) to indicate it's the outline, not the final text. Also, each block may cover multiple "topics" that may require multiple topic sentences (e.g., confidence sets, multiple testing).

Reporting results: when you discuss tables/figures of results in the text, you should (at minimum) discuss the following.

1. How they were constructed.
2. What do the numbers or lines mean?
3. Why are you presenting the table or figure; what are we learning?

Don't claim too much. Empirically, you cannot conclusively prove or demonstrate anything; you can show results that are consistent with some model, and you can estimate values and quantify your degree of uncertainty. Theoretically, there is almost always a trade-off: stronger assumptions are less realistic but allow stronger conclusions. Readers appreciate you being honest and transparent about the assumptions; if you try to hide your assumptions, they will be suspicious and not appreciate having to work harder to understand your assumptions.

Use present tense, even for past papers in the literature and things you already did in the past. ("I find," "This shows," "They propose," etc.)

Be concise: "because" instead of "based on the fact that," "now" instead of "at the present time," etc.

Change negatives to affirmatives: "similar" instead of "not different"; "different" instead of "not the same"; "prevent" instead of "not allow"; etc.

Be precise.

"Use quoted material accurately and sparingly" (Ni, 2020). Usually, you should refer

[6] https://www.sciencedirect.com/browse/journals-and-books?contentType=HB&subject=economics-econometrics-and-finance

to others' work in your own words instead of directly quoting them. In the rare case when you want to use their words verbatim (exactly as written), make sure to put it in quotes. You need to cite the other paper in either case; ideally, put the page number, partly for your own reference (in case you later wonder, "Where did those authors talk about that particular result again?").

"Read what you have written out loud. If it sounds bad it probably is." (Ni, 2020)

"When in doubt, look it up" (Ni, 2020): personally, I (Dave) use the Google dictionary frequently to check if a word means what I think. That said, if you aren't sure of the meaning, then possibly many readers also would not know the meaning; if there is a simpler word with basically the same meaning, then you should probably use that.

Citations and references: I use LaTeX, specifying style `\bibliographystyle{jpe}` before the `\bibliography{_bib}` at the end of my paper, where the file `_bib.bib` contains all the bibliographic information; you can see examples of .bib entries in the Overleaf projects linked in Section 1.1. I suggest including the URL for your own reference, so you (and other readers) can easily click to the paper from your own paper. If you need to cite something other than an article (e.g., a chapter within a *Handbook*), you can ask me or Google it; you can see some examples of other entry types at https://verbosus.com/bibtex-style-examples.html

1.3 Plagiarism

Plagiarism is a very, very serious offense in academia, even if it is committed unintentionally. Thus, it is your responsibility to understand it. The MU library website has resources to help you understand and avoid plagiarism.[7]

1.4 Common Minor Mistakes

Here are some common minor mistakes I've seen in students' research papers. But, when you are starting, don't worry too much about these; with writing, for now it is much better to have high quantity and low quality than high quality and low quantity. It is inefficient to worry about the small details when you are just starting a project and don't even know what your main results will be; even when you do, you will end up revising many times (for other reasons), so you can wait until the end to really "polish" your paper and perfect the details of grammar and spelling and everything.

1. Typos in authors' names. If you use a bibliography manager (like BibTeX or BibLaTeX), you only type the authors' names once (e.g., in the .bib file), so you are much less likely to make a typo, and even if you do, you can easily fix it (just change the .bib entry).

[7] https://libraryguides.missouri.edu/plagiarism

2. It is simply "University of Missouri" and not "University of Missouri–Columbia" as stated in the official MU style guide.[8]

3. Always put a space before acronyms or other things in parentheses, like "ordinary least squares (OLS)" instead of "ordinary least squares(OLS)".

4. If you have an abbreviation with periods in it followed by a space, then you need to put a backslash in the LaTeX code after the last period, otherwise it thinks you're starting a new sentence and inserts too much space. Comparing `Dr.\ K` to `Dr. K`: with backslash is Dr. K and without is Dr. K.

5. Double quotation marks: the opening one is ``` `` ``` (two backticks) and the closing one is `''` (two apostrophes). If you use `"` (double quote character) then it looks different, and using `''` for the opening one is backwards: "wrong" "wrong" "right" from code `"wrong" ''wrong'' ``right''`.

6. Percent and percentage point are different units; be careful.

7. Numerical ranges: use an "en dash" like 5–8 (made by typing two hyphens `--` in the .tex file) instead of a hyphen like 1-3. This applies to calendar years, too.

8. If you have an acronym in an equation, don't just type it, or it gets interpreted by LaTeX as the product of variables; use something like `\mathrm` or `\textup`. Example: instead of $FWER = \alpha$, write $\mathrm{FWER} = \alpha$; note the spacing is more even in the second example (the first one has too much space between W and E).

9. It can be confusing when to use "that" instead of "which" (and when to have a comma).[9]

10. Citations: pretend the year isn't even there, and you are just referring to the authors; and use present tense. So, write things like "Kaplan and Blei (2007) analyze poetry" (not "analyzed" or "analyzes"), or "Kaplan (2015) establishes an Edgeworth expansion" (not "established" or "establish" or "contains"). But sometimes the present tense feels really weird and I use past tense. Like, "Well over a century before more sophisticated analysis like that of Banks, Blundell, and Lewbel (1997), the idea originally was explored by Engel (1857)"; it would not make sense to say, "Long before this, the idea is explored by Engel (1857)." But when in doubt, use present tense.

11. Plurality of "data": whatever. Either is fine (if you are consistent with your choice). "The data say…" or "The data shows…," well, unless you interview with NERA, then always treat it as plural =)

[8] https://styleguide.missouri.edu/term/university-identification
[9] http://blog.apastyle.org/apastyle/2012/01/that-versus-which.html

12. An elipsis is written by `\ldots` not `...`, otherwise the spacing is wrong, whether in text or math mode. Compare (right then wrong): this. . . and...that; $1, \ldots, k, ..., n$.

13. In math mode, I'd suggest just using `\dots` instead of guessing whether to use `\ldots` or `\cdots`, since it usually gets it right automatically; using only `\dots`: $1, \dots, n$ and $1 + \cdots + n$.

14. Not using BibTeX (or BibLaTeX) always leads to typos and other problems. But even with BibTeX, double-check the capitalization and such in your references. Google Scholar's .bib entries are usually close but often slightly wrong. For example, journal titles should always be title case, like "Journal of Health Economics" instead of "Journal of health economics". And usually the leading "The" should be omitted from journal titles (but isn't on Scholar), like *Review of Economic Studies* instead of *The Review of Economic Studies*.

15. I suggest writing probabilities like $\mathrm{P}(\cdot)$ instead of $P(\cdot)$, since upright P looks like an operator while slanted P looks like a variable. Similarly for $\mathrm{E}(\cdot)$ and $\mathrm{Q}_\tau(\cdot)$. But CDFs and quantile functions are functions, not operators on random variables, hence $F_Y(\cdot)$ and $Q_Y(\cdot)$.

16. Don't use $*$ for multiplication. If you really need an explicit symbol (e.g., if you have a product continued over multiple lines), use $\times$.

17. It's "et al." (not et al, et. al, or et. al.); "et" means "and" in Latin, and "al." is an abbreviation of alia/alii/aliae. But: you should almost never be typing this yourself anyway, because the `\citet` and `\citep` commands will do it for you.

18. You should put punctuation around (and in) math as if you had written the math out in words. For example, sentences end with periods, so even if your sentence ends with math, it should always have a period. For inline math, the period should be outside the inline math environment (otherwise LaTeX thinks it's a decimal point and the spacing is wrong); for "display math," the period goes inside the equation environment (or align or gather or whatever environment). Example: this properly ends with $x = 0$. Incorrectly: $x = 0$. Note the different spacing. Other example: if when reading your paper you'd say, "The equation 'y equals x' is interesting," then you should not put a comma or colon or anything after the word "equation" even if you write $y = x$ in an equation environment; e.g., you shouldn't write "the equation: $y = x$ is interest" or "the equation, $y = x$ is interesting."

19. Periods always go inside quotation marks. (This is not 100% true, but probably at least 99% for economics writing.) So, "Inside here." Not, "Outside".

20. Never start a sentence with math. (I don't think this is a great rule, but some people care deeply about it.)

21. The Latin abbreviation for "exempli gratia" ("for example") is e.g., and it should usually be followed by a comma, e.g., like this. The Latin abbreviation for "id est" ("that is") is i.e., and the same comment applies, i.e., it should usually be followed by a comma. At the beginning of a sentence, it's better to write out the English "For example" or "That is." For example, this sentence.

22. After a colon, do *not* capitalize the next word: just lowercase since it's the same sentence.

Exercises

Exercise E1.1. At https://www.overleaf.com/read/gtfzpkwrzhhw get the template `JMC_cv.tex`. Use it to create a CV for yourself.

Exercise E1.2. At https://www.overleaf.com/read/gtfzpkwrzhhw get the template `_paper.tex`. Use it to type up one day of your lecture notes from ECON 9474 or ECON 9477 (if you've taken it yet) or any other ECON 9xxx class that does not already have typed lecture notes.

Exercise E1.3. At https://www.overleaf.com/read/gtfzpkwrzhhw get the template `_talk.tex`. Use it to create slides based on one day of lecture for any ECON 9xxx class you've had that did not have lecture slides (i.e., professor just wrote on blackboard).

Chapter 2
R: Some Basics

Unit learning objectives for this chapter

2.1. Download, run, and maintain R software [TLO 4]

2.2. Write/run/save new data analysis with .csv or .dta data [TLO 4]

2.3. Write/run/save new Monte Carlo simulations [TLO 4]

2.4. Learn new things on your own [TLO 4]

Warning: this chapter has lots of simplifications, which generally I dislike, but you can always look at the help file for any function I mention to learn more details, or Google any topic.

Optional resources for this chapter

- Chapter 1 of Kaplan (2022b), especially for details about getting started (including linked video)

2.1 Getting Help

At first, it may help to have some quick reference "cheat sheets."[1,2]

Eventually you'll just Google to learn, but one of the following free tutorials may help you get started.

1. Section 2.3 ("Lab: Introduction to R") in James, Witten, Hastie, and Tibshirani (2013)

[1] https://www.rstudio.com/resources/cheatsheets/
[2] https://cran.r-project.org/doc/contrib/Short-refcard.pdf

2. Section 1.1 in Hanck, Arnold, Gerber, and Schmelzer (2018)
3. Sections 1.1–1.3 in Heiss (2016)
4. Sections 2.1–2.5 in Kleiber and Zeileis (2008) [Chapter 2 is available free on their website]
5. Cyclismo[3]
6. CRAN[4]
7. No longer free beyond first chapter: courses at `datacamp.com` like Introduction to R.[5]

Help within R is usually helpful. For example, type `help(lm)` or `?lm` to learn about the `lm` function.

2.2 Getting Started

2.2.1 Running R

In a web browser: currently (Fall 2020) the best option seems to be RStudio Cloud.[6] A free account is required to sign in. However, there is a limit to how many hours you can use it for free each month. There are also other free online options like CoCalc.[7]

Download R for Windows: Google "r windows" and try the first result.[8]

Download R for Mac: Google "r mac" and try the first result to find the newest .pkg.[9]

Download RStudio (free, nicer interface): Google "rstudio download" and try the first result.[10]

2.2.2 Packages

In addition to "base" (or "core") R, there are freely downloadable **packages** for additional functionality. These are like Matlab toolboxes (but free), or like the Stata commands that you download with `ssc install`. You can download/install/update R packages easily through RStudio (in the Tools menu) or the `install.packages()` function.

Both the base and (many) packages are being constantly updated (every month?). Updating is usually not critical, but one time I sent a silly email to one of the package owners (sorry Jeff Racine!) about a "bug" that was simply due to my not having fully updated both the base and package; please learn from my mistake.

Even after you download/install a package, you must still explicitly load it in each R script in which you want to use it. You can also load code (say, function definitions) from

[3] `http://www.cyclismo.org/tutorial/R`
[4] `http://cran.r-project.org/doc/manuals/r-release/R-intro.html`
[5] `https://www.datacamp.com/courses/free-introduction-to-r`
[6] `https://rstudio.cloud`
[7] `https://cocalc.com`
[8] Currently `https://cran.r-project.org/bin/windows/base`
[9] Currently `https://cran.r-project.org/bin/macosx`
[10] `https://rstudio.com/products/rstudio/download`

another .R file. For example, imagine you have already installed the `quantreg` package with the command `install.packages('quantreg')`. To load the `quantreg` package as well as the functions in file `ivqr_see.R`:

```
library(quantreg)
source("ivqr_see.R")
```

This assumes the .R file is in the "working directory"; you can check the current working directory with `getwd()` and set it with `setwd()`. If you double-click a .R file to open RStudio, I think RStudio sets the working directory to wherever that .R file is.

2.2.3 RStudio Interface

When you open RStudio, you should see a few **panes** within the window. The **console** should show some basic info on your version of R, and have a **command prompt** below that, which is a single `>` symbol. If you type a statement here and hit enter, then R will do something in response. There should be another two panes that have multiple uses, like showing graphs (**plots**) and help. You can customize in the RStudio options what these display.

You can also open an **editor pane** for editing .R files (like .do or .m or .sas files), by going to File–New File–R Script. You can also run commands from the editor in the console, by highlighting one or multiple lines and hitting Control-Enter (in the menus: Code–Run Line(s)). So when you are first writing a .R file, you can test each new line of code this way (or just copy-paste into the console if you wish). There are keyboard shortcuts to toggle the focus across different panes (e.g., Windows Control-1 puts the cursor in the editor pane, Control-2 puts it in the console), which I find very helpful. You can also set an RStudio option for whether or not to toggle the focus to the console after running code from the editor pane.

Following convention, I usually show the command prompt when showing R code and results. If you copy-paste code to run yourself, then don't copy the command prompt. For example, if I show

```
> ?help
```

then just type `?help` into the console and hit Enter. Incidentally, if you do this, then you should see a help file on "help" itself appear in one of the panes.

2.2.4 Readability

Making your code "readable" is important, for multiple reasons. It will help you structure your code better, and help you debug more easily. If you're working with somebody else, it helps them if it's easier to understand what your code does (or, is supposed to do). Even if you're working alone, academic research projects take a very long time to complete, so you're basically still working with "somebody else": your future self! Especially as a beginner, you should use lots of comments to remind your future self of what you were trying to do with each piece of code. For example: http://xkcd.com/1421

There are some basic ways to improve readability. The best way is to add **comments** to your code (like in the linked xkcd comic). The symbol `#` makes the rest of a line (after the symbol) into a comment that is ignored by R. This allows you to write notes to yourself about what a line/block of code is supposed to do, or what assumptions you're making, etc.

Structuring your code visually can also help. You can put multiple expressions in the same line if they are separated by a semicolon. This has potential to improve readability if you have lots of very short lines consecutively, since then you can see more of the code in one screen. But, this may also make it harder to see certain "lines" of code; there is a tradeoff. In fact, it is often helpful to do the opposite: insert blank lines to divide sections of code. Another perk of RStudio is that it automatically indents code inside loops (and such), which improves readability. Putting spaces after commas (in comma-separated lists, e.g., arguments to a function) can help. You can also break long lines into multiple lines as long as it's "obvious" to R that the line isn't finished; e.g., if the first line is `x <- rbeta(n=5,` then R knows it's continued on the next line because you haven't closed the parentheses yet.

Finally, it can help to write your own functions. For example, imagine you need code to load and prepare raw data, run regressions, and then save results. You could just write this all into one long script. Alternatively, you could define three new functions, say `load.prep.data()`, `run.regressions()`, and `save.results()`. Then, your script would call these three functions (after setting directories and such), so the high-level structure would be clear. The functions would then be defined below, or possibly even in other files, which could be loaded with `source()`. However you choose to do it, explicitly clarifying the high-level structure makes it easier to understand each individual line of code.

2.3 Data Types

See also:

- Cyclismo: data types[11]
- CRAN: commands, case sensitivity[12]
- CRAN: simple numerical manipulations[13]
- CRAN: arrays and matrices[14]
- CRAN: lists and data frames[15]
- Cyclismo: vector indexing[16]

[11] http://www.cyclismo.org/tutorial/R/types.html

[12] http://cran.r-project.org/doc/manuals/r-release/R-intro.html#R-commands_003b-case-sensitivity-etc

[13] http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Simple-manipulations-numbers-and-vectors

[14] http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Arrays-and-matrices

[15] http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Lists-and-data-frames

[16] http://www.cyclismo.org/tutorial/R/vectorIndexing.html

You can define **variables** that can store different types of data. For example, the value 4 can be assigned to variable `x`. Since 4 is a number, R infers that `x` should be a "numeric" variable. Specifically, the **data type** of `x` is **double**:

```
> x <- 4
> x
[1] 4
> typeof(x)
[1] "double"
> is.numeric(x)
[1] TRUE
```

If we assign a different value, R will switch the data type accordingly; or we can **coerce** a variable to a particular data type.

The `<-` is the assignment operator. It looks like a left-pointing arrow. As its shape suggests, it assigns the value on the right-hand side to the variable on the left-hand side. In RStudio, you may hold down the Alt key and hit the hyphen key - to insert this operator (padded by a single space on each side). Historically, `<-` was the only assignment operator, but now `=` also works. Although `=` has other meaning in other contexts, I think the ambiguity is minimal.

Variable naming is similar to Stata/Matlab/etc. except that names can contain periods. (Historically, they could not contain underscores, but now they can.) To improve readibility for people more familiar with other languages, you could consider only using names that are valid in Stata/Matlab, too. Variable names are case sensitive, must begin with a letter or period (but not period followed by number), and can contain letters and numbers (and period and underscore).[17] It is helpful (to your forgetful future self) to give variables descriptive names. For example, `state.abbrev.lookup` may be easier to understand than `STlk`; the time you save by typing 15 fewer characters may be lost later trying to remember what `STlk` is.

There are three kinds of special values for numeric types: `NA`, `Inf`, and `NaN`. `NA` means "missing data" like the `.` value in Stata. `Inf` is positive infinity, while `NaN` stands for "not a number"; type `?NA` or `?NaN` for more.

Another special value is `NULL`. It is the ultimate nothing, beyond even `NA` and `NaN`. It is mostly helpful for error handling. There is an `is.null()` function:

```
> is.null(NULL)
[1] TRUE
> is.null(NA)
[1] FALSE
> is.null(4)
[1] FALSE
```

[17] http://cran.r-project.org/doc/manuals/r-release/R-intro.html#R-commands_003b-case-sensitivity-etc

There are other types of variables we can have. A variable can store text; either double-quotes or single-quotes are fine for text expressions (`'this'` or `"that"`). A **logical** variable stores `TRUE` or `FALSE` (or `NA`) values. You can store calendar dates or times. You can have a vector (like a vector in math) of any of these, or a two-dimensional matrix, or a higher-dimensional array; see `?matrix` or `?array` for help. You can also have a **list** containing elements with different data types; see `?list`. A **data frame** is similar to a Stata dataset: like a matrix, where each element within a column has the same data type, but different columns can have different data types, and you can refer to each column (i.e., each variable in your dataset) by its name, like `dataset$var1` or `dataset[,'var1']`. Built-in functions that load data usually return a data frame.

Square brackets `[]` are used to **index** vectors, matrices, arrays, and data frames, i.e., to extract a subset of the elements. You can find many examples online, but for example `m[3,2]` returns the row 3, column 2 element of matrix `m`, whereas `m[,2]` extracts the entire second column. Logicals can also be used for indexing; e.g., `x[c(FALSE,TRUE,TRUE)]` returns the 2nd and 3rd elements of `x`. You can extract multiple named columns from a data frame with something like `d[,c('age','edu','wage')]`.

2.4 Basic Data Manipulation

See also: http://www.cyclismo.org/tutorial/R/basicOps.html

2.4.1 Numerical Operations

Most of the numerical operators are relatively intuitive. For example, `2+2`, `4/2`, `2*2`, `2^2`, `sqrt(4)`, `exp(1)`, `log(2.71)`, `log10(100)`, `abs(-2)`, `floor(2.9)`, `ceiling(1.1)`, `round(2.49)`. A few less intuitive things: `%%` for modulo/remainder; `round(2.5)` is actually 2 (numbers ending in 0.5 are rounded to the nearest even integer); matrix multiplication is `%*%` whereas element-wise matrix multiplication is simply `*` (unlike Matlab).

You can generate consecutive integers with a colon like `1:3`, or just write out `c(1,2,3)`, or use `seq(from=1,to=3,by=1)`. You can repeat values/sequences like `rep(1:3,each=2)` or `rep(1:3,times=2)` (these results differ; try them). You can fill a matrix like `matrix(1:6,nrow=3)` or equivalently `matrix(1:6,ncol=2)`.

2.4.2 Combining Data

For combining vectors and matrices, `rbind()` (appending rows) and `cbind()` (appending columns) are helpful. Matrix transpose is `t()`. E.g., compare the results of `cbind(1:3,4:6)` versus `rbind(1:3,4:6)` versus `t(cbind(1:3,4:6))`.

2.4.3 String Manipulation

In principle you could just Google all this, too, but personally I've found it more difficult to Google things related to string manipulation, so I've included more examples in this

particular subsection.

To combine strings, I suggest `paste0()`:

```
> paste0("Hello, ","world")
[1] "Hello, world"
> paste0(c("Hello,","world"))
[1] "Hello," "world"
> paste0(c("Hello,","world"),collapse=" ")
[1] "Hello, world"
```

Substring:

```
> substr("abcdefghij",start=3,stop=6)
[1] "cdef"
```

Substitution:

```
> sub(pattern=".txt",replacement=".pdf",x="filename.txt")
[1] "filename.pdf"
```

Length:

```
> nchar("abcde")
[1] 5
> length("abcde")
[1] 1
```

The `sprintf()` function is very helpful. It helps you construct strings using values from variables computed in your code. If you just have single numbers (not vectors), use `%d` as a placeholder for integers and `%g` for decimal numbers, followed by the variables (in the same order):

```
> x <- 41;  y <- 5.2
> sprintf("x=%d and y=%g", x, y)
[1] "x=41 and y=5.2"
```

If you are trying to align things, you can specify, for example, that an integer be padded with whitespace to take up 5 characters (even if it's only two characters) by `%5d`. To pad with zeros instead, `%05d`. You can also specify for decimal numbers the total number of characters and how many should come after the decimal, like `%5.2f` for five total and two after the decimal. Continuing from above:

```
> sprintf("x=%5d and y=%5.2f", x, y)
[1] "x=   41 and y= 5.20"
> sprintf("x=%05d and y=%5.1f", x, y)
[1] "x=00041 and y=  5.2"
```

Alternatively, you can just have R print lots of decimals and take care of rounding in LaTeX with the `S` column type from the `siunitx` package.

You can also insert strings:

Table 2.1: Table generated from R output.

	Bias			Variance		
Method	DGP1	DGP2	DGP3	DGP1	DGP2	DGP3
OLS	−1.48	−1.39	1.13	2.05	−0.86	−0.51
GMM	0.73	−0.42	0.38	−1.40	0.91	−0.52

```
> w="world"; sprintf("Hello, %s", w)
[1] "Hello, world"
```

Strings also support the fixed-width specification, and you can add a minus sign to add the space padding to the right instead of left:

```
> sprintf("%11s","what")
[1] "       what"
> sprintf("%-11s","what")
[1] "what       "
```

You can also pass vectors to `sprintf()`, in which case the output is a vector of type character:

```
> sprintf("Hello, %s",c("world","Dave"))
[1] "Hello, world" "Hello, Dave"
```

These can in turn be combined into a single text string with `paste0`:

```
> paste0(sprintf("Hello, %s",c("world","Dave")),collapse="; ")
[1] "Hello, world; Hello, Dave"
```

Table 2.1 is generated partly from the following R output. The R output is formatted to be pasted directly into the .tex file:

```
> set.seed(112358)
> head0 <- "\\begin{tabular}{lSSSlSSS}\n\\toprule"
> head1 <- paste0(" & \\multicolumn{3}{c}{Bias}",
+                 " && \\multicolumn{3}{c}{Variance} \\\\")
> head2 <- "\\cmidrule{2-4}\\cmidrule{6-8}"
> s <- "{DGP1} & {DGP2} & {DGP3}"
> head3 <- sprintf("Method & %1$s\n      && %1$s \\\\", s)
> head4 <- "\\midrule"
> bias <- list(OLS=rnorm(3), GMM=rnorm(3))
> SE   <- list(OLS=rnorm(3), GMM=rnorm(3))
> OLSstr1 <- paste0(sprintf("%5.2f", bias$OLS), collapse=" & ")
> OLSstr2 <- paste0(sprintf("%5.2f",   SE$OLS), collapse=" & ")
> OLSstr <- paste0(OLSstr1, " && ", OLSstr2)
> body1 <- paste0("OLS & ", OLSstr, " \\\\")
```

```
> GMMstr1 <- paste0(sprintf("%5.2f", bias$GMM), collapse=" & ")
> GMMstr2 <- paste0(sprintf("%5.2f",   SE$GMM), collapse=" & ")
> GMMstr <- paste0(GMMstr1, " && ", GMMstr2)
> body2 <- paste0("GMM & ", GMMstr, " \\\\")
> cat(paste0(c(head0,head1,head2,head3,head4,body1,body2,
+              "\\bottomrule"),collapse="\n"))
\begin{tabular}{lSSSlSSS}
\toprule
 & \multicolumn{3}{c}{Bias} && \multicolumn{3}{c}{Variance} \\
\cmidrule{2-4}\cmidrule{6-8}
Method & {DGP1} & {DGP2} & {DGP3}
      && {DGP1} & {DGP2} & {DGP3} \\
\midrule
OLS & -0.47 &  1.15 &  0.53 && -0.36 &  0.38 & -0.86 \\
GMM & -0.39 &  1.26 & -1.84 &&  0.74 & -1.33 & -0.32 \\
\bottomrule
```

The + starting two of the lines (rather than >) indicates that the prior line is not a complete statement and is continued on the next line. R knows it is not complete because we have not closed all the open parentheses (three open, two closed). Function `cat()` treats backslash as an **escape character**, so we need two in order for it to print one, or four to print two. While this seems annoying, by the same token we can write `\n` and `cat()` inserts a newline character; the `collapse="\n"` tells it to print each text string on a new line, to make it more readable (to humans; LaTeX wouldn't care either way).

2.5 Functions

See also: http://www.cyclismo.org/tutorial/R/scripting.html

A function takes input (**arguments** or **parameters**), does something(s), and might **return** output.

You can define your own function and store it in a variable for later use. You can call `abs.sqrt.fn(-4)` after defining

```
abs.sqrt.fn <- function(x) {
  tmp <- abs(x)
  return(sqrt(tmp))
}
```

The variable named `abs.sqrt.fn` is a function. Using `return()` explicitly helps readability.

Some functions affect more than just the return value. For example, the `library()` function loads a package.

Arguments can be passed by name, as well as by order. Using names is generally better for readability (and avoiding errors). For example:

```
> log(x=100, base=10)
[1] 2
> log(base=10, x=100)
[1] 2
> log(100, 10)
[1] 2
> log(10, 100)
[1] 0.5
```

Arguments may have default values. If the user does not specify a particular argument, the default value is used. When writing your own function, it sounds tempting to give everything a default value (partly just because it seems sophisticated), but it may be better to tell the user that they forgot an argument rather than proceeding with a default value (that may not be desired). For example:

```
> power.fn <- function(base,power=2) { return(base^power) }
> power.fn(3)
[1] 9
> power.fn <- function(base,power) { return(base^power) }
> power.fn(3)
Error in power.fn(3) : argument "power" is missing, with no default
```

2.6 Data File Input

See also: http://www.cyclismo.org/tutorial/R/input.html

I'll cover input of two common file formats; others are supported (you can Google it).

First we need to know in which directory R is looking. We can see what the **working directory** is by

```
> getwd()
[1] "C:/Users/kaplandm/Documents"
```

If we want to be in a different directory, we can use `setwd()`:

```
> setwd('C:\\Users\\kaplandm\\Google Drive\\Teaching\\9476')
> getwd()
[1] "C:/Users/kaplandm/Google Drive/Teaching/9476"
```

Note again the double backslashes to get single backslashes, due to escaping.

For comma-separated values (CSV) files, you can use the function `read.csv()`. For example, try downloading the file `OXY_daily_data_no_holidays.csv` into your working directory from my website.[18] Then,

```
> VaR.data.raw <- read.csv("OXY_daily_data_no_holidays.csv")
```

[18] https://drive.google.com/file/d/0B-_LUSJVBv20anFLdTZpQlJfbms/view

The return variable is a data frame. We can see what the different columns are named, and a snippet of the data:

```
> names(VaR.data.raw)
[1] "Date"       "Year"       "Close"      "DailyLnRet" "LagLnRet"
> head(VaR.data.raw)
       Date Year Close  DailyLnRet    LagLnRet
1 31-Aug-12 2012 85.01  0.00862431 -0.02287356
2 30-Aug-12 2012 84.28 -0.02287356 -0.01404906
3 29-Aug-12 2012 86.23 -0.01404906 -0.00057159
4 28-Aug-12 2012 87.45 -0.00057159 -0.00456101
5 27-Aug-12 2012 87.50 -0.00456101  0.00730764
6 24-Aug-12 2012 87.90  0.00730764 -0.01783226
```

Much economic data is available in Stata .dta format. Over the years, different R packages have figured out how to read .dta files, but there is always a lag after a new Stata version comes out. As of Spring 2020, the function `read_dta()` in package `haven` supports through Stata version 15. If you have access to Stata, you could just work with the raw data in Stata, then export to .csv or save a .dta in version 15 (or earlier) format. For example, with some census (Current Population Survey) data,[19]

```
> library(haven)
> cps80 <- read_dta("census80.dta")
> head(cps80)
# A tibble: 6 x 7
    age  educ logwk perwt exper exper2 black
  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
1    47    16  6.97 1.00     25    625     0
2    42    12  6.17 1.00     24    576     0
3    47    16  6.44 1.00     25    625     0
4    40    12  7.06 1.00     22    484     0
5    40    19  7.07 1.00     15    225     0
6    44    12  7.06 0.991    26    676     0
```

2.7 Basic Statistics

See also: Cyclismo tutorials on basic operations[20] and OLS.[21]

R has a lot of statistical functions. Generally, these take a vector as input. If you pass them a matrix, they (usually) treat the matrix as a big vector; to operate row-by-row or column-by-column, see `apply()` below.

[19] http://economics.mit.edu/faculty/angrist/data1/data/angchefer06
[20] http://www.cyclismo.org/tutorial/R/basicOps.html
[21] http://www.cyclismo.org/tutorial/R/linearLeastSquares.html

For example, there is `mean()`, `median()`, `sd()`, `quantile()`, `sum()`, `min()`, `max()`, etc. Note if you want a "parallel" (element-wise) min or max, then use `pmin()` or `pmax()`:

```
> min(matrix(1:6,ncol=2))
[1] 1
> pmin(1:3,4:2)
[1] 1 2 2
```

See also `which.min()` and `which.max()`, which return the index of the minimum and maximum. There's also a `which()` that returns the indices of `TRUE` elements in a logical vector.

One thing to be aware of is the treatment of missing data (`NA` values). If you have missing data, you should probably think about why there is missing data; see Chapter 21. But, sometimes you just want to remove all the `NA`. Compare:

```
> mean(c(1:5,NA),na.rm=FALSE)
[1] NA
> mean(c(1:5,NA),na.rm=TRUE)
[1] 3
```

See Section 21.4 for more.

To apply statistical functions to matrices, `apply()` is useful. The application can be row-by-row (`MARGIN=1`) or column-by-column (`MARGIN=2`):

```
> (m <- matrix(1:6, nrow=2))
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> apply(m, MARGIN=1, FUN=sum)
[1]  9 12
> apply(m, MARGIN=2, FUN=sum)
[1]  3  7 11
```

OLS is run with `lm()` (which stands for "linear model"). Unfortunately, the default SE are not even robust to heteroskedasticity, but there are packages for that (and cluster-robust SE, etc.). But if you just want to run OLS, I'd suggest using Stata.

2.8 Basic Plotting (Graphs)

See also: plot tutorials from Cyclismo[22] and CRAN.[23]

In RStudio, you can just call `plot()` and a plot appears:

```
> plot(x=1:10,y=11:20)
```

[22] http://www.cyclismo.org/tutorial/R/plotting.html

[23] http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Graphics

Without RStudio, you have to first open a graphics device by calling `x11()` or something, or on a Mac, `quartz()`.

This is fine for exploring, but for a paper, always save your graphs into a PDF file. It is even better than **lossless** (e.g., bitmap, PNG, GIF) let alone **lossy** compression (e.g., JPEG) because it stores the underlying lines in your graph rather than pixel-by-pixel (as a simplification). So even if you (later) increase its size or zoom in a lot, it will still look very nice. You can include .pdf images easily in (pdf)LaTeX using `\includegraphics{}`. In R, you need to first call `pdf()` to tell R to start drawing to a PDF file, then when you're done call `dev.off()`.

You can look at many examples including my preferred formatting styles on my website, e.g., the .R file that generates all the plots in this text.

2.9 Saving Text Output

I find it helpful to save text output/results directly to a .txt file rather than copy-pasting from the console. You can do this with `cat()`, specifying the output filename:

```
> (OUTFILE <- paste0(format(Sys.time(),"%Y_%m_%d"), "_out",".txt"))
[1] "2020_04_21_out.txt"
> cat("results from regression: 12",
+     file=OUTFILE,sep="\n",append=TRUE)
```

2.10 Probability Distributions and Random Numbers

See also: Cyclismo tutorial.[24]

R has four types of functions for a variety of probability distributions, and you may create your own. The four types correspond to four letters: p for a CDF (confusing? perhaps), d for PDF ("density"), q for quantile, and r for random. So the CDF functions are `punif()` for a (continuous) uniform distribution, `pnorm()` for a normal, etc.; see `?Distributions`. For a normal CDF, for example, we need to pass arguments for the point of evaluation, the mean of the distribution μ, and the standard deviation σ, like `pnorm(0,mean=0,sd=1)`. For the PDF, the first argument is also the point of evaluation; for the quantile function, the first argument is the probability (e.g., 0.5 for the median); and for the random generation, the first argument is the number of numbers desired. Be aware that most distributions have default parameters if you don't specify them, e.g., $\mathrm{N}(0,1)$ (standard normal) is the default normal.

Another important function for randomization is `sample()`. My most common use of it is the special case of drawing (with or without replacement; e.g., for bootstrap or subsampling) subsets of integers from $1,\ldots,n$. For example,

[24] http://www.cyclismo.org/tutorial/R/probability.html

```
> set.seed(112358)
> sample(x=1:6, size=6, replace=FALSE)
[1] 2 4 6 5 3 1
> sample(x=1:6, size=6, replace=TRUE)
[1] 5 3 6 3 2 3
```

For replicability, you should always call `set.seed()` before using any randomization code. This starts the (pseudo) random number generator at a particular point, so that somebody else can run your file and get the same random numbers that you did. You should (in my opinion) pick a single seed number that you *always* use, so that you do not try different seeds to make your results look better (which is dishonest and unscientific). For example, I always use 112358; if you ever see a file of mine (which I always post on my website, too, for added accountability) with a different seed, you should ask me why it's not 112358, and alert someone if you're not satisfied by my response! Example:

```
> set.seed(112358)
> runif(3)
[1] 0.3187551 0.7404076 0.8741024
> set.seed(112358); runif(3)
[1] 0.3187551 0.7404076 0.8741024
> set.seed(112358); rnorm(1)
[1] -0.4711828
> set.seed(112358); qnorm(runif(1))
[1] -0.4711828
```

You should see these same exact numbers if you run this code, even if you are using a different computer/operating system/R version/etc. (As long as you haven't changed the default random number generator.)

2.11 Control Flow: If, Loops, Errors

See also: tutorials from Cyclismo[25] and CRAN.[26]

2.11.1 If-Else Statements

Sometimes we want to run one piece of code if some condition is true, but a different piece of code if it's false. An **if-else statement** executes the first block if the condition is true, and the latter if not. Beware if the condition is neither true nor false, but `NULL` or `NA`. The condition needs to be inside parentheses (unlike Matlab, etc.). You can also insert any number of **else if** blocks. Some examples:

[25] http://www.cyclismo.org/tutorial/R/scripting.html

[26] http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Loops-and-conditional-execution

```
> animal <- 'dog'
> if (animal=='cat') {
+     cat("meow\n")
+ } else if (animal=='dog') {
+     cat("woof\n")
+ } else {
+     stop("unknown animal")
+ }
woof
> if (TRUE) cat("meow\n")
meow
> if TRUE cat("meow\n")
Error: unexpected numeric constant in "if TRUE"
> if (NULL) cat("meow\n")
Error in if (NULL) cat("meow\n") : argument is of length zero
> if (NA) cat("meow\n")
Error in if (NA) cat("meow\n") : missing value where TRUE/FALSE needed
```

The more complicated part is often constructing the appropriate condition, which may involve logical functions of many variables' values.[27] R includes all of the usual logical operators like "and" and "or," and numerical comparisons like "less than." Note R has separate elementwise "and" and "or" for vectors, similar to the difference between `min()` and `pmin()`. Some examples:

```
> 2!=2 #"not equal to"
[1] FALSE
> 2==2 #"equal to"
[1] TRUE
> 2<=2 #"less than or equal to"
[1] TRUE
> 2<2 #"strictly less than"
[1] FALSE
> (1:3)<c(2,2,2)
[1]  TRUE FALSE FALSE
> TRUE && FALSE #"and"
[1] FALSE
> TRUE || FALSE #"or"
[1] TRUE
> c(TRUE,TRUE,TRUE) & c(FALSE,TRUE,FALSE) #elementwise
[1] FALSE  TRUE FALSE
> c(TRUE,TRUE,TRUE) && c(FALSE,TRUE,FALSE) #oops!
[1] FALSE
```

[27] See also http://www.cyclismo.org/tutorial/R/types.html#logical

```
> 4 %in% 1:5
[1] TRUE
```

2.11.2 For and While Loops

I don't use **while loops** much, but they're simple: as long as some condition is true, keep evaluating some block of code. Of course, there is a danger if the condition never becomes false: your code will never finish running!

I commonly use **for loops** for simulations and for looping through elements in a vector or list. A for loop has a **counter** variable whose value is different in each iteration of the loop. The counter iterates over a set of specified (by you) values. Inside a for loop, the value of an expression is not printed unless you do so explicitly with `cat()` or `print()`. Examples:

```
> for (i in 1:3) { cat(sprintf("%g",i)) }; cat('\n')
123
> x <- data.frame(a=1:3,b=4:6)
> for (ivar in c("a","b")) {
+   cat(sprintf("%g ",x[[ivar]]),"\n")
+ }
1  2  3
4  5  6
> for (ix in 1:length(x$b)) {
+   cat(sprintf("x$b[%d]=%d",ix,x$b[ix]),'\n')
+ }
x$b[1]=4
x$b[2]=5
x$b[3]=6
```

The keyword `next` skips directly to the (top of the) next iteration in a loop.

The keyword `break` instead breaks out of the loop completely.

2.11.3 Try-Catch, Warnings, Errors

There are **warnings** and **errors** in R. Errors are more severe and prevents any further commands from running. Warnings are displayed, but the code continues executing. Usually you'll just look at warnings/errors from functions you're using, but you can instigate them yourself, too:

```
> for (i in 1:4) {
+     if (i%%2 == 0) { warning(sprintf("Even: i=%d",i)) }
+     cat(sprintf("%g ",i))
+ }
1 2 3 4
```

```
Warning messages:
1: Even: i=2
2: Even: i=4
> for (i in 1:4) {
+     if (i%%2 == 0) { stop(sprintf("Even: i=%d",i)) }
+     cat(sprintf("%g ",i))
+ }
1
Error: Even: i=2
```

Errors and warnings are both displayed in red in RStudio (at least by default).

R supports **try-catch statements**, in which you can "try" to run a block of code, and run additional code if you "catch" an error or warning. This allows your code to keep running even if there's an error (which otherwise stops all code), or allows your code to stop running (or make modifications) if there's a warning. You'll probably never need this unless you're writing functions for other people to use. See `?tryCatch`.

2.12 Time and Timing

See also: Cyclismo tutorial.[28]

You can get the current date/time or time differences like:

```
> (x <- Sys.time())
[1] "2020-04-21 19:50:41 CDT"
> format(x,"%A, %X")
[1] "Tuesday, 7:50:41 PM"
> Sys.time() - x
Time difference of 0.02297997 secs
```

You can also time how long a certain block of code runs:

```
> set.seed(112358); x <- rbinom(n=1e8, size=100, prob=0.5)
> system.time(expr=sort(x,method='radix'))
   user  system elapsed
   2.20    0.11    2.31
> system.time(expr=sort(x,method='quick'))
   user  system elapsed
   3.37    0.05    3.42
> system.time(expr=sort(x,method='shell'))
   user  system elapsed
   5.36    0.09    5.45
```

[28] http://www.cyclismo.org/tutorial/R/time.html

2.13 Parallel Computing (On Your Laptop)

R supports parallel computing. These days, even your laptop (probably) has multiple CPUs. In my personal experience, using parallel computing has sped things up by a factor of two. This is not worth it if your code only takes a few minutes to run. But if it takes 24 hours to run, then cutting this to 12 hours lets you run it overnight and see results the next morning. Also: you will end up running your code many, many more times than you anticipate.

There are different ways you can do this. Depending how you set it up, even the random numbers are fully replicable:

```
> library(doRNG); library(doParallel)
> library(foreach); library(parallel)
> workers <- makeCluster(detectCores()) #use everything available
> registerDoParallel(workers)
> on.exit(stopCluster(workers),add=TRUE)
> N <- 5;  res <- matrix(data=NA,nrow=2,ncol=N)
> # %dorng%: replicable
> for (i in 1:2) {
+   set.seed(112358)
+   res[i,] <- foreach(i=1:N, .combine=rbind, .inorder=TRUE) %dorng% {
+     rnorm(1)
+   }
+ }
> print(res)
          [,1]       [,2]      [,3]       [,4]       [,5]
[1,] 0.4143905 -0.49700764  1.286977 -1.29213296  0.5669822
[2,] 0.4143905 -0.49700764  1.286977 -1.29213296  0.5669822
> stopCluster(workers)
```

If you want to output to a file (e.g., some log message to a .txt to see your code's progress), see `?sink`.

2.14 Simulation: Example #1

If you have never run a simulation (in R), the following code may be helpful. It is a simple simulation looking at the sample average as an estimator of the mean of a distribution, in terms of the bias and standard error (i.e., standard deviation) and RMSE of the estimator. The DGP is normal. You should be able to run it as-is and see output (not shown below).

```
# Simulation example for
# "Distributional and Nonparametric Econometrics"
# by Dave Kaplan
```

```
# Set seed for replicability (important!)
set.seed(112358)

# Constant parameters
OUTFILE <- "" #set this to something like "out.txt"
              #to save output, otherwise outputs to console
NREP <- 1000 #start w/ something small (10?) to debug/time
             #then increase to improve accuracy
N <- 10 #sample size
MU <- 0.2; SIGMA <- 1  #DGP parameters
#output the parameter values
cat(sprintf(paste0("NREPLIC=%d, N=%d, MU=%g, ","SIGMA=%g"),
    NREP, N, MU, SIGMA),
    file=OUTFILE, sep="\n", append=TRUE)

# Replication loop
start.time <- Sys.time() #get the current time
mu.hats <- rep(NA,NREP)
for (irep in 1:NREP) {
  X <- rnorm(N,mean=MU,sd=SIGMA)  # Generate data
  mu.hats[irep] <- mean(X)  # Compute estimator and store
}

# Compute results and save to OUTFILE
mu.hat.bias <- mean(mu.hats) - MU
mu.hat.sd <- sd(mu.hats)
cat(sprintf(paste0("bias=%6.4f, sd=%6.4f, RMSE=%g"),
    mu.hat.bias, mu.hat.sd, sqrt(mu.hat.bias^2+mu.hat.sd^2)),
    file=OUTFILE, sep="\n", append=TRUE)

# Output time elapsed to console
cat(sprintf("Time elapsed: %s",format(Sys.time()-start.time)),
    sep="\n", append=TRUE)
```

Discussion Question 2.1 (R: simulation example 1)**.** Carefully examine the R simulation code in Section 2.14. Run it to see the results.

a) Inside the for-loop, why does the code have `mu.hats[irep]` instead of just `mu.hats`? What would happen if it just said `mu.hats`?
b) Is `mu.hat.bias` a scalar or vector? How can you tell?
c) If you run the code a second time, will you get the same output? Why?
d) What would happen if inside the for-loop, just before the `X <-` line, you put `set.seed(1)`? Explain why that would be worse.

e) Are the observations iid? How can you tell?
f) What is the simulated bias, standard deviation, and RMSE?
g) Explain why/how you would expect the bias and standard deviation to change (or not change) if you: increase `NREP` to 2000? increase `N` to 20? increase `MU` to 0.4? increase `SIGMA` to 2? change the seed to `set.seed(2246)`?

2.15 Simulation: Example #2

Here is another simulation example. It examines the coverage probability of the standard confidence interval for OLS regressors (using the default standard error) when there are different numbers of regressors. Various parameters are stored in UPPERCASE variables; nothing is **hard coded** (e.g., I always write `ALPHA` instead of 0.1 so that I can easily change it to 0.05, or add a for loop over different values).

```
# Simulation example for
# "Distributional and Nonparametric Econometrics"
# by Dave Kaplan
# Coverage probability with many regressors

NREP <- 1e3; n <- 40; ks <- c(1,20,30:38)
BETA0 <- 0 #same for all X
ALPHA <- 0.10;  CV <- qnorm(1-ALPHA/2)
START.TIME <- Sys.time()
OUTFILE <- "" #print to console if empty

CPs <- rep.int(NA,length(ks))
for (ik in 1:length(ks)) {
  k <- ks[ik]
  set.seed(112358) #for replicability
  beta <- matrix(BETA0,k) #k-by-1 vector
  Xs <- array(rnorm(NREP*n*k), c(n,k,NREP))
  Us <- matrix(rnorm(NREP*n),NREP)
  CIs <- matrix(NA,NREP,2)
  for (irep in 1:NREP) {
    X <- Xs[,,irep];  U <- Us[irep,]
    Y <- BETA0 + X%*%beta + U
    ret.lm <- lm(Y~X)
    ret.sum.lm <- summary(ret.lm)
    est <- ret.sum.lm$coef[2,1]
    SE <- ret.sum.lm$coef[2,2]
    CIs[irep,] <- c(est-CV*SE,est+CV*SE)
  }
  CPs[ik] <- mean(CIs[,1]<BETA0 & BETA0<CIs[,2])
```

```
}
cat(paste0(sprintf("CP(k=%2d)=%5.3f", ks, CPs),
    collapse='\n'), '\n',
    file=OUTFILE, sep="", append=TRUE)

tmpt <- as.numeric(Sys.time()-START.TIME,units="secs")
tmps <- sprintf(paste0("Total time elapsed=%g seconds\n"), tmpt)
cat(tmps,file=OUTFILE,sep="",append=T)
```

The above code produces the below output.

```
CP(k= 1)=0.884
CP(k=20)=0.878
CP(k=30)=0.897
CP(k=31)=0.850
CP(k=32)=0.881
CP(k=33)=0.846
CP(k=34)=0.824
CP(k=35)=0.813
CP(k=36)=0.804
CP(k=37)=0.751
CP(k=38)=0.646
Total time elapsed=17 seconds
```

Discussion Question 2.2 (R: simulation example 2)**.** Carefully examine the R simulation code in Section 2.15. Run it to replicate the results shown.

a) What does `NREP` represent? (Where is it used?)
b) Where is sample size `n` used? Why is `Y` a vector of length `n`?
c) Would results change if `set.seed` were moved above the first `for`? Why/not?
d) Try changing `ks` to have `29:38` instead of `30:38`, and re-run everything. Does the simulated CP change for the already-existing k values? Why/not?
e) Modify the `cat` to include the sample size in each line of output. (Refer to the variable `n`, don't just hard code 40.)
f) Modify the code to use upper one-sided CIs instead of two-sided CIs.
g) Does the OLS model estimated with `lm` include an intercept term? How do you know?
h) What does the code `ret.sum.lm$coef[2,1]` mean?

Exercises

Exercise E2.1. The following is similar in spirit to Section 2.15. Design, code, and run a simulation exploring the sensitivity of IV estimates and standard confidence intervals to the strength of the instrument. Use sample size $n = 100$ observations per dataset. Let $Y_i = \beta_0 + \beta_1 X_i + U_i$, $X_i = \gamma_0 + \gamma_1 Z_i + V_i$. Let (U, V) be bivariate normal, with variances both equal to one, and correlation ρ. So, ρ controls the degree of endogeneity (with $\rho = 0$ for exogenous X), and γ_1 controls the degree of relevance of the instrument, or the "strength" of the instrument (with γ_1 implying the instrument is not relevant and thus not valid). Let $\rho = 0.5$. Try different values of γ_1, seeing how small it must be to start affecting the properties of the IV estimator and CIs. In each simulation replication, compute the IV estimator; recall in matrix notation $\hat{\boldsymbol{\beta}} = (\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{Z}}'\boldsymbol{Y}$, where $\underline{\boldsymbol{Z}}$ contains a column of ones and a column of Z_i, and $\underline{\boldsymbol{X}}$ contains a column of ones and a column of X_i, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. Compute the estimated covariance matrix like in (5.34) of Wooldridge (2010), $(\hat{\underline{\boldsymbol{X}}}'\hat{\underline{\boldsymbol{X}}})^{-1}\left(\sum_{i=1}^n \hat{U}_i^2 \hat{\boldsymbol{X}}_i \hat{\boldsymbol{X}}_i'\right)(\hat{\underline{\boldsymbol{X}}}'\hat{\underline{\boldsymbol{X}}})^{-1}$ where $\hat{\boldsymbol{X}}_i = (1, \hat{X}_i)'$ with $\hat{X}_i = \hat{\gamma}_0 + \hat{\gamma}_1 Z_i$ (estimated linear projection); the standard error for β_1 is the square root of the $(2,2)$ entry in that matrix. Compute a "95%" CI for β_1 as $\hat{\beta}_1 \pm 1.96\,\mathrm{SE}(\hat{\beta}_1)$. In 100 (or 1000) replications, store the $\hat{\beta}_1$ and CI from each replication. Afterward, make a histogram of the $\hat{\beta}_1$ values; see if it seems concentrated around the true β_1. Also compute the proportion of replications in which the true β_1 was inside your computed CI, to get the simulated coverage probability; compare this to 95%.

Exercise E2.2. Design, code, and run a simulation exploring the sensitivity of another econometric technique to violations of an assumption. Similar to E2.1, but you get to choose what to simulate. Please ask me (in person or email) about your idea *before* you start to code it, to make sure your time is well spent.

Chapter 3
Logic

Unit learning objectives for this chapter

3.1. Define and apply basic logic terms and relationships [TLO 1]

Some basic logic is useful for understanding certain parts of econometrics. First, logic is useful for understanding the relationship among different conditions. Often these conditions are assumptions used in various theorems. Second, logic is useful for understanding what a theorem actually claims. Third, logic is helpful for interpreting results. The following may not be fully technically correct from a philosopher's perspective, e.g., perhaps I conflate logical implication with the material conditional, but it suffices for econometrics.

Optional resources for this chapter

- Section 6.1 of Kaplan (2022b) is very similar

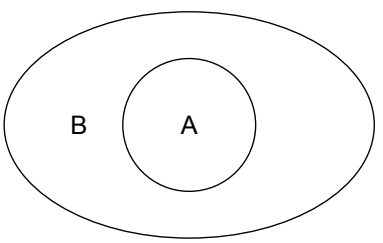
3.1 Terminology

Many words and notations can refer to the same logical relationship. Let A and B be two statements that can be either true or false. For example, maybe A is "$Y \geq 10$" and B is "$Y \geq 0$." Or, A is "this animal is a cat," and B is "this animal is a mammal." The following ways of describing the logical relationship between A and B all have the same meaning.

1. If A (is true), then B (is true)
2. $A \implies B$

3. A **implies** B
4. $B \impliedby A$
5. B is **implied by** A
6. B is true **if** A is true
7. A is true **only if** B is true
8. A is a sufficient condition (or just **sufficient**) for B
9. B is a necessary condition (or just **necessary**) for A
10. A is **stronger** than B
11. B is **weaker** than A
12. It is impossible for B to be false when A is true (but it is fine if both are true, or both are false, or A is false and B is true)
13. The truth table (T=true, F=false):

A	B	$A \implies B$
T	T	T
T	F	F
F	T	T
F	F	T

14. 

To state equivalence of A and B, opposite statements can be combined. Specifically, any of the following have the same meaning:

1. $A \iff B$ (meaning both $A \implies B$ and $A \impliedby B$)
2. A is true **if and only if** B is true (meaning A is true if B is true *and* A is true only if B is true)
3. A is necessary and sufficient for B (or equivalently, B is necessary and sufficient for A)
4. A is equivalent to B
5. It is impossible for A to be false when B is true, and impossible for A to be true when B is false.
6. The truth table (T=true, F=false):

A	B	$A \iff B$
T	T	T
T	F	F
F	T	F
F	F	T

Variations of $A \implies B$ have the following names. Read $\neg A$ as "not A": $\neg A$ is false

when A is true, and $\neg A$ is true when A is false.

- $\neg A \implies \neg B$ is the **inverse** of $A \implies B$.
- $B \implies A$ is the **converse** of $A \implies B$.
- $\neg B \implies \neg A$ is the **contrapositive** of $A \implies B$.

Interestingly, the statement $A \implies B$ is logically equivalent to its contrapositive. That is, statements "$A \implies B$" and "$\neg B \implies \neg A$" can be both true or both false, but it's impossible for one to be true and the other false. The statement $A \implies B$ is not logically equivalent to either its inverse or converse. (The inverse and converse are equivalent to each other: the inverse is the contrapositive of the converse.)

Discussion Question 3.1 (logic)**.** Let A be "$X \leq 0$" and let B be "$X \leq 10$."

a) Explain why $A \implies B$.
b) State the contrapositive in terms of X, and explain why it is also true.
c) State the converse in terms of X, and explain why it is not true.
d) State the inverse in terms of X, and explain why it is not true.

3.2 Assumptions

To compare assumptions, the terms "stronger" and "weaker" are most commonly used. Instead of assumption A and conclusion B, let A and B denote different assumptions. For example, let A be $\mathrm{E}(Y^4) < \infty$, and let B be $\mathrm{E}(Y^2) < \infty$. Any random variable Y with finite $\mathrm{E}(Y^4)$ also has finite $\mathrm{E}(Y^2)$, but some have finite $\mathrm{E}(Y^2)$ and infinite $\mathrm{E}(Y^4)$. Logically, $A \implies B$. Thus, people say "$\mathrm{E}(Y^4) < \infty$ is a stronger assumption than $\mathrm{E}(Y^2) < \infty$," or equivalently, "$\mathrm{E}(Y^2) < \infty$ is weaker than $\mathrm{E}(Y^4)$."

As another example, consider the linear projection and linear CEF models. Consider the linear model $Y = \beta_0 + \beta_1 X + U$. Let assumption A be $\mathrm{E}(U \mid X) = 0$, and let B be $\mathrm{E}(U) = 0$ and $\mathrm{Cov}(X, U) = 0$; i.e., A says U is a CEF error, whereas B says U is a linear projection error. Here, $A \implies B$, so A is a stronger assumption than B, and B is weaker than A. Seen another way, the linear projection model is more general than the linear CEF model: if the CEF is $\beta_0 + \beta_1 x$, then so is the linear projection, but if the linear projection is $\beta_0 + \beta_1 x$, it is still possible to have a nonlinear CEF.

All else equal, weaker assumptions are better because then the theorem applies to more settings (the results are "more general").

3.3 Theorems

Theorems all have the same logical structure: if assumption A is true, then result (conclusion) B is true. Sometimes A and B have multiple parts, like the four parts of Assumption 7.1 of Hansen (2020a, §7.1, p. 170) and the five conclusions in Theorem 7.1 of Hansen

(2020a, §7.2, p. 172), but the logical structure of a theorem is always the same. The theorem claims that if we can verify that A is true, then we know that B is also true. But what if we don't know about A, or we think it's false? Then, B could be false, or it could be true. This may be seen most readily from the picture version of the A and B relationship. We could be somewhere inside B (where B is true) but outside A (where A is false); or we could be outside both, where both are false. The theorem is not equivalent to, "If A is false, then B is false" (the "inverse"). However, it is equivalent to the **contrapositive**: "If B is false, then A is false." Again, this is probably seen most easily in the picture.

Discussion Question 3.2 (median theorem logic)**.** Consider the statement, "If sampling is iid, then the sample median consistently estimates the population median."

a) What does this tell us about consistency of the sample median when sampling is not iid?
b) What does this tell us about sampling when the sample median is not consistent?

Hint: draw a picture.

Discussion Question 3.3 (mean theorem logic)**.** Consider the statement, "If sampling is iid and the population mean is well-defined, then the sample mean consistently estimates the population mean."

a) What does this tell us about consistency of the sample mean when sampling is not iid?
b) What does this tell us about sampling when the sample mean is not consistent?

Hint: draw a picture with A1 (iid), A2 (well-defined), and B (consistency).

Discussion Question 3.4 (logic with feathers)**.** Consider two theorems. Theorem 1 says, "If X is an eagle, then it has feathers." Theorem 2 says, "If X is a bird, then it has feathers."

a) Describe each theorem logically: what's the assumption (A), what's the conclusion (B), what's the relationship?
b) State Theorem 1's contrapositive; is it true?
c) Compare: does Theorem 1 or Theorem 2 have a stronger assumption? Why?
d) Compare: which theorem is more useful? (Which applies to more situations?)

Part II

Quantile Methods

Introduction

This part concerns econometric and statistical methods that look beyond the mean to other features of (conditional) distributions. Specifically, features involving (conditional) quantiles are considered. Depending on the setting and method, these methods may be useful for descriptive, predictive, and/or causal analysis.

Chapter 4

Unconditional Quantiles: Description and Prediction

Unit learning objectives for this chapter

4.1. Interpret quantiles in terms of both optimal prediction and distributional description [TLOs 1 and 2]

4.2. Understand differences between the mean and median in terms of estimation efficiency, sensitivity to outliers and censoring, and statistical inference [TLOs 2 and 3]

Quantiles can be useful for both description and prediction. As description, they capture distributional features beyond the mean, especially features related to inequality and heterogeneity. For prediction, although the mean is optimal for quadratic loss, quantiles are optimal for alternative loss functions that allow asymmetry (over-prediction is worse than under-prediction, or vice-versa). This chapter considers the unconditional distribution of Y to introduce concepts. Chapter 5 extends this to the conditional (on $\boldsymbol{X} = \boldsymbol{x}$) distribution of Y.

Optional resources for this chapter

- Koenker (2005), `http://laurel.lso.missouri.edu/record=b5328718~S1`

4.1 Description

The cumulative distribution function (CDF) is a complete but complex description of the probability distribution of Y. Because it is difficult (for humans) to discuss and compare entire functions, usually the CDF is summarized by certain features.

Conversely, summary features are convenient but lose information. Two popular features are the mean and standard deviation. If $Y \sim \mathrm{N}(\mu, \sigma^2)$, then the mean and standard deviation fully describe the distribution. However, most variables are not Gaussian, so such a summary loses potentially valuable information.

Quantiles complement the mean in summarizing a distribution. They can help capture skewness, spread, tails, and other important aspects of the distribution's shape.

4.2 Formal Definitions

The τ-quantile is the same as the 100τth percentile: the value for which τ proportion of the population has a smaller value. (There are some caveats; see below.)

The τ-**quantile**'s formal definition and notation follow. Let $\tau \in [0, 1]$ denote the **quantile index** (or **quantile level**). Let $\mathrm{Q}_\tau(Y)$ denote the τ-quantile of random variable Y, analogous to notation $\mathrm{E}(Y)$ for the mean of Y. The informal definition above suggests $\mathrm{Q}_\tau(Y)$ satisfies $\mathrm{P}(Y \leq \mathrm{Q}_\tau(Y)) = \tau$, or equivalently $F_Y(\mathrm{Q}_\tau(Y)) = \tau$. This further suggests $\mathrm{Q}_\tau(Y) = F_Y^{-1}(\tau)$ if the CDF $F_Y(\cdot)$ is invertible. More generally,

$$\mathrm{Q}_\tau(Y) \equiv \inf\{y : F_Y(y) \geq \tau\}. \tag{4.1}$$

The **quantile function** $Q_Y(\cdot)$ more explicitly expresses the quantiles of Y as a function of τ. That is,

$$Q_Y(\tau) \equiv \inf\{y : F_Y(y) \geq \tau\}, \quad 0 \leq \tau \leq 1. \tag{4.2}$$

If the CDF $F_Y(\cdot)$ is invertible, then $Q_Y(\cdot) = F_Y^{-1}(\cdot)$. If $F_Y(\cdot)$ has a flat spot, then $Q_Y(\cdot)$ has a jump discontinuity. If $F_Y(\cdot)$ has a discontinuity (e.g., if Y is discrete), then $Q_Y(\cdot)$ has a corresponding flat spot. Whereas CDFs are right-continuous (with left limits), quantile functions are left-continuous (with right limits).

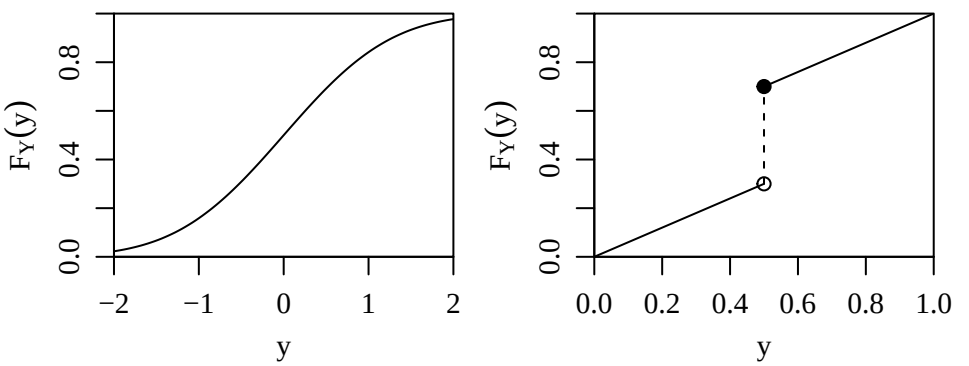


Figure 4.1: CDFs for DQ 4.1.

Discussion Question 4.1 (quantiles from CDF)**.** For each of the CDFs shown in Figure 4.1, do each of the following.

a) Verbally describe the distribution of Y.
b) Visually locate $\mathrm{Q}_{0.5}(Y)$.
c) Visually locate $\mathrm{Q}_{0.8}(Y)$.
d) Sketch $Q_Y(\cdot)$.

4.3 Prediction

Like the mean, quantiles are optimal predictors under certain "loss functions."

Discussion Question 4.2 (population minimization: quadratic and absolute loss)**.** Let Y be a discrete rv with $\mathrm{P}(Y=1)=\mathrm{P}(Y=2)=\mathrm{P}(Y=99)=1/3$.
a) Compute $\theta_1 = \arg\min_{t\in\mathbb{R}} \mathrm{E}[(Y-t)^2]$.
b) Compute $\theta_2 = \arg\min_{t\in\mathbb{R}} \mathrm{E}\big(|Y-t|\big)$.
c) What are the common names for θ_1 and θ_2?

Some vocabulary is useful. In the frequentist framework, we repeatedly guess the same value g for repeated random draws of Y, and we see "how bad" our guess is on average in the long run. A **loss function** $L(y,g)$ quantifies how bad it is to guess g when the true value is y. The (infinitely) long-run average loss given fixed g is thus the **expected loss** (also called **risk**), where the expectation is wrt the distribution of Y: $\mathrm{E}[L(Y,g)]$. Given this framework and a particular loss function, the optimal predictor minimizes risk.

Definition 4.1 (loss, risk, optimal prediction)**.** Loss function $L(y,g)$ quantifies how bad it is to guess (predict) g when the true value is y. Given loss function L, the optimal frequentist predictor minimizes risk (expected loss):

$$g_L^* \equiv \arg\min_g \mathrm{E}[L(Y,g)]. \tag{4.3}$$

Recall from Hansen (2020a, §2.11) that the population mean is the "best" unconditional predictor of Y given a certain definition of "best." Specifically, consider the quadratic loss function

$$L_2(y,g) = \rho_2(y-g) = (y-g)^2. \tag{4.4}$$

Then, the mean is optimal in that

$$\mathrm{E}(Y) = \arg\min_g \mathrm{E}[L_2(Y,g)]. \tag{4.5}$$

Equivalently, the mean $\mathrm{E}(Y)$ minimizes the mean squared prediction error, where $y-g$ is the prediction error, $(y-g)^2$ is the squared prediction error, and $\mathrm{E}[(Y-g)^2]$ is the **mean squared prediction error** (MSPE). This can be derived from the first-order condition:

$$0 = \frac{\mathrm{d}}{\mathrm{d}g}\mathrm{E}[L_2(Y,g)]\bigg|_{g=g_2^*} = \frac{\mathrm{d}}{\mathrm{d}g}\mathrm{E}[(Y-g)^2]\bigg|_{g=g_2^*} = 2\,\mathrm{E}[Y-g_2^*],$$

so $g_2^* = \mathrm{E}(Y)$.

There is no data here. Everything is in the population. More arguments would be required to say that the *sample* mean is the optimal predictor. (I imagine somebody has studied that, but I am ignorant of such results.) The weaker argument here is that the sample mean consistently estimates the population mean, which in turn is the optimal predictor for Y under quadratic loss.

As seen in DQ 4.2, replacing $L_2(y, g)$ with the alternative loss function $L_1(y, g) = |y - g|$ yields a different optimal predictor, specifically the population median:

$$\mathrm{Q}_{0.5}(Y) = \arg\min_g \mathrm{E}[L_1(Y, g)]. \tag{4.6}$$

A broader class of loss functions characterizes all quantiles over $\tau \in (0, 1)$. Given τ, the **check function** or **tick function** is

$$\rho_\tau(u) \equiv u(\tau - \mathbb{1}\{u < 0\}). \tag{4.7}$$

With $\tau = 0.5$, actually $\rho_{0.5}(u) = |u|/2$, not $|u|$. However, scaling by a constant does not affect minimization:

$$\mathrm{Q}_{0.5}(Y) = \arg\min_g \mathrm{E}[L_1(Y, g)] = \arg\min_g (1/2)\,\mathrm{E}[L_1(Y, g)] = \arg\min_g \mathrm{E}[\rho_{0.5}(Y - g)].$$

More generally,

$$\mathrm{Q}_\tau(Y) = \arg\min_g \mathrm{E}[\rho_\tau(Y - g)]. \tag{4.8}$$

That is, given Definition 4.1, the τ-quantile of Y is the optimal unconditional predictor of Y under loss function $L(y, g) = \rho_\tau(y - g)$.

Discussion Question 4.3 (check function)**.** Consider Figure 4.2.

a) Which function corresponds to which τ?
b) Does $\tau = 0.95$ penalize over-prediction ($g > y$) or under-prediction ($g < y$) more heavily?
c) Given the asymmetry in (b), explain intuitively why it makes sense that $\mathrm{Q}_{0.95}(Y)$ is a better predictor than $\mathrm{Q}_{0.5}(Y)$ given the $\tau = 0.95$ loss function $\rho_{0.95}(\cdot)$.

4.4 Estimation and Sample Quantiles

Discussion Question 4.4 (sample minimization: quadratic and absolute loss)**.** Consider a dataset with $n = 3$: $Y_1 = 1$, $Y_2 = 2$, $Y_3 = 99$.

a) Compute $\hat{\theta}_1 = \arg\min_{t \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (Y_i - t)^2$.
b) Compute $\hat{\theta}_2 = \arg\min_{t \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |Y_i - t|$.
c) What names do we call $\hat{\theta}_1$ and $\hat{\theta}_2$?

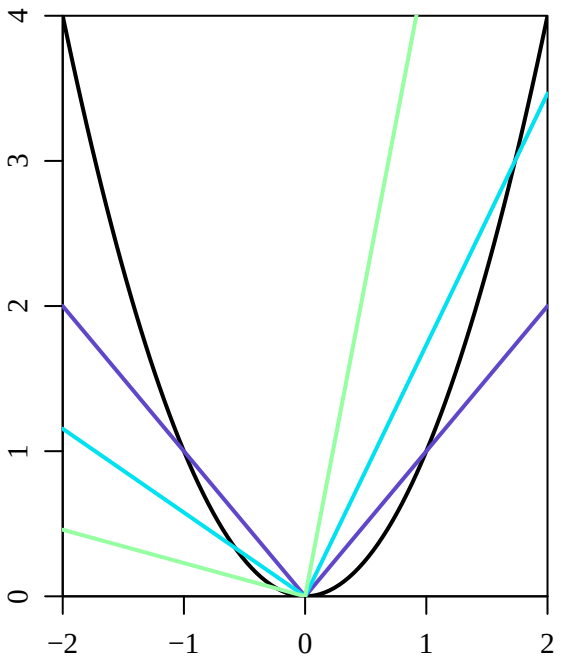


Figure 4.2: Check functions $\rho_\tau(\cdot)$ scaled by $1/\sqrt{\tau(1-\tau)}$, i.e., $\rho_\tau(\cdot)/\sqrt{\tau(1-\tau)}$, for $\tau \in \{0.5, 0.75, 0.95\}$, along with quadratic $\rho_2(\cdot)$.

In DQ 4.4, we could solve an FOC to get an explicit formula for $\hat{\theta}_1$, but not for $\hat{\theta}_2$. This hints at some of the computational difficulties of quantile estimators. Despite such difficulties, there are functions in R and Stata to estimate a wide variety of quantile models.

As with the mean, there are two approaches to estimating quantiles. First, related to description: we could "plug in" the estimated CDF into a CDF-based definition. With iid data, the "empirical CDF" is

$$\hat{F}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y\}, \quad \forall y \in \mathbb{R}, \tag{4.9}$$

i.e., the sample proportion of Y_i below the point of evaluation y. This is the CDF for a discrete distribution with probability $1/n$ on each observed Y_i value (if values are unique). The population mean is

$$\mathrm{E}(Y) = \int_{\mathbb{R}} y \, \mathrm{d}F_Y(y). \tag{4.10}$$

The **plug-in principle** or **analogy principle** suggests "plugging in" $\hat{F}_Y(\cdot)$ for $F_Y(\cdot)$ to get the **sample analog** of $\mathrm{E}(Y)$:

$$\widehat{\mathrm{E}}(Y) = \int_{\mathbb{R}} y \, \mathrm{d}\hat{F}_Y(y) = \sum_{i=1}^{n} Y_i(1/n) = \bar{Y}_n, \tag{4.11}$$

the familiar sample mean. For $\mathrm{Q}_\tau(Y)$, we can replace $F_Y(\cdot)$ in (4.1) to get

$$\widehat{\mathrm{Q}}_\tau(Y) = \inf\{y : \hat{F}_Y(y) \geq \tau\}, \tag{4.12}$$

often called the **sample τ-quantile**.

The second quantile estimation approach relates to prediction: solve the sample version of the population minimization problem. For the mean,

$$\mathrm{E}(Y) = \arg\min_g \mathrm{E}[(Y-g)^2] \tag{4.13}$$

in the population. Replacing the population expectation $\mathrm{E}[\cdot]$ with the sample expectation $\widehat{\mathrm{E}}[\cdot]$ (i.e., sample average),

$$\widehat{\mathrm{E}}(Y) = \arg\min_g \widehat{\mathrm{E}}[(Y-g)^2] = \arg\min_g \frac{1}{n}\sum_{i=1}^n (Y_i - g)^2. \tag{4.14}$$

This is the familiar "least squares" approach, minimizing the sum of squared residuals. For quantiles, replacing $\mathrm{E}[\cdot]$ with $\widehat{\mathrm{E}}[\cdot]$ in (4.8) yields

$$\widehat{\mathrm{Q}}_\tau(Y) = \arg\min_g \widehat{\mathrm{E}}[\rho_\tau(Y-g)] = \arg\min_g \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - g). \tag{4.15}$$

There are other approaches and variations; in R, I use the `quantile()` function with argument `type=6`.

4.5 Censoring

Quantiles (and quantile regression) are useful when observations are **censored**. Censoring means we do not always observe the true value. More specifically, the observed value is a function of the true value, but this function is not injective (not one-to-one), so the true values cannot be recovered exactly from the censored values.

One example of censoring is **top-coding** of earnings data, as in the Survey of Income and Program Participation (SIPP), the National Longitudinal Survey(s) of Youth (NLSY), and Current Population Survey (CPS). The simplest version of top-coding replaces earnings values above some threshold (like $150,000.00/yr) with the threshold value.

One approach to top-coding is to **impute** values, i.e., guess the true values. After imputation, the resulting dataset is treated like any other dataset. MU econ PhD alum Li Tan published a paper on imputation that exploits the repeated observations in a panel dataset, which seems to outperform methods developed with cross-sectional datasets in mind; see Tan (2021).

Another approach to top-coding is to ask economic questions that don't rely on the very upper tail; such questions often involve quantiles. Consider income inequality. The

qualitative idea of "inequality" can be quantified in many possible ways. One is the standard deviation, but it depends crucially on the very upper tail that we can't observe with top-coding. Alternatively, the difference between the 0.9-quantile and the 0.1-quantile (the 0.9–0.1 interquantile range) does not require any knowledge about the top 10% of the distribution, so top-coding is no problem.

Discussion Question 4.5 (inequality measures)**.** Consider an income distribution's variance and 0.9–0.1 interquantile range (IQR).

a) Which aspect(s) of income inequality can the variance capture that the IQR cannot?
b) Which aspect(s) of income inequality can the IQR capture that the variance cannot?

Before thinking about learning means and quantiles from top-coded data, a formal definition of **identification** is given. The definition and subsequent examples are similar to those of Hansen (2020a, §2.32). This definition has a microeconometric flavor since it implicitly assumes that we can learn about the joint distribution of observable variables (e.g., we can consistently estimate the population distribution). In a time series setting, this may not make sense. The rough idea of identification is: assuming we can learn the population distribution of observables, is that sufficient to learn about the parameter of interest?

Definition 4.2 (identification)**.** Let $\mathcal{F}$ be a set of possible joint distributions of observable variables. Parameter $\theta \in \mathbb{R}$ is **identified** on $\mathcal{F}$ if F uniquely determines θ for all $F \in \mathcal{F}$.

Consider the following form of top-coding. An individual's true earnings are Y^*. Constant c is the top-coding threshold. The observed Y is

$$Y = \begin{cases} Y^* & \text{if } Y^* \leq c \\ c & \text{if } Y^* > c. \end{cases} \tag{4.16}$$

Since $\mathrm{P}(Y = c) = \mathrm{P}(Y^* \geq c)$, the distribution of Y may have a mass point at c even if Y^* is continuous. This means the observable CDF $F(\cdot)$ may jump discontinuously at c since $F(c) = 1$. More generally, the CDF of the observed Y is

$$F(y) = \begin{cases} F^*(y) & \text{if } y < c \\ 1 & \text{if } y \geq c. \end{cases} \tag{4.17}$$

Discussion Question 4.6 (identification with top-coding: mean)**.** Consider the top-coding of (4.16). Show that the mean is *not* identified. Hint: a counterexample suffices to disprove identification. Provide a counterexample where Y^* CDFs $F_1^*(\cdot)$ and $F_2^*(\cdot)$ have different means but imply the same top-coded $F(\cdot)$, i.e., $F(\cdot)$ does not uniquely determine the parameter of interest $\mathrm{E}(Y^*)$.

Discussion Question 4.7 (identification with top-coding: median)**.** Continue from DQ 4.6.

a) Draw a graph of a pair of CDFs for possible Y_1^* and Y_2^*, say $\{F_1^*(\cdot), F_2^*(\cdot)\}$, with the following properties: same top-coded CDF; different mean; same median.

b) Repeat (a) but where Y_1^* and Y_2^* have different medians.
c) How can $\mathcal{F}$ be restricted to ensure they always have the same median?

Extending DQ 4.7, there are conditions under which $\mathrm{Q}_\tau(Y^*)$ is identified for $0 \leq \tau \leq b$ for some constant b. If $b \geq 0.9$, then both $\mathrm{Q}_{0.9}(Y^*)$ and $\mathrm{Q}_{0.1}(Y^*)$ are identified, and thus the 0.9–0.1 IQR $\mathrm{Q}_{0.9}(Y^*) - \mathrm{Q}_{0.1}(Y^*)$ is also identified.

This idea can be extended to quantile regression (and extensions like quantile duration models), too.

4.6 Robustness and Efficiency

You may hear that the median is more "robust" than the mean. Any time you hear "robust," you should ask: robust to what? Here, people would say, "robust to outliers." But that begs the question: what's an "outlier"?

The median is well defined for any probability distribution, whereas the mean is not. For example, a Cauchy distribution has median zero, but its mean is undefined.

Even if the mean is defined, "fat tails" may make the sample mean's variance much larger than the sample median's variance (because the sample mean is more sensitive to a single very large value than the sample median). That is, the median could be preferred because of better estimation efficiency (i.e., smaller standard error).

In both the population and sample, the median is less sensitive than the mean to very large but unlikely values. For example, imagine a discrete distribution with

$$\mathrm{P}(Y = j) = 1/99 \text{ for } j = 1, 2, \ldots, 98, J. \tag{4.18}$$

As $J \to \infty$, the median remains 50, but $\mathrm{E}(Y) = (1/99)(1 + 2 + \cdots + 98 + J) \to \infty$.

Discussion Question 4.8 (robustness to outliers: population)**.** Consider (4.18) as a population income distribution, with very large J.

a) What does the mean capture that the median doesn't?
b) What does the median capture that the mean doesn't?

We can also interpret (4.18) as a sample distribution based on $Y_i = i$ for $i = 1, \ldots, 98$ and $Y_{99} = J$. As $J \to \infty$, the sample mean $\widehat{\mathrm{E}}(Y) = \bar{Y}_n \to \infty$ for the same reason as before. In contrast, the sample median remains 50. If we are worried that sample "outliers" may be due to bad data (measurement error), then we may prefer an estimator like the median that's less sensitive to outliers.

However, for regression, quantile regression is only robust to outliers in Y, not X. As with OLS, the quantile regression slope estimate can be made arbitrarily large (or small) by changing just a single point $(Y_i, \boldsymbol{X}_i)$. There are more "robust" regression methods like the "least median of squares" (Rousseeuw, 1984), and Koenker (2005, §8.5) discusses proposals for higher-breakdown quantile regression.

4.7 Inference

For the population mean, the usual procedure to construct a confidence interval (CI) is: 1) show the sample mean is asymptotically normal, $\sqrt{n}(\bar{Y}_n - \mathrm{E}(Y)) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$, 2) estimate the unknown σ^2 by $\hat{\sigma}^2$, 3) use a formula like $\bar{Y}_n \pm 1.96\hat{\sigma}/\sqrt{n}$ for a 95% CI.

In principle, the same can be done for a population quantile. With iid sampling,

$$\sqrt{n}(\widehat{\mathrm{Q}}_\tau(Y) - \mathrm{Q}_\tau(Y)) \xrightarrow{d} \mathrm{N}(0, \sigma^2), \quad \sigma^2 = \tau(1-\tau)/[f_Y(\mathrm{Q}_\tau(Y))]^2, \tag{4.19}$$

where $f_Y(\cdot)$ is the PDF of Y. Thus, given $\hat{\sigma} \xrightarrow{p} \sigma$, the CI $\widehat{\mathrm{Q}}_\tau(Y) \pm 1.96\hat{\sigma}/\sqrt{n}$ has coverage probability approaching 95% as $n \to \infty$.

However, the PDF estimate in $\hat{\sigma}$ may be inaccurate in finite samples, so many alternatives for quantile CIs have been explored. One approach is to explicitly account for the estimation error in the nonparametric $\hat{\sigma}$ to improve accuracy, as in Kaplan (2015). Various bootstraps have been studied. For example, a variant of the Bayesian bootstrap (as in Chapter 14 and Section 13.1) produces very accurate quantile CIs; see Kaplan and Hofmann (2020). CIs based on order statistics are also very accurate; see Goldman and Kaplan (2017) and Goldman and Kaplan (2018b).

Plot twist: despite the added difficulty, you can (sometimes) nonparametrically compute CIs with known *exact finite-sample* coverage probability (using order statistics), which is impossible for the mean!

Chapter 5

Quantile Regression: Description and Prediction

Unit learning objectives for this chapter

5.1. Interpret quantile regression in terms of description and prediction, including under misspecification [TLO 1]

5.2. Develop intuition about conditional quantile functions and prediction with asymmetric loss functions [TLO 2]

5.3. Evaluate the (dis)advantages of quantile regression compared to OLS [TLO 3]

Like unconditional quantiles, **quantile regression** (QR) aids both description and prediction. For description, QR captures more of the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ than just the mean. For prediction (guessing Y given $\boldsymbol{X} = \boldsymbol{x}$), although the conditional expectation function is optimal for quadratic loss, conditional quantile functions are optimal for "check function" loss that allows asymmetry. Some results in this chapter for QR are analogous to results in Chapter 2 of Hansen (2020a) for mean regression. Parallel to how the word "regression" is used with multiple meanings (the population conditional mean, the estimation procedure, the sample results, etc.), the phrase "quantile regression" is also used with multiple meanings, so beware.

Optional resources for this chapter

- Koenker (2005), `http://laurel.lso.missouri.edu/record=b5328718~S1`
- *Handbook of Quantile Regression*
- Survey: `http://www.econ.uiuc.edu/~roger/research/QR40/QR40.pdf`
- Angrist, Chernozhukov, and Fernández-Val (2006)

- R: package `quantreg` by Koenker (2019)
- R: other code strewn about, like function `npqreg` in package `np`
- R and Stata code on Blaise Melly's website
- Additional R code: `https://kaplandm.github.io`

Discussion Question 5.1 (context for QR)**.** Before learning more about QR, recall what you know about non-quantile regression.

a) What does "conditional mean" or "conditional expectation function" (CEF) mean?
b) How can we estimate a CEF?
c) Why do economists estimate CEFs?

Discussion Question 5.2 (motivation for quantile regression)**.** Explain why we might care about anything besides the CEF, in terms of each of the following.

a) Description (of the joint distribution of observable variables)
b) Prediction (guessing Y based on $\boldsymbol{X}$)
c) Causality

5.1 Description

Consider the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$. That is, within the overall population, there is a subpopulation with $\boldsymbol{X} = \boldsymbol{x}$, and there is some distribution of Y within that subpopulation. For example, the subpopulation could be individuals with a certain education level, age, and occupation; or firms of a certain size in a particular industry; etc.

5.1.1 Conditional Quantile Function

Previously, you learned about the **conditional expectation function** (CEF), $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$. The CEF tells us the mean of the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$, for any $\boldsymbol{x}$, showing how the mean of Y varies with $\boldsymbol{x}$.

Complementing the CEF, the **conditional quantile function** (CQF) captures other features of the conditional distributions. The conditional τ-quantile of Y given $\boldsymbol{X} = \boldsymbol{x}$ is

$$\mathrm{Q}_\tau(Y \mid \boldsymbol{X} = \boldsymbol{x}) \equiv \inf\{y : F_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X} = \boldsymbol{x}) \geq \tau\}, \tag{5.1}$$

parallel to (4.1) but now conditioning on $\boldsymbol{x}$. When specifying τ explicitly is important, I write τ-CQF. Also, parallel to (4.2), the quantile function of Y conditional on $\boldsymbol{X} = \boldsymbol{x}$ is

$$Q_{Y|\boldsymbol{X}}(\tau \mid \boldsymbol{X} = \boldsymbol{x}) \equiv \inf\{y : F_{Y|\boldsymbol{X}}(y \mid \boldsymbol{X} = \boldsymbol{x}) \geq \tau\}, \quad 0 \leq \tau \leq 1. \tag{5.2}$$

Like the CEF, the CQFs describe how the distribution of Y varies with $\boldsymbol{x}$, without overwhelming us with the full conditional CDF. The conditional CDF is essentially a

function-valued function of $\boldsymbol{x}$, i.e., a different function (CDF) at each possible $\boldsymbol{x}$. More simply, the CEF and CQFs are scalar-valued functions of $\boldsymbol{x}$. With scalar x, these are easily plotted. With vector $\boldsymbol{x}$, usually either (average) partial derivatives are reported, or the function is structured to be summarized by a vector $\boldsymbol{\beta}$. As with unconditional distributions, quantiles capture features of conditional distributions that the mean alone does not: skewness, spread, tails, and other aspects of the conditional distribution's shape.

5.1.2 CQF Models

As with CEF models, there are different ways to write a CQF model. (These could be called "quantile regression" models, but CQF is less ambiguous.) Without specifying a functional form, we can characterize the CQF $q_\tau(\boldsymbol{x})$ by any of these:

$$q_\tau(\boldsymbol{x}) = \mathrm{Q}_\tau(Y \mid \boldsymbol{X} = \boldsymbol{x}), \tag{5.3}$$

$$Y = q_\tau(\boldsymbol{X}) + V, \quad \mathrm{Q}_\tau(V \mid \boldsymbol{X}) = 0, \tag{5.4}$$

$$\tau = \mathrm{P}(Y \leq q_\tau(\boldsymbol{X}) \mid \boldsymbol{X}). \tag{5.5}$$

If a linear model is (optimistically) specified, then

$$q_\tau(\boldsymbol{x}) = \mathrm{Q}_\tau(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}(\tau), \tag{5.6}$$

$$Y = \boldsymbol{X}'\boldsymbol{\beta}(\tau) + V, \quad \mathrm{Q}_\tau(V \mid \boldsymbol{X}) = 0, \tag{5.7}$$

$$\tau = \mathrm{P}(Y \leq \boldsymbol{X}'\boldsymbol{\beta}(\tau) \mid \boldsymbol{X}). \tag{5.8}$$

5.1.3 Monotonicity

Without further assumptions, the CQFs could have any shape with respect to $\boldsymbol{x}$, but they must obey a certain monotonicity in τ. Let $0 < s < t < 1$. In the unconditional case, by definition, $\mathrm{Q}_s(Y) \leq \mathrm{Q}_t(Y)$. This remains true conditional on any $\boldsymbol{X} = \boldsymbol{x}$:

$$\mathrm{Q}_s(Y \mid \boldsymbol{X} = \boldsymbol{x}) \leq \mathrm{Q}_t(Y \mid \boldsymbol{X} = \boldsymbol{x}).$$

That is, the s-CQF lies weakly below the t-CQF.

Alternatively, monotonicity can be written in terms of (conditional) quantile functions, which increase monotonically. Again let $0 < s < t < 1$. Unconditionally, by definition, $Q_Y(s) \leq Q_Y(t)$. Conditionally, $Q_{Y|\boldsymbol{X}}(s \mid \boldsymbol{X} = \boldsymbol{x}) \leq Q_{Y|\boldsymbol{X}}(t \mid \boldsymbol{X} = \boldsymbol{x})$ for any $\boldsymbol{x}$. This is analogous to CDF monotonicity: for $c < d$, $F_Y(c) \leq F_Y(d)$ unconditionally, and $F_{Y|\boldsymbol{X}}(c \mid \boldsymbol{X} = \boldsymbol{x}) \leq F_{Y|\boldsymbol{X}}(d \mid \boldsymbol{X} = \boldsymbol{x})$ for any $\boldsymbol{x}$.

Monotonicity plays a large role in Chapter 6.

Discussion Question 5.3 (QR monotonicity)**.** Consider the model $\mathrm{Q}_\tau(Y \mid X = x) = \alpha(\tau) + x\beta(\tau)$, for all $0 < \tau < 1$. That is, the intercept and slope can be different for different τ, according to the functions $\alpha(\cdot)$ and $\beta(\cdot)$.

a) Does quantile monotonicity imply $\beta(0.25) < \beta(0.5)$? Why/not?
b) Draw a picture (of conditional quantile functions) where $\beta(0.25) > \beta(0.5)$; does it look wrong? Why/not?
c) What is the "economic" interpretation of $\beta(0.25) > \beta(0.5)$?

5.2 Prediction

Recall from Hansen (2020a, §2.11) that the conditional mean provides the "best" predictor under quadratic loss. If we imagine repeated draws of Y given a single fixed $\boldsymbol{X} = \boldsymbol{x}$, then we are essentially in the unconditional setting: we simply treat the subpopulation with $\boldsymbol{X} = \boldsymbol{x}$ as the population and apply (4.5). If instead we imagine repeated draws of $(Y, \boldsymbol{X}')$ from the joint population distribution, then as shown by Hansen (2020a, §2.11),

$$\begin{aligned} \mathrm{E}(Y \mid \boldsymbol{X}) &= \underset{g(\boldsymbol{X})}{\arg\min}\, \mathrm{E}[L_2(Y, g(\boldsymbol{X}))] = \underset{g(\boldsymbol{X})}{\arg\min}\, \mathrm{E}[\rho_2(Y - g(\boldsymbol{X}))] \\ &= \underset{g(\boldsymbol{X})}{\arg\min}\, \mathrm{E}[(Y - g(\boldsymbol{X}))^2]. \end{aligned} \tag{5.9}$$

Here, $\mathrm{E}(Y \mid \boldsymbol{X})$ is a random variable rather than a function of $\boldsymbol{x}$: it conditions on the random variable $\boldsymbol{X}$, not on a particular value $\boldsymbol{X} = \boldsymbol{x}$. The "best" predictor $g(\cdot)$ gets the random variable $g(\boldsymbol{X})$ "closest" to Y in the stochastic sense of minimizing mean squared prediction error, $\mathrm{E}[(Y - g(\boldsymbol{X}))^2]$.

Parallel to the unconditional result in (4.8), replacing the quadratic loss function in (5.9) with the check function $\rho_\tau(\cdot)$ makes the τ-CQF the optimal predictor:

$$\mathrm{Q}_\tau(Y \mid \boldsymbol{X}) = \underset{g(\boldsymbol{X})}{\arg\min}\, \mathrm{E}[\rho_\tau(Y - g(\boldsymbol{X}))]. \tag{5.10}$$

The check function allows asymmetry in how bad it is to over-predict versus under-predict. Like before, when τ is closer to 1, it is very bad to under-predict ($g < y$), so it is optimal to guess relatively high values, i.e., high conditional quantiles. Conversely, when τ is near 0, over-prediction is very bad, so low conditional quantiles are better because they more often avoid over-prediction.

5.3 QR with Misspecification

With mean regression, even if the CEF is misspecified, the OLS estimator's probability limit has some meaningful interpretations: OLS estimates the linear projection of Y on $\boldsymbol{X}$, and the the linear projection is both the "best" linear approximation of the CEF as well as the "best" linear predictor of Y given $\boldsymbol{X}$. QR has analogous properties.

5.3.1 "Best" Linear Predictor

Recall from Hansen (2020a, §2.18) that even if the CEF is not of the form $\boldsymbol{x}'\boldsymbol{b}$, OLS still consistently estimates the "best" linear predictor, meaning the predictor of the form $\boldsymbol{X}'\boldsymbol{b}$ that minimizes expected quadratic loss.

QR also has a best linear predictor interpretation. From Theorem 5.1, even if the CQF is not linear in $\boldsymbol{x}$, QR is consistent for the population vector

$$\boldsymbol{\beta}(\tau) = \underset{\boldsymbol{b}}{\arg\min}\, \mathrm{E}[\rho_\tau(Y - \boldsymbol{X}'\boldsymbol{b})].$$

That is, among predictors of the form $\boldsymbol{X}'\boldsymbol{b}$, the predictor $\boldsymbol{X}'\boldsymbol{\beta}(\tau)$ minimizes risk (expected loss) under the loss function $L(y, g) = \rho_\tau(y - g)$. Thus, $\boldsymbol{X}'\boldsymbol{\beta}(\tau)$ is the "best" linear predictor if we define "best" according to $\rho_\tau(\cdot)$, just as the OLS plim is the "best" linear predictor if we define "best" according to $\rho_2(u) = u^2$.

5.3.2 "Best" Linear Approximation

Recall from Hansen (2020a, §2.25) that the OLS plim (the linear projection) is also the "best" linear approximation of the CEF (in terms of mean squared error). Writing the CEF as $m(\boldsymbol{x}) \equiv \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ and the linear projection coefficient as $\boldsymbol{\beta} = [\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\mathrm{E}(\boldsymbol{X}Y)$,

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}} \mathrm{E}[(\boldsymbol{X}'\boldsymbol{b} - m(\boldsymbol{X}))^2], \tag{5.11}$$

where the expectation is wrt the distribution of random vector $\boldsymbol{X}$.

Angrist, Chernozhukov, and Fernández-Val (2006, Thm. 1) provide a similar result for QR. Given the QR plim $\boldsymbol{\beta}(\tau)$, $\boldsymbol{x}'\boldsymbol{\beta}(\tau)$ is the "best" linear approximation of the true CQF in terms of a weighted mean squared error. Skipping the (complicated) definition of the weight,

$$\boldsymbol{\beta}(\tau) = \arg\min_{\boldsymbol{b}} \mathrm{E}\{\text{weight}_{\tau,\boldsymbol{X},\boldsymbol{b}} \times [\mathrm{Q}_\tau(Y \mid \boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{b}]^2\}. \tag{5.12}$$

This is not as easy to interpret as an unweighted mean squared error, but it's something.

5.4 Estimation

The same approaches from Section 4.4 can be used for QR. R function `rq()` in package `quantreg` (Koenker, 2019) is essentially the quantile analog of `lm()`.

Analogous to (4.15), the standard QR estimator is

$$\hat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{b}} \widehat{\mathrm{E}}[\rho_\tau(Y - \boldsymbol{X}'\boldsymbol{b})] = \arg\min_{\boldsymbol{b}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_i'\boldsymbol{b}). \tag{5.13}$$

This is OLS but with $\rho_\tau(\cdot)$ replacing the function $\rho_2(u) = u^2$. Instead of minimizing the sum of squared residuals, (5.13) minimizes the sum of "checked" residuals.

Computationally, (5.13) is more difficult than OLS. For OLS, the first-order condition leads to a closed-form expression for $\hat{\boldsymbol{\beta}}$. This is not possible for QR: $\rho_\tau(\cdot)$ is not differentiable at zero. However, some clever algorithms make QR very fast in practice.

Discussion Question 5.4 (quantile crossing problem)**.** Let Y and X be scalars. You estimate quantile regressions of the form $\beta_0(\tau) + X\beta_1(\tau)$ for $\tau = 0.5$ and $\tau = 0.75$. You estimate $\hat{\beta}_1(0.75) = 2$ and $\hat{\beta}_1(0.5) = 1$ for the slope coefficients. Picking any $\hat{\beta}_0(0.75)$ and $\hat{\beta}_0(0.5)$, can you draw the two estimated functions such that monotonicity is preserved, i.e., $\hat{\beta}_0(0.75)+x\hat{\beta}_1(0.75) > \hat{\beta}_0(0.5)+x\hat{\beta}_1(0.5)$ for all values x in the support of X? Explain why or why not, or any other considerations.

Similar to OLS being consistent for the LP/BLP/BLA rather than the CEF, QR is generally consistent for the population objects in Section 5.3 rather than the CQF. To estimate the true CEF or CQF, nonparametric methods are best; see Chapters 16 and 17 (in Part V) for general nonparametric estimation approaches and R packages (e.g., `np` has function `npqreg` for nonparametric QR).

5.5 Asymptotic Properties

Angrist, Chernozhukov, and Fernández-Val (2006, Thm. 3, p. 549) establish consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}(\tau)$ (for the corresponding population minimizer), uniformly over a continuum of τ, under certain assumptions (sufficient conditions). I comment on some of the assumptions and state the results, but refer to Angrist, Chernozhukov, and Fernández-Val (2006) for details.

Let $\mathcal{T} = [\epsilon, 1-\epsilon]$ for some $\epsilon > 0$. (To learn about "extreme" quantiles like $\tau = 1/n$, different methods are required.)

Condition (i) in Theorem 3 of Angrist, Chernozhukov, and Fernández-Val (2006) is iid sampling. This is sufficient, but not necessary (time series QR results also exist).

Condition (ii) is about the smoothness of the conditional PDF of Y, $f_Y(y \mid X = x)$. This PDF's very existence excludes discrete Y.

Condition (iii) is a rank condition, similar to $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')$ being invertible for OLS. Here, the conditional PDF of Y is also involved: $\mathrm{E}[f_{Y|\boldsymbol{X}}(\boldsymbol{X}'\boldsymbol{\beta}(\tau) \mid \boldsymbol{X})\boldsymbol{X}\boldsymbol{X}']$ must be invertible.

Condition (iv) requires finite variance (slightly stronger) for $\boldsymbol{X}$, but not for Y. Unlike with OLS, here it is fine if Y does not even have a well-defined mean.

Theorem 5.1 (Theorem 3 of Angrist, Chernozhukov, and Fernández-Val (2006))**.** *Let* $\boldsymbol{\beta}(\tau) = \arg\min_{\boldsymbol{b}} \mathrm{E}[\rho_\tau(Y - \boldsymbol{X}'\boldsymbol{b})]$. *Under conditions (i)–(iv) in Theorem 3 of Angrist, Chernozhukov, and Fernández-Val (2006),*

$$\hat{\boldsymbol{\beta}}(\tau) \xrightarrow{p} \arg\min_{\boldsymbol{b}} \mathrm{E}[\rho_\tau(Y - \boldsymbol{X}'\boldsymbol{b})]. \tag{5.14}$$

More strongly than the pointwise consistency of (5.14)*, there is uniform consistency:*

$$\sup_{\tau \in \mathcal{T}} \|\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\| \xrightarrow{p} 0. \tag{5.15}$$

For intuition about uniform consistency, recall $W_n \xrightarrow{p} w$ is equivalent to $W_n - w \xrightarrow{p} 0$. Similarly, $\hat{\boldsymbol{\beta}}(\tau) \xrightarrow{p} \boldsymbol{\beta}(\tau)$ is equivalent to $\|\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\| \xrightarrow{p} 0$. Theorem 5.1 extends this by taking a supremum over $\tau \in \mathcal{T}$.

Angrist, Chernozhukov, and Fernández-Val (2006, Thm. 3) also establish asymptotic normality of the QR estimator under misspecification. This includes "pointwise" asymptotic normality of the vector $\hat{\boldsymbol{\beta}}(\tau)$ for a single τ. They also show that the random function $\hat{\boldsymbol{\beta}}(\cdot)$ over $\tau \in \mathcal{T}$ is "asymptotically normal," i.e., when centered and scaled it converges to a (multivariate) **Gaussian process**. A (scalar) Gaussian process is a random function

$G(\cdot)$ whose finite-dimensional marginals follow multivariate normal (Gaussian) distributions, i.e., $(G(t_1), \ldots, G(t_k))$ is multivariate normal. Besides being fancy, this allows us to quantify our statistical uncertainty about the relationship among $\boldsymbol{\beta}(\tau)$ for different τ. For example, we could construct a **uniform confidence band** that includes the true function $\boldsymbol{\beta}(\cdot)$ with $1-\alpha$ probability (asymptotically), or test a hypothesis involving multiple τ.

5.6 Inference

One option for inference (confidence intervals, hypothesis testing) is to use the Gaussian limit distribution. However, the asymptotic covariance matrix is difficult to estimate accurately due to the conditional PDF term.

There are many other approaches to QR inference in the literature, although many of them historically have assumed homoskedasticity, which economists usually avoid.

Angrist, Chernozhukov, and Fernández-Val (2006) suggest subsampling for inference on the function $\boldsymbol{\beta}(\cdot)$; see their Section 3 (and see my Section 13.4 for a basic introduction).

For pointwise (single τ at a time) inference, Bayesian bootstrap is one possibility; see Chapter 14 and Section 13.1 and Hahn (1997), for example.

Chernozhukov, Hansen, and Jansson (2009) offer a clever approach that's exact even in finite samples. However, it relies on having a properly specified conditional quantile function. The general idea is: if $q_\tau(x) = \mathrm{Q}_\tau(Y \mid X = x)$, then $\mathrm{P}(Y \leq q_\tau(X)) = \tau$, so (with iid sampling) $\mathbb{1}\{Y \leq q_\tau(X)\}$ are iid Bernoulli(τ).

5.7 Censoring

The ideas in Section 4.5 extend to QR. In Stata, try the `cqiv` command available in SSC.

Chapter 6
Quantile Regression: Causality

Unit learning objectives for this chapter

6.1. Interpret the structural and treatment effect parameters that can be estimated with quantile methods [TLO 1]

6.2. Develop intuition for random coefficients models and potential outcomes [TLO 1]

6.3. Judge whether conditional or unconditional quantile regression better answers a particular economic question [TLO 3]

There are two primary frameworks for learning about causality with quantiles. First, the quantile treatment effect extends the average treatment effect, within the potential outcomes framework. Second, QR can estimate a structural random coefficients model under certain assumptions. Both approaches can allow endogeneity as in Chapter 7.

Optional resources for this chapter

- *Handbook of Quantile Regression*
- R and Stata code on Blaise Melly's website

6.1 Background: Potential Outcomes and ATE

The following is a very brief review; see Section 4.4 of Kaplan (2022b) for details.

Sometimes, there is a binary "treatment" that only affects the treated individual (or firm, or county, or whatever unit) and nobody else. This makes sense for something like a medical intervention (e.g., knee surgery), but it is often unrealistic in economics since it excludes peer effects, general equilibrium effects, spillovers, etc. Nonetheless, economists

study many areas, and sometimes it's plausible. This assumption (that one individual's treatment does not affect anyone else) is sometimes called "no interference" and is part of the stable unit treatment value assumption (SUTVA).

Let Y_1 denote an individual's treated **potential outcome** and Y_0 her untreated potential outcome. These refer to the individual's outcome in two parallel universes: one in which the individual is treated, and another in which the individual is not treated, but where everything else is identical between the two parallel universes.

Different population objects can be formed from these potential outcomes (Y_1, Y_0). The **treatment effect** for an individual is $Y_1 - Y_0$. (Or, the treatment effect for individual i is $Y_{1i} - Y_{0i}$.) The **average treatment effect** (ATE) takes the population mean of the individual treatment effect: $\text{ATE} = \mathrm{E}(Y_1 - Y_0)$. By linearity of expectation, $\mathrm{E}(Y_1 - Y_0) = \mathrm{E}(Y_1) - \mathrm{E}(Y_0)$, the treatment effect on the mean. That is, $\mathrm{E}(Y_1)$ is the mean outcome in the parallel universe where everyone is treated, $\mathrm{E}(Y_0)$ is the mean outcome in the parallel universe where nobody is treated, and $\mathrm{E}(Y_1) - \mathrm{E}(Y_0)$ shows how the mean outcome changes (the effect on the mean) when we move from the all-untreated universe to the all-treated universe. Even though the interpretation differs, this is often just called the ATE since it is mathematically equivalent.

The difficulty is: for any individual, usually we only observe one potential outcome or the other, not both; we cannot travel to the parallel universe. Thus, we cannot observe $Y_1 - Y_0$ for any individual. Thus, we cannot estimate $\mathrm{E}(Y_1 - Y_0)$ by $\widehat{\mathrm{E}}(Y_1 - Y_0)$.

However, we can get more traction on the equivalent formulation $\mathrm{E}(Y_1) - \mathrm{E}(Y_0)$. We can take the sample average outcome of treated individuals, and subtract the sample average outcome of untreated individuals. But, more assumptions are required for this to work well.

Assumptions are required for the ATE to be identified. There are multiple ways to think about identification in this case. Let $X = 1$ if the individual is treated and $X = 0$ otherwise. The observed outcome is $Y = Y_0 + X(Y_1 - Y_0)$. This could be seen as a simple regression model with random intercept Y_0 and random slope $Y_1 - Y_0$. If the random coefficients $(Y_0, Y_1 - Y_0)$ are independent of the regressor X, then OLS can estimate $\mathrm{E}(Y_0)$ and $\mathrm{E}(Y_1 - Y_0)$; e.g., see Theorem 2.11 in Hansen (2020a, §2.29).

Alternatively, if $Y_0, Y_1 \perp\!\!\!\perp X$, then

$$\mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0) = \mathrm{E}(Y_1 \mid X = 1) - \mathrm{E}(Y_0 \mid X = 0) = \mathrm{E}(Y_1) - \mathrm{E}(Y_0), \quad (6.1)$$

where the first equality uses $Y = Y_0 + X(Y_1 - Y_0)$ (which implies $Y = Y_1$ if $X = 1$, and $Y = Y_0$ if $X = 0$), and the second equality uses independence (so conditioning on X does not change the mean of Y_1 or Y_0; X "has no information" about Y_0 or Y_1). The ATE is "identified" because it is equal to an expression that depends only on the population joint distribution of the observable (Y, X). That is, given the independence assumption $Y_0, Y_1 \perp\!\!\!\perp X$, the distribution of (Y, X) uniquely determines the ATE because the ATE equals the difference of observable conditional means, as seen in (6.1).

Thus, given independence, the ATE can be estimated by $\widehat{\mathrm{E}}(Y \mid X = 1) - \widehat{\mathrm{E}}(Y \mid X = 0)$, the difference of the subsample averages.

Table 6.1: Potential outcomes example for DQ 6.1.

Y_0	Y_1	$Y_1 - Y_0$	Probability
0	1	1	0.25
1	2	1	0.25
2	4	2	0.25
3	0	-3	0.25

6.2 Quantile Treatment Effects

Discussion Question 6.1 (ATE, QTE)**.** Table 6.1 describes a population with four types of individuals, each with probability 0.25. Each "type" has a different (Y_0, Y_1) potential outcome pair.

a) Compute $\mathrm{E}(Y_0)$.
b) Compute $\mathrm{E}(Y_1)$.
c) Compute $\mathrm{E}(Y_1) - \mathrm{E}(Y_0)$.
d) Compute $\mathrm{E}(Y_1 - Y_0)$.
e) Compute $\mathrm{Q}_{0.4}(Y_1) - \mathrm{Q}_{0.4}(Y_0)$.
f) Compute $\mathrm{Q}_{0.4}(Y_1 - Y_0)$.

Hint: here, $\mathrm{Q}_{0.4}$ is simply the second-smallest value, per (4.1).

As DQ 6.1 illustrates, $\mathrm{Q}_\tau(Y_1 - Y_0) \neq \mathrm{Q}_\tau(Y_1) - \mathrm{Q}_\tau(Y_0)$. Unlike the expectation operator, the quantile operator is nonlinear. Thus, the τ-quantile of the population distribution of individual treatment effects $Y_1 - Y_0$ differs from the treatment effect on the τ-quantile of the outcome distribution.

For the same reasons as in Section 6.1, it is difficult to learn about $\mathrm{Q}_\tau(Y_1 - Y_0)$ because we never observe $Y_1 - Y_0$. In fact, even if we know the population marginal distributions of Y_1 and Y_0, we can only learn bounds for $\mathrm{Q}_\tau(Y_1 - Y_0)$; see Fan and Park (2010, 2012).

This is one of two reasons to focus on $\mathrm{Q}_\tau(Y_1) - \mathrm{Q}_\tau(Y_0)$, called the **quantile treatment effect** (QTE), or more specifically the τ-QTE. The other reason is that QTEs describe how the treatment affects quantiles of the population outcome distribution. If we have a social welfare function whose input is the population distribution of Y, and we wish to learn the effect of a treatment on social welfare, then it is more relevant to look at QTEs than quantiles of treatment effects.

QTE identification parallels (6.1) given $Y_0, Y_1 \perp\!\!\!\perp X$:

$$\begin{aligned} \mathrm{Q}_\tau(Y \mid X = 1) - \mathrm{Q}_\tau(Y \mid X = 0) &= \mathrm{Q}_\tau(Y_1 \mid X = 1) - \mathrm{Q}_\tau(Y_0 \mid X = 0) \\ &= \mathrm{Q}_\tau(Y_1) - \mathrm{Q}_\tau(Y_0) \equiv \tau\text{-QTE}. \end{aligned} \tag{6.2}$$

Thus, the τ-QTE can be estimated by $\widehat{\mathrm{Q}}_\tau(Y \mid X = 1) - \widehat{\mathrm{Q}}_\tau(Y \mid X = 0)$, the difference of treated and untreated sample τ-quantiles.

More generally, independence identifies the full marginal distributions of Y_1 and Y_0, so any summary of these distributions is also identified. Beyond the mean and quantiles, this includes how treatment affects the standard deviation, interquantile ranges, upper tail, lower tail, etc.

Discussion Question 6.2 (effect heterogeneity)**.** For the following, try to define a relevant object of interest in terms of Y_1, Y_0, X (treatment dummy), and possibly other variables. That is: ideally, what do we want to learn? Also: is this related to QTEs at all, and if so, how? Hint: who is actually affected by the policy change?

a) The Missouri state legislature is considering increasing funding to increase the number of college scholarships (to increase college degree attainment); they want to know the effect of such a policy change on individuals' annual earnings.
b) The Missouri state legislature is considering lowering the income threshold for Medicaid (health insurance for low-income individuals and families) so fewer people are eligible; they want to know the effect on total annual emergency room visits in Missouri.
c) Expanding public pre-school: they want to know the effect on 5th-grade math scores.

As DQ 6.2 suggests, there are different types of heterogeneity in treatment effects. QTEs capture more heterogeneity than the ATE, but there is also (for example) heterogeneity along the dimension of propensity to be treated; e.g., see the (conditional, average) marginal treatment effect (MTE) of Heckman and Vytlacil (2001, 2007).

6.3 Background: Random Coefficients

Previously, you've seen the structural model $Y = \boldsymbol{X}'\boldsymbol{\beta}+V$, in which the uppercase letters denote random variables, whereas the coefficient vector $\boldsymbol{\beta}$ is non-random. The model is also written with subscripts as $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta}+V_i$, where the random variables have individual i subscripts but the coefficient vector $\boldsymbol{\beta}$ does not. "Random" essentially means that each individual i has their own $(Y_i, \boldsymbol{X}_i, V_i)$ drawn from the population distribution of random vector $(Y, \boldsymbol{X}, V)$. In contrast, $\boldsymbol{\beta}$ is a constant, the same for all individuals.

Alternatively, different individuals may have their own different coefficients. For example, some individuals may have a higher "return to education" than others, or firms may have different parameters in their production functions. Such individual-specific coefficients are usually called **random coefficients**. To model this, the constant $\boldsymbol{\beta}$ can be replaced with random vector $\boldsymbol{B}$.

The resulting structural model is $Y = \boldsymbol{X}'\boldsymbol{B}$. An additive error V would be redundant if $\boldsymbol{X}$ includes an intercept; e.g., if $Y = \tilde{B}_0 + B_1 X + V$, then equivalently $Y = (B_0, B_1)(1, X)'$ with $B_0 \equiv \tilde{B}_0 + V$. The population is now the joint distribution of $(Y, \boldsymbol{X}, \boldsymbol{B})$.

Exogeneity here means $\boldsymbol{B}$ is unrelated to $\boldsymbol{X}$, like $\boldsymbol{B} \perp\!\!\!\perp \boldsymbol{X}$ (independence) or $\mathrm{E}(\boldsymbol{B} \mid \boldsymbol{X}) = \mathrm{E}(\boldsymbol{B})$ (mean independence). The idea is the same as usual: regressors ($\boldsymbol{X}$) are unrelated to unobserved determinants of Y (here $\boldsymbol{B}$; previously V).

Given exogeneity, $\mathrm{E}(\boldsymbol{B})$ is identified by the CEF slope:

$$\mathrm{E}(Y \mid \boldsymbol{X}=\boldsymbol{x}) = \mathrm{E}(\overbrace{\boldsymbol{X}'\boldsymbol{B}}^{Y} \mid \boldsymbol{X}=\boldsymbol{x}) = \boldsymbol{x}'\overbrace{\mathrm{E}(\boldsymbol{B} \mid \boldsymbol{X}=\boldsymbol{x})}^{\text{use mean } \perp\!\!\!\perp} = \boldsymbol{x}'\,\mathrm{E}(\boldsymbol{B}). \tag{6.3}$$

That is, the CEF is $\mathrm{E}(Y \mid \boldsymbol{X}=\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$ with $\boldsymbol{\beta} = \mathrm{E}(\boldsymbol{B})$. Thus, if we regress Y on $\boldsymbol{X}$, OLS consistently estimates $\boldsymbol{\beta}$, which we can interpret as the mean of the structural random coefficient vector $\boldsymbol{B}$. The estimator (OLS) is the same as usual; the interpretation is new.

Discussion Question 6.3 (random coefficient exogeneity: wage)**.** Previously, you've (probably) thought about why there is endogeneity in the structural model $Y = \beta_0 + \beta_1 X + U$, where Y is log wage, X is years of education, β_0 and β_1 are fixed constant parameters, and U is other determinants of Y. Now, consider the same Y and X, but in the structural random coefficients model $Y = B_0 + B_1 X$.

a) What does it mean that B_0 is "random"? Explain why you find this realistic or not.
b) What does it mean that B_1 is "random"? Explain why you find this realistic or not.
c) Explain why B_0 and X might be correlated (and in which direction).
d) Explain why B_1 and X might be correlated (and in which direction).

So, can we learn anything else about $\boldsymbol{B}$ besides its mean? Section 6.4 considers how to link the structural random coefficients model to the CQFs instead of the CEF.

6.4 A Random Coefficients Model for QR

To link the structural random coefficients model to conditional quantiles, additional restrictions are imposed beyond Section 6.3.

6.4.1 The Model

To be concrete, imagine Y is log wage and $X = 1$ for "high education" and $X = 0$ for "low education." A general structural random coefficients model is $Y = B_0 + B_1 X$, where B_0 is the individual's log wage when $X = 0$ and B_1 is the change in the individual's log wage cause by the change from low to high education. That is, B_1 is the individual's return to schooling. The coefficients are "random" in that each individual is allowed to have a different log wage given low education (B_0) as well as a different return to schooling (B_1).

Now, assume the heterogeneity in both the intercept and slope can be represented by a scalar random variable U. That is, instead of (Y, X, B_0, B_1), each individual has their own (Y, X, U), and then U determines both the intercept and slope. To be concrete, imagine U represents "ability." Specifically, there are (non-random) functions $\beta_0(\cdot)$ and $\beta_1(\cdot)$ such that the random intercept is $B_0 = \beta_0(U)$ and the random slope is $B_1 = \beta_1(U)$. Thus, the structural random coefficients model is

$$Y = \beta_0(U) + \beta_1(U)X. \tag{6.4}$$

Model (6.4) is more restrictive than the general random coefficients model $Y = B_0 + B_1X$, but less restrictive than $Y = \beta_0 + \beta_1X + U$... sort of. Just as (6.4) is a special case of $Y = B_0 + B_1X$, $Y = \beta_0 + \beta_1X + U$ is a special case of (6.4): set $\beta_1(U) = \beta_1$ and $\beta_0(U) = \beta_0 + U$. That said, if $Y = \beta_0 + \beta_1X + U$ is a CEF model, then (given independence) the constant slope β_1 can be interpreted as $\mathrm{E}(B_1)$ in the more general structural model $Y = B_0 + B_1X$.

The U is normalized to $U \sim \mathrm{Unif}(0, 1)$, without loss of generality. For example, if $B_0 \sim \mathrm{N}(0, 1)$, then let $\beta_0(\cdot) = \Phi^{-1}(\cdot)$, the inverse CDF of $\mathrm{N}(0, 1)$, so $\beta_0(U) = \Phi^{-1}(U) \sim \mathrm{N}(0, 1)$.

To develop intuition, you could "slice" (6.4) two ways: fix $X = x$, or fix $U = u$. Fixing $X = x$, we see the relationship between Y and U for a particular education subpopulation: $Y = \beta_0(U)$ for the low-education subpopulation ($X = 0$), $Y = \beta_0(U) + \beta_1(U)$ for the high-education subpopulation ($X = 1$). So you can think of the functions $\beta_0(\cdot)$ and $\beta_1(\cdot)$ as describing the wage–ability relationship. Alternatively, fixing $U = u$, we see the relationship between Y and X for a particular ability level: $Y = \beta_0(u) + \beta_1(u)X$. For example, for individuals with median ability $U = 0.5$, $Y = \beta_0(0.5) + \beta_1(0.5)X$, where $\beta_0(0.5)$ and $\beta_1(0.5)$ are constants (not random). For individuals with upper quartile ability $U = 0.75$, $Y = \beta_0(0.75) + \beta_1(0.75)X$. So you can also think of the functions $\beta_0(\cdot)$ and $\beta_1(\cdot)$ as describing the wage–education relationship at different ability levels.

Discussion Question 6.4 (structural random coefficients wage model)**.** Consider (6.4), where Y is log wage and $X = 1$ if high education ($X = 0$ if low); you can think of U as "ability." Let $u_2 > u_1$. Based on economic theory or your intuition, what do you think is the relationship between the objects in each of the following pairs? Explain.

a) Between $\beta_0(u_2)$ and $\beta_0(u_1)$?
b) Between $\beta_1(u_2)$ and $\beta_1(u_1)$?
c) Between $\beta_0(u_2) + \beta_1(u_2)$ and $\beta_0(u_1) + \beta_1(u_1)$?

6.4.2 Monotonicity and Identification

Although there is no quantile analog of $\mathrm{E}(\boldsymbol{X}'\boldsymbol{B} \mid \boldsymbol{X}) = \boldsymbol{X}'\mathrm{E}(\boldsymbol{B} \mid \boldsymbol{X})$, certain features of the structural model can be identified (linked to conditional quantiles) under exogeneity and another assumption called **monotonicity**. A crude interpretation of A6.1 in the log wage model would be: given either level of education ($X = 0$ or $X = 1$), log wage (Y) is strictly increasing in "ability" (U).

Assumption A6.1 (structural QR monotonicity)**.** In the structural random coefficients model $Y = \boldsymbol{X}'\boldsymbol{\beta}(U)$, Y is strictly increasing in U given any $\boldsymbol{X} = \boldsymbol{x}$. That is, given any $\boldsymbol{x}$ in the support of $\boldsymbol{X}$, the function $\boldsymbol{x}'\boldsymbol{\beta}(u)$ is strictly increasing in u over $0 \leq u \leq 1$.

To get started on DQ 6.5, consider the following proof that $\mathrm{Q}_{0.5}(Y \mid X = 0) = \beta_0(0.5)$ given (6.4) and A6.1 and independence ($U \perp\!\!\!\perp X$). By (6.4), if $X = 0$, then $Y = \beta_0(U)$. By independence, the median of $\beta_0(U)$ is independent of X. By monotonicity, the median of $\beta_0(U)$ is $\beta_0(\cdot)$ evaluated at the median of U, which is 0.5 by the normalization $U \sim$

$\text{Unif}(0,1)$. (Quantile **equivariance** refers to the property that $\text{Q}_\tau(f(W)) = f(\text{Q}_\tau(W))$ if $f(\cdot)$ is strictly increasing; this has been used to simplify estimation of quantile Euler equations in de Castro, Galvao, Kaplan, and Liu (2019, §6.3), for example.) Altogether, with the first two equalities analogous to (6.3),

$$\text{Q}_{0.5}(Y \mid X = 0) = \text{Q}_{0.5}(\beta_0(U) \mid X = 0) = \text{Q}_{0.5}(\beta_0(U)) = \beta_0(\text{Q}_{0.5}(U)) = \beta_0(0.5). \quad (6.5)$$

Discussion Question 6.5 (random coefficients model conditional quantiles)**.** Consider (6.4) with $X \perp\!\!\!\perp U$, Assumption A6.1, and the normalization $U \sim \text{Unif}(0,1)$. Similar to (6.5), express the following statistical objects (i.e., features of the joint distribution of observables (Y, X)) in terms of the functions $\beta_0(\cdot)$ and $\beta_1(\cdot)$ (i.e., the structural parameters).

a) The median log wage in the high-education subpopulation, $\text{Q}_{0.5}(Y \mid X = 1)$.
b) The difference between the two prior objects, $\text{Q}_{0.5}(Y \mid X = 1) - \text{Q}_{0.5}(Y \mid X = 0)$.
c) Other quantiles of log wage in the low-education subpopulation: $\text{Q}_{0.25}(Y \mid X = 0)$ and $\text{Q}_{0.75}(Y \mid X = 0)$.

Finally:

d) What is the return to education for an individual with median ability? (Recall the ability distribution is $U \sim \text{Unif}(0,1)$; what's the median?)

Discussion Question 6.6 (QR monotonicity 1)**.** For each of the following, construct an example $\beta_0(\cdot)$ and $\beta_1(\cdot)$ that satisfy the stated requirements while still satisfying monotonicity (A6.1). Or, if you think it is impossible, explain why. Hint: drawing may help.

a) Education increases the log wage of every individual, but the increase is larger when "ability" (U) is higher.
b) "Ability" (U) affects log wage when $X = 0$, but everyone has very similar wage for $X = 1$.
c) Some individuals have a lower log wage with high education than with low education (i.e., education lowers their wage).

Discussion Question 6.7 (QR monotonicity 2)**.** Continue DQ 6.6.

a) Of the different possible conditions, which do you think is the most realistic, and why?
b) How/would $X \perp\!\!\!\perp U$ affect any of your answers?

6.4.3 Heteroskedasticity

Discussion Question 6.8 (random coefficients model: heteroskedasticity)**.** Consider the random coefficients model $Y = \beta_0(U) + \beta_1(U)X$. Construct an example with heteroskedasticity that still satisfies $U \perp\!\!\!\perp X$ (and A6.1). Hint: for simplicity, let $\beta_0(u) = 0$ for all u, and let $X \in \{0,1\}$; compute $\text{Var}(Y \mid X = 0)$ and pick $\beta_1(\cdot)$ such that $\text{Var}(Y \mid X = 1) > \text{Var}(Y \mid X = 0)$.

As DQ 6.8 shows, there can be heteroskedasticity even if $U \perp\!\!\!\perp X$. This is different than with an **additively separable** error, like $Y = \boldsymbol{X}'\boldsymbol{\beta} + U$: $U \perp\!\!\!\perp \boldsymbol{X}$ implies

$$\operatorname{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \operatorname{Var}(\boldsymbol{X}'\boldsymbol{\beta} + U \mid \boldsymbol{X} = \boldsymbol{x}) = \overbrace{\operatorname{Var}(U \mid \boldsymbol{X} = \boldsymbol{x}) = \operatorname{Var}(U)}^{\text{by } U \perp\!\!\!\perp \boldsymbol{X}}, \tag{6.6}$$

homoskedasticity. The model $Y = \boldsymbol{X}'\boldsymbol{\beta}(U)$ is **nonseparable**, so heteroskedasticity can arise through $\boldsymbol{\beta}(\cdot)$ even if $U \perp\!\!\!\perp \boldsymbol{X}$.

6.5 Unconditional Quantile Regression

Consider the following way to evaluate "how good" is a population's distribution of some outcome Y. For example, Y is income, or a composite measure of well-being. Let $w(\cdot)$ be a social welfare function, like a utility function but for the whole society (population), not just an individual. If Y is a random variable representing the income or well-being of an individual from the population of interest, then $w(Y)$ provides a scalar summary of "how good" is the distribution of Y. This is similar to computing your expected utility for a lottery, to summarize "how good" that lottery is.

My point is simply to motivate the policy interest in the overall (unconditional) population distribution of Y. In certain conditions, QTEs can help us learn about how a certain policy affects the distribution of Y; see Section 6.2. In other cases, the potential policy change is not binary, and we may need to condition on other variables for it to be exogenous.

The goal of **unconditional quantile regression** (UQR) is to see the "effect" of changes in $\boldsymbol{X}$ to the unconditional distribution (quantiles) of Y. This goal aligns with the social welfare approach to policy analysis. The output is often more directly relevant for policy than the coefficients of a structural model like $\boldsymbol{\beta}(u)$ for various $0 < u < 1$. The "change" is a change in the marginal distribution of $\boldsymbol{X}$. The "effect" assumes the conditional distributions of Y given any $\boldsymbol{X} = \boldsymbol{x}$ are invariant to (unaffected by) the policy.

For different approaches to UQR estimation, see Firpo, Fortin, and Lemieux (2009) and Chernozhukov, Fernández-Val, and Melly (2013). The former is simpler but only applies to infinitesimal policy changes. See also Sasaki, Ura, and Zhang (2020) for UQR with high-dimensional data (many regressors). For more discussion of when UQR can be used to estimate policy effects in practice, see the paragraph after Proposition 1 of Rothe (2010).

Exercises

Exercise E6.1. a. Find a paper that runs a randomized experiment (and makes its data publicly available) but does not look at quantile effects; provide a link to the paper. The paper must be either published in a respectable economics journal[1] or be unpublished but have an author who has previously published in such a journal (like any Mizzou econ professor); or if you really want, you can use an example from an econometrics textbook.

b. Replicate (at least, reasonably close) one particular average treatment effect estimate from the paper, or intention-to-treat estimate if treatment assignment is randomized but not treatment itself (i.e., estimate the average effect of treatment assignment, rather than the average effect of the treatment itself). If code is provided with the paper, feel free to use it (just say so).

c. For $\tau = 0.1, 0.2, \ldots, 0.9$ (or more, if you want), estimate the τ-QTE (or τ-quantile ITT effect).

d. Describe the pattern of the effect estimates over τ (e.g., roughly constant, increasing, decreasing, etc.), and compare the values with the average effect estimate.

e. "Economically": interpret the pattern/comparison you just described, and explain a reason why that pattern may exist (like "The pattern of the effect increasing with τ could be explained by…").

Exercise E6.2. a. Find a paper that uses conditional independence (a.k.a. unconfoundedness or selection-on-observables) for its identification strategy and then runs OLS. The paper must be either published in a respectable economics journal[2] or be unpublished but have an author who has previously published in such a journal (like any Mizzou econ professor); or if nobody else takes it then https://doi.org/10.1177/2332858417690511 is ok, too; or if you really want, you can use an example from an econometrics textbook.

b. Replicate (at least, reasonably close) one particular estimate of interest (the coefficient on a particular regressor). If code is provided with the paper, feel free to use it (just say so).

c. For $\tau = 0.1, 0.2, \ldots, 0.9$, run unconditional quantile regression, and report the coefficient estimates for the regressor of interest. (Stata: I think if you install the command `rifreg` or `xtrifreg` or maybe `rifhdreg` it can work, but maybe there are better options?)

d. For the regressor of interest, describe the pattern of the UQR coefficient estimates over τ (e.g., roughly constant, increasing, decreasing, etc.), and compare the values with the OLS coefficient estimate.

[1] For example, in top 500 of https://ideas.repec.org/top/top.journals.all.html
[2] For example, in top 500 of https://ideas.repec.org/top/top.journals.all.html

e. "Economically": interpret the pattern/comparison you just described, and explain a reason why that pattern may exist (like "The pattern of the effect increasing with τ could be explained by...").

f. For the same τ, run the usual quantile regression (Stata: `qreg`), and report the coefficient estimates for the regressor of interest.

g. Interpret the QR coefficients, explaining explicitly how the interpretation differs from the UQR coefficients.

Chapter 7
Quantile Regression: Endogeneity

Unit learning objectives for this chapter

7.1. Develop intuition for different approaches to endogeneity in quantile models [TLO 2]

7.2. Compare quantile and mean structural models with endogeneity [TLO 3]

This chapter discusses identification and estimation of the models in Chapter 6 under endogeneity.

Optional resources for this chapter

- *Handbook of Quantile Regression*, especially Chernozhukov, Hansen, and Wüthrich (2017) and Melly and Wüthrich (2017)
- R and Stata code for IVQR: `https://kaplandm.github.io`
- R and Stata code on Blaise Melly's website
- Stata: `cqiv` can be installed from SSC for control function estimation (with or without censoring)

7.1 Instrumental Variables Quantile Regression

Chernozhukov and Hansen (2005) establish identification results for the **instrumental variables quantile regression** (IVQR) model. (Previous attempts did not really succeed in extending IV/2SLS to QR.)

7.1.1 Reminder: Usual IV Regression

In the standard IV/2SLS identification argument, the instruments must satisfy two conditions: exogeneity and relevance. Exogeneity ensures that the true structural parameter vector value satisfies certain moment conditions. Relevance ensures that no other value satisfies the moment conditions.

For example, consider structural model $Y = \boldsymbol{X}'\boldsymbol{\beta}+U$ with endogeneity so $\mathrm{E}(\boldsymbol{X}U) \neq \mathbf{0}$ but the full instrument vector $\boldsymbol{Z}$ satisfies $\mathrm{E}(\boldsymbol{Z}U) = \mathbf{0}$ (exogeneity), where $\boldsymbol{Z}$ includes both exogenous regressors and excluded instruments (that do not appear in the structural model). Substituting $U = Y - \boldsymbol{X}'\boldsymbol{\beta}$ from the structural model, the moment condition becomes $\mathrm{E}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{b})] = \mathbf{0}$, which is satisfied by $\boldsymbol{b} = \boldsymbol{\beta}$. Given exact identification, $\boldsymbol{\beta} = [\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')]^{-1}\,\mathrm{E}(\boldsymbol{Z}Y)$ is the unique solution if $\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')$ is invertible (relevance).

Given the moment conditions and identification, GMM consistently estimates the structural parameters. (GMM with weighting matrix $\widehat{\mathrm{E}}(\boldsymbol{Z}\boldsymbol{Z}')$ is 2SLS.)

7.1.2 IVQR Identification

Chernozhukov and Hansen (2005) show how the structural parameters in the random coefficients model in Section 6.4 satisfy certain moment conditions given exogenous instruments. The relevance condition is qualitatively similar to the usual IV relevance, but more technically complicated; see Chernozhukov and Hansen (2005) for details. Chernozhukov and Hansen (2005) allow a more general functional form (and discuss a more complicated potential outcomes model), but the core intuition is the same as below.

As in Section 6.4, consider structural model

$$Y = \boldsymbol{X}'\boldsymbol{\beta}(U), \quad U \sim \mathrm{Unif}(0,1). \tag{7.1}$$

Let $\boldsymbol{Z}$ be the full vector of instruments, including both exogenous regressors and excluded instruments. Here, "exogeneity" means $\boldsymbol{Z} \perp\!\!\!\perp U$. (Recall from Section 6.4.3 that this still allows heteroskedasticity.) Monotonicity (A6.1) is again assumed.

The following derivation is similar to (6.5) and DQ 6.5. Independence and monotonicity are both crucial. Independence implies that probabilities involving (only) U are unaffected by conditioning on $\boldsymbol{Z}$. Monotonicity implies that, given $\boldsymbol{X} = \boldsymbol{x}$, the relative value of Y depends on U. The conclusion is that the structural parameter $\boldsymbol{\beta}(\tau)$ for some $0 < \tau < 1$ solves a particular conditional probability that involves only observable variables.

Formalizing the above verbal arguments,

$$\begin{aligned}
\mathrm{P}(Y \leq \boldsymbol{X}'\boldsymbol{\beta}(\tau) \mid \boldsymbol{Z}) &= \mathrm{P}(\boldsymbol{X}'\boldsymbol{\beta}(U) \leq \boldsymbol{X}'\boldsymbol{\beta}(\tau) \mid \boldsymbol{Z}) && \text{by (7.1)} && (7.2)\\
&= \mathrm{P}(U \leq \tau \mid \boldsymbol{Z}) && \text{by A6.1} && (7.3)\\
&= \mathrm{P}(U \leq \tau) && \text{by } U \perp\!\!\!\perp \boldsymbol{Z} && (7.4)\\
&= \tau && \text{by (7.1), } U \sim \mathrm{Unif}(0,1). && (7.5)
\end{aligned}$$

That is, the structural $\boldsymbol{\beta}(\tau)$ satisfies a particular conditional quantile restriction. However, unlike with 2SLS, it does not correspond to a second-stage regular QR where the regressors are fitted values from a first-stage regression.

The conditional quantile restriction can be written as a conditional moment restriction, which then implies unconditional moments conditions. Recall $\mathrm{P}(A) = \mathrm{E}[\mathbb{1}\{A\}]$. Thus, $\mathrm{P}(Y \leq \boldsymbol{X}'\boldsymbol{\beta}(\tau) \mid \boldsymbol{Z}) = \tau$ becomes $\mathrm{E}[\mathbb{1}\{Y \leq \boldsymbol{X}'\boldsymbol{\beta}(\tau)\} \mid \boldsymbol{Z}] = \tau$ and then

$$\mathrm{E}[\mathbb{1}\big\{Y \leq \boldsymbol{X}'\boldsymbol{\beta}(\tau)\big\} - \tau \mid \boldsymbol{Z}] = 0. \tag{7.6}$$

Unconditional moments can be generated by multiplying the main part by any function of $\boldsymbol{Z}$. Although not theoretically optimal, a common choice is $\boldsymbol{Z}$ itself. This is similar to using unconditional moments $\mathrm{E}(\boldsymbol{Z}U) = \mathbf{0}$ given conditional moment $\mathrm{E}(U \mid \boldsymbol{Z}) = 0$. Here,

$$\mathbf{0} = \mathrm{E}[\boldsymbol{Z}(\mathbb{1}\big\{Y \leq \boldsymbol{X}'\boldsymbol{\beta}(\tau)\big\} - \tau)]. \tag{7.7}$$

Discussion Question 7.1 (IVQR model)**.** Consider the setting of Card (1995). Let Y be log wage. Let $\boldsymbol{X}$ include an intercept, years of education, years of experience (and its square), and various other demographic and geographic characteristics. Let $Z = 1$ if the individual grew up near a 4-year college and $Z = 0$ if not. Let U be a scalar unobservable variable normalized to $U \sim \mathrm{Unif}(0, 1)$. Consider the model $Y = \boldsymbol{X}'\boldsymbol{\beta}(U)$.

a) First assume (unrealistically) $\boldsymbol{X} \perp\!\!\!\perp U$, and show how there can still be heteroskedasticity. Start with the simpler model $Y = \beta_0 + X\beta_1(U)$ with $X > 0$ (years of education); show how you can have $\mathrm{Var}(Y \mid X = x)$ increasing in x even though $X \perp\!\!\!\perp U$ (and still maintaining monotonicity).
b) Explain what the "monotonicity" assumption says in this case, including how you might interpret U economically.
c) Explain one reason you may doubt monotonicity.
d) Doubts aside, interpret the coefficient on years of education for $\tau = 0.5$ and $\tau = 0.75$, and explain whether you think the coefficient is higher, lower, or the same with $\tau = 0.75$.

7.1.3 IVQR Estimation

Since we have moment conditions from (7.7), GMM estimation seems natural, but there are computational challenges. The parameter is "stuck" inside the indicator function $\mathbb{1}\{\cdot\}$. Consequently, searching numerically over possible parameter values, slight changes result in either no change, or a discontinuous jump. Further, the GMM criterion function turns out to be non-convex.

Because of this, there have been many different approaches to estimation. Of course, I highly recommend the smoothed approach of Kaplan and Sun (2017), de Castro et al. (2019), and Kaplan (2022c), which seems fast, reliable, scalable, and has code on my website (both R and Stata). The basic idea is to replace the discontinuous indicator function $\mathbb{1}\{\cdot\}$ with a smoothed version, which smooths the moment conditions enough that standard numerical solvers can be used. Changing the moments introduces bias, but

it also decreases variance, so smoothing also decreases mean squared error as a secondary benefit. Stata command `ivqreg2` (on SSC) based on Machado and Santos Silva (2019) also seems to work well in practice, and there are yet other methods I have not yet tried.

7.1.4 IVQR Inference

There are different approaches to IVQR inference, including some that are robust to weak identification or even lack of identification. With strong identification, something like a Bayesian bootstrap (Chapter 14 and Section 13.1) should work fine. The exact finite-sample approach of Chernozhukov, Hansen, and Jansson (2009) (noted in Section 5.6) applies here, too, and it is robust to weak or partial identification (lack of point identification). Other approaches that apply to IVQR with weak/partial identification include Chernozhukov and Hansen (2008), Jun (2008), and Andrews and Mikusheva (2016). See also Chernozhukov, Hansen, and Wüthrich (2017, §9.3.3–9.3.4) for a brief overview of all of these.

7.2 Other Approaches to Endogeneity

7.2.1 Triangular Model

The following is a brief summary of the summary in Chernozhukov, Hansen, and Wüthrich (2017, §9.2.5), which contains other references.

The triangular model's structural equation of interest is $Y = g(D, \boldsymbol{\epsilon})$, where scalar continuous D is endogenous and modeled by $D = h(\boldsymbol{Z}, \eta)$, in which one of the instruments in $\boldsymbol{Z}$ must also be continuous. (Other exogenous regressors can be added, too.) Identification follows from monotonicity of $h(\boldsymbol{z}, \eta)$ in η as well as the exogeneity condition $\boldsymbol{Z} \perp\!\!\!\perp (\boldsymbol{\epsilon}, \eta)$.

Generally, the triangular model restricts the selection equation (for D) more than IVQR, but restricts the structural (outcome) equation less than IVQR. For example, $\boldsymbol{\epsilon}$ can be a vector here, whereas IVQR had scalar U. Conversely, IVQR does not restrict the (implicit) selection equation; e.g., it can handle simultaneous equations like supply and demand, which the triangular model here cannot. The triangular model also requires the instruments to be independent of the unobservables in both the structural equation and the selection equation, instead of only the structural equation (as in IVQR).

The triangular model also requires the endogenous regressor to be continuous, and seemingly there can only be one endogenous regressor(?). With IVQR, any type or number of endogenous variables is allowed, as long as there are enough instruments.

7.2.2 Local Quantile Treatment Effect

Melly and Wüthrich (2017) provide an excellent survey of the **local quantile treatment effect** (LQTE) model. The idea is similar to the local average treatment effect (LATE) of Imbens and Angrist (1994): "local" refers to "compliers," and identification follows from

SUTVA, instrument independence, instrument relevance, no defiers, and an exclusion restriction; see Assumption 3 in Melly and Wüthrich (2017). As with LATE, "complier" is defined by the relationship between the binary instrument Z and the binary treatment status X. Considering both possible universes ($Z = 0$ or $Z = 1$), compliers receive treatment ($X = 1$) in the universe with $Z = 1$, but they do not ($X = 0$) in the universe with $Z = 1$. The LQTE is then the QTE (see Section 6.2) for the subpopulation of compliers.

Melly and Wüthrich (2017) also compare LQTE and IVQR. Like the triangular model, the LQTE model restricts the selection equation more than the outcome equation (compared to IVQR). Also, under LQTE assumptions, even if some IVQR assumptions fail, the IVQR estimand is the LQTE at shifted τ values.

With similar motivation as unconditional quantile regression (Section 6.5), Melly and Wüthrich (2017, §10.2.3) also discuss unconditional LQTE when covariates are required to satisfy (conditional) independence.

Discussion Question 7.2 (LQTE)**.** Consider a randomized experiment with imperfect compliance, like the Job Training Partnership Act (JTPA): the offer of training is made randomly, but some individuals do not take the offer. Let $Z_i = 1$ if the offer is made to individual i, and $Z_i = 0$ otherwise, with Z_i randomized. Treatment $X_i = 1$ if the individual actually does the job training, and $X_i = 0$ if not. Assume you cannot take the training without the offer, so $Z_i = 0$ implies $X_i = 0$ (thus there are no "defiers" or "always-takers"). However, you can decline the training if offered, so individuals with $Z_i = 1$ can choose either $X_i = 1$ (compliers) or $X_i = 0$ (never-takers). The outcome Y_i is labor earnings (in dollars) over the two years following the training period.

a) First consider intention-to-treat (ITT) quantile effects, i.e., the causal effect of offering the training (regardless of whether or not it's taken). Ignore general equilibrium effects, so that (given randomization) the τ-ITTQE equals $\mathrm{Q}_\tau(Y \mid Z = 1) - \mathrm{Q}_\tau(Y \mid Z = 0)$, the difference in τ-quantile between the "offered" and "not-offered" subpopulations. Explain one scenario in which the true τ-ITTQE increases with τ, and another in which it decreases with τ. For each scenario, explain your assumptions about the population studied, the type of training, and anything else important.

b) Assume individuals choose X_i rationally, based on their anticipated benefit. For a given τ, do you expect the overall QTE is less than, greater than, or equal to the LQTE (i.e., the QTE for compliers)? Explain any assumptions you make, including any assumptions about τ.

7.3 Panel Data with Fixed Effects

The first QR model with panel data and fixed effects (FE) seems to be from Koenker (2004). Much of the following literature proceeded in a similar vein, modeling the FE by including as regressors a dummy variable for each "individual" in the data. Although this is equivalent to the usual (mean) FE regression, which can be simplified computationally

by demeaning (partialling out the individual dummies), it cannot be simplified due to the nonlinearity of the quantile operator; e.g., $\mathrm{Q}_\tau(Y_2 - Y_1) \neq \mathrm{Q}_\tau(Y_2) - \mathrm{Q}_\tau(Y_1)$. This leaves n parameters to estimate (for the n dummies), which requires large T and some penalization or such to deal with. However, most attention was given to the computational and statistical issues rather than the structural interpretation.

Recall the random coefficients model of Section 6.4: $Y = \boldsymbol{X}'\boldsymbol{\beta}(U)$. With panel data, each individual has not just Y but $Y_1, \ldots, Y_T$, where T is the number of time periods. In the usual FE model, the unobserved component is split into a time-invariant term U_i and an idiosyncratic term V_{it}. It seems most natural (though "natural" is not always correct) to replace U in the cross-sectional model with $U_i + V_{it}$, yielding

$$Y_t = \boldsymbol{X}_t'\boldsymbol{\beta}(U + V_t). \tag{7.8}$$

In contrast, the original panel QR models had only V_t as the rank variable, and added individual heterogeneity through a dummy regressor: adding the i subscript explicitly,

$$Y_{it} = \boldsymbol{X}_{it}'\boldsymbol{\beta}(V_{it}) + \boldsymbol{\eta}_i'\boldsymbol{\gamma}(V_{it}), \tag{7.9}$$

as in (2.4) of Arellano and Bonhomme (2016), where $\boldsymbol{\gamma}(\cdot)$ is a function like $\boldsymbol{\beta}(\cdot)$ and $\boldsymbol{\eta}_i$ is a vector of time-invariant, possibly-unobserved variables. Another option is the very general nonseparable model $Y_t = q(\boldsymbol{X}_t, U, V_t)$. There are yet more options, like

$$Y_t = \boldsymbol{X}_t'\boldsymbol{\beta}(V_t) + U, \tag{7.10}$$

$$Y_t = \boldsymbol{X}_t'\boldsymbol{\beta}(U) + V_t, \tag{7.11}$$

etc. Most (not all) of these have the standard FE model $Y_t = \boldsymbol{X}_t'\boldsymbol{\beta} + U + V_t$ as a special case.

For more comparison of possible structural models, as well as an approach to estimating a certain class of them, see Arellano and Bonhomme (2016), especially Sections 2.1–2.2. See also the new estimators for alternative structural models with nonseparable FE from Liu (2020) and Powell (2020), implemented in Stata command `qregpd` (in SSC).

Discussion Question 7.3 (panel FE QR: airfare)**.** Let Y_{it} be the (average) airfare (plane ticket price) for route i at time t, where i is defined by the departure airport and arrival airport. Let X_{it} measure the (lack of) competition on route i at time t: $X_{it} = 1$ if it is a monopoly (only one airline flies route i), and values closer to zero indicate more competition.

a) Among (7.8)–(7.11), which do you think is most appropriate here? Why?
b) Is there anything you think is missing from even your preferred model?

Exercises

Exercise E7.1. a. Find a published paper that runs a cross-sectional IV regression (and makes its data publicly available); provide a link to the paper.[1] The paper must be either published in a respectable economics journal[2] or be unpublished but have an author who has previously published in such a journal (like any Mizzou econ professor); or if you really want, you can use an example from an econometrics textbook. (I think the `wooldridge` package in R may have some possible datasets.) Even if it's not purely cross-sectional, but it just runs standard 2SLS or IV (i.e., no FE or anything), it should be fine, but you're welcome to check with me first.

b. Replicate (at least, reasonably close) one particular IV estimate from the paper. (Meaning, don't do all 12 variations they try, just pick one specification.) If code is provided with the paper, feel free to use it (just say so).

c. Using the same specification, run an IV *quantile* regression for a variety of quantile levels τ (0.5, 0.25, etc.), using either the R code (file `ivqr_see.R` has the main `ivqr.see()` function; `gmmq.R` contains helper functions) or Stata code (`sivqr`) available on my website,[3] or using the Stata `ivqte` command if you have a binary treatment variable.[4] Note: I think my Stata code is much easier to use than the R code. Provide the code you write/run. Try both the plug-in bandwidth as well as a very small bandwidth (for which you can just set the bandwidth argument to zero). Note the warnings in the comments at the top of `ivqr_see.R`, or read the help file in Stata. If you have computational problems (like it's taking multiple hours or something), you can take a random sample from the original data sample and/or omit some control regressors (if it doesn't affect the 2SLS estimate too much), but please say so explicitly. (As always, you can also just ask me for advice, if you start enough before the submission deadline.)

d. Discuss any similarities and differences across quantile levels (τ) and between the "mean" and median. Are any of the differences economically significant and/or interesting?

Note: functions in files `gmmq.R` and `ivqr_gmm.R` can estimate more general (nonlinear-in-variables) models if you need it; but it's probably easiest to find an example where you don't.

Exercise E7.2. Like Exercise E7.1, but find a published paper reporting (usual) FE results and now use `qregpd`, which you can install in Stata by `ssc install qregpd`. Note: you must also install `moremata` if you have not already, by running `ssc install moremata` in Stata.

---

[1] E.g., if nobody's claimed it yet, https://doi.org/10.1016/j.jhealeco.2016.08.002

[2] For example, in top 500 of https://ideas.repec.org/top/top.journals.all.html

[3] https://kaplandm.github.io

[4] https://sites.google.com/site/blaisemelly/home/computer-programs/estimation-of-quantile-treatment-effects-in-stata

Part III

Distributional Methods

Chapter 8

Distributional Inference: One-Sample, Two-Sided

Unit learning objectives for this chapter

8.1. Develop intuition about the empirical CDF, goodness-of-fit testing, and test inversion for functions [TLO 2]

8.2. Interpret the results of goodness-of-fit tests and uniform confidence bands [TLO 1]

The goal here is to learn about a single continuous population CDF from a single iid sample ("one-sample"). "Inference" includes both hypothesis testing and a uniform confidence band for the CDF (here "two-sided" in both cases). Most importantly, this chapter introduces ideas useful in one-sided and/or two-sample extensions that are more economically interesting.

8.1 Warning: Weights

In practice, many economic datasets are not iid (like any big survey) and contain survey weights (or "sampling weights"). Essentially, certain groups (strata) are over-represented in the sample (compared to the population of interest), while other groups are under-represented, and the weights help adjust the sample to better represent the population. For some econometric analysis, weights can be ignored, but not for learning about the population distribution itself.

For example, imagine we observe height, but females have been over-sampled and make up 90% of our sample, so the "sampling weight" for each female is 5/9. (Imagine a sample of 100 people: the sum of weights for the 90 females is $(90)(5/9) = 50$.) If we want to learn about the CEF of height conditional on sex, then there's no problem; we'll have a much smaller standard error for female height, but everything will be valid; i.e.,

we can safely ignore the weights. However, if we want to learn about the unconditional distribution of height, we cannot ignore the weights: the mostly-female sample tends to have lower values of height than the half-female population.

The focus on iid in this chapter is to help develop intuition, but in practice, iid-based methods can be misleading for unconditional distributions.

8.2 Discrete and Categorical Distributions

The focus of this chapter (and subsequent chapters) is continuous CDFs, but this section has some notes on discrete and categorical distributions.

Discrete and categorical distributions are characterized by their probability mass function (PMF). Writing the possible values/categories as $v_1, \ldots, v_J$, vector $\boldsymbol{p} = (p_1, \ldots, p_J)$ with $p_j \equiv \mathrm{P}(Y = v_j)$ fully describes the Y distribution.

Thus, standard results for finite-dimensional parameters apply. Assuming J is not too big compared to n, the estimators $\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i = v_j\}$ work well and are jointly asymptotically normal.

8.3 Preliminary Results for Continuous Distributions

Many approaches are based on the **empirical CDF** (ECDF), also known as the empirical distribution function (EDF). Recall $\mathrm{P}(A) = \mathrm{E}[\mathbb{1}\{A\}]$, so the population CDF of Y is

$$F(y) \equiv \mathrm{P}(Y \leq y) = \mathrm{E}[\mathbb{1}\{Y \leq y\}]. \tag{8.1}$$

Given iid sampling, by the analogy principle, the ECDF is

$$\hat{F}(y) \equiv \widehat{\mathrm{E}}[\mathbb{1}\{Y \leq y\}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y\}, \quad y \in \mathbb{R}. \tag{8.2}$$

The ECDF is a nonparametric estimator of the population CDF $F(\cdot)$. The ECDF starts at zero at $y = -\infty$, and increases in steps of size $1/n$ at each of the n data points Y_i. (If instead you have a dataset with weights, then you can normalize the weights to sum to 1 and take steps of size w_i, the weight for observation i.)

At least with iid data, the asymptotic properties of the ECDF are well understood (Section 8.A), and even some finite-sample properties are known.

8.4 Goodness-of-Fit Testing

Consider the **goodness-of-fit** (GOF) null hypothesis

$$H_0 \colon F(\cdot) = F_0(\cdot), \tag{8.3}$$

where $F(\cdot)$ is the true unknown population CDF of Y, and $F_0(\cdot)$ is a specified distribution. The methods in this chapter may (mostly) be adapted to the one-sided version $H_0\colon F(\cdot) \leq F_0(\cdot)$, or $\geq$, but intuition is easier with (8.3).

The name "goodness-of-fit" refers (roughly) to whether our guess $F_0(\cdot)$ is a good fit for the data sampled from $F(\cdot)$. This is not usually a concern of modern economics. (E.g., at least in serious economics, nobody is testing if regression errors are normal.) However, it develops intuition that carries over to more complex settings.

The general approach to GOF testing is to define a measure of distance from $\hat{F}(\cdot)$ to $F_0(\cdot)$, and then approximate the distribution of that distance measure under H_0.

Different distance measures underlie different tests. The **Kolmogorov–Smirnov** (KS) approach (Section 8.5) uses

$$D_n \equiv \sup_{r\in\mathbb{R}} |\hat{F}(r) - F_0(r)|, \tag{8.4}$$

the biggest vertical distance between the ECDF and $F_0(\cdot)$. Alternatively, the **Cramér–von Mises** (CvM or CM) approach integrates squared differences:

$$W_n^2 \equiv n \int_{\mathbb{R}} [\hat{F}(r) - F_0(r)]^2 \, \mathrm{d}F_0(r). \tag{8.5}$$

The **Anderson–Darling** (AD) test (Anderson and Darling, 1952, 1954) usually refers to a normalized version of CvM,

$$A_n^2 \equiv n \int_{\mathbb{R}} \frac{[\hat{F}(r) - F_0(r)]^2}{F_0(r)[1 - F_0(r)]} \, \mathrm{d}F_0(r), \tag{8.6}$$

but it can also mean a similarly normalized KS statistic.

Which approach is best? All can control size; other considerations are power and extensibility. KS can generate uniform confidence bands and detect *where* the two CDFs differ, but it can have poor power. Yet another approach based on the Dirichlet distribution retains the KS advantages while improving power; see Goldman and Kaplan (2018a) and Kaplan (2019), with R and Stata code. However, the Dirichlet approach seems more difficult than KS to extend to non-iid sampling.

8.5 Kolmogorov–Smirnov Test

To control size, the distribution of D_n in (8.4) must be approximated. An asymptotic approximation and corresponding critical values were initially provided by Kolmogorov (1933) and Smirnov (1948). Given $F(\cdot) = F_0(\cdot)$, the asymptotic distribution is

$$\sqrt{n} D_n \xrightarrow{d} K \equiv \sup_{t\in[0,1]} |B(t)|, \tag{8.7}$$

where $B(\cdot)$ is a standard Brownian bridge (don't worry about it). To control asymptotic size, the KS test rejects H_0 when $\sqrt{n}D_n$ exceeds $K_{1-\alpha}$, the $(1-\alpha)$-quantile of the Kolmogorov distribution in (8.7).

In fact, size can be controlled in finite samples. Finite-sample critical values are available in `ks.test()` in R, for example. Continuity of $F(\cdot)$ implies $F(Y) \sim \text{Unif}(0,1)$. Thus, $H_0\colon F(\cdot) = F_0(\cdot)$ is equivalent to $H_0\colon F_0(Y) \sim \text{Unif}(0,1)$. That is, we can use the transformed data $Z_i = F_0(Y_i)$ and test $H_0\colon Z_i \overset{iid}{\sim} \text{Unif}(0,1)$. To simulate the distribution of D_n: draw many random samples of $Z_i \overset{iid}{\sim} \text{Unif}(0,1)$, computing D_n for each. The critical value $K_{1-\alpha}/\sqrt{n}$ is the $(1-\alpha)$-quantile among these simulated D_n values.

Nonetheless, DQs 8.1 and 8.2 show a significant deficiency.

Discussion Question 8.1 (KS tail power 1)**.** Let $H_0\colon Y_i \overset{iid}{\sim} \text{Unif}(0,1)$. Let $n = 20$. Use critical value $K_{1-\alpha}/\sqrt{n} = 0.26$, for $\alpha \approx 10\%$, so the KS rejects H_0 when $D_n > 0.26$, where D_n is the distance between $\hat{F}(\cdot)$ and $F_0(\cdot)$ defined in (8.4). Imagine a dataset with $Y_i = i/20$ for $i = 1, \ldots, k$, and $Y_i = i + 1000$ for $i > k$.

a) Let $k = 20$ and graph $F_0(\cdot)$, the $\text{Unif}(0,1)$ CDF, along with $\hat{F}(\cdot)$. Recall $F_0(r) = r$ over $r \in [0,1]$, and $\hat{F}(Y_i) = i/20$ in this example.
b) With $k = 20$, explain why $D_n = 0.05$.
c) With $k = 19$, explain why $D_n = 0.05$.
d) Mathematically (or visually), why is D_n the same for $k = 20$ and $k = 19$?
e) Intuitively, do you agree with KS that $\hat{F}(\cdot)$ with $k = 19$ is equally "far" from $F_0(\cdot)$ as $\hat{F}(\cdot)$ with $k = 20$?

Discussion Question 8.2 (KS tail power 2)**.** Continue from DQ 8.1.

a) With $k = 18$, explain why $D_n = 0.1$.
b) With $k < 18$, explain why $D_n = (20 - k)/20$.
c) How many "large" observations ($Y_i > 1000$) are needed for the KS to reject? That is, what's the largest k for which $D_n > 0.26$ and the KS test rejects?
d) Only using your intuition, what's the largest k for which you personally would reject H_0? Why?

In R, you can try replacing `k=20` with other values in the expression

```
with(data=list(k=20),expr=ks.test(c(1:k/20,1000+(k+1):20),punif,exact=T))
```

8.6 Uniform Confidence Band

8.6.1 Test Inversion: Scalar

See Section 9.19 "Confidence Intervals by Test Inversion" of Hansen (2020a), for example.

You may have seen how for a scalar parameter θ, a hypothesis test can be "inverted" into a confidence interval (CI). Specifically, for any value $t \in \mathbb{R}$, define null hypothesis $H_t\colon \theta = t$. The CI from **test inversion** with confidence level $1 - \alpha$ is

$$\widehat{\text{CI}}_{1-\alpha} \equiv \{t : H_t \text{ not rejected at level } \alpha\}. \tag{8.8}$$

The justification follows from the size control of the hypothesis test. The following can be made asymptotic, but for simplicity I leave it finite-sample. Size control at level α means

that H_θ is rejected with probability α (or less), where θ is the true value. For simplicity, assume $\mathrm{P}(H_\theta \text{ rejected at level } \alpha) = \alpha$ exactly. Then,

$$\mathrm{P}(\theta \in \widehat{\mathrm{CI}}_{1-\alpha}) = \mathrm{P}(H_\theta \text{ not rejected at level } \alpha) = 1 - \overbrace{\mathrm{P}(H_\theta \text{ rejected})}^{=\alpha} = 1 - \alpha. \tag{8.9}$$

8.6.2 Test Inversion: Vectors and Functions

Nothing in the argument of (8.9) is special to the dimension of θ. The argument goes through with vector $\boldsymbol{\theta}$. It also works if the parameter is a function ("infinite-dimensional"), like the population CDF.

Consider the set $\hat{\mathcal{F}}_{1-\alpha}$ of all CDFs that a level-α KS test does not reject. That is,

$$\hat{\mathcal{F}}_{1-\alpha} = \{F_0(\cdot) : \text{KS does not reject } H_0\colon F(\cdot) = F_0(\cdot) \text{ at level } \alpha\}. \tag{8.10}$$

Assume exact finite-sample critical values are used, so the probability KS rejects the true $F(\cdot)$ is exactly α. Then,

$$\begin{aligned}\mathrm{P}(F(\cdot) \in \hat{\mathcal{F}}_{1-\alpha}) &= \mathrm{P}(\text{true } F(\cdot) \text{ not rejected by KS at level } \alpha) \\ &= 1 - \overbrace{\mathrm{P}(\text{KS rejects true CDF at level } \alpha)}^{=\alpha} \\ &= 1 - \alpha.\end{aligned}$$

Thus, $\hat{\mathcal{F}}_{1-\alpha}$ is a confidence set: it contains the true $F(\cdot)$ with probability $1 - \alpha$.

8.6.3 Uniform Confidence Band

A **uniform confidence band** consists of two data-dependent functions that make a "band" that contains the true function with high probability. The true function could be a CDF, CEF, CQF, hazard function, etc. Let $\hat{L}(\cdot)$ and $\hat{U}(\cdot)$ be functions computed from data (hence the "hats"), with L for "lower" and U for "upper." Let $1-\alpha$ be the confidence level. Let $F(\cdot)$ be the true function. Then, at least asymptotically, a uniform confidence band satisfies

$$1 - \alpha = \mathrm{P}(\hat{L}(\cdot) \leq F(\cdot) \leq \hat{U}(\cdot)) = \mathrm{P}(\hat{L}(r) \leq F(r) \leq \hat{U}(r) \text{ for all } r \in \mathbb{R}). \tag{8.11}$$

A one-sided band sets $\hat{L}(r) = -\infty$ or $\hat{U}(r) = \infty$ for all $r \in \mathbb{R}$.

The uniform confidence band contrasts with a **pointwise confidence band**. The latter promises $\mathrm{P}(\hat{L}(r) \leq F(r) \leq \hat{U}(r)) = 1 - \alpha$ (or $\to 1 - \alpha$) for any individual $r \in \mathbb{R}$. That is, the pointwise band just aggregates all the individual confidence intervals for the different $F(r)$. Whether this is "better" or "worse" depends on the research question; either way, it is very different. Discussion Question 8.3 tries to provide some intuition.

Discussion Question 8.3 (pointwise vs. joint)**.** Consider parameters θ_1 and θ_2. (You could imagine $\theta_1 = F(r_1)$ and $\theta_2 = F(r_2)$, although the following would not quite be right.) For simplicity, assume $\hat{\theta}_1 \sim \mathrm{N}(\theta_1, \mathrm{SE}_1^2)$ and $\hat{\theta}_2 \sim \mathrm{N}(\theta_2, \mathrm{SE}_2^2)$.

a) Briefly explain why $\widehat{\text{CI}}_1 = \hat{\theta}_1 \pm 1.64\,\text{SE}_1$ is a 90% pointwise CI for θ_1, and similarly $\widehat{\text{CI}}_2 = \hat{\theta}_2 \pm 1.64\,\text{SE}_2$ for θ_2.
b) Assuming $\hat{\theta}_1 \perp\!\!\!\perp \hat{\theta}_2$, what is the joint coverage probability of the pointwise CIs, i.e., $\text{P}(\theta_1 \in \widehat{\text{CI}}_1 \text{ and } \theta_2 \in \widehat{\text{CI}}_2)$?
c) Without the independence assumption, is $\text{P}(\theta_1 \in \widehat{\text{CI}}_1 \text{ and } \theta_2 \in \widehat{\text{CI}}_2) = 90\%$ possible? How/why?
d) In general, to get 90% joint coverage probability, would the individuals CIs have to be longer or shorter than the 90% pointwise CIs considered above? Why?
e) Why does this suggest that a uniform confidence band is wider than a pointwise confidence band?

One uniform confidence band for the CDF is the KS confidence set of all functions not rejected by KS at level α, as in (8.10). In DQ 8.4, we'll try to picture this band.

Discussion Question 8.4 (KS band)**.** Consider $\hat{\mathcal{F}}_{1-\alpha}$ from (8.10). Recall that the KS rejects $H_0\colon F(\cdot) = F_0(\cdot)$ at level α when $D_n > c_{n,\alpha}$, with D_n defined in (8.4) and $c_{n,\alpha} \equiv K_{1-\alpha}/\sqrt{n}$. The particular α and critical value are not important in the following; just think about the shape.

a) Consider a particular $r \in \mathbb{R}$. What do we know about $F_0(r)$ if $F_0(\cdot)$ is not rejected by KS?
b) Consequently, argue that any $F_0(\cdot) \in \hat{\mathcal{F}}_{1-\alpha}$ satisfies $\hat{L}(r) \le F_0(r) \le \hat{U}(r)$ if $\hat{L}(r) = \hat{F}(r) - c_{n,\alpha}$ and $\hat{U}(r) = \hat{F}(r) + c_{n,\alpha}$.
c) Further argue that this applies at any $r \in \mathbb{R}$, so any $F_0(\cdot) \in \hat{\mathcal{F}}_{1-\alpha}$ satisfies $\hat{L}(\cdot) \le F_0(\cdot) \le \hat{U}(\cdot)$.
d) Draw an example $\hat{F}(\cdot)$ and the corresponding $\hat{L}(\cdot)$ and $\hat{U}(\cdot)$.
e) For the true $F(\cdot)$, what is $\text{P}(\hat{L}(\cdot) \le F(\cdot) \le \hat{U}(\cdot))$? Why? Hint: recall $\text{P}(F(\cdot) \in \hat{\mathcal{F}}_{1-\alpha}) = 1 - \alpha$.

Appendix to Chapter 8

8.A ECDF: Asymptotic Properties

With iid sampling, the ECDF is uniformly consistent:

$$\sup_{r\in\mathbb{R}}|\hat{F}(r) - F(r)| \xrightarrow{p} 0. \tag{8.12}$$

This also holds with $\xrightarrow{\text{a.s.}}$ and is called the Glivenko–Cantelli Theorem.

A **functional central limit theorem** (FCLT) called Donsker's Theorem says

$$\sqrt{n}(\hat{F}(\cdot) - F(\cdot)) \xrightarrow{d} B(F(\cdot)), \tag{8.13}$$

where $B(\cdot)$ is a standard Brownian bridge on the unit interval $[0, 1]$.

This $B(\cdot)$ is a type of Gaussian process, a random function whose marginal distributions are Gaussian. That is, the vector $(B(t_1), \ldots, B(t_k))$ is a k-dimensional Gaussian vector, with mean $\mathbf{0}$ and $\text{Cov}(B(s), B(t)) = s(1-t)$ for $s < t$. Figure 8.1 shows some **sample paths** (realizations) of $B(\cdot)$, along with KS test critical values.

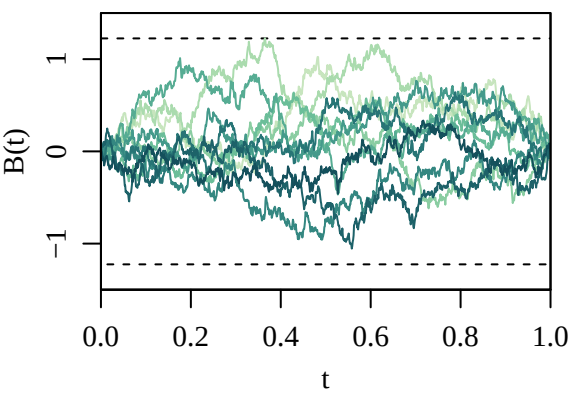


Figure 8.1: Brownian bridge sample paths, with two-sided $\alpha = 0.1$ critical values.

Chapter 9

Distributional Inference: Two-Sample, Two-Sided

Unit learning objectives for this chapter

9.1. Develop intuition for ways to achieve exact inference in finite samples [TLO 2]

9.2. Interpret two-sample goodness-of-fit tests [TLO 1]

The focus of this chapter is goodness-of-fit (GOF) testing; additional two-sample, two-sided approaches are in Chapter 11. Again, iid sampling is assumed for simplicity but may not be appropriate in practice. With iid sampling in practice, the Dirichlet approach again has power advantages over KS, although it cannot form a uniform confidence band for the true CDF difference function.

Optional resources for this chapter

- R: `ks.test()`
- R: code for "Comparing distributions by multiple testing across quantiles or CDF values" at `https://kaplandm.github.io`
- Stata: command `distcomp` described by Kaplan (2019)
- Stata: `ksmirnov`

9.1 Setup

The following describes the two-sample, two-sided GOF setup. The null is

$$H_0 \colon F_1(\cdot) = F_2(\cdot), \tag{9.1}$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are both unknown continuous CDFs. ECDFs $\hat{F}_1(\cdot)$ and $\hat{F}_2(\cdot)$ are computed from iid samples of size m and n, respectively. The samples are also independent of each other.

The same KS, CvM, AD, and Dirichlet approaches can be applied to two-sample testing, with similar tradeoffs. In particular, the two-sample KS statistic is

$$D_{m,n} \equiv \sup_{r \in \mathbb{R}} |\hat{F}_1(r) - \hat{F}_2(r)|. \tag{9.2}$$

Discussion Question 9.1 (two-sample GOF)**.** (Inspired by Gneezy and List, 2006) You have data on productivity Y_i of individual i, collected for the same job on the same day. Individuals were randomized into either the control or treatment group. The control group was paid the advertised \$10.00/hr wage, whereas the treatment group was surprised with a \$20.00/hr wage. You're curious if there's any evidence of "gift exchange:" does productivity increase in response to the "gift" of higher wage? You test the GOF null hypothesis H_{0d} that the treatment and control productivity distributions are identical. You also test the null hypothesis H_{0m} that the treatment and control population means are identical.

a) You fail to reject H_{0d}. What do you learn?
b) You reject H_{0d}. What do you learn?
c) Is it possible for H_{0d} to be true if H_{0m} is false? Why/not? If yes, give an example.
d) Is it possible for H_{0m} to be true if H_{0d} is false? Why/not? If yes, give an example.
e) Explain how it's possible to have H_{0m} rejected but not H_{0d}, and how you interpret such a result.
f) Explain how it's possible to have H_{0d} rejected but not H_{0m}, and how you interpret such a result.

9.2 Exact Finite-Sample Testing

The approach from Section 8.5 does not work here because H_0 does not determine either true CDF, only that they are the same.

However, finite-sample size control is possible using a **randomization test** or **permutation test** with the KS statistic. Assuming H_0: $F_1(\cdot) = F_2(\cdot)$ holds, let $F(\cdot) = F_1(\cdot) = F_2(\cdot)$, so the Y_i in both samples are iid from $F(\cdot)$. Combine both samples into Y_i for $i = 1, \ldots, m+n$, where the first m ($i = 1, \ldots, m$) are from the first sample and the next n ($i = m+1, \ldots, m+n$) are from the second sample. Then, $Y_i \overset{iid}{\sim} F(\cdot)$ for all $i = 1, \ldots, m+n$. Thus, under H_0, the joint distribution of all Y_i is the same even if we switch which i corresponds to which sample. For example, if $m = n = 1$, the distribution of pooled vector (Y_1, Y_2) is the same as that of (Y_2, Y_1) under H_0, whereas if $F_1 \neq F_2$ then these vectors have different distributions (e.g., maybe the mean of the second component is higher when you switch the order). Thus, under H_0, the sampling distribution of $D_{m,n}$ is the same even if we first "permute" the observations (i.e., switch which observations are in which of the two samples) before computing $D_{m,n}$.

With a permutation test, the value of $D_{m,n}$ is computed under all possible permutations of the observed data. The p-value is the proportion of such permutations that have $D_{m,n}$ at least as big as the original sample's $D_{m,n}$. This approach applies much more broadly than just the KS test, and new tests applying the approach continue to be developed in statistics and econometrics.

In R, specify the argument `exact=TRUE` in `ks.test()`.

Discussion Question 9.2 (permutations)**.** Consider sample sizes $m = 19$ and $n = 1$.

a) If the second sample's value is larger than all $m = 19$ values in the first sample, then what is $D_{m,n}$ in (9.2)?
b) Is there any other permutation that gives the same $D_{m,n}$ value? That is, if we use the same 20 values but pretend a different one came from the second sample, is it possible that $D_{m,n}$ remains the same?
c) Besides the original data, there are 19 other possible permutations: each of the $m = 19$ values can be treated as the second sample. Explain how many of these have: the same $D_{m,n}$ as the original data; strictly smaller $D_{m,n}$; strictly larger $D_{m,n}$.
d) The p-value is the proportion of permutations (including the original data, so 20 total) that have $D_{m,n}$ greater than or equal to the original data's $D_{m,n}$; what is the p-value? Does it seem reasonable? Why/not?

Consider a related but different permutation question, assuming both samples are iid from the same continuous population distribution.

e) What's the probability of randomly sampling a dataset where the second sample's value is larger than all $m = 19$ first sample values?
f) Smaller than all $m = 19$?
g) What does this suggest the two-sided p-value should be for a dataset in which the second sample's value is either larger or smaller than all $m = 19$ first sample values?

9.3 Asymptotic KS

The asymptotic argument for one-sample KS extends readily to the two-sample case. The two samples are independent, so it is straightforward to take a difference of ECDFs. Assuming $\sqrt{n/m} \to \lambda \in (0, \infty)$, under H_0, $\sqrt{n}(\hat{F}_1(\cdot) - \hat{F}_2(\cdot))$ converges to a certain Gaussian limit, which implies the asymptotic distribution of $\sqrt{n}D_{m,n}$.

Unlike the finite-sample permutation-type approach, which depends critically on $F_1 = F_2$, this asymptotic derivation allows non-zero $\Delta(\cdot) \equiv F_1(\cdot) - F_2(\cdot)$. Letting $\hat{\Delta}(\cdot) = \hat{F}_1(\cdot) - \hat{F}_2(\cdot)$,

$$\begin{aligned}
\sqrt{n}(\hat{\Delta}(\cdot) - \Delta(\cdot)) &= \sqrt{n}[\hat{F}_1(\cdot) - \hat{F}_2(\cdot) - (F_1(\cdot) - F_2(\cdot))] \\
&= \sqrt{n}[\hat{F}_1(\cdot) - F_1(\cdot)] - \sqrt{n}[\hat{F}_2(\cdot) - F_2(\cdot)] \\
&= \overbrace{\sqrt{n/m}}^{\to\lambda} \overbrace{\sqrt{m}[\hat{F}_1(\cdot) - F_1(\cdot)]}^{\text{Gaussian limit}} - \overbrace{\sqrt{n}[\hat{F}_2(\cdot) - F_2(\cdot)]}^{\text{Gaussian limit}}.
\end{aligned}$$

This allows a uniform confidence band for $\Delta(\cdot)$ by test inversion.

Chapter 10

Distributional Inference: Stochastic Dominance

Unit learning objectives for this chapter

10.1. Interpret hypotheses and results of stochastic dominance tests, both economically and statistically [TLO 1]

10.2. Judge whether a null of dominance or non-dominance is more appropriate in a given setting [TLO 3]

In addition to learning whether or not two distributions are equal (GOF), we may want to know which distribution is "better." One approach is to just compare means (or quantiles); here, stochastic dominance is considered.

Here, the variables are assumed to have cardinal meaning (like dollars). If they are ordinal, then see Chapter 23.

Optional resources for this chapter

- Davidson and Duclos (2013): bootstrap test, null of non-SD
- Dirichlet approach: Goldman and Kaplan (2018a)
- Stata: command `distcomp` described by Kaplan (2019)

10.1 First-Order Stochastic Dominance

In economics, **first-order stochastic dominance** (SD1) means one distribution is unequivocally better. Letting $Y_1 \sim F_1(\cdot)$ and $Y_2 \sim F_2(\cdot)$, SD1 can be characterized as

$$Y_1 \text{ SD}_1 Y_2 \iff \text{E}[u(Y_1)] \geq \text{E}[u(Y_2)] \text{ for all } u(\cdot) \in \mathcal{U}, \tag{10.1}$$

where $\mathcal{U}$ is the set of all (non-decreasing) utility functions. (Given the eventual statistical focus, the difference between $\geq$ and $>$ is ignored.) That is, Y_1 yields higher expected utility than Y_2 for any possible utility function. So even if we all have different utility functions, we all agree Y_1 is better.

Notationally, let Y_1 *non*SD$_1$ Y_2 mean "Y_1 does not first-order stochastically dominate Y_2." For any (Y_1, Y_2), either Y_1 SD$_1$ Y_2 or Y_1 *non*SD$_1$ Y_2 (but not both).

SD1 has been used in various settings: Y could represent returns from a financial portfolio, agricultural productivity, consumption, auction bids, etc.

SD1 is useful economically because it provides such a strong (unequivocal) assessment, but this stringency is also a drawback. First, SD1 provides only a **partial ordering** of distributions: for many distribution pairs, there is not SD1 in either direction. Second, it is difficult to find strong statistical evidence in favor of SD1.

SD1 can also be characterized in terms of the CDFs $F_1(\cdot)$ and $F_2(\cdot)$. Specifically,

$$Y_1 \text{ SD}_1 Y_2 \iff F_1(\cdot) \leq F_2(\cdot). \tag{10.2}$$

It's often confusing that the "better" CDF is lower. More intuitively, in terms of quantile functions,

$$Y_1 \text{ SD}_1 Y_2 \iff Q_1(\cdot) \geq Q_2(\cdot). \tag{10.3}$$

That is, for any $\tau \in [0, 1]$, $Q_1(\tau) \geq Q_2(\tau)$, i.e., every τ-quantile is higher for Y_1 than Y_2.

Discussion Question 10.1 (SD1 and non-SD1)**.** For each of the following, draw a picture of $F_1(\cdot)$ and $F_2(\cdot)$, or explain why it's impossible.

a) Y_1 SD$_1$ Y_2 and Y_2 *non*SD$_1$ Y_1
b) Y_1 *non*SD$_1$ Y_2 and Y_2 SD$_1$ Y_1
c) Y_1 SD$_1$ Y_2 and Y_2 SD$_1$ Y_1
d) Y_1 *non*SD$_1$ Y_2 and Y_2 *non*SD$_1$ Y_1

10.2 Null of Dominance

Most of the literature (with continuous Y_1 and Y_2) considers

$$H_0 \colon Y_1 \text{ SD}_1 Y_2 \text{ or equivalently } H_0 \colon F_1(\cdot) \leq F_2(\cdot). \tag{10.4}$$

Implicitly, the alternative hypothesis is Y_1 *non*SD$_1$ Y_2.

Discussion Question 10.2 is about how we interpret SD1 tests in practice, considering both the "economic" relationships (like in DQ 10.1) as well as the statistical difference

between type I and type II errors. Consider DQ 10.2(a). Rejecting H_0: Y_1 SD_1 Y_2 means either that Y_1 does not dominate Y_2 or that we made a type I error and Y_1 actually does dominate Y_2. Non-rejection of H_0: Y_2 SD_1 Y_1 means either that Y_2 dominates Y_1 or that we made a type II error and Y_2 actually does not dominate Y_1. From DQ 10.1(b), it is possible that both tests are correct, i.e., that Y_2 SD_1 Y_1 and Y_1 *non*SD_1 Y_2. But we may also suspect the non-rejection is a type II error, especially if the sample size is small and/or the ECDFs do not look like Y_2 SD_1 Y_1. It's also possible that the first test made a type I error, but frequentist tests control the type I error rate at a low level (here 5%), so we feel fairly sure that the rejection is correct, if not 100% sure. (Although we may feel less sure if this is the 100th such test we ran; see Section 11.1.)

Discussion Question 10.2 (null of SD1)**.** You have a dataset and test (10.4). Explain how each of the following possible results would affect your beliefs about SD1, where the significance level is 5%. Hint: consider both DQ 10.1 and the difference between type I and type II errors (and remember which error type 5% refers to).

a) Reject H_0: Y_1 SD_1 Y_2; do not reject H_0: Y_2 SD_1 Y_1
b) Do not reject H_0: Y_1 SD_1 Y_2; reject H_0: Y_2 SD_1 Y_1
c) Reject both H_0: Y_1 SD_1 Y_2 and H_0: Y_2 SD_1 Y_1
d) Do not reject either H_0: Y_1 SD_1 Y_2 or H_0: Y_2 SD_1 Y_1

10.2.1 KS Test

The KS test readily extends to one-sided testing. For two-sample KS, instead of the $|\hat{F}_1(r) - \hat{F}_2(r)|$ in $D_{m,n}$ in (9.2), either $\hat{F}_1(r) - \hat{F}_2(r)$ or $\hat{F}_2(r) - \hat{F}_1(r)$ is used, again taking the supremum over $r \in \mathbb{R}$. Large $\hat{F}_1(r) - \hat{F}_2(r)$ provides evidence against H_0: $F_1(\cdot) \leq F_2(\cdot)$, whereas large $\hat{F}_2(r) - \hat{F}_1(r)$ provides evidence against H_0: $F_1(\cdot) \geq F_2(\cdot)$.

In R, specify argument `alternative='less'` or `alternative='greater'`. The former is for H_0: $F_x(\cdot) \geq F_y(\cdot)$, the latter for H_0: $F_x(\cdot) \leq F_y(\cdot)$; the names can be confusing, so it's good to sanity-check your results.

10.2.2 Dirichlet Test

The Dirichlet approach also extends to one-sided testing. As with two-sided testing, it spreads power more evenly across the distribution, whereas KS has low power in the tails.

10.3 Null of Non-Dominance

Discussion Question 10.3 (null vs. alternative)**.** Forget about SD1 for now. You have experimental data and estimate population parameter θ (e.g., the ATE). You estimate $\hat{\theta} > 0$ but know that might be due to random sampling variation even if really $\theta \leq 0$. Let H_0: $\theta \leq 0$ (against H_1: $\theta > 0$).

a) How would you interpret rejection of H_0?
b) How would you interpret non-rejection of H_0?

c) Can you reject H_0 even if $\hat{\theta} \leq 0$, or must $\hat{\theta} > 0$ to reject? Explain.
d) Can you fail to reject H_0 even if $\hat{\theta} > 0$, or must $\hat{\theta} \leq 0$ to not reject? Explain.
e) Why do people test this H_0 instead of $H_0 \colon \theta > 0$, when they're trying to show evidence of $\theta > 0$?

Davidson and Duclos (2013) argue that SD1 should be the alternative, not the null:

$$H_0 \colon Y_1 \; non\text{SD}_1 \; Y_2, \quad H_1 \colon Y_1 \; \text{SD}_1 \; Y_2. \tag{10.5}$$

Discussion Question 10.4 (null of non-SD1)**.** You have a dataset and test (10.5). Explain how each of the following possible results would affect your beliefs about SD1, where the significance level is 5%. Hint: consider both DQ 10.1 and the difference between type I and type II errors.

a) Reject $H_0 \colon Y_1 \; non\text{SD}_1 \; Y_2$; do not reject $H_0 \colon Y_2 \; non\text{SD}_1 \; Y_1$
b) Do not reject $H_0 \colon Y_1 \; non\text{SD}_1 \; Y_2$; reject $H_0 \colon Y_2 \; non\text{SD}_1 \; Y_1$
c) Reject both $H_0 \colon Y_1 \; non\text{SD}_1 \; Y_2$ and $H_0 \colon Y_2 \; non\text{SD}_1 \; Y_1$
d) Do not reject either $H_0 \colon Y_1 \; non\text{SD}_1 \; Y_2$ or $H_0 \colon Y_2 \; non\text{SD}_1 \; Y_1$

Unfortunately, if Y is continuous with unbounded support, it is impossible to truly test (10.5) and control size at level α. This is essentially because $Y_1 \; \text{SD}_1 \; Y_2$ can be violated by even a very small probability on a very high value of Y_2. That is, even if it looks like $Y_1 \; \text{SD}_1 \; Y_2$ in nearly all datasets, it may in fact be $Y_1 \; non\text{SD}_1 \; Y_2$; so no matter how strong the evidence seems, we can never be sure enough to reject $non\text{SD}_1$ and control size.

There are (at least) three responses. First, we can give up and go back to the null of dominance. But then we can never get very convincing positive evidence of $Y_1 \; \text{SD}_1 \; Y_2$; the best we can say is that we don't reject that claim. (Remember how everyone cringed at the seminar where the speaker said, "And since I can't reject $H_0 \colon \beta = 0$, my conclusion is that there is zero effect"?)

Second, we can weaken SD1 to something that's statistically tractable as the alternative hypothesis. Davidson and Duclos (2000, 2013) suggest "restricted" SD1, from Condition I of Atkinson (1987, p. 751). Restricted SD1 weakens the CDF characterization of (10.2): the CDF ordering only needs to hold on the interval $[a, b] \subset \mathbb{R}$, i.e., $F_1(r) \leq F_2(r)$ for all $r \in [a, b]$. Restricted SD1 is "weaker" in the sense that it is implied by SD1, but it does not imply SD1. For example, if the CDFs then cross in the very lower or upper tail, there can be restricted SD1 but not SD1.

Definition 10.1 (restricted SD1; Atkinson (1987))**.** For random variables Y_1 and Y_2 with respective CDFs $F_1(\cdot)$ and $F_2(\cdot)$, there is restricted first-order stochastic dominance of Y_1 over Y_2 on interval $[a, b]$ when $F_1(r) \leq F_2(r)$ for all $r \in [a, b]$.

Third, we can weaken the expected utility characterization in (10.1). Instead of defining $\mathcal{U}$ as the set of all utility functions, a restricted subset is used. For example, we could restrict attention to a certain parameteric class of utility functions (like CRRA) with

parameters restricted to a compact set. Again, the tradeoff is that we are better able to find positive statistical evidence of "dominance" for weaker notions of dominance. In the extreme, we could simply assume a single utility function, say linear utility $u(x) = x$, in which case the problem reduces to comparing means. See Kaplan (2022a) for details and examples.

The CDF-based "restriction" is easier to approach statistically, but more difficult to interpret economically.

Kaplan (2022a) also constructs confidence sets for utility functions satisfying higher expected utility for Y_1. Let $\mathcal{U}$ be the (restricted) set of all utility functions under consideration. Let $\mathcal{D}$ be the subset of all utility functions for which $\mathrm{E}[u(Y_1)] \geq \mathrm{E}[u(Y_2)]$ for all $u(\cdot) \in \mathcal{D}$. Then, Kaplan (2022a) constructs $\hat{\mathcal{D}}_1$ and $\hat{\mathcal{D}}_2$, respectively called "inner" and "outer" confidence sets for $\mathcal{D}$, such that $\mathrm{P}(\hat{\mathcal{D}}_1 \subseteq \mathcal{D}) \to 1 - \alpha$ and $\mathrm{P}(\hat{\mathcal{D}}_2 \supseteq \mathcal{D}) \to 1 - \alpha$.

Exercises

Exercise E10.1. Apply the methodology from Goldman and Kaplan (2018a) and/or Kaplan (2022a) for assessing first-order stochastic dominance. Code in R and Stata is available at `https://kaplandm.github.io`. Note: for Goldman and Kaplan (2018a), I think the Stata code is much easier to use than the R code.

a. Find a paper (with publicly available data) that involves either stochastic dominance testing or a randomized experiment with a continuous outcome variable (like earnings); provide a link to the paper. The paper must be either published in a respectable economics journal[1] or be unpublished but have an author who has previously published in such a journal (like any Mizzou econ professor); or if you really want, you can use an example from an econometrics textbook.

b. Replicate one of the original paper's stochastic dominance tests (if it did one), or (if it's a randomized experiment) run two-sample one-sided Kolmogorov–Smirnov tests in both directions (i.e., alternative of treatment CDF above control, then alternative of treatment CDF below control CDF), using the post-treatment outcomes for the treatment group and control group.

c. Make a graph with the two empirical CDFs involved in the two-sample test in the previous step.

d. With the same data, run the new methodology from Goldman and Kaplan (2018a) or Kaplan (2022a) to assess stochastic dominance.

e. Compare your new results to the original results, both statistically and economically.

[1] For example, in top 500 of `https://ideas.repec.org/top/top.journals.all.html`

Chapter 11

Distributional Inference: Multiple Testing

Unit learning objectives for this chapter

11.1. Develop intuition and vocabulary for multiple testing, as opposed to testing a single joint hypothesis [TLO 2]

11.2. Interpret results of multiple testing generally and for distributional comparisons [TLO 1]

11.3. Judge whether multiple testing or joint testing is more appropriate for a particular economic question [TLO 3]

The distributional tests in Chapters 8–10 can only report one of two results: reject, or don't reject. This simplicity contrasts the complexity of the parameters themselves (CDFs). Simplicity is a tradeoff: easy to communicate, but possibly missing important information. For example, the NBER time series of US recessions is easy to communicate and can be very helpful, but reducing the entire economy to a single binary variable may lose important information for certain applications.

This chapter considers more informative statistical inference for distributions.

Optional resources for this chapter

- Goldman and Kaplan (2018a): CDF comparison across values
- Kaplan (2022a): expected utility comparison across utility functions

11.1 Multiple Testing: Concepts and Terms

Discussion Question 11.1 (FWER)**.** Consider two hypothesis tests, respectively testing null hypotheses H_{0a} and H_{0b}. Each test individually has 10% type I error rate. Let $f = \text{P}(\text{reject } H_{0a} \text{ and/or } H_{0b} \mid \text{both true})$. That is, f is the probability that at least one test rejects (i.e., that either or both tests reject) when both null hypotheses are true.

a) If the tests are statistically independent, then what is the probability f?
b) If the dependence is unknown, then what is a lower bound on the probability f?
c) If the dependence is unknown, then what is an upper bound on f?
d) If the dependence is unknown, then how can we adjust each test's α so that $f \leq 0.1$? (This is called a **Bonferroni adjustment** or **Bonferroni corrrection**.)

Discussion Question 11.2 (jelly beans)**.** Consider the comic at https://xkcd.com/882/. Explain the problem with running so many hypothesis tests with $\alpha = 0.05$ and claiming 95% confidence as in the newspaper article at the end.

Imagine testing multiple hypotheses simultaneously. Instead of a single H_0, we test H_{0h} for $h = 1, \ldots, H$. Or there could even be an infinite number of hypotheses: H_{0h} for $h = 1, 2, \ldots$, or for $h \in [0, 1]$, or $h \in \mathbb{R}$.

A **multiple testing procedure** (MTP) decides whether or not to reject each hypothesis under consideration. This is different than testing a single joint hypothesis like $H_0 \colon \beta_1 = \beta_2 = 0$. Although a joint hypothesis has multiple components (like $\beta_1 = 0$ and $\beta_2 = 0$), it requires only one single decision: reject H_0 or not. In contrast, multiple testing requires a different decision for each hypothesis. If we have two hypotheses, then the MTP must make two decisions, so there are four possible outcomes: reject/reject, reject/not, not/reject, not/not. If we have 10 hypotheses, then the MTP must make 10 decisions, with $2^{10} = 1024$ possible outcomes.

An MTP provides more information than a single joint hypothesis test; are the results more difficult to communicate? Yes, but often only slightly. For example, testing three hypotheses yields three reject/not decisions, which is still pretty simple. Even an infinite number of decisions may be simple to visualize. For example, if we have H_{0h} for $h \in [0, 1]$ and the MTP rejects for $h \in [0, 0.2]$ and $h \in [0.75, 1]$, we can draw those two line segments within the unit interval and immediately see which H_{0h} were rejected.

11.1.1 Familywise Error Rate

When looking across a family of hypotheses, how can we quantify the "false positive rate" that we want to control? Definition 11.1 offers one possibility. The following multiple testing terms are defined following Lehmann and Romano (2005b, §9.1).

Definition 11.1 (familywise error rate (FWER))**.** For a family of null hypotheses H_{0h} indexed by h, let $\mathcal{T} \equiv \{h : H_{0h} \text{ is true}\}$ be the set of indices of true hypotheses. The **familywise error rate** (FWER) is

$$\text{FWER} \equiv \text{P}(\text{reject any } H_{0h} \text{ with } h \in \mathcal{T}).$$

Definition 11.2 (weak and strong control of FWER)**.** Given Definition 11.1, **weak control of FWER** at level α requires FWER $\leq \alpha$ if each H_{0h} is true; **strong control of FWER** requires FWER $\leq \alpha$ for any $\mathcal{T}$.

Weak control of FWER can be useful theoretically, but it is not particularly useful in practice because nothing is guaranteed if even a single H_{0h} is false.

11.1.2 Interpretation as Confidence Set

An MTP with strong control of FWER also generates an "inner" confidence set (CS) for the true set $\mathcal{F} = \{h : H_{0h} \text{ is false}\}$ of false hypotheses (the complement of $\mathcal{T}$). That is, the set of rejected hypotheses $\hat{\mathcal{F}} = \{h : H_{0h} \text{ is rejected}\}$ is a conservative "estimate" of the true $\mathcal{F}$ in the sense that

$$\mathrm{P}(\hat{\mathcal{F}} \subseteq \mathcal{F}) \geq 1 - \alpha. \tag{11.1}$$

Thus, if $h \in \hat{\mathcal{F}}$, then we can feel relatively confident that H_{0h} is indeed false. This would not be true if only the pointwise type I error rates P(reject $H_{0h} \mid H_{0h}$ true) were controlled. See Kaplan (2022a).

Discussion Question 11.3 (expected utility MTP/CS)**.** Let $\mathcal{U}$ be a set of utility functions. Let X and Y represent two different consumption distributions (or income, or asset returns, or productivity, etc.). Assume that given utility function $u(\cdot)$, X is preferred over Y if $\mathrm{E}[u(X)] > \mathrm{E}[u(Y)]$, i.e., higher expected utility. Define $H_{0u}\colon \mathrm{E}[u(X)] \leq \mathrm{E}[u(Y)]$ over $u(\cdot) \in \mathcal{U}$.

a) Economically describe/interpret the set $\mathcal{F} \equiv \{u(\cdot) : H_{0u} \text{ is false}\}$.
b) Given this $\mathcal{F}$, economically and statistically describe/interpret an inner CS that satisfies (11.1).
c) For a given $\tilde{u}(\cdot)$, consider the usual one-sided t-test of $H_{0\tilde{u}}\colon \mathrm{E}[\tilde{u}(X)] \leq \mathrm{E}[\tilde{u}(Y)]$ at level α. Also consider an MTP of all H_{0u} over $u \in \mathcal{U}$ that has strong control of FWER at the same level α. Compared to the t-test, is the MTP more, less, or equally likely to reject $H_{0\tilde{u}}$? Why?
d) Imagine instead we reverse the inequality and run an MTP on $H_{0u}\colon \mathrm{E}[u(X)] \geq \mathrm{E}[u(Y)]$ over $u(\cdot) \in \mathcal{U}$, and collect the $u(\cdot)$ of the rejected H_{0u} into $\hat{\mathcal{T}}$. What is the economic and statistical interpretation of this $\hat{\mathcal{T}}$?

11.1.3 Alternatives to FWER

There are alternative false positive rates besides FWER. Their primary motivation is a concern that FWER is too strict when there are many hypotheses, i.e., that the corresponding type II error rate is too high.

The k-FWER is the probability of making at least k familywise errors:

$$\begin{aligned} k\text{-FWER} &\equiv \mathrm{P}(\text{reject at least } k \text{ null hypotheses } H_{0h} \text{ with } h \in \mathcal{T}) \\ &= \mathrm{P}(k \text{ or more false positives}). \end{aligned} \tag{11.2}$$

The original FWER is the special case $k = 1$. This concept and related procedures were introduced by Lehmann and Romano (2005a).

Discussion Question 11.4 (k-FWER)**.** Consider testing H_{0h} for $h = 1, 2, \ldots, 10$. You run one test with strong control of (1-)FWER at level 10%. You run a second test with strong control of 4-FWER at level 10%.

a) For any particular H_{0h}, which test is more likely to reject? Why?
b) What's your interpretation if the second test rejects 3 of the 10 hypotheses?
c) Imagine the second test rejects H_{0h} for $h = 2, 4, 6, 8, 10$. How does this affect your belief about H_{02}?
d) How do you interpret the results if the first test rejects H_{03} and the second test rejects H_{0h} for $h = 1, 3, 5, 7$?

The **false discovery proportion** (FDP) is the proportion of rejected hypotheses that were actually true. For example, if there are 100 total hypotheses, and 20 are rejected, of which 2 were actually true, then the FDP is $2/20 = 0.1$. If nothing is rejected, then FDP is defined to be zero. One way to use FDP is for an MTP to control FDP below γ with high probability $1 - \alpha$: $\mathrm{P}(\mathrm{FDP} \leq \gamma) \geq 1 - \alpha$. See Lehmann and Romano (2005a).

Alternatively, the **false discovery rate** (FDR) is the expected value of the FDP: $\mathrm{FDR} = \mathrm{E}(\mathrm{FDP})$. An MTP could be designed to control FDP $\leq \alpha$. See Benjamini and Hochberg (1995).

11.1.4 Other Ways to Improve Power

Stepdown and pre-test procedures improve power while maintaining strong control of FWER.

A **stepdown procedure** works iteratively: run the MTP once, and then if any H_{0h} are rejected, adjust critical values appropriately and re-run the MTP, etc. Stepdown procedures improve power without sacrificing FWER control, although they result in more false positives (but only in cases that do not affect FWER). Stepdown procedures trace back to Holm (1979); see also Lehmann and Romano (2005b, Ch. 9). (There are also step-up procedures, but at least the most basic one requires independence, which often fails in economic applications.)

A **pre-test procedure** is usually for one-sided MTPs, to determine which H_{0h} are "clearly" true and thus do not "need" to be examined. By reducing the number of hypotheses, the power is improved. The pre-test is usually run at a level much smaller than α, to guarantee negligible impact to FWER.

11.2 One-Sample, Two-Sided

The family of null hypotheses is

$$H_{0r} \colon F(r) = F_0(r), \quad r \in \mathbb{R}. \tag{11.3}$$

This uncountably infinite number of hypotheses sounds daunting, but the monotonicity of $F(\cdot)$ and the discreteness of the sample data simplify the structure.

In contrast, the GOF $H_0\colon F(\cdot) = F_0(\cdot)$ warranted only one single decision. The MTP provides more detailed information, at least when the GOF rejects.

11.2.1 KS and Dirichlet

The KS and Dirichlet approaches can generate a uniform confidence band (Section 8.6). The corresponding MTP rejects any H_{0r} for which $F_0(r)$ lies outside the band. This MTP has strong control of FWER.

There is a general equivalence between uniform confidence bands and MTPs that control FWER. Let $F(\cdot)$ be any function on $\mathbb{R}$ (like a CDF). Assume there exists a uniform confidence band $[\hat{L}(\cdot), \hat{U}(\cdot)]$ such that

$$\mathrm{P}(\hat{L}(\cdot) \le F(\cdot) \le \hat{U}(\cdot)) = 1 - \alpha. \tag{11.4}$$

Consider the MTP that rejects $H_{0r}\colon F(r) = F_0(r)$ when $F_0(r)$ lies outside the band. The true hypotheses are H_{0r} for $r \in \mathcal{T}$. Then,

$$\begin{aligned}
\mathrm{FWER} &= \mathrm{P}(\text{reject at least one true } H_{0r}) \\
&= \mathrm{P}(F_0(r) \notin [\hat{L}(r), \hat{U}(r)] \text{ for some } r \in \mathcal{T}) \\
&= 1 - \mathrm{P}(\hat{L}(r) \le F_0(r) \le \hat{U}(r) \text{ for all } r \in \mathcal{T}) \\
&= 1 - \mathrm{P}(\hat{L}(r) \le F(r) \le \hat{U}(r) \text{ for all } r \in \mathcal{T}) \\
&\le 1 - \mathrm{P}(\hat{L}(r) \le F(r) \le \hat{U}(r) \text{ for all } r \in \mathbb{R}) \\
&= 1 - (1 - \alpha) = \alpha.
\end{aligned}$$

Thus, the MTP has strong control of FWER at level α.

11.3 Two-Sample and/or One-Sided

Instead of a two-sample GOF test that says only "they're different" or "I can't reject that they're identical," an MTP tests $H_{0r}\colon F_1(r) = F_2(r)$ for each $r \in \mathbb{R}$. The two-sample results can show which particular part(s) of a distribution are affected by an experimental treatment, or differ across regions, or change over time, etc.

The one-sided variant relates to first-order stochastic dominance (SD1), in particular restricted SD1 (Definition 10.1). Let $H_{0r}\colon F_1(r) \ge F_2(r)$, so rejecting H_{0r} contributes evidence toward restricted SD1 of $F_1(\cdot)$ over $F_2(\cdot)$. Then $\mathcal{F}$, the set of false hypotheses, is the set over which there is restricted SD1 because $\mathcal{F} = \{r : F_1(r) < F_2(r)\}$. Thus, the MTP with strong FWER control at level α also generates an inner $1-\alpha$ CS for $\mathcal{F}$ as described in Section 11.1.2. That is, the set of rejected hypotheses $\hat{\mathcal{F}}$ is the set of points for which we are confident that $F_1(\cdot)$ is below $F_2(\cdot)$, in the sense that $\mathrm{P}(\hat{\mathcal{F}} \subseteq \mathcal{F}) \ge 1 - \alpha$.

Discussion Question 11.5 (GOF vs. MTP)**.** Consider $H_{0r}\colon F_1(r) = F_2(r)$ for $r \in \mathbb{R}$. Consider an MTP with strong control of FWER at level α. Consider the GOF test that rejects $H_0\colon F_1(\cdot) = F_2(\cdot)$ if and only if at least one H_{0r} is rejected by the MTP.

a) Explain why this GOF test controls size at level α.
b) Prove that the MTP is more informative by showing that the GOF test's result is uniquely determined by the MTP results, but the MTP results are not uniquely determined by the GOF test's result.

Discussion Question 11.6 (MTP for SD1)**.** Imagine you want to learn about the (causal) effects of an unconditional cash transfer program in Kenya, in which certain households are given a one-time $1000.00 gift (not just a loan). For now, don't worry about possible issues like spillovers and non-iid sampling. For both the treatment and control group, household consumption is measured five years later, to see if there is any persistent effect. Let $F_T(\cdot)$ and $F_C(\cdot)$ denote the corresponding treatment and control CDFs. You want to learn about (restricted) SD1 of the treatment distribution over the control distribution. All tests/MTPs use level $\alpha = 0.05$.

a) A global test does not reject $H_0\colon F_T(\cdot) \leq F_C(\cdot)$. What do you learn about SD1?
b) A global test rejects $H_0\colon F_T(\cdot) \geq F_C(\cdot)$. What do you learn about SD1?
c) An MTP rejects $H_{0r}\colon F_T(r) \geq F_C(r)$ for all r between the sample 5th and 95th percentiles of the control distribution. What do you learn about SD1?
d) An MTP rejects $H_{0r}\colon F_T(r) \geq F_C(r)$ for all r between the sample 5th and 65th percentiles of the control distribution, and it rejects $H_{0r}\colon F_T(r) \leq F_C(r)$ for r between the sample 80th and 90th control percentiles, with no other rejections in either direction. What do you learn about SD1?

Discussion Question 11.7 (MTP for distributional treatment effects)**.** Consider again the randomized experiment setup of DQ 9.1, inspired by Gneezy and List (2006). Let $F_T(\cdot)$ and $F_C(\cdot)$ denote the treatment and control CDFs (for the productivity variable). The FWER level is $\alpha = 0.05$.

a) An MTP for $H_{0r}\colon F_T(r) = F_C(r)$ rejects (only) for all r between the 5th and 25th sample percentiles of the control distribution. What does this suggest about the gift exchange treatment effect?
b) An MTP for $H_{0r}\colon F_T(r) = F_C(r)$ rejects (only) for all r between the 75th and 95th sample percentiles of the control distribution. What does this suggest about the gift exchange treatment effect?
c) Would you have come to the same conclusion as in the previous part if instead of the MTP results you (only) knew that a GOF test rejected $H_0\colon F_T(\cdot) = F_C(\cdot)$? Why/not?
d) An MTP for $H_{0r}\colon F_T(r) \geq F_C(r)$ rejects (only) for all r between the 5th and 25th sample percentiles of the control distribution, and an MTP for $H_{0r}\colon F_T(r) \leq F_C(r)$ rejects (only) for all r between the 75th and 95th sample percentiles of the control distribution. What does this suggest about the gift exchange treatment effect?
e) Would you have come to the same conclusion as in the previous part if instead of

the MTP results you (only) knew that a global test rejected both $H_0\colon F_T(\cdot) \geq F_C(\cdot)$ and $H_0\colon F_T(\cdot) \leq F_C(\cdot)$? Why/not?

Exercises

Exercise E11.1. Apply the two-sample method from Goldman and Kaplan (2018a), similar to the example in their Section 8.1. R and Stata code is available.[1] Note: the Stata code (`distcomp` command) is much easier to use than the R code.

a. Find a paper that includes (publicly available) data from a randomized experiment; provide a link to the paper. The paper must be either published in a respectable economics journal[2] or be unpublished but have an author who has previously published in such a journal (like any Mizzou econ professor); or if you really want, you can use an example from an econometrics textbook.

b. Using only the post-treatment outcomes (e.g., wages) for the treatment and control groups: 1) compute the difference between the treated and untreated mean outcomes, 2) compute differences between various quantiles of the treated and untreated outcome distributions, 3) run the distributional method to get an overall 2-sided p-value as well as the ranges of values (if any) where equality is rejected at a 10% level. Compare the three results.

c. Compare your new results to whatever main result was in the original paper.

d. Discuss any reasons that comparing the post-treatment outcomes in the data may not estimate the true treatment effect (e.g., attrition, general equilibrium concerns, etc.).

[1] https://kaplandm.github.io

[2] For example, in top 500 of https://ideas.repec.org/top/top.journals.all.html

Part IV

Bootstrap and Friends

Chapter 12
Bootstrap: Basics

Unit learning objectives for this chapter

12.1. Develop intuition about the "bootstrap world" and how frequentist measures of uncertainty can be computed from it [TLO 2]

12.2. Compare among the different ways to use the bootstrap approach to quantify uncertainty [TLO 3]

Notation: I use scalars because it's easier, but almost everything applies to vectors, too.

This chapter assumes iid sampling to develop intuition; Chapter 13 has non-iid extensions.

Optional resources for this chapter

- Textbook: Efron and Tibshirani (1993) is more applied-focused and starts from basic probability theory, though doesn't contain newer methods from the past few decades (I've heard good things about the book, and generally concur); MU library: `http://merlin.lib.umsystem.edu/record=b2432078~S1`
- Textbook: Shao and Tu (1995) is good for basic theory (I read it); MU library: `http://merlin.lib.umsystem.edu/record=b2717057~S1`
- Textbook: Davison and Hinkley (1997)? MU: `http://merlin.lib.umsystem.edu/record=b3774612~S1`
- Survey papers: MacKinnon (2002) (he has some helpful slides I found on Google, too), and MacKinnon (2006), "Bootstrap methods in econometrics."
- R: try package `boot` (Canty and Ripley, 2019; Davison and Hinkley, 1997)

12.1 Introduction

Bootstrap techniques are popular in economics. Although people say "the bootstrap," there are actually many different types of bootstrap, and multiple ways to construct a confidence interval for each type. Although bootstrap methods work for a wide variety of estimators and models, they are not magic and can fail. When bootstrap fails, sometimes the related technique of subsampling can work. However, subsampling is not strictly better: subsampling involves an additional smoothing parameter and often has lower power.

"The" bootstrap was introduced by Efron (1979).[1] It improved upon the Quenouille–Tukey jackknife,[2] though jackknife methods are still used in some cases. As computation power has grown since 1979, bootstrap methods have become more convenient in practice, although with complex estimators and/or large datasets they may still be computationally demanding.

The bootstrap is primarily frequentist, but see Chapter 14 for Bayesian interpretation.

12.2 Preliminaries: The Plug-in Principle

Notationally, let $F(\cdot)$ denote the population joint distribution of all observable variables. Let $\theta = \theta(F(\cdot))$ be the population parameter of interest. That is, θ can be interpreted as a feature of the population distribution $F(\cdot)$, where the function $\theta(\cdot)$ describes how to compute that feature given a distribution. Let $\hat{F}(\cdot)$ be the empirical distribution, as in Section 8.3.

The **plug-in principle** (or **analogy principle**) suggests replacing $F(\cdot)$ with $\hat{F}(\cdot)$ to estimate θ:

$$\hat{\theta} = \theta(\hat{F}(\cdot)) \quad \text{estimates} \quad \theta = \theta(F(\cdot)). \tag{12.1}$$

12.2.1 Example: Mean

Consider the population mean. In terms of $F(\cdot)$, the mean is

$$\theta(F(\cdot)) = \int_{\mathbb{R}} y \, \mathrm{d}F(y). \tag{12.2}$$

For example, if the population is standard normal, so $F(\cdot) = \Phi(\cdot)$, then applying $\theta(\cdot)$ to $\Phi(\cdot)$ yields $\theta(\Phi(\cdot)) = 0$, since $\int_{\mathbb{R}} x \, \mathrm{d}\Phi(x) = \int_{\mathbb{R}} x\phi(x) \, \mathrm{d}x = 0$.

[1] My favorite part is admittedly the acknowledgements section, which ends with, "I also wish to thank the many friends who suggested names more colorful than *Bootstrap*, including *Swiss Army Knife*, *Meat Axe*, *Swan-Dive*, *Jack-Rabbit*, and my personal favorite, the *Shotgun*, which, to paraphrase Tukey, 'can blow the head off any problem if the statistician can stand the resulting mess.'"

[2] The jackknife is a linear approximation of the bootstrap. They are similar when applied to linear statistics/estimators like the mean, but the jackknife can be much worse for nonlinear statistics.

Analogous to (12.2), using the same $\theta(\cdot)$ but plugging in $\hat{F}(\cdot)$ for $F(\cdot)$:

$$\theta(\hat{F}(\cdot)) = \int_{\mathbb{R}} y \, \mathrm{d}\hat{F}(y) = \sum_{i=1}^{n} (Y_i)(1/n) = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}, \tag{12.3}$$

the usual sample mean. That is, the sample mean can be interpreted as the mean of the empirical distribution. (The empirical distribution is a discrete distribution with probability mass $1/n$ on each Y_i value, so the integral simplifies to a sum over the Y_i values with probability weight $1/n$ each.)

12.2.2 Example: OLS

Consider the linear projection coefficient vector in terms of population CDFs:

$$\boldsymbol{\beta}(F(\cdot)) = \left[\int_{\mathbb{R}^k} \boldsymbol{x}\boldsymbol{x}' \, \mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x}) \right]^{-1} \int_{\mathbb{R}^{k+1}} \boldsymbol{x} y \, \mathrm{d}F_{\boldsymbol{X},Y}(\boldsymbol{x}, y) = [\mathrm{E}(\boldsymbol{X}_i \boldsymbol{X}_i')]^{-1} \, \mathrm{E}(\boldsymbol{X}_i Y_i). \tag{12.4}$$

Applying the same function $\boldsymbol{\beta}(\cdot)$ to $\hat{F}(\cdot)$ instead of $F(\cdot)$,

$$\begin{aligned} \boldsymbol{\beta}(\hat{F}(\cdot)) &= \left[\int \boldsymbol{x}\boldsymbol{x}' \, \mathrm{d}\hat{F}_{\boldsymbol{X}}(\boldsymbol{x}) \right]^{-1} \int \boldsymbol{x} y \, \mathrm{d}\hat{F}_{\boldsymbol{X},Y}(\boldsymbol{x}, y) = \left[\sum_{i=1}^{n} (\boldsymbol{X}_i \boldsymbol{X}_i') \frac{1}{n} \right]^{-1} \sum_{i=1}^{n} (\boldsymbol{X}_i Y_i) \frac{1}{n} \\ &= \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right]^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i Y_i, \end{aligned}$$

the OLS estimator $\hat{\boldsymbol{\beta}}$. This is the same as replacing the E (population expectation operator) with $\widehat{\mathrm{E}}$ (sample expectation operator):

$$\boldsymbol{\beta}(F(\cdot)) = [\mathrm{E}(\boldsymbol{X}_i \boldsymbol{X}_i')]^{-1} \, \mathrm{E}(\boldsymbol{X}_i Y_i) \implies \boldsymbol{\beta}(\hat{F}(\cdot)) = [\widehat{\mathrm{E}}(\boldsymbol{X}_i \boldsymbol{X}_i')]^{-1} \, \widehat{\mathrm{E}}(\boldsymbol{X}_i Y_i). \tag{12.5}$$

12.2.3 Other Types of Parameters

Some parameters are not defined as functions of $F(\cdot)$, like treatment effects and other causal effects. However, if they are identified, then they can be written as functions of $F(\cdot)$.

Like most statistical inference, the bootstrap helps us quantify uncertainty about the feature of $F(\cdot)$ that (maybe) has a causal interpretation, but it does not quantify uncertainty about identification. Thus, when identification fails, even if the bootstrap confidence interval is valid for the statistical parameter $\theta(F(\cdot))$, the causal parameter may be far away.

12.3 The Real World and the Bootstrap World

Discussion Question 12.1 (frequentist bias)**.** In the frequentist framework, if you actually knew the true population $F(\cdot)$, how could you figure out the bias of an estimator? Explain. (Hint: what is the definition of bias? What in the definition does $F(\cdot)$ help us figure out?)

Discussion Question 12.2 (frequentist SE and CI)**.** Imagine you knew the true population $F(\cdot)$. (Hint: what are the definitions/properties of SE and CI?)

a) How could you figure out the true standard error $\text{SE}(\hat{\theta})$ of an estimator $\hat{\theta}$? Explain. (Not the estimated standard error, which you can just compute from a single dataset like usual.)
b) How could you figure out the coverage probability of a proposed confidence interval for θ? Explain.

Although bootstraps usually involve simulation, the bootstrap idea itself is not intrinsically computational. Indeed, some bootstrap methods are purely analytic (no simulation), like that of Hutson (2007). (That said, in the Chapter 6 introduction in Efron's own book, he calls the bootstrap a "computer-based method," so perhaps my point is unimportant.)

(I worry the following intuition does not capture the full depth and beauty of bootstrap theory, but I hope it provides an anchor for a bootstrap novice (. . . like me?).)

Extending the plug-in principle of Section 12.2, the bootstrap idea is to learn about the true sampling distribution of an estimator (under true $F(\cdot)$) by using the sampling distribution under $\hat{F}(\cdot)$. From Section 8.3, $\hat{F}(\cdot)$ is close to $F(\cdot)$; hopefully, the corresponding sampling distribution is also close to the true sampling distribution. Computationally, to estimate the frequentist properties of estimator $\hat{\theta}$, we can repeatedly draw random samples from $\hat{F}(\cdot)$, and see how estimator $\hat{\theta}$ varies across these samples.

12.3.1 The Real World

Consider an example of the frequentist sampling framework. Let X be a person's height (in meters), with population distribution $F(\cdot)$. There is iid sampling, $X_i \overset{iid}{\sim} F(\cdot)$. After taking a sample, we just have numbers (non-random), like $X_1 = 1.68$. But the frequentist view imagines all the possible samples that could have been drawn, treating X_1 as a random variable (before its value is observed). We could have drawn $X_1 = 1.43$, or $X_1 = 0.89$ (a child, perhaps), etc. For more review, see Sections 2.1, 3.1, and 3.5 of Kaplan (2022b).

The purpose of frequentist statistical inference is essentially to get a sense of how different our sample could have been. Our sample of size n could have contained very different heights than the sample we actually drew, and some of these samples may have an ECDF $\hat{F}(\cdot)$ very different from $F(\cdot)$.

Operators like expectation $\text{E}(\cdot)$ are usually implicitly defined wrt $F(\cdot)$. For example, consider the expected value of estimator $\hat{\theta} = \theta(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1, \ldots, X_n)$ is the full

dataset, and $X_i \overset{iid}{\sim} F(\cdot)$. Let $\tilde{F}(\cdot)$ be the distribution of $\boldsymbol{X}$ (with support $\mathcal{X}$), which is determined by $F(\cdot)$. The expectation of $\hat{\theta}$ means the weighted average over all possible samples we could have drawn from $\tilde{F}(\cdot)$:

$$\mathrm{E}(\hat{\theta}) = \mathrm{E}[\theta(\boldsymbol{X})] = \int_{\mathcal{X}} \theta(\boldsymbol{x}) \, \mathrm{d}\tilde{F}(\boldsymbol{x}).$$

In the bootstrap literature, this mechanism is often called the **real world**. In the real world, the population is $F(\cdot)$, with parameter of interest $\theta(F(\cdot))$. In the real world, sampling is $X_i \overset{iid}{\sim} F(\cdot)$, generating dataset $\boldsymbol{X}$ or equivalently $\hat{F}(\cdot)$. The real-world estimator is computed from the real-world sample: $\hat{\theta} = \theta(\boldsymbol{X})$, or $\hat{\theta} = \theta(\hat{F}(\cdot))$.

12.3.2 The Bootstrap World

Table 12.1 compares the real world with the parallel **bootstrap world**.[3] In the bootstrap world, the population is $\hat{F}(\cdot)$, with parameter of interest $\theta(\hat{F}(\cdot))$. In the bootstrap world, sampling is $X_i^* \overset{iid}{\sim} \hat{F}(\cdot)$, generating dataset $\boldsymbol{X}^*$ or equivalently $\hat{F}^*(\cdot)$. The bootstrap-world estimator is computed from the bootstrap-world sample: $\hat{\theta}^* = \theta(\boldsymbol{X}^*)$, or $\hat{\theta}^* = \theta(\hat{F}^*(\cdot))$. That is, we treat the ECDF as if it were the population, and define other objects accordingly.

Importantly, we can take repeated samples from the population in the bootstrap world, which we can't do in the real world. The hope is that $\hat{F}(\cdot)$ is close enough to $F(\cdot)$ that the sampling distribution using $\hat{F}(\cdot)$ is a good approximation of the true sampling distribution using $F(\cdot)$.

Discussion Question 12.3 (bootstrap world)**.** Cover up the Bootstrap World column in Table 12.1, except the first row. Try to figure out what the bootstrap world analogs are in the other rows, given that the "population distribution" is $\hat{F}(\cdot)$.

Table 12.1: The Real World and the parallel Bootstrap World.

Object	Real World	Bootstrap World
pop. distribution	$F(\cdot)$	$\hat{F}(\cdot)$
pop. parameter	$\theta(F(\cdot))$	$\theta(\hat{F}(\cdot))$
sample data	$X_i \overset{iid}{\sim} F(\cdot)$, $i = 1, \ldots, n$	$X_i^* \overset{iid}{\sim} \hat{F}(\cdot)$, $i = 1, \ldots, n$
sample dist/ECDF	$\hat{F}(\cdot)$	$\hat{F}^*(\cdot)$
estimator	$\hat{\theta} = \theta(\hat{F}(\cdot))$	$\hat{\theta}^* = \theta(\hat{F}^*(\cdot))$
root	$\hat{\theta} - \theta$	$\hat{\theta}^* - \hat{\theta}$

[3] David Freedman's term, according to Efron and Tibshirani (1993, p. 86); and see their Figure 8.3 on page 87.

12.4 Empirical Bootstrap

This section discusses one specific bootstrap approach to estimating the sampling distribution of an estimator, assuming iid sampling. Section 12.6 discusses how to use this to construct a confidence interval.

The **empirical bootstrap** uses the idea from Section 12.3: treat the empirical distribution $\hat{F}(\cdot)$ as if it were the population distribution $F(\cdot)$. The empirical bootstrap is also known as the **multinomial bootstrap**, for reasons seen in Section 13.1. It is also known as the **nonparametric bootstrap**, to contrast the parametric bootstrap in Section 13.2.1. It is also known as the **pairs bootstrap** because it samples "pairs" $(Y_i, \boldsymbol{X}_i)$ from the empirical distribution, instead of keeping the $\boldsymbol{X}_i$ fixed and only resampling residuals like in Section 13.2.2.

Method 12.1 details the empirical bootstrap steps.

Method 12.1 (empirical bootstrap)**.** Assume (in the real world) $\boldsymbol{W}_i \overset{iid}{\sim} F(\cdot)$; e.g., perhaps $\boldsymbol{W}_i = (Y_i, \boldsymbol{X}_i')'$. Let n denote the sample size (in both the real world and bootstrap world). Let B denote the number of bootstrap replications. Let $\boldsymbol{W}_i^{*b}$ denote observation i in bootstrap sample b, $i = 1, \ldots, n$, $b = 1, \ldots, B$.

To generate the bootstrap samples, for $i = 1, \ldots, n$ and $b = 1, \ldots, B$, draw $\boldsymbol{W}_i^{*b}$ randomly from the original $\{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n\}$; i.e., $\boldsymbol{W}_i^{*b} = \boldsymbol{W}_S$ with $\mathrm{P}(S = j) = 1/n$ for $j = 1, \ldots, n$. Do this independently and **with replacement**, i.e., it is fine to get the same S twice for a given b and thus have $\boldsymbol{W}_i^{*b} = \boldsymbol{W}_j^{*b}$ for $i \neq j$. This could also be written as $\boldsymbol{W}_i^{*b} \overset{iid}{\sim} \hat{F}(\cdot)$, where $\hat{F}(\cdot)$ is the empirical distribution.

Let θ denote the population parameter. Let $\hat{\theta}$ denote an estimator computed from the original $\boldsymbol{W}_i$, $i = 1, \ldots, n$. Let $\hat{\theta}^{*b}$ denote the same estimator but computed from bootstrap sample b of observations $\boldsymbol{W}_i^{*b}$, $i = 1, \ldots, n$. The empirical bootstrap estimates the real-world sampling distribution of $\hat{\theta} - \theta$ by the bootstrap-world distribution of $\hat{\theta}^{*b} - \hat{\theta}$. More specifically, the above procedure provides B random draws of $\hat{\theta}^{*b} - \hat{\theta}$ from its probability distribution (conditional on the original sample). □

Although Method 12.1 consistently estimates the sampling distribution for a wide variety of estimators, it can fail in some cases; see Sections 13.3–13.6.

Discussion Question 12.4 (empirical bootstrap 1)**.** Let $n = 2$ with $Y_1 = 0$ and $Y_2 = 1$. Let $\theta = \mathrm{E}(Y)$.

a) What is the "population" mean in the bootstrap world?
b) What are the possible bootstrap samples (Y_1^*, Y_2^*)?
c) What's the probability of drawing each of the possible samples in (b)? (Sanity check: should sum to 100% probability.)

Discussion Question 12.5 (empirical bootstrap 2)**.** Continue from DQ 12.4.

a) What are the possible values of the bootstrap-world estimator $\hat{\theta}^* = (Y_1^* + Y_2^*)/2$?
b) For each of the values in (a), what's the probability of drawing such a value? (Sanity check: should sum to 1.)

c) If we take B bootstrap samples and compute $\hat{\theta}^{*b}$ in each sample ($b = 1, \ldots, B$), then what's the bootstrap approximation of the bootstrap-world probability that $\hat{\theta}^* = 0$?

d) For large B, explain why the approximation in (c) will be close to your exact probability in (b).

12.5 Standard Errors

The standard error is a feature of the sampling distribution, so it can also be estimated by bootstrap. The following describes how to estimate the standard error based on B draws of $\hat{\theta}^{*b} - \hat{\theta}$, whether from Method 12.1 or any other bootstrap.

The standard error of $\hat{\theta}$ in the real world is simply the standard deviation of its sampling distribution. Because θ is a constant, the standard error also equals the standard deviation of the sampling distribution of $\hat{\theta} - \theta$.

Method 12.2 is essentially Algorithm 6.1 of Efron and Tibshirani (1993).

Method 12.2 (empirical bootstrap: SE)**.** Compute B values of $\hat{\theta}^{*b}$ using Method 12.1 or another bootstrap. Then the estimated standard error is

$$\widehat{\mathrm{SE}}(\hat{\theta}) = \sqrt{(B-1)^{-1} \sum_{b=1}^{B} (\hat{\theta}^{*b} - \overline{\hat{\theta}^*})^2}, \quad \overline{\hat{\theta}^*} \equiv \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b}. \qquad \square$$

Method 12.3 provides an alternative bootstrap standard error suggested by Chernozhukov, Fernández-Val, and Melly (2013, p. 2222–2223). If the sampling distribution is (asymptotically) normal, then its standard deviation equals its interquartile range divided by the standard normal interquartile range. Thus, a bootstrap estimate of the sampling distribution interquartile range can be used, which may be less sensitive to outliers.

Method 12.3 (empirical bootstrap: SE for normal)**.** First compute B values of $\hat{\theta}^{*b}$ using Method 12.1 or another bootstrap. Let q^*_τ denote the sample τ-quantile among $\hat{\theta}^{*b}$ over $b = 1, \ldots, B$. Let z_τ denote the τ-quantile of a $\mathrm{N}(0, 1)$ distribution. Then the estimated standard error is

$$\widehat{\mathrm{SE}}(\hat{\theta}) = \frac{q^*_{0.75} - q^*_{0.25}}{z_{0.75} - z_{0.25}} \qquad \square$$

As $B \to \infty$, the bootstrapped standard error estimator approaches the bootstrap world "population" standard error, which hopefully is near the real-world standard error.

12.6 Confidence Intervals

There are multiple ways to construct a CI for θ based on B draws of $\hat{\theta}^{*b} - \hat{\theta}$.

12.6.1 CI Properties

Consider the two-sided CI $[\hat{L}, \hat{U}]$, where the hats remind us that the lower and upper endpoints are random variables from the frequentist perspective (they can have different values in different datasets).

If the CI has $1-\alpha$ **coverage probability** (CP), then $\mathrm{P}(\hat{L} \leq \theta \leq \hat{U}) = 1-\alpha$. (In practice, these probabilities are often asymptotic, like $\mathrm{P}(\hat{L} \leq \theta \leq \hat{U}) \to 1-\alpha$.) That is, given the population/DGP and given our procedure for computing $\hat{L}$ and $\hat{U}$, there is a $1-\alpha$ probability of randomly sampling a dataset in which the CI contains the true (non-random) population parameter θ. For example, if $1-\alpha = 0.90$ and we randomly sampled 100 datasets from the same population, then we'd expect around 90 of the corresponding CIs to contain the true θ.

There are an infinite number of possible two-sided CIs that have correct CP, so sometimes other properties are desired.

An **equal-tailed** CI satisfies

$$\mathrm{P}(\hat{L} > \theta) = \mathrm{P}(\hat{U} < \theta), \tag{12.6}$$

i.e., there is equal probability of the CI being "too low" or "too high." To satisfy the overall CP, this implies $\mathrm{P}(\hat{L} > \theta) = \mathrm{P}(\hat{U} < \theta) = \alpha/2$. This further implies

$$\mathrm{P}(\hat{L} \leq \theta) = 1-\alpha/2, \quad \mathrm{P}(\theta \leq \hat{U}) = 1-\alpha/2, \tag{12.7}$$

i.e., $[\hat{L}, \infty)$ and $(-\infty, \hat{U}]$ are one-sided $1-\alpha/2$ CIs. Thus, an equal-tailed $1-\alpha$ CI can be constructed as the intersection of two one-sided $1-\alpha/2$ CIs.

A **symmetric** CI satisfies $\hat{U} - \hat{\theta} = \hat{\theta} - \hat{L}$. Equivalently, a symmetric CI can be written as $\hat{\theta} \pm \hat{c}$, like $\hat{c} = 1.96\,\widehat{\mathrm{SE}}(\hat{\theta})$.

If the distribution of estimator $\hat{\theta}$ is normal with mean equal to the true θ, then the equal-tailed CI is also symmetric. The normal distribution allows some convenient simplifications in the formulas that we're all familiar with, but that can actually make it more difficult for us to understand the fundamental properties themselves. (And bootstrap distributions are not normal.) To help us try to understand the properties themselves, DQ 12.6 considers a non-normal CI.

Discussion Question 12.6 (CI properties)**.** Consider the CI $[\hat{L}, \hat{U}]$ with $\mathrm{P}(\hat{L} = \hat{U} = 3) = \alpha$ and $\mathrm{P}(\hat{L} = -\infty, \hat{U} = \infty) = 1-\alpha$. That is, with probability α we set the CI to $[3, 3]$, and otherwise we set it to $(-\infty, \infty)$. (So if $\alpha = 0.1$ and we randomly sample 100 datasets, the CI is $[3, 3]$ in around 10 of the datasets and the CI is $(-\infty, \infty)$ in all the remaining datasets.)

a) Starting from the definition of coverage probability, compute the coverage probability given $\theta = 3$.
b) Starting from the definition of coverage probability, compute the coverage probability given $\theta = 7$.
c) Starting from the definition of coverage probability, compute the coverage probability given general $\theta \neq 3$.

d) Let $\theta = 7$; is the CI equal-tailed? Why/not?
e) Let $\theta = 7$; is the CI symmetric? Why/not?

12.6.2 Normal CI, Bootstrapped SE

One approach is to use a bootstrapped standard error from Section 12.5 along with asymptotic normality. Often,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma^2). \tag{12.8}$$

Given asymptotic variance estimator $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$, defining $\widehat{\mathrm{SE}}(\hat{\theta}) = \hat{\sigma}/\sqrt{n}$,

$$\hat{Z}_n \equiv \frac{\hat{\theta} - \theta}{\widehat{\mathrm{SE}}(\hat{\theta})} = \overbrace{\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma}}^{\xrightarrow{d}\sigma^{-1}\mathrm{N}(0,\sigma^2)} \overbrace{\frac{\sigma}{\hat{\sigma}}}^{\xrightarrow{p}1} \xrightarrow{d} Z \sim \mathrm{N}(0,1). \tag{12.9}$$

The left-hand side has been **Studentized** since the estimator $\hat{\theta}$ was "centered" (at the true θ) and "scaled" using the estimated standard error. The right-hand side shows that the Studentized estimator has an asymptotically **pivotal** distribution: it does not depend on any unknown parameters. A statistic with an (asymptotic) pivotal distribution can be called an (asymptotic) pivot.

Consider an equal-tailed 95% CI for θ. With Z in (12.9), $\mathrm{P}(-1.96 < Z < 1.96) = 0.95$. Consequently, with $\doteq$ meaning we drop asymptotically negligible terms,

$$\begin{aligned}
0.95 &\doteq \mathrm{P}(-1.96 < \frac{\hat{\theta} - \theta}{\widehat{\mathrm{SE}}(\hat{\theta})} < 1.96) \\
&= \mathrm{P}(-1.96\,\widehat{\mathrm{SE}}(\hat{\theta}) - \hat{\theta} < -\theta < 1.96\,\widehat{\mathrm{SE}}(\hat{\theta}) - \hat{\theta}) \\
&= \mathrm{P}(\hat{\theta} - 1.96\,\widehat{\mathrm{SE}}(\hat{\theta}) < \theta < \hat{\theta} + 1.96\,\widehat{\mathrm{SE}}(\hat{\theta})),
\end{aligned}$$

which is the familiar CI $\hat{\theta} \pm 1.96\,\widehat{\mathrm{SE}}(\hat{\theta})$.

More generally, let Z follow any continuous distribution. Let z_τ denote the τ-quantile of Z, so $\mathrm{P}(Z \le z_\tau) = \tau$ for $0 < \tau < 1$. For any α, $\mathrm{P}(z_{\alpha/2} \le Z \le z_{1-\alpha/2}) = 1 - \alpha$. Thus,

$$\begin{aligned}
1 - \alpha &\doteq \mathrm{P}(z_{\alpha/2} \le \frac{\hat{\theta} - \theta}{\widehat{\mathrm{SE}}(\hat{\theta})} \le z_{1-\alpha/2}) \\
&= \mathrm{P}(z_{\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta}) - \hat{\theta} \le -\theta \le z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta}) - \hat{\theta}) \\
&= \mathrm{P}\big((-1)(z_{\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta}) - \hat{\theta}) \ge (-1)(-\theta) \ge (-1)(z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta}) - \hat{\theta})\big) \\
&= \mathrm{P}(\hat{\theta} - z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta}) \le \theta \le \hat{\theta} - z_{\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta})),
\end{aligned} \tag{12.10}$$

the CP of the CI $[\hat{\theta} - z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta}), \hat{\theta} - z_{\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta})]$.

This CI structure looks unfamiliar and intially wrong. But if $Z \sim \mathrm{N}(0,1)$, then $z_{\alpha/2} = -z_{1-\alpha/2}$ and (12.10) becomes $\hat{\theta} \pm z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta})$.

Method 12.4 (bootstrap CI: SE)**.** Let $\widehat{\mathrm{SE}}(\hat{\theta})$ be a bootstrap standard error estimator as in Section 12.5. An equal-tailed $1-\alpha$ asymptotic CI is

$$\left[\hat{\theta} - z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta}), \hat{\theta} - z_{\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta})\right],$$

where z_τ is the τ-quantile of the Z in (12.9). For example, if $Z \sim \mathrm{N}(0,1)$, then $z_{0.025} = -1.96$ or $z_{0.95} = 1.64$. □

12.6.3 Root Method

Method 12.5 is sometimes called the **root method** because it is based on the bootstrap sampling distribution of the **root**, $\hat{\theta} - \theta$. Method 12.5 is also called the basic bootstrap or standard bootstrap. As with Method 12.4, Method 12.5 can be used to construct a CI given any bootstrapped sampling distribution; it is not specific to Method 12.1. Before describing the method itself, some motivation is provided.

As seen in (12.10), a CI's lower endpoint actually comes from the upper $(1-\alpha/2)$-quantile of the asymptotic distribution of the Studentized estimator, whereas the upper endpoint comes from the lower $(\alpha/2)$-quantile. This may initially seem strange since we're so used to Gaussian distributions, which are symmetric, so we usually replace $-z_{\alpha/2}$ by $+z_{1-\alpha/2}$.

Besides the math in (12.10), a simple example may help intuition. Imagine $\hat{\theta} \sim \mathrm{N}(\theta, 1)$, so $\mathrm{P}(\theta - 1.96 < \hat{\theta} < \theta + 1.96) = 0.95$. If we happen to sample a dataset with $\hat{\theta} = \theta + 1.96$, then $\theta = \hat{\theta} - 1.96$, i.e., the true θ is below our estimate by 1.96. This characterizes the lower endpoint: we'll include values as much as 1.96 below our estimate, but exclude values even farther below our estimate. The true value θ is farthest below our estimated $\hat{\theta}$ when $\hat{\theta}$ is drawn from the upper quantiles of its sampling distribution, which is why an upper quantile determines the lower endpoint. Similarly, if we sample a dataset with $\hat{\theta} = \theta - 1.96$, then $\theta = \hat{\theta} + 1.96$, i.e., the true θ is considerably above our estimate. This provides the upper endpoint: we'll include values up to 1.96 above our estimate, but exclude values even higher.

Discussion Question 12.7 (deriving CI from sampling distribution)**.** Consider a standard exponential sampling distribution for the root: $\hat{\theta} - \theta \sim \mathrm{Exp}(1)$, whose CDF is $F(x) = 1 - e^{-x}$. The τ-quantile is $F^{-1}(\tau) = -\ln(1-\tau)$. Imagine a two-sided 90% CI for θ. With $\alpha = 0.1$, (12.10) suggests the equal-tailed CI $[\hat{\theta} - F^{-1}(0.95), \hat{\theta} - F^{-1}(0.05)]$. Here, you'll consider why other variations would not work well. Draw an example with each of these; try drawing the PDF of $\hat{\theta}$ on a graph first, with the true θ labeled. (Note: $-\ln(0.05) \approx 3$ and $-\ln(0.95) \approx 0.05$, but you shouldn't need to use any numbers if you make good drawings.)

a) What's the CP of the CI $[\hat{\theta} - F^{-1}(0.05), \hat{\theta} - F^{-1}(0.95)]$?
b) What's the CP of $[\hat{\theta} - F^{-1}(0.95), \hat{\theta} + F^{-1}(0.95)]$?
c) What's the CP of $[\hat{\theta} + F^{-1}(0.05), \hat{\theta} + F^{-1}(0.95)]$?

The bootstrap estimates not only the standard error, but the entire sampling distribution of $\hat{\theta}$. In Section 12.6.2, we presumed to know the pivotal asymptotic distribution of a Studentized estimator, like $Z \sim \mathrm{N}(0,1)$ in (12.9). Put differently, we used an approximation like $\hat{\theta} - \theta \sim \mathrm{N}(0, \widehat{\mathrm{SE}}^2)$ and then used the fact that the τ-quantile of the $\mathrm{N}(0, \widehat{\mathrm{SE}}^2)$ distribution is $z_\tau \widehat{\mathrm{SE}}$.

Instead of imposing asymptotic normality, the bootstrap could estimate the shape of the sampling distribution. That is, instead of plugging in quantiles from a known (asymptotic) distribution like $\mathrm{N}(0,1)$, the bootstrap can estimate the relevant quantiles of the distribution of $\hat{\theta}$ directly.

Consider an example similar to (12.10) but based on the root's sampling distribution. Imagine we know the true sampling distribution of the root $\hat{\theta} - \theta$, and that its τ-quantiles are $r_\tau \equiv \mathrm{Q}_\tau(\hat{\theta} - \theta)$. Unlike in (12.10), the standard errors are implicitly captured by r_τ. Then,

$$\begin{aligned} 1 - \alpha &\doteq \mathrm{P}(r_{\alpha/2} < \hat{\theta} - \theta < r_{1-\alpha/2}) \\ &= \mathrm{P}(r_{\alpha/2} - \hat{\theta} < -\theta < r_{1-\alpha/2} - \hat{\theta}) \\ &= \mathrm{P}\big((-1)(r_{\alpha/2} - \hat{\theta}) > (-1)(-\theta) > (-1)(r_{1-\alpha/2} - \hat{\theta})\big) \\ &= \mathrm{P}(\hat{\theta} - r_{1-\alpha/2} < \theta < \hat{\theta} - r_{\alpha/2}), \end{aligned} \tag{12.11}$$

the coverage probability of the CI $[\hat{\theta} - r_{1-\alpha/2}, \hat{\theta} - r_{\alpha/2}]$.

The r_τ in (12.11) can be replaced by bootstrap estimates. That is, r_τ can be replaced by r^*_τ, the τ-quantile of the bootstrap-world distribution of $\hat{\theta}^* - \hat{\theta}$. Since $\hat{\theta}$ is a constant in the bootstrap world, this is equivalent to taking the τ-quantile of $\hat{\theta}^*$ first and then subtracting $\hat{\theta}$. That is, if q^*_τ is the bootstrap τ-quantile of $\hat{\theta}^*$, then $r^*_\tau = q^*_\tau - \hat{\theta}$. Thus, the CI $[\hat{\theta} - r_{1-\alpha/2}, \hat{\theta} - r_{\alpha/2}]$ can be estimated by the bootstrap version $[\hat{\theta} - r^*_{1-\alpha/2}, \hat{\theta} - r^*_{\alpha/2}]$ or equivalently

$$[\hat{\theta} - (q^*_{1-\alpha/2} - \hat{\theta}), \hat{\theta} - (q^*_{\alpha/2} - \hat{\theta})] = [2\hat{\theta} - q^*_{1-\alpha/2}, 2\hat{\theta} - q^*_{\alpha/2}]. \tag{12.12}$$

This is summarized in Method 12.5.

Method 12.5 (bootstrap CI: basic/standard/root method)**.** Given B bootstrapped estimates $\hat{\theta}^{*b}$ for $b = 1, \ldots, B$, an equal-tailed $1 - \alpha$ CI for θ is

$$\big[2\hat{\theta} - q^*_{1-\alpha/2}, 2\hat{\theta} - q^*_{\alpha/2}\big], \tag{12.13}$$

where q^*_τ denotes the sample τ-quantile of $\hat{\theta}^{*b}$. This is equivalent to $[\hat{\theta} - r^*_{1-\alpha/2}, \hat{\theta} - r^*_{\alpha/2}]$, where r^*_τ is the sample τ-quantile of the bootstrapped roots, $\hat{\theta}^{*b} - \hat{\theta}$. □

If the finite-sample distribution is skewed, this may be more accurate in finite samples than restricting ourselves to a symmetric CI based on normality. However, I'm not aware of any theoretical results establishing any such improved accuracy.

12.6.4 Percentile Bootstrap CI

The **percentile bootstrap** in Method 12.6 is easy to describe and seemingly intuitive.

Method 12.6 (percentile bootstrap)**.** Assume we have computed a set of B bootstrapped estimators $\hat{\theta}^{*b}$ for $b = 1, \ldots, B$. Then, the equal-tailed percentile bootstrap $1 - \alpha$ confidence interval for θ is

$$[q^*_{\alpha/2}, q^*_{1-\alpha/2}], \tag{12.14}$$

where q^*_τ is the sample τ-quantile of $\hat{\theta}^{*b}$. □

Despite the simplicity and intuition, Method 12.6 is more difficult to rationalize. It is very different than the CI in Method 12.5 that clearly derives from the bootstrap estimate of the sampling distribution of $\hat{\theta}^{*b}$. Indeed, in some cases the percentile bootstrap is worse, but in other cases (like for population quantiles) it performs better; see Kaplan and Hofmann (2020) and Falk and Kaufmann (1991, p. 487).

The percentile CI is the most closely related to Bayesian approaches (Chapter 14).

12.6.5 Studentized Bootstrap CI

Instead of estimating the sampling distribution of the estimator $\hat{\theta}$ or the root $\hat{\theta} - \theta$, the bootstrap could estimate the sampling distribution of the Studentized $(\hat{\theta} - \theta)/\widehat{\mathrm{SE}}(\hat{\theta})$.

With Studentization, both the estimator $\hat{\theta}$ and the standard error estimator $\widehat{\mathrm{SE}}(\hat{\theta})$ must be computed in each bootstrap sample. Notationally, let $\hat{Z} = (\hat{\theta} - \theta)/\widehat{\mathrm{SE}}$ in the real world and $\hat{Z}^{*b} = (\hat{\theta}^{*b} - \hat{\theta})/\widehat{\mathrm{SE}}^{*b}$ in bootstrap sample b, for $b = 1, \ldots, B$. The standard error estimator $\widehat{\mathrm{SE}}^{*b}$ could use a formula or a bootstrap estimator. However, computationally, bootstrapping the standard error requires a bootstrap loop nested inside the original loop bootstrap (**double bootstrap**), which may take a long time to compute.

The benefit of the added complication is that Studentization increases accuracy, at least in theory, as long as the estimator is "smooth" enough (e.g., excluding quantile regression). In practice, it may require variance stabilization as in Algorithm 12.1 of Efron and Tibshirani (1993), although I am not familiar with this aspect myself.

Method 12.7 is similar to Method 12.4, but the latter uses quantiles z_τ from the asymptotic distribution of the Studentized estimator, whereas Method 12.7 uses bootstrap estimates of those quantiles, $\hat{z}_\tau$.

Method 12.7 (bootstrap-t/percentile-t/Studentized bootstrap CI)**.** Given bootstrapped Studentized estimators $\hat{Z}^{*b} = (\hat{\theta}^{*b} - \hat{\theta})/\widehat{\mathrm{SE}}^{*b}$, $b = 1, \ldots, B$, an equal-tailed $1 - \alpha$ CI is

$$[\hat{\theta} - \hat{z}_{1-\alpha/2}\,\widehat{\mathrm{SE}}, \hat{\theta} - \hat{z}_{\alpha/2}\,\widehat{\mathrm{SE}}], \tag{12.15}$$

where $\hat{z}_\tau$ is the sample τ-quantile of the B values of $\hat{Z}^{*b}$. □

Chapter 13
More Bootstrap and Subsampling

Unit learning objectives for this chapter

13.1. Develop intuition for how the bootstrap world can be structured to capture dependence among observations in the real world [TLO 2]

13.2. Judge whether or not a particular bootstrap is appropriate in a given empirical setting [TLO 3]

This chapter describes alternatives to the empirical bootstrap (Section 12.4). Some alternatives work in cases where the empirical bootstrap fails (e.g., due to non-iid data). Other topics are briefly mentioned, like bias correction, choice of B, and model selection.

Optional resources for this chapter

- Textbook: Shao and Tu (1995, Ch. 9) at MU library: `http://merlin.lib.umsystem.edu/record=b2717057~S1`
- Textbook: Davison and Hinkley (1997, Ch. 8), at MU: `http://merlin.lib.umsystem.edu/record=b3774612~S1`
- R: try package `boot` (Canty and Ripley, 2019; Davison and Hinkley, 1997)

13.1 Exchangeable Weights Bootstrap

The empirical bootstrap of Section 12.4 is actually a special case of **exchangeable weights bootstrap** (or "exchangeable bootstrap"). Instead of resampling observations in each bootstrap sample, observations are reweighted each time using randomly drawn weights. This can also be interpreted as reweighting $\hat{F}(\cdot)$. There are general theoretical

results on exchangeable bootstrap consistency; e.g., see Theorem 3.6.13 of van der Vaart and Wellner (1996, p. 355).

The empirical bootstrap is also called the **multinomial bootstrap** because it can be recast as an exchangeable bootstrap with multinomial weights. Let $\boldsymbol{W}$ be a random vector of weights, independent of the data, with

$$\boldsymbol{W} = (W_1, \ldots, W_n) \sim \text{Multinomial}(n; 1/n, \ldots, 1/n), \tag{13.1}$$

i.e., n "trials" in which each "category" ($i = 1, \ldots, n$) has the same "success" probability $1/n$. Recall that the empirical distribution assigns $1/n$ probability mass to each observation $i = 1, \ldots, n$. Instead, the $\boldsymbol{W}$-weighted empirical distribution assigns probability mass W_i/n to observation i. (Sanity check: since $\sum_{i=1}^n W_i = n$, the sum of probabilities equals $n/n = 1$.) The bootstrap-world estimator $\hat{\theta}^*$ then applies the function $\theta(\cdot)$ to the weighted empirical distribution. In R and Stata, many estimation functions have something like a `weights=` argument that allows you to compute the weighted $\hat{\theta}^*$ readily. In the case of multinomial weights in which the W_i are nonnegative integers, it is equivalent to use the W_i as frequency weights (i.e., how many times observation i appears in the bootstrap sample) or to use the W_i/n-weighted empirical distribution. As before, this process of randomly drawing $\boldsymbol{W}$ and recomputing $\hat{\theta}^*$ is done B times to generate the $\hat{\theta}^{*b}$.

More generally, other weights can be used if they are nonnegative and have an **exchangeable** distribution. This property is similar to iid, but weaker. For example, the multinomial weights W_i are not independent: the last weight W_n is fully determined by the first $n-1$ weights because $W_i = n - \sum_{i=1}^{n-1} W_i$ to ensure the weights sum to n. Exchangeability means that any permutation of the weights vector has the same joint distribution as the original vector. For example, (W_1, W_2, W_3) has the same joint distribution as (W_3, W_1, W_2) or as (W_2, W_1, W_3). If the weights don't sum to n like the multinomial W_i, then more generally the weighted empirical distribution puts $W_i / \sum_{i=1}^n W_i$ probability on observation i, which ensures the probabilities sum to 1.

There are many possible examples of exchangeable bootstrap; the following are the most notable. The ***m*-out-of-*n* bootstrap** is the same as the multinomial bootstrap but with $\sum_{i=1}^n W_i = m$ for some $m < n$. However, for the same reasons as in Section 13.4.2, the computed standard errors have to be scaled by $\sqrt{m/n}$ to give standard errors for $\hat{\theta}_n$ instead of just $\hat{\theta}_m$. This can be confusing, so be careful. "Subsampling" m out of n observations *without* replacement can also be written as exchangeable weights; see Section 13.4. Taking $E_i \overset{iid}{\sim} \text{Exp}(1)$ and $W_i = E_i / \sum_{i=1}^n E_i$ makes $\boldsymbol{W} \sim \text{Dir}(1, 1, \ldots, 1)$, called the **Bayesian bootstrap** (Chapter 14), although from this perspective it is a valid frequentist bootstrap.

Any of these can replace Method 12.1 to generate the $\hat{\theta}^{*b}$, and then any of the methods from Section 12.6 can be applied as before.

13.2 Other Bootstraps

13.2.1 Parametric Bootstrap

In principle, the nonparametric $\hat{F}(\cdot)$ could be replaced by a parametric (maximum likelihood) estimator. This is the **parametric bootstrap**. For example, if $F(\cdot)$ is assumed to be Gaussian, then $\hat{F}(x) = \Phi((x - \hat{\mu})/\hat{\sigma})$, and bootstrap samples can be drawn iid from $\mathrm{N}(\hat{\mu}, \hat{\sigma}^2)$. As you might guess, this is rarely used in economics.

13.2.2 Residuals Bootstrap

We could also take bootstrap samples of the residuals, $\hat{\epsilon}_i$, if we have some regression model $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i$. The $\boldsymbol{X}_i$ remain the same in the boostrap world, but $Y_i^* = \boldsymbol{X}_i'\hat{\boldsymbol{\beta}} + \hat{\epsilon}_i^*$, where $\hat{\epsilon}_i^*$ is a random sample from $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$. More directly analogous would be sampling from the error terms ϵ_i, but of course they are unobserved. As-is, this is less robust than pairs bootstrap because it requires assumptions about the regression model and implicitly assumes homoskedasticity.

13.2.3 Bias-Corrected Bootstrap

There is a bias-corrected bootstrap, and a **bias-corrected and accelerated bootstrap** (BC_a) that seems popular, although I am not familiar with its inner workings. It has an approximation that is less computationally demanding, the **approximate bootstrap confidence** (ABC) interval; see Efron and Tibshirani (1993, §14,22) for an introduction to both.

13.2.4 Wild Bootstrap

The **wild bootstrap** is similar to a residuals bootstrap (Section 13.2.2) but generates the bootstrap world residuals differently. The original proposal is due to Wu (1986). Instead of drawing $\hat{\epsilon}_i^*$ from among all n residuals, only $\hat{\epsilon}_i$ is used, but it is multiplied by a random weight. Specifically, $\hat{\epsilon}_i^* = V_i^*\hat{\epsilon}_i$, where V_i^* is drawn randomly in each bootstrap replication, like $V_i^* \overset{iid}{\sim} \mathrm{N}(0,1)$ or $\mathrm{P}(V_i^* = -1) = \mathrm{P}(V_i^* = 1) = 1/2$. This accounts for heteroskedasticity, unlike the residuals bootstrap.

13.2.5 Smoothed and Iterated Bootstraps

Various smoothed bootstraps and iterated bootstraps have been proposed due to improved theoretical properties, though there are practical tradeoffs in terms of computation time and smoothing parameter selection.

13.3 Bootstrap Failure

The bootstrap can fail for multiple reasons. One reason is ignoring non-iid sampling; see Sections 13.5 and 13.6. Another reason is a lack of "smoothness" of an estimator wrt the data, as seen below.

Discussion Question 13.1 (bootstrap max)**.** Let $X_i \overset{iid}{\sim} \text{Unif}(0, \theta)$, where $\theta > 0$ is the upper bound of the distribution. Let $X_{n:n}$ denote the original sample maximum, and let $X^{*b}_{n:n}$ denote the bth bootstrap sample maximum.

a) What's $\text{P}(X_{n:n} < \theta)$?
b) What's $\text{P}(X_{n:n} = \theta)$?
c) What's $\text{P}(X^{*b}_{n:n} < X_{n:n})$? Hint: all the X_i are unique wp1, and the X^*_i are resampled with replacement and independently, and $\text{P}(A \text{ and } B) = \text{P}(A)\,\text{P}(B)$ if $A \perp\!\!\!\perp B$.
d) What's $\text{P}(X^{*b}_{n:n} = X_{n:n})$? How does this bootstrap world probability compare with the real world probability in (b)?
e) Does the discrepancy disappear as $n \to \infty$? Hint: $\lim_{n\to\infty}(1 + cn^{-1})^n = e^c$.

Another example is the standard "matching" estimator of average treatment effects; see Abadie and Imbens (2008), "On the Failure of the Bootstrap for Matching Estimators."

13.4 Subsampling

Much of the bootstrap's appeal is the ability to compute a CI without deriving an analytic formula. However, there are cases like DQ 13.1 where the bootstrap does not perform as desired, even asymptotically.

Subsampling (Politis and Romano, 1994a) is valid under weaker conditions than bootstrap. The theory is elegant. The drawback is that an additional smoothing parameter is introduced, and power may be reduced.

The intuition for subsampling is essentially the same as for bootstrap: repeatedly sample from $\hat{F}(\cdot)$, compute the estimator, and use the resulting sampling distribution. The difference is that instead of bootstrap samples of size n drawn with replacement, we take subsamples of size $m < n$ without replacement.

13.4.1 Subsampling Consistency

The convergence rate can be any n^r; I use $r = 1/2$ for simplicity.

Assumption A13.1 (subsampling limit distribution)**.** With iid sampling,

$$J_n(\cdot) \equiv \text{P}\{\sqrt{n}(\hat{\theta}_n - \theta) \le \cdot\} \to J(\cdot)$$

for some CDF $J(\cdot)$, where population parameter θ is estimated by $\hat{\theta}_n$.

Assumption A13.2 (subsample size)**.** As $n \to \infty$, $m \to \infty$ and $m/n \to 0$.

Method 13.1 (subsampling)**.** Let Q_m be the set of all subsamples of size m from the data, labeled in some way. The cardinality of Q_m is

$$|Q_m| = \binom{n}{m} = \frac{n!}{m!(n-m)!}.$$

Let $\boldsymbol{X}_{m,i}$ be the ith subsample of size m, and let $\hat{\theta}_{m,i}$ be the estimator calculated from $\boldsymbol{X}_{m,i}$. To estimate $J_n(\cdot)$, the subsampling distribution is

$$L_{m,n}(\cdot) \equiv \frac{1}{|Q_m|} \sum_{i=1}^{|Q_m|} \mathbb{1}\Big\{\sqrt{m}(\hat{\theta}_{m,i} - \hat{\theta}_n) \le \cdot\Big\}.$$

In practice, if $|Q_m|$ is prohibitively large, a randomly selected subset of Q_m is used. Using $m = n^{2/3}$ is one often reasonable option (Politis and Romano, 1994a, Rmk. 2.1 and §2.4) but is not always optimal. □

Computationally, the jackknife is a special case of subsampling with $m = n - 1$, and the delete-d jackknife is subsampling with $m = n - d$. However, if d is fixed as $n \to \infty$ (rather than $d \to \infty$), then both of these violate A13.2: $(n-1)/n \to 1$ and $(n-d)/n \to 1$, violating $m/n \to 0$.

Theorem 13.1 (subsampling)**.** *Under A13.1 and A13.2, for all x that are points of continuity of $J(\cdot)$, $L_{m,n}(x) \xrightarrow{p} J(x)$.*

13.4.2 Standard Errors

You probably shouldn't compute standard errors using subsampling. First, the standard deviation (of which the standard error is a special case) is most helpful when the distribution is Gaussian and thus characterized by its mean and standard deviation. However, if the asymptotic distribution is normal, you probably don't need to resort to subsampling. Second, consistently estimating the CDF $J(\cdot)$ does not imply consistent estimation of the standard deviation. For both reasons, it's better to compute a confidence interval directly (e.g., root method) than estimate standard errors.

Nonetheless, if you must, then read on.

For standard errors, you cannot just use the standard deviation of subsampled estimates since $\sqrt{n} \neq \sqrt{m}$. We want the standard error of the original sample's estimator, $\hat{\theta}_n$. Since $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} W$, defining W as a random variable with $W \sim J(\cdot)$, then

$$\mathrm{Var}\big(\sqrt{n}(\hat{\theta}_n - \theta)\big) \doteq \mathrm{Var}(W) \implies \mathrm{Var}(\hat{\theta}_n) \doteq \mathrm{Var}(W)/n.$$

Now, $\sqrt{m}(\hat{\theta}_m - \theta)$ has the same limit W, which means

$$\mathrm{Var}\big(\sqrt{m}(\hat{\theta}_m - \theta)\big) \doteq \mathrm{Var}(W) \implies \mathrm{Var}(\hat{\theta}_m) \doteq \mathrm{Var}(W)/m = \overbrace{[\mathrm{Var}(W)/n]}^{\mathrm{Var}(\hat{\theta}_n)}(n/m).$$

Thus, if we can estimate $\text{Var}(\hat{\theta}_m)$ from subsampling, the standard error estimate for $\hat{\theta}_n$ should be

$$\widehat{\text{SE}}(\hat{\theta}_n) = \sqrt{\text{Var}(\hat{\theta}_n)} \doteq \sqrt{(m/n)\,\text{Var}(\hat{\theta}_m)} = \sqrt{m/n}\,\widehat{\text{SE}}(\hat{\theta}_m).$$

That is, take the standard deviation of your subsampled $\hat{\theta}_{m,i}$ (i.e., your estimate of $\text{SE}(\hat{\theta}_m)$) and multiply by $\sqrt{m/n}$.

13.5 Clustered Data

Discussion Question 13.2 (bootstrap world correlation: pairs)**.** Let Y_i be annual labor income and X_i years of education. Consider a sample of (Y_i, X_i) for $i = 1, \ldots, n$.

a) In the real world, how/are X_i and Y_i correlated?
b) Imagine that in the bootstrap world, Y_i^* is drawn randomly from among the observed $(Y_1, \ldots, Y_n)$, and independently X_i^* is drawn randomly from among the observed $(X_1, \ldots, X_n)$. What's the correlation between Y_i^* and X_i^* in the bootstrap world?
c) How does the usual empirical bootstrap fix this problem and preserve the real-world correlation in the bootstrap world?

Discussion Question 13.3 (bootstrap world correlation: time)**.** Let $Y_{it} = 1$ if individual i is employed in time period t, and $Y_{it} = 0$ if not employed. Consider panel data with $n = 1000$, $T = 2$. Each time period is one week; e.g., $t = 1$ is last week, $t = 2$ is this week. Assume individuals i are sampled randomly from the population, then observed for two consecutive weeks each.

a) In the real world, how/are Y_{i1} and Y_{i2} correlated?
b) Imagine that in the bootstrap world, each of the $nT = 2000$ bootstrap sample values Y_{it}^* is drawn randomly with replacement from among the original $nT = 2000$ sample values. What's the correlation between Y_{i1}^* and Y_{i2}^* in the bootstrap world?
c) Propose an alternative bootstrap procedure to fix this problem and preserve the real-world correlation in the bootstrap world.

See also Cameron, Gelbach, and Miller (2008) and a variety of papers since then.

Clustered data can refer to many situations, but you can just think of panel data for now. In this case, each "cluster" is an individual i. Since by design we sample the same individual over time, there is usually dependence in the time dimension even if the individuals are all independent.

The general solution it to resample (or reweight) individuals. That is, treat i as the unit for resampling, rather than (i, t).

The biggest difficulty (and need for fancier methods) is when the number of clusters is "small" (say below 30, or even below 10). This is usually not a problem in panel data clustered by i (since n is usually big), but it can be a problem with clustering by geography, institution, etc. For example, there could be individual-level data from only

10 randomly chosen villages in India, or student-level data from 10 randomly selected schools. It seems like the wild bootstrap generally does best in such settings.

13.6 Time Series Data

Discussion Question 13.4 (time series bootstrap)**.** Let Y_t for $t = 1, \ldots, T$ be a sample of stationary time series data. Specifically, let Y_t denote the quarterly unemployment rate in the U.S. (And please correct me if that's not thought to be stationary.)

a) In the real world, how/are Y_t and Y_{t-1} correlated?
b) Imagine that in the bootstrap world, Y_t^* is drawn randomly from among the observed $(Y_1, \ldots, Y_T)$. What's the resulting correlation between Y_t^* and Y_{t-1}^* in the bootstrap world?
c) How well will the corresponding bootstrap CI perform? Why?

With time series data, we want the dependence structure in the real world to be replicated in the bootstrap world. Here, iid sampling in the bootstrap world is not sufficient. However, we can still use the same methods like Method 12.5 once we figure out an appropriate replacement for Method 12.1.

The following overviews are rather cursory. I think all methods are valid for stationary data (under certain assumptions), and the moving blocks bootstrap may also be valid for certain types of nonstationary data.

13.6.1 Moving Blocks Bootstrap

The **moving blocks bootstrap** tries to preserve dependence by sampling blocks of consecutive observations instead of individual observations.

For example, consider block length $\ell = 7$ for sample size $T = 42$. In the sample, there are 36 blocks of seven consecutive observations: $(X_1, X_2, \ldots, X_7)$, $(X_2, \ldots, X_8)$, $\ldots$, $(X_{36}, \ldots, X_{42})$. We draw our bootstrap sample by sampling from the 36 blocks (with replacement) rather than the 42 individual observations. For example, we might draw the six blocks starting with time (t) indices $\{4, 35, 20, 21, 2, 11\}$, so that our bootstrap sample has indices 4–10, 35–41, 20–26, 21–27, 2–8, 11–17, i.e.,

$$(Y_1^*, \ldots, Y_T^*) = (Y_4, Y_5, \ldots, Y_{10}, Y_{35}, Y_{36}, \ldots, Y_{41}, Y_{20}, Y_{21}, \ldots, Y_{26}, Y_{21}, Y_{22}, \ldots, Y_{27}, Y_2, Y_3, \ldots, Y_8, Y_{11}, Y_{12}, \ldots, Y_{17}). \quad (13.2)$$

Method 13.2 (moving blocks bootstrap)**.** Let T denote sample size. Given block length ℓ, the moving blocks bootstrap samples $k = T/\ell$ blocks of length ℓ from the original sample, with replacement. There are $T - \ell + 1$ blocks to choose from, where a block consists of consecutive observations $(X_t, X_{t+1}, \ldots, X_{t+\ell-1})$. This is equivalent to picking k indices from $\{1, 2, \ldots, T - \ell + 1\}$ with replacement for the starting index of each block, and then filling in the rest of each block with consecutive indices. If the k leading indices

are I_j for $j = 1, \ldots, k$, then the bootstrap sample is

$$(X_{I_1}, X_{I_1+1}, \ldots, X_{I_1+\ell-1}, X_{I_2}, \ldots, X_{I_2+\ell-1}, \ldots, X_{I_k}, \ldots, X_{I_k+\ell-1}). \qquad \square$$

Discussion Question 13.5 (bootstrap block length)**.** Consider the choice of ℓ.

a) Describe the problem with using $\ell = 1$.
b) Describe a setting where $\ell = 2$ is probably too small.
c) Describe the problem with using $\ell = T$.

If the block length ℓ is too small or too large, then moving blocks bootstrap will perform poorly. The special case $\ell = 1$ means the bootstrap world sampling is iid, which is usually not appropriate. However, at the other extreme, choosing $\ell = T$ is also bad: the bootstrap sample is simply the original sample, so there is no variation among bootstrap samples. The optimal ℓ is not too small or too big. The block length ℓ must be chosen large enough so that the dependence between X_t and $X_{t+\ell}$ is negligible, but small enough that there are enough possible blocks in the original sample to get enough variation. Unfortunately, I do not have a good reference for optimal selection of ℓ.

13.6.2 Circular Block Bootstrap

A related method is the **circular block bootstrap** (Politis and Romano, 1992). The only difference in implementation is that the time series is made "circular" so that $X_{T+1} \equiv X_1$, or generally $X_{T+j} \equiv X_j$. Thus, one may have blocks like

$$(X_{T-1}, X_T, X_1, X_2, \ldots).$$

The benefit is that now there are T possible blocks.

13.6.3 Stationary Bootstrap

Another related method is the **stationary bootstrap** (Politis and Romano, 1994b). The main difference is that block length is no longer a constant ℓ but randomly drawn (from some distribution) for each block. Thus, a single bootstrap sample could contain a block of length 7, a block of length 4, a block of length 5, etc. This seems to improve accuracy. See the resampling algorithm just before Proposition 1 of Politis and Romano (1994b, p. 1304) with tuning parameter value $p = n^{-1/3}$ (that determines the distribution of block length), which is the p rate they suggest on page 1306.

13.7 Other Bootstrap Uses and Considerations

Bootstrap can estimate an estimator's bias, but it is not always best to bias-correct (subtract the estimated bias from your estimator). Bias-correction reduces bias but increases variance, so it can decrease or increase mean squared error.

How big should B be? It depends. And in statistical software, often the default B is too small. For long-term research projects, I use a small B when exploring data, and increase B as I become more confident in the model specification. For the very final run, I use as large a B as I have the patience for, which may take overnight (or longer) to run. There are also more formal suggestions from Andrews and Buchinsky (2000, 2001, 2002), and Davidson and MacKinnon (2000).

Bootstrap can also be used for model selection; see Shao and Tu (1995) and Efron and Tibshirani (1993, §17.6–18), and Liu (2019) for an application to IVQR.

Exercises

Exercise E13.1. a. Find a published paper with replication materials (i.e., data and code) available.

b. Replicate one standard error estimate or one CI from the paper (related to their regressor of interest). It can be any type of estimate (IV, probit, QR, etc.).

c. Estimate the standard error using at least 2 different bootstrap or subsampling methods. Either way, make sure your methods are appropriate for the type of data; e.g., use a time series bootstrap for time series data, use a cluster bootstrap if the original SE are clustered (e.g., with panel data), etc. You may use either Method 12.2 or Method 12.3, but explain which you use. You may use existing code/packages/functions/commands, as long as you fully understand how they work and you describe them enough (probably with quotations from the documentation/help file) to convince me that they indeed do what you hope.

d. Alternatively, instead of part (c), use a single bootstrap/subsampling approach, but compute at least 3 different types of two-sided CI, choosing among Methods 12.4–12.7. Again, you may use existing code as long as you understand and describe it sufficiently well.

e. Qualitatively discuss your results (compared to each other and to the original paper's result).

Exercise E13.2. a. Find a paper that uses a bootstrap to get a CI for their main parameter of interest; provide a link to the paper. The paper must be either published in a respectable economics journal[1] or be unpublished but have an author who has previously published in such a journal (like any Mizzou econ professor); or if you really want, you can use an example from an econometrics textbook. Get their data, and replicate one such CI (you can use their code if they provide it).

b. Construct a DGP based on the empirical joint distribution of the variables in the data (making a reasonable guess about a structural error term distribution, if necessary). You can make small changes to simplify the DGP, but it should be reasonable that the observed data came from the DGP.

c. With your DGP, run 1000 simulation replications. In each replication, draw a new dataset from the DGP, and then compute the paper's bootstrap CI given the simulated dataset.

d. Compute the simulated coverage probability, i.e., the number of replications in which the CI contained the true parameter value (that you chose) divided by 1000.

e. Report and discuss the results.

[1] For example, in top 500 of https://ideas.repec.org/top/top.journals.all.html

Chapter 14
Bayesian Bootstrap

Unit learning objectives for this chapter

14.1. Develop intuition about the Bayesian perspective generally, including differences with the frequentist perspective [TLO 2]

14.2. Interpret priors and posteriors in various models [TLO 1]

This chapter is the long answer to, "What's the Bayesian bootstrap?" Although the Bayesian bootstrap has a frequentist interpretation (Section 13.1), I focus on the Bayesian interpretation.

Sampling is assumed iid, but there are variations allowing for more complex sampling, like Dong, Elliott, and Raghunathan (2014).

Optional resources for this chapter

- Textbook: Kaplan (2022b) Section 3.1
- R package `bayesboot` (Bååth, 2018)
- Textbook: Berger (1985)
- Chamberlain and Imbens (2003)

14.1 Bayesian Basics

Generally, the Bayesian approach helps us update our beliefs based on observed data. Specifically, the beliefs are about population parameters, and the updating requires a model that connects parameters with data. The **prior** represents our beliefs about parameters before seeing the data. The **likelihood** is the model of how data is generated

depending on the parameters. The **posterior** represents our beliefs about parameters after seeing the data and updating our prior. Very roughly speaking, the posterior is computed by multiplying the prior by the likelihood.

14.1.1 Beliefs and Data

A **belief** is quantified as a probability distribution. Let β be the parameter of interest, an unknown constant. Let random variable B represent our beliefs about β. For example, the belief that there's a 50% chance of negative β is expressed as $\mathrm{P}(B \leq 0) = 0.5$. We could keep asking ourselves what we believe to be the chance that β is below b, for more and more b, to trace out the CDF $\mathrm{P}(B \leq b)$. Or, we could take a shortcut and say $B \sim \mathrm{N}(\mu, \sigma^2)$ and pick (μ, σ^2) to best match our beliefs. This process of quantifying real-world beliefs is called **prior elicitation**.

To confuse you, instead of writing $B \sim \mathrm{N}(\mu, \sigma^2)$ to describe our beliefs, we write $\beta \sim \mathrm{N}(\mu, \sigma^2)$. This looks like we don't believe in a single true value of β. However, it is just a notational difference: in Bayesian analysis, the "parameter" actually represents our beliefs about the parameter, which are naturally expressed as a random variable.

Another main difference with the frequentist approach is how the data are treated. In the frequentist framework, observations like Y_i are treated as random variables whose distribution depends on the population distribution. In the Bayesian framework, the data are treated as non-random. That is, we condition on the actually observed values in the actual dataset. To emphasize this, I'll generally write observations as y_i (lowercase) instead of Y_i.

14.1.2 Model: The Likelihood

"Likelihood" sounds like maximum likelihood, which sounds like fully parametric models that assume independent Gaussian error terms and such. Indeed, a common first example is learning about the population mean after specifying a Gaussian likelihood. Similarly, basic Bayesian linear regression specifies Gaussian error terms.

However, such parametric assumptions are not always required. Even in the frequentist world, there is such thing as a "nonparametric likelihood," more commonly called **empirical likelihood** (EL); e.g., see Kiefer and Wolfowitz (1956) and Owen (1988, 2001). One nonparametric Bayesian approach is the (eventual) focus of this chapter.

14.1.3 Bayes' Theorem

You've probably seen **Bayes' Theorem** (or "law" or "rule"):

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(B)\,\mathrm{P}(A \mid B)}{\mathrm{P}(A)}. \tag{14.1}$$

There's also a version with PDFs. Abstracting somewhat, think of A as the data, and B as the parameter (or, the parameter being in some interval). The LHS is like the

posterior: what do we believe about the parameter, conditional on the data? The RHS says this equals our prior $\mathrm{P}(B)$ times the likelihood $\mathrm{P}(A \mid B)$, normalized by something that doesn't depend on the parameter, $\mathrm{P}(A)$.

Discussion Question 14.1 (are you sick?)**.** Let $\theta = 0$ if you're healthy and $\theta = 1$ if you're sick; this is the parameter of interest. Let $X = 0$ if the test says you're healthy, and $X = 1$ if it says you're sick. Assume the type I error rate is $\mathrm{P}(X = 1 \mid \theta = 0) = \alpha = 5\%$. Assume the type II error rate is zero. Your doctor says the test reports that you're sick. What do you believe? What do you think of the frequentist versus Bayesian approach here? Hint: does the prior $\mathrm{P}(\theta = 1)$ matter here? Hint: make a table of the *joint* probability distribution of (X, θ), and see the effect of changing various values.

This linked cookie example is also good.

14.1.4 Strengths and Weaknesses

The Bayesian approach has strengths and weaknesses compared with the frequentist approach. (Some of these are vaguely reminiscent of the comparison of structural and reduced-form approaches.)

- Strength: having a full posterior distribution for the parameter is much more helpful for making decisions under uncertainty than just a point estimate and confidence interval.
  - But: there is a large frequentist literature on "statistical decision theory."
    - But...
- Strength: the ability to incorporate prior beliefs may be valuable if there is important prior knowledge and/or the data don't say much (but in the real world a decision is required, so we can't just say "I don't know"), like in macro.
- Weakness: Bayesian analysis is not "objective" due to the influence of the prior over the posterior.
  - But: there are "objective" or "uninformative" priors that may be used, which make the "objectivity" even more transparent than frequentist methods.
    - But: which prior is truly objective?
      - But the frequentist estimate is equivalent to a Bayesian estimate with a certain prior...
  - But: asymptotically, Bayesian and frequentist estimates often agree (Bernstein–von Mises theorems), and frequentist properties are mostly asymptotic anyway.
    - But: this assumes a fixed prior asymptotically; in practice, given any sample size (no matter how large), the prior can have arbitrarily large influence over the posterior.

        - But...

- Weakness: misspecification error may be big since a parametric likelihood is required.
    - But: a parametric likelihood is not required.
        - But: in infinite-dimensional parameter space...
    - But: lots of people use frequentist maximum likelihood (probit, etc.), which can also be interpreted under misspecification.
        - But...

As you can see, these are all fruitful discussions, but there are no simple answers. Also, both Bayesian and frequentist methods can be misused (intentionally or not).

14.2 Beta–Binomial Model

My notation is not all conventional but hopefully helpful. The true population parameter is the non-random θ, while beliefs about θ are represented by random variable P (stands for "parameter"). Further, p is a dummy variable (in the calculus sense, not the econometrics sense), representing any possible value of θ but not necessarily the true value, e.g., for integrating a PDF. Also, variables Y_i, N_0, and N_1 are uppercase when treating them as random variables but lowercase y_i, n_0, and n_1 when conditioning on observed variables. (Elsewhere, θ might be used for θ, P, and p alike, and the uppercase/lowercase distinction may not be made.) When reading other Bayesian material, you can practice your understanding by trying to infer in each instance whether θ (or whatever variable) refers to the true θ, the belief P, or the dummy p.

14.2.1 Likelihood, Prior, and Posterior

Likelihood

First, the likelihood: given the parameter, what's the distribution of the data? Let $Y \in \{0, 1\}$, $\mathrm{P}(Y = 1) = \theta$, so $\mathrm{P}(Y = 0) = 1 - \theta$; Y has a **Bernoulli distribution** with parameter θ.

If the likelihood is Bernoulli, why is it called "binomial"? Define

$$N_1 \equiv \sum_{i=1}^{n} \mathbb{1}\{Y_i = 1\}, \quad N_0 \equiv \sum_{i=1}^{n} \mathbb{1}\{Y_i = 0\} = n - N_1. \tag{14.2}$$

Given iid Bernoulli Y_i, the sampling distribution of N_1 given θ is $N_1 \sim \mathrm{Binomial}(n, \theta)$, a binomial distribution with parameters n and θ (i.e., how many "successes" out of n, where the "success" probability is θ). It turns out only N_1 is needed to update the prior, not the individual Y_i.

When treated as non-random, lowercase is used:

$$n_1 \equiv \sum_{i=1}^{n} \mathbb{1}\{y_i = 1\}, \quad n_0 \equiv \sum_{i=1}^{n} \mathbb{1}\{y_i = 0\} = n - n_1. \tag{14.3}$$

Prior

Second, the prior: how can we quantify our beliefs about θ? We know $0 \leq \theta \leq 1$, so we should use a distribution with support on $[0, 1]$. One convenient option is the beta distribution; the prior is

$$P \sim \text{Beta}(a, b). \tag{14.4}$$

The beta distribution is restrictive, though often reasonable. For example, it cannot be bimodal, nor does it allow $\text{P}(P = 0.2) = 0.5$.

Posterior

Third, the posterior: given the prior and likelihood, what do we believe about θ after seeing the data? Magically, the posterior is also a beta distribution, an example of conjugacy (Section 14.3). Specifically, the posterior is

$$P \mid \boldsymbol{y} \sim \text{Beta}(a + n_1, b + n_0), \tag{14.5}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the observed data, (a, b) are from the prior in (14.4), and (n_1, n_0) are from (14.3). That is, given our prior belief $P \sim \text{Beta}(a, b)$ and the iid Bernoulli sampling (likelihood), our belief about θ after seeing the data $\boldsymbol{y}$ is now described by the updated distribution $\text{Beta}(a + n_1, b + n_0)$.

The posterior's form suggests an interpretation of the prior. If $a = b = 0$ in the prior, then the posterior is $\text{Beta}(n_1, n_0)$. So, prior $P \sim \text{Beta}(a, b)$ is like having previously seen data with a observations of $y_i = 1$ and b observations of $y_i = 0$. If a and b are big, then we are more confident in our prior, so more data (n_0, n_1) are needed to change our beliefs. Conversely, if $a = b = 0$, then the posterior seems entirely driven by the data, which seems "objective" (but there are other considerations we won't discuss).

Posterior Mean

The **posterior mean** is an important feature of the posterior. It is often reported as the "point estimate." Given quadratic loss function $L_2(\cdot)$, the posterior mean minimizes **posterior expected loss**:

$$\text{E}(P \mid \boldsymbol{y}) = \underset{g \in [0,1]}{\arg\min}\, \text{E}[L_2(P, g) \mid \boldsymbol{y}] = \underset{g \in [0,1]}{\arg\min}\, \text{E}[(P - g)^2 \mid \boldsymbol{y}], \tag{14.6}$$

where the (conditional) expectation is wrt the posterior of P (given data $\boldsymbol{y}$). Thus, in a Bayesian sense, the posterior mean is our "best guess" of θ under quadratic loss. Under

other loss functions, it may be optimal to report the posterior median (or τ-quantile) or mode.

The mean of a Beta(α, β) distribution is $\alpha/(\alpha+\beta)$, so the posterior mean of (14.5) is

$$\mathrm{E}(P \mid \boldsymbol{y}) = \frac{a + n_1}{a + b + n}. \tag{14.7}$$

If $n = 0$ (so $n_1 = 0$), then we simply have the prior mean, $a/(a + b)$. If a and b are big relative to n (strong prior), then the posterior mean remains close to the prior mean. Conversely, if a and b are small relative to n, then the posterior mean is driven primarily by the data.

Discussion Question 14.2 (beta–binomial vs. frequentist 1)**.** The usual frequentist point estimator of $\theta = \mathrm{P}(Y = 1)$ is $\hat{\theta} = N_1/n$.

a) What is the frequentist justification of $\hat{\theta}$ as a "good" estimator? (Hint: recall $\mathrm{P}(Y = 1) = \mathrm{E}[\mathbb{1}\{Y = 1\}]$.)

b) Is there any prior (i.e., any a and b) that makes the posterior mean equal the frequentist estimator, i.e., $\mathrm{E}(P \mid \boldsymbol{y}) = \hat{\theta}$?

Discussion Question 14.3 (beta–binomial vs. frequentist 2)**.** Continue from DQ 14.2.

a) If $\hat{\theta} = 0.5$ but $a/(a + b) = 1$, then is the posterior mean $\mathrm{E}(P \mid \boldsymbol{y})$ above, below, or equal to $\hat{\theta}$?

b) If $a/(a + b) \in [0, 1]$ is fixed but a and b increase, then does $\mathrm{E}(P \mid \boldsymbol{y})$ get closer to or farther from $\hat{\theta}$, or does it not change? Why?

c) Let $N_1 = 750$ and $n = 1000$, so $\hat{\theta} = 0.75$. Is there any (a, b) that makes $\mathrm{E}(P \mid \boldsymbol{y}) = 0.5$? What values/why?

Discussion Question 14.4 (posterior mean consistency)**.** Consider the beta–binomial posterior mean in (14.7) as an estimator of the true population parameter θ. Given a fixed prior Beta(a, b), as $n \to \infty$, does $\mathrm{E}(P \mid \boldsymbol{Y}) \xrightarrow{p} \theta$? (Note uppercase $\boldsymbol{Y}$; alternatively, one could ask about convergence for almost-all sequences $y_1, y_2, \ldots$.)

14.3 Conjugacy

The nice property in (14.26) where the prior and posterior are in the same distributional family is called **conjugacy**. The beta distribution is called the **conjugate prior** of the binomial (or Bernoulli) likelihood because it results in conjugacy. If a Gaussian prior had been used instead, the posterior would not have been Gaussian (or beta).

Generalizing the beta–binomial model, the multinomial likelihood's conjugate prior is a **Dirichlet distribution**. Just as the binomial distribution is a special case of the multinomial distribution, the beta distribution is a special case of the Dirichlet distribution. The Dirichlet distribution is a continuous, usually-unimodal (or flat) distribution over $(p_1, \ldots, p_J)$ with all $p_j \geq 0$ and $\sum_{j=1}^{J} p_j = 1$. (That is, the Dirichlet distribution's

support is the unit J-simplex.) If vector $\boldsymbol{P} = (P_1, \ldots, P_J)'$ follows a Dirichlet distribution, then the marginal distribution of each P_j is a beta distribution. Details are on Wikipedia, for example.[1]

The other common example of conjugacy is with Gaussian distributions. For example, a Gaussian likelihood (with known variance) and Gaussian prior (on the unknown mean parameter) lead to a Gaussian posterior. This can be extended to unknown variance, too. These and other examples of conjugate priors can be found on Wikipedia, for example.[2]

14.4 Dirichlet–Multinomial Model

The beta–binomial model extends to variables with J possible values in the **Dirichlet–multinomial model**.

Toward this generalization, it helps to rewrite the beta–binomial in different notation. Instead of $\mathrm{P}(Y = 1) = \theta$ and $\mathrm{P}(Y = 0) = 1 - \theta$, let $\theta_1 = \mathrm{P}(Y = 0)$ and $\theta_2 = \mathrm{P}(Y = 1)$. More generally, $Y = 0$ and $Y = 1$ could be replaced by $Y = v_1$ and $Y = v_2$, with respectively n_1 and n_2 such observations in the sample. Clearly $\theta_1 + \theta_2 = 1$ and $\theta_1, \theta_2 \geq 0$. Let vector $\boldsymbol{P} = (P_1, P_2)$ describe our belief about $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The prior can be written

$$\boldsymbol{P} \sim \mathrm{Dir}(a_1, a_2). \tag{14.8}$$

This is the same as $P_2 \sim \mathrm{Beta}(a_2, a_1)$ and $P_1 = 1 - P_2$ (and yes, the parameter order (a_2, a_1) is correct). The posterior is

$$\boldsymbol{P} \mid \boldsymbol{y} \sim \mathrm{Dir}(a_1 + n_1, a_2 + n_2). \tag{14.9}$$

More generally, let $Y \in \{v_1, \ldots, v_J\}$. Let

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J), \quad \theta_j \equiv \mathrm{P}(Y = v_j),\ j = 1, \ldots, J. \tag{14.10}$$

Let $\boldsymbol{P} = (P_1, \ldots, P_J)$ represent the belief about $\boldsymbol{\theta}$. The Dirichlet prior is

$$\boldsymbol{P} \sim \mathrm{Dir}(a_1, \ldots, a_J). \tag{14.11}$$

In the observed data, define

$$n_j \equiv \sum_{i=1}^{n} \mathbb{1}\{y_i = v_j\}, \quad j = 1, \ldots, J. \tag{14.12}$$

Then, the posterior is

$$\boldsymbol{P} \mid \boldsymbol{y} \sim \mathrm{Dir}(a_1 + n_1, \ldots, a_J + n_J). \tag{14.13}$$

For the posterior mean, one can apply the Dirichlet mean formula:

$$\boldsymbol{P} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_J) \implies \mathrm{E}(\boldsymbol{P}) = (\alpha_1, \ldots, \alpha_J) / \sum_{j=1}^{J} \alpha_j. \tag{14.14}$$

[1] https://en.wikipedia.org/wiki/Dirichlet_distribution

[2] https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions

Discussion Question 14.5 (Dirichlet posterior mean)**.** Consider the notation, prior, posterior, and properties in (14.11)–(14.14).

a) For any j, what's the posterior mean of P_j? That is, what's the jth component of (14.14) when you plug in the Dirichlet parameters from the posterior in (14.13)?
b) Which prior makes the posterior means all equal the corresponding frequentist estimators? That is, which $\boldsymbol{a} = (a_1, \ldots, a_J)$ makes $\mathrm{E}(P_j \mid \boldsymbol{y}) = n_j/n$ for all $j = 1, \ldots, J$? Hint: recall DQ 14.2(b).
c) Verbally describe the prior from (b).

14.5 Improper Priors

In DQs 14.2, 14.3, and 14.5, the Bayesian posterior mean is identical to the usual frequentist estimator given a particular prior. This is not necessarily the best definition of an "objective" prior, but it should reassure anybody who thinks the frequentist estimator is more "objective." It is a type of **matching prior**, although usually that refers to matching frequentist coverage probability (rather than the point estimate).

However, in both cases, the required "prior" is not actually a real distribution. That is, it is not a **proper prior**, but rather an **improper prior**. The required $\mathrm{Beta}(a, b)$ "prior" had $a = b = 0$, i.e., a $\mathrm{Beta}(0, 0)$ "distribution." Even Wikipedia knows that beta distributions need $a > 0$ and $b > 0$; there is no such thing as a $\mathrm{Beta}(0, 0)$ distribution. Similarly, $\mathrm{Dir}(0, 0, \ldots, 0)$ is not a real distribution.

An improper prior can be interpreted as the limit of a sequence of proper priors. Consider prior $\mathrm{Beta}(a, a)$ with $a \downarrow 0$. There is a corresponding sequence of $\mathrm{Beta}(a + n_1, a+n_0)$ posteriors. As $a \downarrow 0$, the posterior limit is $\mathrm{Beta}(n_1, n_0)$. Similarly, the improper Dirichlet prior is the limit when taking a sequence of $\mathrm{Dir}(a, a, \ldots, a)$ priors as $a \downarrow 0$; the limit of the corresponding sequence of posteriors is $\mathrm{Dir}(n_1, n_2, \ldots, n_J)$.

There are other examples of improper priors. If $P \sim \mathrm{N}(\mu, \tau^2)$, then one could take $\tau \to \infty$. If $P \sim \mathrm{Unif}(-a, a)$, then one could take $a \to \infty$.

14.6 Nonparametric Bayes

A popular "nonparametric" approach comes from Ferguson (1973, 1974). ("Nonparametric" meaning: not assuming distributions are known up to a finite-dimensional parameter vector.) A nice, more recent summary with examples from economics is given by Chamberlain and Imbens (2003). Instead of considering (belief) distributions over possible values of finite-dimensional parameter vector $\boldsymbol{\theta}$, this approach uses a **Dirichlet process** to describe a probability distribution over possible values of the population CDF. (A distribution of distributions.) Conjugacy makes the Dirichlet process easy to update.

With an improper prior, the posterior Dirichlet process simplifies to a Dirichlet distribution. Specifically, the posterior only includes discrete distributions of Y with the observed y_i as the only possible values (i.e., the support). This is convenient, but a little weird; but, recall the empirical bootstrap world's "population" distribution $\hat{F}(\cdot)$ is also

discrete, and that seemed fine. If n is small and either the continuity of the true CDF is important or acknowledging possible values outside $(y_{n:1}, y_{n:n})$ is important, then this may be of particular concern. Otherwise, it may not actually be much of a practical disadvantage, especially if the parameter of interest is something like a population mean.

14.7 Bayesian Bootstrap

For simplicity, imagine the y_j values are unique. This occurs with probability 1 if the distribution of Y is continuous.

The Bayesian bootstrap (Rubin, 1981) is the Dirichlet process model with improper prior, which leads to the posterior

$$\boldsymbol{P} \mid \boldsymbol{y} \sim \text{Dir}(1, \ldots, 1), \tag{14.15}$$

where the parameter vector is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ with $\theta_j = \text{P}(Y = y_j)$. (If some y_j values are repeated, then the Dirichlet distribution is adjusted accordingly, computing the Dirichlet parameter by adding together the 1s for each repeated observation of the corresponding y_j value.) Below, properties of this posterior are explored.

14.7.1 Population Mean

Consider the implied posterior for the population mean, $\mu \equiv \text{E}(Y)$. Let random variable M represent our belief about the non-random value μ. The posterior of M follows from the posterior of $\boldsymbol{P}$ in (14.15) because a specific value of $\boldsymbol{P} = \boldsymbol{p}$ uniquely determines the value $M = m$. Specifically, recall that the posterior only includes discrete distributions on values $y_1, \ldots, y_n$, with respective probabilities $\text{P}(Y = y_j) = P_j$. Thus,

$$M = \text{E}(Y \mid \boldsymbol{\theta} = \boldsymbol{P}) = \sum_{i=1}^{n} y_i P_i. \tag{14.16}$$

Computationally, the posterior distribution of M is approximated by repeatedly drawing $\boldsymbol{P}$ from its posterior and computing the corresponding M.

The expression for M in (14.16) can be used to compute its posterior mean, $\text{E}(M \mid \boldsymbol{y})$. Using (14.16),

$$\text{E}(M \mid \boldsymbol{y}) = \text{E}\left(\sum_{i=1}^{n} y_i P_i \mid \boldsymbol{y}\right) = \sum_{i=1}^{n} y_i \, \text{E}(P_i \mid \boldsymbol{y}) = \sum_{i=1}^{n} y_i (1/n) = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}, \tag{14.17}$$

where $\text{E}(P_i \mid \boldsymbol{y})$ is from (14.14) with each $\alpha_j = 1$ because $(P_1, \ldots, P_n) \mid \boldsymbol{y} \sim \text{Dir}(1, \ldots, 1)$. That is, with this particular nonparametric Bayesian model and improper prior, the posterior mean of (our belief about) the population mean is the same as the basic nonparametric frequentist estimator.

However, the Bayesian approach provides more than just a point estimate. It provides an entire posterior distribution describing our belief about μ, which is arguably more

useful than a sampling distribution. For example, imagine we need to make a decision whose consequences depend on μ. Given an appropriate loss function that captures such consequences, along with our posterior for M, we can choose the decision that minimizes posterior expected loss. A frequentist sampling distribution cannot be used for this purpose. That said, often the sampling distribution and posterior distribution are asymptotically equivalent. Such equivalence results are called **Bernstein–von Mises theorems**.

If the y_i are not unique, then the Dirichlet is modified. Assume the sample contains values $v_1, \ldots, v_J$ for some $J \leq n$. Assume value v_j is observed f_j times; that is, there are f_j observations whose y equals v_j. Let $\boldsymbol{P} = (P_1, \ldots, P_J)'$ refer to the probabilities $\mathrm{P}(Y = v_j)$. Then,

$$\boldsymbol{P} \mid \boldsymbol{y} \sim \mathrm{Dir}(f_1, \ldots, f_J). \tag{14.18}$$

As a special case, when the y_i are unique, then $v_j = y_j$ $(j = 1, \ldots, n)$ and $f_j = 1$, which reduces (14.18) to (14.15).

14.7.2 Other Population Features

Not only μ, but any feature of the distribution has an easily simulated posterior distribution. For example, Chamberlain and Imbens (2003) consider quantile regression as well as an IV estimator of returns to schooling.

Kaplan and Hofmann (2020) show higher-order frequentist accuracy of Bayesian bootstrap confidence intervals for population quantiles.

14.7.3 Population CDF

Discussion Question 14.6 (Bayesian bootstrap CDF 1)**.** Let $y_i = i$, $i = 1, 2, 3$. Let (P_1, P_2, P_3) be a random vector representing our belief about the population probabilities $\mathrm{P}(Y = 1)$, $\mathrm{P}(Y = 2)$, and $\mathrm{P}(Y = 3)$. The Bayesian bootstrap posterior here is $\boldsymbol{P} \mid \boldsymbol{y} \sim \mathrm{Dir}(1, 1, 1)$, as usual, with marginal distributions $P_j \mid \boldsymbol{y} \sim \mathrm{Beta}(1, 2)$.

a) Why is P_1 a random variable? E.g., where does the "randomness" come from, and what does it represent?
b) If we sampled a different dataset with different y_i, how/would the meaning of $\boldsymbol{P}$ differ?
c) Draw the CDF of Y evaluated at point y, $F_Y(y)$, and label important values in terms of $\mathrm{P}(Y = 1)$, $\mathrm{P}(Y = 2)$, and $\mathrm{P}(Y = 3)$ (which sum to 1).
d) What is the CDF corresponding to a particular value of $\boldsymbol{P} = \boldsymbol{p}$?

Discussion Question 14.7 (Bayesian bootstrap CDF 2)**.** Continue from DQ 14.6.

a) What's the mean of our posterior belief about $F_Y(y_1)$? Hint: recall the mean of $\mathrm{Beta}(a, b)$ is $a/(a + b)$.
b) What's the mean of our posterior belief about $F_Y(y_2)$? Hint: see previous hint, and $\mathrm{E}(A + B) = \mathrm{E}(A) + \mathrm{E}(B)$.
c) What's the mean of our posterior belief about $F_Y(\cdot)$?

Appendix to Chapter 14

14.A Technical Details: Posterior Derivation

This section shows the technical details for deriving the beta posterior in (14.5) and the Dirichlet posterior in (14.13).

First, the PDF of the Beta(a, b) prior is $\pi(\cdot)$:

$$\pi(p) = f_{a,b}(p) = \text{constant} \times p^{a-1}(1-p)^{b-1}. \tag{14.19}$$

The constant does not involve argument p and is not necessary to compute the posterior.

Second, consider the likelihood function. For a single Y_i, $\mathrm{P}(Y_i = 1) = \theta$, so

$$\ell(y_i \mid \theta = p) = p^{\mathbb{1}\{y_i=1\}}(1-p)^{\mathbb{1}\{y_i=0\}}. \tag{14.20}$$

Although for maximum likelihood it is common to write likelihoods with reverse notation $\ell(p \mid \boldsymbol{y})$, I write $\ell(\boldsymbol{y} \mid p)$ to emphasize the use of Bayes' theorem, parallel to (14.1). With iid sampling, the likelihood for the full vector $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the product of the individual likelihoods,

$$\ell(\boldsymbol{y} \mid p) = \prod_{i=1}^{n} p^{\mathbb{1}\{y_i=1\}}(1-p)^{\mathbb{1}\{y_i=0\}} = p^{n_1}(1-p)^{n_0}, \tag{14.21}$$

using n_1 and n_0 defined in (14.3).

Bayes' theorem in (14.1) extends to PDFs. Generally, consider data vector $\boldsymbol{W}$ and parameter vector $\boldsymbol{\theta}$, with $f_{\boldsymbol{W}}(\cdot)$ the marginal PDF of $\boldsymbol{W}$, $\pi(\cdot)$ the prior on $\boldsymbol{\theta}$, $\ell(\boldsymbol{W} \mid \boldsymbol{\theta} = \boldsymbol{t})$ the likelihood, and $\pi(\cdot \mid \boldsymbol{W})$ the posterior for $\boldsymbol{\theta}$. Let $\boldsymbol{t}$ be a possible value of $\boldsymbol{\theta}$, and $\boldsymbol{w}$ a value of $\boldsymbol{W}$. Then, parallel to (14.1),

$$\pi(\boldsymbol{t} \mid \boldsymbol{w}) = \frac{\pi(\boldsymbol{t})\ell(\boldsymbol{w} \mid \boldsymbol{t})}{f_{\boldsymbol{W}}(\boldsymbol{w})}. \tag{14.22}$$

Further, any PDF must integrate to 1. Thus, integrating the posterior (over $\boldsymbol{t}$) must equal 1. Thus, we can ignore any "constant" terms (not depending on $\boldsymbol{t}$) because they

can be determined later, as whichever constant makes the posterior integrate to 1. Often this is written as: the posterior is proportional to the prior times the likelihood,

$$\pi(\boldsymbol{t} \mid \boldsymbol{w}) \propto \pi(\boldsymbol{t})\ell(\boldsymbol{w} \mid \boldsymbol{t}). \tag{14.23}$$

In our beta–binomial example,

$$\pi(p \mid \boldsymbol{y}) = \frac{\pi(p)\ell(\boldsymbol{y} \mid p)}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \propto \pi(p)\ell(\boldsymbol{y} \mid p). \tag{14.24}$$

Determining the constant is straightforward: if $\tilde{f}(\cdot)$ is the unscaled PDF, then the constant must be $1/\int_{\mathbb{R}} \tilde{f}(t)\,\mathrm{d}t$ to ensure

$$\int_{\mathbb{R}} C\tilde{f}(t)\,\mathrm{d}t = C\int_{\mathbb{R}} \tilde{f}(t)\,\mathrm{d}t = 1. \tag{14.25}$$

Ignoring the denominator in Bayes' theorem sounds like cheating, but it actually makes sense. The denominator represents the prior belief about the marginal distribution of the data. This sounds like the likelihood, but it's not. Rather than "conditioning" on the true value θ, it integrates out θ according to the prior. Thus, it does not contain any "new" information.

Other terms that do not depend on the parameter can similarly be removed. The part of the PDF that depends on the parameter is called the **kernel**. (This differs from the "kernel" for nonparametric smoothing.) From (14.22), generally, the kernel of the posterior is proportional to the kernel of the prior times the kernel of the likelihood.

In the beta–binomial model, the kernel approach is used as follows. From (14.19), the kernel of the beta prior is $p^{a-1}(1-p)^{b-1}$. From (14.21), the kernel of the likelihood is actually the entire likelihood, $p^{n_1}(1-p)^{n_0}$. Thus, up to a multiplicative constant, the posterior is

$$\pi(p \mid \boldsymbol{y}) \propto \overbrace{p^{a-1}(1-p)^{b-1}}^{\text{prior kernel}} \times \overbrace{p^{n_1}(1-p)^{n_0}}^{\text{likelihood}} = \overbrace{p^{a+n_1-1}(1-p)^{b+n_0-1}}^{\text{posterior kernel}}. \tag{14.26}$$

By inspection, the posterior kernel has the same form as the beta kernel for the prior. Indeed, it is a beta kernel, from which we can infer the parameters of the corresponding beta distribution: $a+n_1$ and $b+n_0$. Thus, the posterior is $\mathrm{Beta}(a+n_1, b+n_0)$, as stated in (14.5).

More generally, the Dirichlet posterior in (14.13) can be derived formally using the prior kernel and likelihood. The Dirichlet prior's kernel is

$$\prod_{j=1}^{J} p_j^{a_j-1} = p_1^{a_1-1} \times \cdots \times p_J^{a_J-1}. \tag{14.27}$$

The likelihood is

$$\prod_{j=1}^{J} p_j^{n_j} = p_1^{n_1} \times \cdots \times p_J^{n_J}. \tag{14.28}$$

Thus,

$$\pi(\boldsymbol{p} \mid \boldsymbol{y}) \propto (\prod_{j=1}^{J} p_j^{a_j-1})(\prod_{j=1}^{J} p_j^{n_j}) = \prod_{j=1}^{J} p_j^{a_j+n_j-1}, \tag{14.29}$$

which is another Dirichlet PDF kernel. Specifically, it corresponds to the posterior already stated in (14.13).

14.B Dirichlet Process Notes

The Dirichlet process extends the finite-dimensional Dirichlet distribution, analogous to how a Gaussian process extends a finite-dimensional multivariate Gaussian distribution. If you're familiar with Gaussian processes, you may recall that instead of having a vector of means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)$ like a multivariate Gaussian distribution, a Gaussian process has a mean function $\mu(\cdot)$. When parameters are functions like that, they are often called infinite-dimensional parameters. Similarly, instead of having a finite-dimensional vector $\boldsymbol{a} = (a_1, \ldots, a_J)$ like a Dirichlet distribution, a Dirichlet process has an infinite-dimensional parameter, the function $a(\cdot)$. (Sometimes $a(\cdot)$ is factored into a scalar parameter times a probability measure, like $a(\cdot) = \lambda H(\cdot)$.) Similar to how a_j helped capture the probability of $\theta_j = \mathrm{P}(Y = v_j)$ being high relative to the other θ_j, $a(\cdot)$ helps capture the relative probabilities of intervals. For any $J < \infty$, let $-\infty < t_1 < \cdots < t_{J-1} < t_J = \infty$ partition $\mathbb{R}$ into intervals $B_1 = (-\infty, t_1)$ and $B_j = [t_{j-1}, t_j)$ for $j = 2, \ldots, J$. If random probability measure $P(\cdot)$ follows a Dirichlet process with parameter $a(\cdot)$, i.e., if

$$P(\cdot) \sim \mathrm{DP}(a(\cdot)), \tag{14.30}$$

then for any partition with any $J < \infty$, the finite-dimensional vector

$$(P(B_1), \ldots, P(B_J)) \sim \mathrm{Dir}(a(B_1), \ldots, a(B_J)). \tag{14.31}$$

This is analogous to the finite-dimensional marginals of a Gaussian process being multivariate Gaussian.

The Dirichlet process prior is easy to update. The posterior is also a Dirichlet process. After observing value y, you simply add unit probability mass at value y in the Dirichlet process's parameter $a(\cdot)$. With notation $\delta_v(x) = \mathbb{1}\{x = v\}$, the posterior is

$$P(\cdot) \mid \boldsymbol{y} \sim \mathrm{DP}(a(\cdot) + \sum_{i=1}^{n} \delta_{y_i}(\cdot)). \tag{14.32}$$

It is especially easy to use an improper prior. This generalizes the improper Dirichlet prior that took the limit of priors $\mathrm{Dir}(a, a, \ldots, a)$ as $a \downarrow 0$. Here, defining $0(\cdot)$ as the zero function with $0(x) = 0$ for all $x \in \mathbb{R}$, the improper prior takes $a(\cdot) \downarrow 0(\cdot)$. From (14.32), the posterior becomes $\mathrm{DP}(\delta_{y_1}(\cdot) + \cdots + \delta_{y_n}(\cdot))$. Using (14.32), if the y_j are unique, then this is just a finite-dimensional Dirichlet distribution. In fact, it is equivalent to the posterior

from the Dirichlet–multinomial model using $J = n$ with $v_j = y_j$ for all $j = 1, \ldots, n$ and the improper prior. In that case, the parameter vector of interest is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ with $\theta_j = \mathrm{P}(Y = y_j)$, and the posterior is

$$\boldsymbol{P} \mid \boldsymbol{y} \sim \mathrm{Dir}(1, \ldots, 1)$$

as in (14.15). If some y_j values are repeated, then the Dirichlet distribution is adjusted accordingly, adding together the 1s for each repeated value. However, things are trickier in infinite dimensions; the implied prior for the parameter of interest should be verified to also not be informative, as discussed by Chamberlain and Imbens (2003).

Exercises

Exercise E14.1. a. Find and provide a link to a published paper with readily available data, and (approximately) replicate one of its main parameter estimates from a cross-sectional analysis (since we didn't learn about Bayesian bootstrap with dependent data). (Many papers now provide code for replication, too; it may be worth the extra time to find one such paper since it makes this step easier.)

b. Run a Bayesian bootstrap to get a posterior distribution for that same parameter of interest.

c. How does the mean of the posterior compare to the original point estimate?

d. How does the shape of the posterior compare to a normal distribution? (E.g., make a histogram or KDE and just compare visually with a fitted normal distribution. Recall that a formal hypothesis test may fail to reject even if the posterior looks very non-normal but the sample size is small, or it may reject even if the posterior looks very close to normal but the sample size is very large.)

e. How does the posetrior's standard deviation compare to the originally reported standard error?

f. Briefly describe a decision for which the full posterior belief would be more helpful than just a point estimate and CI. (It can be a decision for an individual person, or a firm, or a government, etc.)

g. Submit your code, your results (output/graphs/etc.), and brief qualitative verbal notes on your (re-)analysis, including your answers to the above questions and anything else you find notable.

Part V

Nonparametric Regression

Introduction

This part concerns nonparametric regression. Both kernel and sieve approaches are discussed, as well as model selection. Regression discontinuity is discussed as a popular application.

At a high level, there are three steps for nonparametric regression. First, a family of possible estimated functions must be specified. This is much larger than a single functional form (like quadratic), but it still has a particular structure. This step is especially important with multiple regressors. Second, using the data, the "best" model within the family is chosen ("model selection"); it should be somewhat flexible to avoid bias, but too flexible causes "overfitting" problems. Third, a summary of the CEF estimate is reported. In certain special cases, it may be possible to succinctly describe the full estimated function itself, but often this is not the most efficient way to communicate your results and address your economic research question. Also, even if the CEF is not estimated very precisely, certain summaries may still have small standard errors.

Although not covered, it is straightforward to compute extensions like sieve-type nonparametric instrumental variables and/or quantile regression, and the theory has been established. Model selection is trickier, but there are some suggestions in the literature.

Chapter 15

Nonparametric Methods: Preliminaries

Unit learning objectives for this chapter

15.1. Develop intuition and vocabulary for nonparametric regression [TLO 2]

Optional resources for this chapter

- Textbook: Kaplan (2022b, §8.3) has a very basic intro.

15.1 Motivation

Previously, you learned why the conditional expectation function (CEF) is useful for description, prediction, and causality. The CEF is $m(\boldsymbol{x}) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$.

- Description: the CEF describes a statistical relationship between Y and $\boldsymbol{X}$; i.e., it summarizes the joint distribution of $(Y, \boldsymbol{X})$.
- Prediction: the CEF provides the "best" (under quadratic loss) predictor of Y given $\boldsymbol{X} = \boldsymbol{x}$.
- Causality: under additional identifying assumptions, the CEF is the average structural function, or the CEF partial derivative is a conditional average structural effect.

For example, see Sections 6.3, 6.5, and 10.6.1 of Kaplan (2022b) and Sections 2.5, 2.11, and 2.30 of Hansen (2020a).

The **average structural function** (ASF) is from Blundell and Powell (2003). Essentially, it takes the structural model $Y = h(\boldsymbol{X}, \boldsymbol{U})$, plugs in value $\boldsymbol{X} = \boldsymbol{x}$, then averages

over the unconditional distribution of the unobservable vector $\boldsymbol{U}$:

$$\text{ASF}(\boldsymbol{x}) = \text{E}[h(\boldsymbol{x}, \boldsymbol{U})]. \tag{15.1}$$

The ASF is identified and equals the CEF if $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{U}$:

$$\text{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \text{E}(h(\boldsymbol{X}, \boldsymbol{U}) \mid \boldsymbol{X} = \boldsymbol{x}) = \text{E}(h(\boldsymbol{x}, \boldsymbol{U}) \mid \boldsymbol{X} = \boldsymbol{x}) = \text{E}[h(\boldsymbol{x}, \boldsymbol{U})].$$

The CEF is useful, but previously we only estimated an approximation of it. For example, Chapter 7 of Kaplan (2022b) explains how to interpret what OLS actually estimates: a linear proejction, or "best" linear approximation of the CEF, or "best" linear predictor; see also Chapter 2 of Hansen (2020a). Unfortunately, the "best" linear approximation of the CEF may be a very poor approximation (if the CEF is not approximately linear). For example, imagine a structural model $Y = m(X) + U$, and (lucky us) $U \perp\!\!\!\perp X$, so $m(\cdot)$ is the CEF. But if we estimate the model $Y = \beta_0 + \beta_1 X + U$, then our estimates may be very biased.

Nonparametric regression claims to estimate the true CEF. But, depending on your mood, this is not "really" possible in practice, so these other interpretations are still useful.

15.2 Simple Examples for Intuition

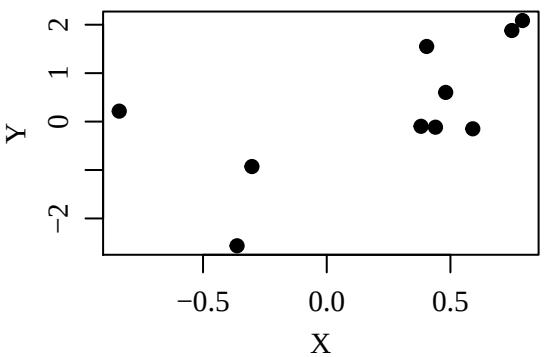


Figure 15.1: A scatterplot, or Rorschach test: what function do you think fits best?

Discussion Question 15.1 (best fit: scatter)**.** Examine the scatterplot in Figure 15.1.

a) Draw what you consider the "best fit" function on the same graph.
b) How do you define "best," either formally or informally? Hint: are we just trying to make a pretty picture, or are we actually trying to learn something (what?) from the data?

Discussion Question 15.2 (best fit: comparison)**.** Consider two CEF estimators: 1) a linear (in X) regression, $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, and 2) a degree $n-1$ polynomial (n is sample size).

a) Come up with one model and dataset where (1) is better.
b) Come up with one model and dataset where (2) is better.
c) What are the important features of the model and dataset that help determine whether CEF estimator (1) or (2) is better?

Figure 15.2 shows a scatterplot of $n = 10$ points with two estimated CEFs. The linear-in-variables estimate doesn't fit any data point exactly, but it looks reasonable. In contrast, the ninth-degree polynomial fits all data points exactly, but it looks unreasonable. The linear functional form may not be flexible enough, but the ninth-degree polynomial here is "too flexible" (overfitting).

Given that even a ninth-degree polynomial doesn't work, if the true CEF is a 20th-degree polynomial, then how can nonparametric regression claim to estimate it? Indeed, it's not magic: we can't estimate the true CEF perfectly in finite samples. Instead, it's about trying to find the optimal amount of "flexibility," which requires admitting we don't know the true CEF's functional form.

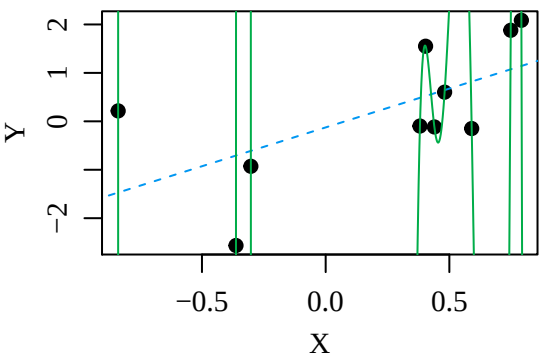


Figure 15.2: An example of overfitting.

Figures 15.3 and 15.4 are inspired by https://xkcd.com/2048. They each show four different CEF estimates given the same dataset.

Discussion Question 15.3 (curve fitting 1)**.** Consider Figure 15.3. Focus on the "partial effect" of X on Y, i.e., the derivative.

a) For the linear, log-linear, and linear-log estimates, say whether each indicates a constant partial effect, increasing partial effect (as X increases), or decreasing partial effect.
b) Could the linear model have estimated an increasing partial effect? Could the log-linear model have estimated a constant partial effect? Is either model "more flexible" than the other? (E.g., is one a special case of the other?)
c) Compare the linear, log-linear, and linear-log estimates. Explain what you can learn about whether the true CEF has constant, increasing, or decreasing partial effect.
d) Overall, decide which CEF estimate looks the "best" to you, and try to explain why you think it looks best (including how you define "best").

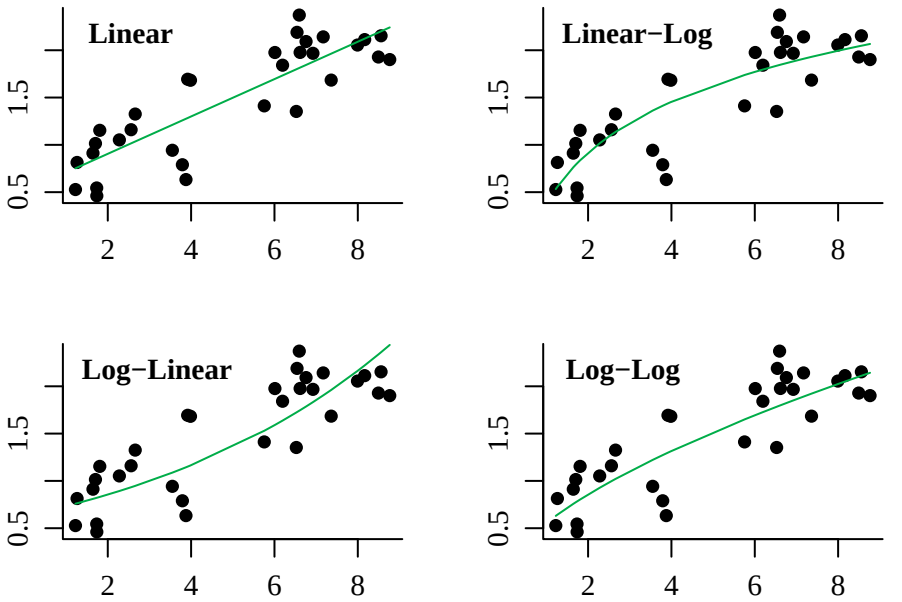


Figure 15.3: Same data, different models.

Discussion Question 15.4 (curve fitting 2)**.** Consider Figure 15.4. Focus on the "partial effect" of X on Y, i.e., the derivative.

a) Qualitatively, describe how the linear and quadratic estimates differ.
b) Consider the linear, quadratic, and cubic models. Is any of these models "more flexible" than any other? (E.g., is one a special case of another?)
c) Compare the four models' estimated partial effects when $X \in [2.8, 3]$.
d) Again for $X \in [2.8, 3]$, which estimate do you think is closest to the true partial effect? Why?
e) Overall, decide which CEF estimate looks the "best" to you, and try to explain why you think it looks best (including how you define "best").

Discussion Question 15.5 (bias–variance tradeoff)**.** Consider Figure 15.5. Focus on $m(0.5)$.

a) Which estimator seems to have larger bias? Why?
b) Which estimator seems to have larger variance? Why?
c) Which aspects of the DGP (sample size, CEF, error term, distribution of X, etc.) could decrease (or increase) the difference in bias? Explain.
d) Which aspects of the DGP could decrease (or increase) the difference in variance? Explain.

The bias–variance tradeoff in Figure 15.5 and DQ 15.5 is one of the central ideas in

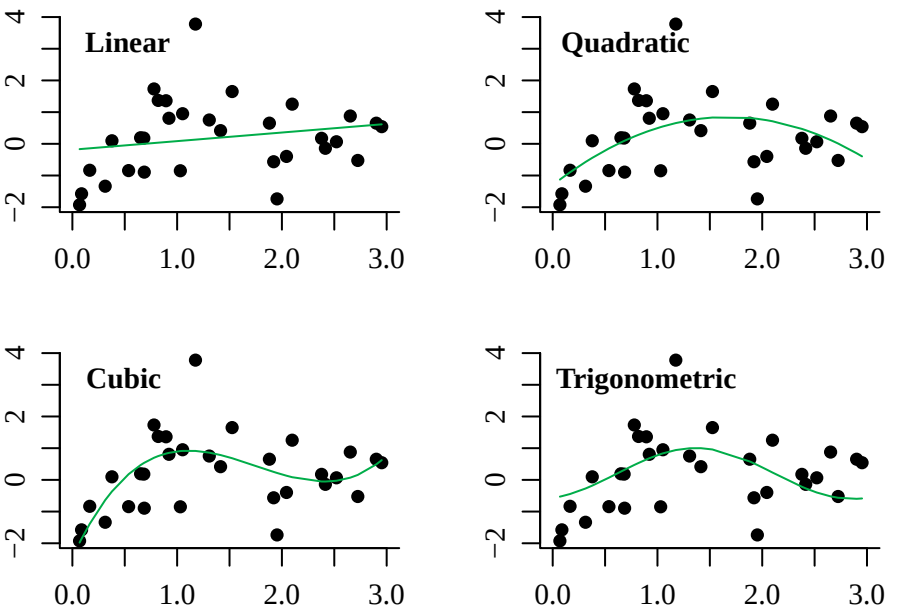


Figure 15.4: Same data, different models.

nonparametric regression and model selection.

15.3 Terminology

Not everyone uses the terms "nonparametric" and "parametric" the same way; I use the definitions in Footnote 1 of Chen (2007), which currently Wikipedia also agrees with.[1]

Definition 15.1 (parametric, nonparametric, semiparametric, semi-nonparametric)**.** A function or model is **parametric** if it is specified up to a finite-dimensional vector of parameters. It is **semiparametric** there are a finite number of parameters of interest (whose values we want to learn), but at least one infinite-dimensional nuisance parameter (whose value we don't care about. It is **nonparametric** if all parameters are infinite-dimensional. It is **semi-nonparametric** if there are both finite-dimensional and infinite-dimensional parameters of interest. Usually **infinite-dimensional parameter** means a function.

Confusion often arises when it's unclear whether a term from Definition 15.1 refers to a "function" or "model." For example, the CEF $m(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$ is parametric because $\boldsymbol{\beta}$ is a finite number of unknown parameters. But the CEF model $Y = \boldsymbol{X}'\boldsymbol{\beta} + U$ with $\mathrm{E}(U \mid \boldsymbol{X}) = 0$ can be called semiparametric because the conditional CDF of U given

[1] http://en.wikipedia.org/wiki/Parametric_model

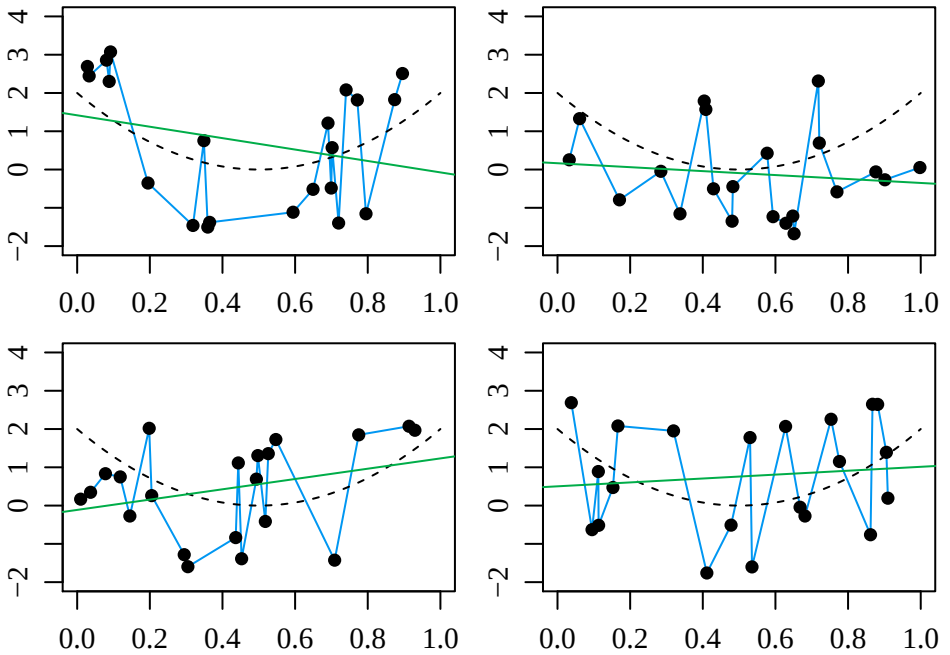


Figure 15.5: Four datasets, linear vs. connect-the-dots; true CEF dashed.

$\boldsymbol{X}$ is an infinite-dimensional nuisance parameter, i.e., an unknown function whose value we don't care about. This "semiparametric regression model" differs from a parametric regression model with a stronger assumption like $U \mid \boldsymbol{X} \sim \mathrm{N}(0, \sigma_U^2)$. So, depending on the setting, people may refer to specifying $\boldsymbol{x}'\boldsymbol{\beta}$ as "parametric estimation" or "semiparametric estimation." The confusion could mostly be avoided by saying "parametric CEF" or "semiparametric regression model."

Parametric models include probit, logit, Poisson regression, and the old "classical linear regression model" (which you may be too young to have ever encountered).

Nonparametric regression does not specify the structure of $m(\cdot)$ up to a finite number of parameters. With scalar X, $m(\cdot)$ is an unknown function. Certain properties of the function may still be specified; e.g., $m(\cdot)$ is twice continuously differentiable, or other "smoothness" properties. With vector $\boldsymbol{X}$, specifying something like $m(x_1, x_2) = g(x_1) + h(x_2)$ adds structure but leaves the model "nonparametric" if $g(\cdot)$ and $h(\cdot)$ are left as unknown functions.

A common semiparametric CEF is $m(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1'\boldsymbol{\beta} + g(\boldsymbol{x}_2)$, where only finite-dimensional $\boldsymbol{\beta}$ is of interest and function $g(\cdot)$ is a nuisance parameter. If $g(\cdot)$ were also of interest, then it would be semi-nonparametric.

Chapter 16
Local (Kernel) Regression

Unit learning objectives for this chapter

16.1. Develop intuition about the bias–variance tradeoff from smoothing and the "local" approach to nonparametric regression [TLO 2]

16.2. Qualitatively, describe how local/kernel regression works and the conditions under which it works well [TLO 1]

This chapter introduces one of the two main approaches to nonparametric regression. Here, X is scalar; for vector $\boldsymbol{X}$, see Chapter 19. Sampling of (Y_i, X_i) is assumed iid unless otherwise stated; this can also be relaxed. The CEF is $m(\cdot)$, where $m(x) \equiv \mathrm{E}(Y \mid X = x)$.

Optional resources for this chapter

- Textbook: Hansen (2020a), Chapter 19 (kernel/local nonparametric regression) and Chapter 23 (model selection).
- Textbook: Pagan and Ullah (1999)
- Textbook: Li and Racine (2007)
- Monograph/book: Racine (2008)
- Textbook: Hastie, Tibshirani, and Friedman (2009) Chapter 7 ("Model Assessment and Selection")
- Bias–variance tradeoff: James et al. (2013, §2.2.2), Hastie, Tibshirani, and Friedman (2009, §§2.9,5.5.2,7.2,7.3)
- R: built-in package `stats` (R Core Team, 2022) has some related functions like `loess` and `ksmooth`, although the latter is not recommended even by its own help file.

- R: package `np` (Hayfield and Racine, 2008) has many kernel methods and a helpful vignette[1]
- R: package `caret` (Kuhn, 2020) helps with model selection.
- R: package `KernSmooth` (Wand, 2019) for its `locpoly` local polynomial regression.
- R: see recommended code in Chapter 5 for nonparametric quantile methods.

16.1 Constant "Regressor"

To build intuition, consider a constant regressor, $X = 1$. The CEF is a single point, and $m(1) = \mathrm{E}(Y)$.

The unconditional mean can be estimated "nonparametrically"[2] by the sample mean, $\bar{Y}$, so $\hat{m}(1) = \bar{Y}$. To match later notation,

$$\hat{m}(1) = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} \mathbb{1}\{X_i = 1\}}, \tag{16.1}$$

where the denominator equals n because $X_i = 1$ for $i = 1, \ldots, n$.

16.2 Binary Regressor

Let $X \in \{0, 1\}$. The CEF consists of $m(0)$ and $m(1)$. Here, $m(0) = \mathrm{E}(Y \mid X = 0)$ is the mean Y value in the $X = 0$ subpopulation, and $m(1) = \mathrm{E}(Y \mid X = 1)$ is the mean Y in the $X = 1$ subpopulation.

16.2.1 Estimation

Just as the population mean can be estimated by the sample mean, the subpopulation means can be estimated by the subsample means. Extending (16.1),

$$\hat{m}(0) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i = 0\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i = 0\}}, \quad \hat{m}(1) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i = 1\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i = 1\}}, \tag{16.2}$$

16.2.2 Bias–Variance Tradeoff

The following appears frivolous now, but it plants an important seed.

Imagine you compute $\hat{m}(0)$ but are sad about a large standard error. You remember $\hat{m}(0)$ only uses the $X_i = 0$ subsample of the data and wonder if you can reduce the

[2] Although people say "nonparametrically," I think it should be "semiparametrically" because there is only one parameter of interest, $m(1)$, with the CDF of Y being an infinite-dimensional nuisance parameter.

standard error by using the full sample average $\bar{Y}$ as an estimate of $m(0)$. But then you worry this is biased, unlike the subsample average.

This shows the **bias–variance tradeoff**. Including observations with $X_i \neq 0$ introduces bias but may reduce variance (squared standard error). The effect on mean squared error (MSE, variance plus squared bias) could go either way. If the bias is negligible because $m(1) \approx m(0)$, then the variance reduction can dominate and decrease MSE, in which case $\bar{Y}$ is a better (lower MSE) estimator of $m(0)$ than the $X_i = 0$ subsample average. However, if $m(1)$ is far from $m(0)$, then the bias may dominate, so the subsample average has lower MSE.

Two additional ideas reappear later. First, to compare the estimators' MSE, we need to know $m(1)$ and $m(0)$, but that's what we want to estimate to begin with. Second, increased model flexibility reduces bias but can increase variance.

16.2.3 Binary Regressor: Small Probability

Discussion Question 16.1 (small probability of conditioning event 1)**.** Let $Y_i = 1$ if individual i is employed and $Y_i = 0$ if not. Let $X_i = 1$ if individual i has a college degree and $X_i = 0$ if not. Consider the estimator $\hat{m}(1)$ in (16.2). Let $N_1 \equiv \sum_{i=1}^{n} \mathbb{1}\{X_i = 1\}$. Let $p_x \equiv \mathrm{P}(X = x) > 0$ for $x = 0, 1$. Assume (Y_i, X_i) are sampled iid.

a) Let $n = 2$ and $N_1 = 1$. What are the possible values of $\hat{m}(1)$?
b) Let $n = 10$ and $N_1 = 1$. What are the possible values of $\hat{m}(1)$?
c) Let $n \to \infty$, but still with $N_1 = 1$. In the limit, what are the possible values of $\hat{m}(1)$?
d) In terms of n and p_1, what's the mean of the distribution of N_1, i.e., what's $\mathrm{E}(N_1)$?
e) If p_1 is fixed as $n \to \infty$, then explain why it's impossible to have $\mathrm{E}(N_1) \to 1$ as $n \to \infty$.
f) Allow p_1 to change with n, so it is a sequence p_{1n} for $n = 1, 2, \ldots$. Then, explain how it's possible to have $\mathrm{E}(N_1) \to 1$ as $n \to \infty$.

In DQ 16.1, N_1 is a **local sample size** or **effective sample size**. Even though there are n observations, only N_1 are used for $\hat{m}(1)$. The tradeoff in Section 16.2.2 is essentially: smaller local sample size avoids bias but can increase variance.

16.3 Discrete Regressor

The ideas of Section 16.2 readily extend to discrete X with more than two possible values.

16.3.1 Estimation

Let x denote any possible value, with $\mathrm{P}(X = x) > 0$. Then,

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i = x\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i = x\}} \xrightarrow{p} m(x). \tag{16.3}$$

Like before, (16.3) is a subsample average, averaging the Y_i for observations with $X_i = x$.

Discussion Question 16.2 (discrete local regression)**.** You observe $n = 100$ observations of wage Y_i and age X_i, from a survey of 18- to 65-year-olds. Is (16.3) a good estimator? Why/not?

16.3.2 Local Sample Size

Consider the asymptotic performance of $\hat{m}(x)$ in (16.3) with J possible values of X. For simplicity, imagine stratified sampling with n/J observations of $X_i = x$ for each possible x, so $\hat{m}(x)$ is the average of n/J iid random variables (the corresponding Y_i). Although $n/J < n$, the local sample size n/J grows proportionally to n (as $n \to \infty$) because J is a fixed constant. So the convergence rate of $\hat{m}(x)$ remains the same as for the unconditional sample mean $\bar{Y}$.

Discussion Question 16.3 (small probability of conditioning event 2)**.** Let $Y_i = 1$ if individual i is employed and $Y_i = 0$ if not. Let X_i be the individual's total consumption (expenditure) over the past year rounded to the nearest dollar. Consider the estimator $\hat{m}(x)$ in (16.3). Let $N_x \equiv \sum_{i=1}^{n} \mathbb{1}\{X_i = x\}$ denote the local sample size. Let $p_x \equiv \mathrm{P}(X = x) > 0$. Assume (Y_i, X_i) are sampled iid. Hint: DQ 16.1 was similar.

a) Let $n = 2$ and $N_x = 1$. What are the possible values of $\hat{m}(x)$?
b) Let $n = 10$ and $N_x = 1$. What are the possible values of $\hat{m}(x)$?
c) Let $n \to \infty$, but still with $N_x = 1$. In the limit, what are the possible values of $\hat{m}(x)$?
d) In terms of n and p_x, what's the mean of the distribution of N_x, i.e., what's $\mathrm{E}(N_x)$?
e) If p_x is fixed as $n \to \infty$, then explain why $\lim_{n\to\infty} \mathrm{E}(N_x) = \infty$.
f) Allow p_x to change with n, so it is a sequence p_{xn} for $n = 1, 2, \ldots$. Then, explain how it's possible to have $\mathrm{E}(N_x) \to 1$ as $n \to \infty$.

Discussion Question 16.4 (local sample size rate)**.** Continue DQ 16.3. Let J_n be the number of possible values of X, which is allowed to change with n.

a) Let $m_n = \min_x N_x$, the smallest local sample size (given a particular dataset). Given J_n, explain why the largest possible value of m_n is $\lfloor n/J_n \rfloor$. (So in the best case scenario, all local sample sizes have at least m_n observations.)
b) If $J_n = J$, a fixed constant as $n \to \infty$, then how does m_n change as $n \to \infty$? Specifically, if $m_n \propto n^r$, what's r?
c) Similar to (b): if $J_n = n$, then what's m_n and r?
d) Similar to (b): if $J_n = n^{1/5}$, then what's m_n and r?

Having $J_n \to \infty$ in DQ 16.4 does not mean literally there are more and more possible values, just as $n \to \infty$ does not mean literally we are collecting more and more data. Both are mathematical approximations to help us better understand finite-sample performance. If you have $n = 100$ and $J = 88$, then $J_n/n \to 0$ is probably a bad asymptotic approximation.

16.3.3 Bias–Variance Tradeoff

There is again a tradeoff like in Section 16.2.2. In the extreme, pooling all the data makes the "local" sample size n, reducing variance but increasing bias.

As a compromise, we could only pool similar values of x. For example, if the possible values are $X = 1, 2, 3, \ldots, J$, we could pool pairs of values together: $\{1, 2\}$, $\{3, 4\}$, etc. Modifying (16.3), for odd x,

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i \in \{x, x+1\}\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i \in \{x, x+1\}\}} \xrightarrow{p} \mathrm{E}(Y \mid X \in \{x, x+1\}). \tag{16.4}$$

Intuitively, this should work well if $m(x) \approx m(x+1)$, so the increase in squared bias is smaller than the decrease in variance.

Discussion Question 16.5 (discrete CEF bias and variance)**.** Let $X \in \{1, 2, 3, \ldots, J\}$. Let $m(x) = bx$ for some constant b. Assume stratified sampling with n/J observations for each possible $X_i = x$. Assume $\mathrm{Var}(Y \mid X = x) = \sigma^2$, a constant (unrelated to x). Consider the estimator

$$\hat{m}_h(1) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{1 \le X_i \le h\}}{\sum_{i=1}^{n} \mathbb{1}\{1 \le X_i \le h\}}.$$

for some integer $h \ge 1$. The first parts concern the bias.

a) Explain why the bias is zero if $h = 1$, i.e., why $\mathrm{E}[\hat{m}_1(1)] = m(1) = b$. (Recall that with stratified sampling, we have n/J iid draws of Y_i from each subpopulation corresponding to each possible x.)
b) Show why the bias is $b/2$ when $h = 2$, i.e., $\mathrm{E}[\hat{m}_2(1)] - m(1) = 3b/2 - b = b/2$.
c) For general h, show the bias is $\mathrm{E}[\hat{m}_h(1)] - m(1) = b(h-1)/2$.

The next parts concern the variance $\mathrm{Var}[\hat{m}_h(1)] = \sum_{j=1}^{h} (1/h)^2 \sigma^2/(n/J)$, where $\sigma^2/(n/J)$ is the variance of a single subsample average with n/J observations.

d) Without using the formula, explain why the variance with $h = 1$ is $\mathrm{Var}[\hat{m}_1(1)] = \sigma^2/(n/J)$.
e) Using the formula, explain why the variance with $h = 2$ is $\mathrm{Var}[\hat{m}_2(1)] = \sigma^2/(2n/J)$.
f) Using the formula, explain why the variance as a function of h is $\mathrm{Var}[\hat{m}_h(1)] = (1/h)[\sigma^2/(n/J)]$.

Discussion Question 16.6 (discrete CEF MSE)**.** Continue from DQ 16.5. The next parts concern the bias–variance tradeoff. Recall MSE equals variance plus squared bias. The "MSE-optimal" estimator has the lowest MSE.

a) Let $n = 10$, $\sigma^2 = 10$, $J = 2$, $b = 1$. Which h is MSE-optimal? That is, which estimator has lower MSE, $\hat{m}_1(1)$ or $\hat{m}_2(1)$?
b) Now let $n = 100$, but still with $\sigma^2 = 10$, $J = 2$, $b = 1$. Which h is MSE-optimal?

Discussion Question 16.7 (discrete CEF tradeoff)**.** Continue from DQ 16.6. If h were real-valued instead of integer-valued, then the first-order condition (FOC) for MSE-minimization (over h) would be $0 = hb/2 - b^2/2 - h^{-2}\sigma^2/(n/J)$. That is, setting zero equal to the derivative of MSE (variance plus squared bias) with respect to h. Ignoring

the $b^2/2$ term, the solution to the FOC is $h^3 = (2/b)\sigma^2/(n/J)$, or $h = n^{-1/3}(2J\sigma^2/b)^{1/3}$. The next parts consider the relationship between h and the other variables. For each of the following, say whether h is increasing or decreasing in the variable, and explain why that does (or doesn't) make intuitive sense to you. The first part shows an example response to help inspire you for the next parts.

a) n: h is decreasing in n. This makes sense because larger n reduces all the subsample average variances, which reduces the incentive to pool data; e.g., if $\hat{m}_1(1)$ already has a very small variance, then there is very little reason to add bias for the purpose of reducing variance.
b) J
c) σ^2
d) b

16.4 Continuous Regressor: Introduction

Consider the "subsample average" approach when X is continuous. Now $\mathrm{P}(X = x) = 0$: the probability of even one observation with $X_i = x$ is zero, so we can't use the $X_i = x$ subsample. Because our subsample must have $X_i \neq x$, bias is unavoidable. There is still a bias–variance tradeoff: including more X_i in a subsample increases bias but decreases variance.

There are three main approaches to defining the subsamples. Each forms a branch within the **local approach** to nonparametric regression: the partitioning, kernel (local polynomial), and k-nearest neighbor approaches.

First, we could partition the range of X into mutually exclusive **bins**. For example, if X is age in (decimal) years, we can make bins for each integer year: somebody age $X = 24.1$ goes in the 24-year-old subsample, as does somebody age $X = 24.8$, or anyone with $X \in [24, 25)$. The 24-year-old subsample average is then our estimated $\hat{m}(x)$ for any $24 \leq x < 25$. Or the bins can be larger, like $[20, 25)$, $[25, 30)$, etc. This is the foundation for **partitioning estimators**; e.g., see Cattaneo and Farrell (2013).

Second, we could use a bin centered at x to estimate $m(x)$. For example, let $h > 0$ be the bin width, called the **bandwidth**. Given x, the bin is $[x - h/2, x + h/2]$. Then, the estimator is

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i \in [x - h/2, x + h/2]\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i \in [x - h/2, x + h/2]\}}, \tag{16.5}$$

where the denominator is the local sample size (which depends on the dataset). This is a crude example of **kernel regression** or **local polynomial regression**.

Third, instead of a bin, we can use the k observations with X_i closest to x, i.e., the smallest $|X_i - x|$. Like before, the corresponding Y_i values are averaged to get $\hat{m}(x)$. This is the core of the **k-nearest neighbor** (kNN) approach.

16.5 Local Constant Regression

For this section (and the rest of the chapter), let X denote a continuous random variable, with x a possible value. As before, Y is the outcome (either discrete or continuous), (Y_i, X_i) are sampled iid, and $m(x) = \mathrm{E}(Y \mid X = x)$.

In practice, the local constant estimator should not be used (see Section 16.6), but it helps develop intuition.

Following (16.5), consider the **local constant** regression estimator with bandwidth $h > 0$:

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}\{x - h/2 \le X_i \le x + h/2\}}{\sum_{i=1}^n \mathbb{1}\{x - h/2 \le X_i \le x + h/2\}}. \tag{16.6}$$

It averages the Y_i for observations whose X_i is close to (within $h/2$ of) x.

The bandwidth h affects both bias and variance. Larger h decreases variance by increasing the local sample size. However, larger h increases bias by including X_i farther from x.

Discussion Question 16.8 (local constant local sample size)**.** Let $X_i \overset{iid}{\sim} \mathrm{Unif}(0, 1)$. Let $N_x = \sum_{i=1}^n \mathbb{1}\{x - h/2 \le X_i \le x + h/2\}$. Hint: $\mathrm{E}[\mathbb{1}\{A\}] = \mathrm{P}(A)$.

a) What's $\mathrm{E}(N_x)$ if $n = 100$, $x = 0.5$, $h = 0.4$?
b) What's $\mathrm{E}(N_x)$ if $n = 100$, $x = 0.5$, $h = 0.2$?
c) What's $\mathrm{E}(N_x)$ as a function of n, x, and h?

The follow DQs consider models with no error terms to focus on the bias.

Discussion Question 16.9 (local constant flexibility)**.** Consider the local constant regression estimator in (16.6). Let $Y_i = \sin(X_i)$ with no error term, so $m(x) = \sin(x)$, $0 \le x \le 2\pi$. Let $n = 101$. Let $X_i = 2\pi(i - 1)/(n - 1)$, $i = 1, \ldots, n$. Consider points of evaluation $x_j = j\pi/2$, $j = 1, 2, 3$, so $m(x_1) = 1$, $m(x_2) = 0$, $m(x_3) = -1$. Hint: draw a picture.

a) Let $h = 4\pi$. Explain why $\hat{m}_h(x_1) = \hat{m}_h(x_2) = \hat{m}_h(x_3) = 0$.
b) Let $h = 2\pi$. Explain why $0 < \hat{m}_h(x_1) < m(x_1)$, $\hat{m}_h(x_2) = m(x_2)$, and $0 > \hat{m}_h(x_3) > m(x_3)$.
c) Let $h = \pi$. For each $j = 1, 2, 3$, explain whether $\hat{m}_\pi(x_j)$ is closer to, farther from, or equally far from $m(x_j)$ compared to $\hat{m}_{2\pi}(x_j)$.
d) For each $j = 1, 2, 3$, explain how $\hat{m}_h(x_j)$ continues to change (or not) as h continues to decrease toward zero.
e) Do any of your answers change if instead of evenly spaced X_i we take randomly sampled $X_i \overset{iid}{\sim} \mathrm{Unif}(0, 2\pi)$? Why/not?
f) Qualitatively, generally: does the estimator $\hat{m}_h(x)$ become more or less flexible as $h \downarrow 0$? Why?

Discussion Question 16.10 (local constant boundary)**.** Consider the same setup of DQ 16.9 but with evaluation points $x_1 = 0$ and $x_2 = 2\pi$, which are **boundary points**. Note $m(x_1) = m(x_2) = 0$. Hint: draw a picture.

a) Let $h = 2\pi$. Explain why $\hat{m}_h(0) > m(0)$ and $\hat{m}_h(2\pi) < m(2\pi)$.
b) Let $h = \pi$. Explain why $\hat{m}_h(0) > m(0)$ and $\hat{m}_h(2\pi) < m(2\pi)$.
c) How do $\hat{m}_h(0)$ and $\hat{m}_h(2\pi)$ change as $h \downarrow 0$?
d) Recall from DQ 16.9 that $\hat{m}_h(\pi) = m(\pi) = 0$ for any bandwidth h. Why was it so different at $x = \pi$ than at $x = 0$ or $x = 2\pi$?
e) Do any of your answers change if instead of evenly spaced X_i we take randomly sampled $X_i \overset{iid}{\sim} \text{Unif}(0, 2\pi)$? Why/not?

Discussion Question 16.11 (local constant smoothness 1)**.** Consider again the estimator in (16.6). Let $X_i \overset{iid}{\sim} \text{Unif}(0, 1)$. Let $x_0 = 0.5$, small $\epsilon > 0$. Consider $Y_i = \mathbb{1}\{x_0 - \epsilon \leq X_i \leq x_0 + \epsilon\}$ with no error term, so $m(x) = \mathbb{1}\{x_0 - \epsilon \leq x \leq x_0 + \epsilon\}$ is the true CEF. Hint: draw a picture.

a) Explain why the true $m(x_0) = 1$.
b) Let $h \geq 1$. Given n and ϵ, what's the probability of sampling a dataset with $\hat{m}_h(x_0) = 0$? Hint: for a single i, compute the probability that X_i is such that $Y_i = 0$; then use independence to compute the joint probability for all $i = 1, \ldots, n$.
c) When $\hat{m}_h(x_0) = 0$ with $h \geq 1$, can using a smaller h help? Why/not?
d) Let $h \geq 1$. Given n and ϵ, what's the probability of sampling a dataset with $\hat{m}_h(0.5) = 1$?
e) Let $n = 10$ and $\epsilon = 0.01$; will $\hat{m}_h(0.5)$ be reasonable?
f) Let $n = 1000$ and $\epsilon = 0.01$; will $\hat{m}_h(0.5)$ be reasonable?

Discussion Question 16.12 (local constant smoothness 2)**.** Let $X_i \overset{iid}{\sim} \text{Unif}(-1, 1)$, $Y_i = |X_i|$, so $m(x) = |x|$. Let x_0 be the point of interest, so $m(x_0)$ is the object of interest, estimated by $\hat{m}_h(x_0)$ as in (16.6). Hint: draw a picture.

a) Let $x_0 = 0.1$. Explain why $\hat{m}_h(x_0)$ is biased if $h = 2$.
b) Let $x_0 = 0.1$. Explain why $\hat{m}_h(x_0)$ is not biased if $h = 0.1$.
c) Let $x_0 = 0$. Explain why $\hat{m}_h(x_0)$ is biased if $h = 2$.
d) Let $x_0 = 0$. Is $\hat{m}_h(x_0)$ biased if $h = 0.1$? Why/not?
e) Let $x_0 = 0.01$. Explain why $\hat{m}_h(x_0)$ is biased if $h = 2$.
f) Let $x_0 = 0.01$. Is $\hat{m}_h(x_0)$ biased if $h = 0.1$? Why/not?

Discussion Question 16.13 (local constant bandwidth 1)**.** Consider Figure 16.1.

a) The four bandwidths used were $h = 0.032, 0.1, 0.7, 2$. Explain which graph you think corresponds to each h.
b) Which of the four estimators looks "best" to you?
c) How are you defining "best"?
d) Are there other types of "best" we may care about?

Some formal assumptions and a theorem are now given.

Assumption A16.1 (iid)**.** Sampling of (Y_i, X_i) is iid.

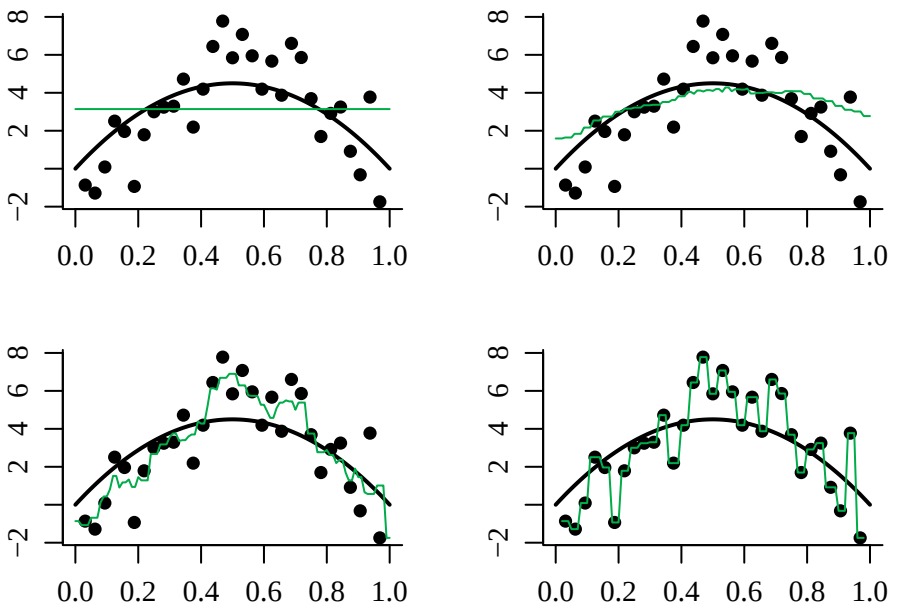


Figure 16.1: Comparison of local constant CEF estimator with four different h, same dataset. Thick black line is true CEF.

Assumption A16.2 (finite variance)**.** For all x in the support of X, $\mathrm{Var}(Y \mid X = x) < \infty$. Slightly different: $\mathrm{E}(|Y|^{2+\delta} \mid X) < \infty$ almost surely for some (small) $\delta > 0$. ("Almost surely" meaning the expectation can be infinite for some values of X, but the set of such values is probability zero.)

Assumption A16.3 (smoothness)**.** The CEF $m(\cdot)$ has two continuous derivatives in a neighborhood of the point of interest x.

Assumption A16.4 (interior point)**.** Point of interest x is in the interior of the support of X, i.e., not a boundary point.

Assumption A16.5 (bandwidth)**.** As $n \to \infty$, $h \downarrow 0$ and $nh \to \infty$. More specifically, $nh^5 \to M \in [0, \infty)$.

Theorem 16.1 (local constant asymptotics)**.** *Consider a given point x, with interest in $m(x) = \mathrm{E}(Y \mid X = x)$. Let A16.1–A16.5 hold. Then,*

$$\begin{aligned}\sqrt{nh}[\hat{m}_h(x) - m(x) - h^2 B(x)] &\xrightarrow{d} \mathrm{N}(0, V(x)),\\ B(x) &= (1/24)[m''(x) + 2m'(x)f_X'(x)/f_X(x)],\\ V(x) &= \mathrm{Var}(Y \mid X = x)/f_X(x),\end{aligned}$$

where $\hat{m}_h(x)$ is the local constant regression estimator in (16.6), $f_X(\cdot)$ is the PDF of X, and $'$ and $''$ indicate first and second derivatives.

The use of A16.3 was seen in DQs 16.11 and 16.12. If the CEF has a jump discontinuity, then the bias can be severe, even with large n. If the CEF is continuous but not differentiable at x, then the bias becomes proportional to h instead of the smaller h^2. (Recall $h \downarrow 0$, so $h^2 < h$ and $h^2 \downarrow 0$ faster than h.) Asymptotically, the theory only requires such smoothness in a neighborhood of x, but in practice, the estimator can still be poor even if the assumption is technically satisfied. For example, the assumption is technically satisfied in DQ 16.11 because $\epsilon > 0$, but the estimator is very bad with small n.

The use of A16.4 was seen in DQ 16.10. Like a non-differentiable (but continuous) CEF, having a boundary point makes the bias larger, but the bias still goes to zero as $h \downarrow 0$. Asymptotically, everything is in an infinitesimal neighborhood of x, so theoretically x can be infinitesimally close to a boundary, but in practice what matters is if x is within h of the boundary.

The presence of $m''(x)$ in the bias was seen in DQ 16.9 for $x = \pi/2$ and $x = 3\pi/2$. There, $m''(\pi/2) < 0$ caused negative bias, while $m''(3\pi/2) > 0$ caused positive bias.

The $m'(x)f'_X(x)/f_X(x)$ part of the bias can actually be removed by a more sophisticated estimator; see Section 16.6. Because of this unnecessary bias, the local constant estimator should not be used in practice.

What does Theorem 16.1 say about the bias–variance tradeoff? As an asymptotic approximation, the variance is proportional to $1/(nh)$, and the bias is proportional to h^2. The variance derives from the convergence rate, similar to how the usual $\sqrt{n}$ rate means variance proportional to $1/n$, or equivalently standard errors proportional to $1/\sqrt{n}$. That is, given asymptotic approximation $\sqrt{nh}(\hat{m}_h(x) - m(x) - h^2 B(x)) \overset{a}{\sim} \mathrm{N}(0, V(x))$, then

$$\hat{m}_h(x) - m(x) - h^2 B(x) \overset{a}{\sim} \mathrm{N}(0, V(x)/(nh)). \tag{16.7}$$

Rearranging further,

$$\hat{m}_h(x) \overset{a}{\sim} \mathrm{N}(m(x) + h^2 B(x), V(x)/(nh)). \tag{16.8}$$

That is, the sampling distribution of $\hat{m}_h(x)$ is approximately normal with mean $m(x) + h^2 B(x)$ and variance $V(x)/(nh)$.

From (16.8), the corresponding asymptotic MSE (AMSE) can be calculated. With variance is $V(x)/(nh)$ and bias $h^2 B(x)$, the AMSE is

$$\mathrm{AMSE}_x(h) = V(x)/(nh) + h^4 [B(x)]^2. \tag{16.9}$$

Using (16.9), given $V(x) > 0$ and $B(x) > 0$, some bandwidth h_* with $0 < h_* < \infty$ minimizes AMSE. This is true because the variance grows to infinity as $h \downarrow 0$, while the squared bias grows to infinity as $h \to \infty$. Writing the AMSE-optimal bandwidth as h_*,

the first-order condition (FOC) is

$$\begin{aligned} 0 = \frac{\partial}{\partial h} \text{AMSE}_x(h)\bigg|_{h=h_*} &= -h_*^{-2} V(x)/n + 4h_*^3 [B(x)]^2 \\ \implies h_*^{-2} V(x)/n = 4h_*^3 [B(x)]^2 &\implies h_*^{-5} = 4n[B(x)]^2/V(x) \\ \implies h_* = n^{-1/5} \left(\frac{V(x)}{4[B(x)]^2} \right)^{1/5} &. \end{aligned} \tag{16.10}$$

Discussion Question 16.14 (local constant bandwidth 2)**.** Consider (16.10). For each of the following, explain whether h_* is increasing or decreasing in the variable, and try to explain the intuition for why. Hint: DQ 16.7 was qualitatively similar.

a) n
b) $V(x)$
c) $B(x)$

Discussion Question 16.15 (local constant bandwidth 3)**.** Equation (16.10) may help here. Explain both mathematically and intuitively.

a) I thought bias was bad, and unbiased estimators were good. Wasn't the BLUE property of OLS really important, where U stands for "unbiased"? So why don't we pick h to make $\hat{m}_h(x)$ unbiased, or at least asymptotically unbiased?
b) I thought big standard errors were bad. Here, the standard error is proportional to $1/\sqrt{nh}$. It seems like making h really close to zero is a bad idea because that makes the standard errors really big. Why don't we just use a fixed bandwidth like $h = 1$ so we can have standard errors proportional to $1/\sqrt{n}$ like with OLS?

The terms **undersmoothing** and **oversmoothing** are relative to the AMSE-optimal bandwidth rate, like $h \propto n^{-1/5}$ in (16.10). "Undersmoothing" means smoothing less, which means smaller h. Given (16.10), $h \propto n^{-r}$ for $r > 1/5$ is undersmoothing. "Oversmoothing" is the opposite: more smoothing than the AMSE-optimal amount. Here, $h \propto n^{-r}$ for $r < 1/5$ is oversmoothing. It can be confusing because the exponent is negative, and because larger h means less flexibility.

Discussion Question 16.16 (local constant CI)**.** Use Theorem 16.1. Consider a confidence interval for $m(x)$. For simplicity, let $V(x) = 1$ and $h = n^{-1/5}$. This is still the AMSE-optimal bandwidth rate, although the constant in (16.10) is omitted. If the bias is ignored, then the conventional 95% CI is roughly $\hat{m}_h(x) \pm 2n^{-2/5}$.

a) Show why this CI is reasonable if $B(x) = 0$. Hint: draw a normal PDF using (16.8), with the horizontal axis in units of standard deviations (standard errors).
b) Approximate this CI's asymptotic coverage probability if $B(x) = 2$. Hint: draw a picture of a normal PDF, with the horizontal axis labeled in units of standard errors (written as a power of n); then draw another normal PDF shifted by the bias.
c) Try to reconcile the following paradox: if $h \downarrow 0$ is required to make the bias disappear asymptotically, then why does the effect of the bias on this CI still seem to remain important even asymptotically, even though $h = n^{-1/5} \to 0$?

d) Would oversmoothing or undersmoothing fix the problem? Why? Hint: how can you get the bias to go to zero at a faster rate than the standard error?

16.6 Local Linear Regression

The idea of **local linear regression** is to regress Y on $(1, X)$ in the local sample. Then, $\hat{m}(x) = \hat{\beta}_0 + x\hat{\beta}_1$.

Discussion Question 16.17 (local linear boundary bias)**.** Consider $Y_i = X_i$, $X_i \overset{iid}{\sim} \text{Unif}(0, 1)$. Hint: draw a picture.

a) For OLS regression of Y_i on X_i (including an intercept), what's the (approximate) bias of $\hat{m}(0) = \hat{\beta}_0$?
b) For the same OLS regression, what's the (approximate) bias of $\hat{m}(1) = \hat{\beta}_0 + \hat{\beta}_1$?
c) For the local constant estimator with bandwidth h, what's the (approximate) bias of $\hat{m}_h(0)$?
d) For the local constant estimator with bandwidth h, what's the (approximate) bias of $\hat{m}_h(1)$?
e) For the local linear estimator with bandwidth h, what's the (approximate) bias of $\hat{m}_h(0)$?
f) For the local linear estimator with bandwidth h, what's the (approximate) bias of $\hat{m}_h(1)$?

Local linear regression is always better than local constant regression. This seems counterintuitive because it seems like the "more flexible" local linear estimator should have smaller bias but larger variance. However, theoretical results show the asymptotic variance is actually the same, even though the bias is indeed "smaller" (see below).

Specifically, instead of the local constant bias $h^2[m''(x)+2m'(x)f'_X(x)/f_X(x)]/24$, the local linear bias is $h^2m''(x)/24$. Technically, the local linear bias is not always closer to zero because the additional terms in the local constant bias could cancel out the $m''(x)$ term. But, most people (like me) prefer to have only one source of bias instead of two, so many people just say the local linear bias is smaller (without quotes). Specifically, the local linear estimator removes the **design bias**, i.e., the term involving $f_X(\cdot)$.

To formally write the local linear estimator is somewhat complicated, even though the intuition is relatively simple. First, for comparison, the local constant estimator in (16.6) is equivalent to

$$\hat{m}_h(x) = \hat{\beta}_0(x), \quad \hat{\beta}_0(x) = \arg\min_{b_0} \sum_{i=1}^{n} \mathbb{1}\{x - h/2 \leq X_i \leq x + h/2\}(Y_i - b_0)^2. \quad (16.11)$$

For local linear regression, the regressor X_i is centered at x, and the "residual" changes

from $Y_i - b_0$ to $Y_i - b_0 - b_1(X_i - x)$:

$$\begin{aligned}(\hat{\beta}_0(x), \hat{\beta}_1(x)) &= \underset{(b_0,b_1)}{\arg\min} \sum_{i=1}^{n} \mathbb{1}\{x - h/2 \le X_i \le x + h/2\}(Y_i - b_0 - b_1(X_i - x))^2, \\ \hat{m}_h(x) &= \hat{\beta}_0(x), \quad \hat{m}_h'(x) = \hat{\beta}_1(x).\end{aligned} \tag{16.12}$$

Sometimes $(X_i - x)/h$ is used instead of $X_i - x$, but it only affects the scaling of $\hat{\beta}_1$.

As seen in (16.12), another advantage of the local linear estimator is easy estimation of the CEF derivative $m'(x)$. This is another advantage over local constant estimation.

16.7 Local Polynomial Regression

Including local linear as a special case, **local polynomial regression** runs a polynomial regression on the local sample. However, local linear remains the most common in practice because the bandwidth already controls flexibility (the bias–variance tradeoff), so increasing the polynomial order to increase flexibility is not necessary. Sometimes local cubic regression is used; analogous to the advantage of local linear over local constant, local cubic has "smaller" asymptotic bias than local quadratic, while maintaining the same asymptotic variance.

16.8 Kernel Regression

The idea of **kernel regression** is to run weighted least squares with weight W_i based on $|X_i - x|$. Local regression is a special case with $W_i = \mathbb{1}\{|X_i - x| \le h/2\}$. The **kernel function** describes the "shape" of the weight as a decreasing function of $|X_i - x|$, while the bandwidth describes how fast the weight goes to zero as X_i deviates from x.

Practically, the weighting scheme is not as crucially important as the bandwidth; you would be fine to just use the "Epanechnikov kernel" and not worry about it.

16.8.1 Local Linear Regression: Uniform Kernel

The local linear estimator in (16.12) can be written in terms of the **uniform kernel**

$$K(u) = \mathbb{1}\{-1/2 \le u \le 1/2\}. \tag{16.13}$$

As $K(u/h) = \mathbb{1}\{-1/2 \le u/h \le 1/2\} = \mathbb{1}\{-h/2 \le u \le h/2\}$, (16.12) becomes

$$\begin{aligned}(\hat{\beta}_0(x), \hat{\beta}_1(x)) &= \underset{(b_0,b_1)}{\arg\min} \sum_{i=1}^{n} \mathbb{1}\{x - h/2 \le X_i \le x + h/2\}(Y_i - b_0 - b_1(X_i - x))^2 \\ &= \underset{(b_0,b_1)}{\arg\min} \sum_{i=1}^{n} K((X_i - x)/h)(Y_i - b_0 - b_1(X_i - x))^2.\end{aligned} \tag{16.14}$$

The uniform kernel $K(\cdot)$ in (16.14) can be replaced by any other kernel.

16.8.2 Other Second-Order Kernels

There are many, many possible kernels; some are more commonly used. The following are symmetric ($K(-u) = K(u)$) **second-order kernels** because with $r = 2$ they satisfy

$$\begin{aligned} 1 &= \int_{\mathbb{R}} K(u)\,\mathrm{d}u, \\ 0 &= \int_{\mathbb{R}} u^j K(u)\,\mathrm{d}u \text{ for all } j = 1, \ldots, r-1, \\ 0 &< \mu_r \equiv \int_{\mathbb{R}} u^r K(u)\,\mathrm{d}u < \infty. \end{aligned} \tag{16.15}$$

Most second-order kernels also satisfy $K(u) \geq 0$ for all $u \in \mathbb{R}$, so $K(\cdot)$ is a PDF.

The **Epanechnikov kernel** is the AMSE-optimal choice and makes you sound fancy:

$$K(u) = \mathbb{1}\{|u| \leq 1\}(3/4)(1 - u^2). \tag{16.16}$$

The **triangle kernel**, also known as the **tent kernel** or **Bartlett kernel**, is implicitly used in the Newey and West (1987) long-run variance estimator:

$$K(u) = \mathbb{1}\{|u| \leq 1\}(1 - |u|). \tag{16.17}$$

The **Gaussian kernel** is simply the standard normal (Gaussian) PDF:

$$K(u) = (2\pi)^{-1/2} \exp(-u^2/2). \tag{16.18}$$

The Gaussian kernel differs from the others because it has $K(u) > 0$ for all $u \in \mathbb{R}$, but $K(4) = 0.0001$ and $K(6) < 10^{-9}$, so the practical effect is negligible.

16.8.3 Effect on AMSE

The local linear kernel regression estimator's asymptotic bias and variance depend on the kernel and bandwidth. Let $\kappa_2 \equiv \int_{\mathbb{R}} [K(u)]^2\,\mathrm{d}u$; μ_2 is from (16.15). AMSE is

$$\begin{aligned} \mathrm{AMSE}_x(h) &= V(x)/(nh) + h^4[B(x)]^2, \\ V(x) &= \frac{\mathrm{Var}(Y \mid X = x)}{f_X(x)}\kappa_2, \quad B(x) = (1/2)m''(x)\mu_2. \end{aligned} \tag{16.19}$$

The AMSE is minimized by h_* solving

$$\begin{aligned} 0 &= 4h_*^3[B(x)]^2 - \frac{V(x)}{nh_*^2} \implies h_*^{-5} = \frac{4n[B(x)]^2}{V(x)} \\ \implies h_* &= n^{-1/5}\left(\frac{V(x)}{4[B(x)]^2}\right)^{1/5}. \end{aligned} \tag{16.20}$$

Table 16.1: Various kernels' μ_2 and κ_2 (*: normalized to $\mu_2 = 1$).

Kernel	μ_2	κ_2
uniform	$1/12$	1
triangle	$1/6$	$2/3$
Epanechnikov	$1/5$	$3/5$
Gaussian	1	$(4\pi)^{-1/2} \approx 0.282$
uniform*	1	$1/(2\sqrt{3}) \approx 0.289$
triangle*	1	$\sqrt{2}/(3\sqrt{3}) \approx 0.272$
Epanechnikov*	1	$3/(5\sqrt{5}) \approx 0.268$

Plugging h_* back into the AMSE in (16.19), the best possible AMSE is

$$\begin{aligned}
\text{AMSE}_x(h_*) &= V(x)/(nh_*) + h_*^4[B(x)]^2 \\
&= [V(x)/n]n^{1/5}V(x)^{-1/5}4^{1/5}[B(x)]^{2/5} \\
&\quad + n^{-4/5}[V(x)]^{4/5}4^{-4/5}[B(x)]^{-8/5}[B(x)]^2 \\
&= n^{-4/5}4^{1/5}[B(x)]^{2/5}[V(x)]^{4/5} + n^{-4/5}4^{-4/5}[B(x)]^{2/5}[V(x)]^{4/5} \\
&= n^{-4/5}\{B(x)[V(x)]^2\}^{2/5}(4^{1/5} + 4^{-4/5}) \\
&= n^{-4/5}\{(1/2)m''(x)\mu_2\kappa_2^2[\text{Var}(Y \mid X = x)/f_X(x)]^2\}^{2/5}(4^{1/5} + 4^{-4/5}) \\
&= n^{-4/5}(\mu_2\kappa_2^2)^{2/5}C(x),
\end{aligned}$$

where $C(x)$ gathers the other terms. The main points are a) AMSE is proportional to $n^{-4/5}$ and b) AMSE depends on the kernel $K(\cdot)$ through $\mu_2\kappa_2^2$.

Thus, the AMSE-optimal second-order kernel minimizes $\mu_2\kappa_2^2$ subject to $K(u) \geq 0$, $\int_{\mathbb{R}} K(u)\,\mathrm{d}u = 1$, $K(-u) = K(u)$. This minimization problem is solved by the Epanechnikov kernel; e.g., see Pagan and Ullah (1999, p. 27).

Table 16.1 shows the μ_2 and κ_2 of aformentioned kernels. The entires with * have been normalized to $\mu_2 = 1$ to facilitate comparison. Recall that the AMSE depends on $\mu_2\kappa_2^2$, so if we normalize kernels to have $\mu_2 = 1$, then the AMSE (given h_*) ranking is the same as the κ_2 ranking. Table 16.1 shows that the Epanechnikov has the smallest κ_2 when normalized to $\mu_2 = 1$, as claimed earlier. Nonetheless, its κ_2 is not that much smaller than the worst κ_2 (less than 10% better), that of the uniform kernel.

16.8.4 Higher-Order Kernels

For especially smooth CEFs, **higher-order kernels** reduce bias. Higher-order kernels satisfy (16.15) with $r > 2$. With $r = 2$, the bias is proportional to h^2. With $r = 4$, this drops to h^4, which is smaller than h^2 because $h \downarrow 0$. However, this benefit requires the CEF to have four derivatives instead of just two. The trick is that higher-order $K(\cdot)$ cancel out higher-order terms in a Taylor expansion of $m(\cdot)$ around x. However, even if

technically $m(\cdot)$ is smooth enough at x, in finite samples the magic may fail if the Taylor approximation is poor over the local sample. This is especially important because the promised small bias leads to a larger AMSE-optimal bandwidth.

Alternatively, instead of worry about which order kernel to use, you could let your model selection procedure (Chapter 18) search over both h and r.

16.9 Linear Smoother

Local and kernel regression estimators belong to an important class of estimators called **linear smoothers**. This is true even with vector $\boldsymbol{X}$. They are so called because (for any $\boldsymbol{x}$) the estimated $\hat{m}(\boldsymbol{x})$ is a linear combination of the Y_i.

Definition 16.1 (linear smoother)**.** Estimator $\hat{m}(\tilde{\boldsymbol{x}})$ is a linear smoother if it can be expressed as a linear combination of the Y_i with linear combination weights $W_i(\tilde{\boldsymbol{x}})$:

$$\hat{m}(\tilde{\boldsymbol{x}}) = \sum_{i=1}^{n} W_i(\tilde{\boldsymbol{x}}) Y_i, \tag{16.21}$$

where $W_i(\tilde{\boldsymbol{x}})$ may depend on $\tilde{\boldsymbol{x}}$, i, and all the observed $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, but may not depend on any Y_i.

Discussion Question 16.18 ($\bar{Y}$ as a linear smoother)**.** Show that the sample mean $\bar{Y}$ is a linear smoother. That is, because there is no x here, determine the W_i such that $\bar{Y} = \sum_{i=1}^{n} W_i Y_i$.

Discussion Question 16.19 (OLS as a linear smoother)**.** Show that OLS is a linear smoother. That is, determine $W_i(\tilde{\boldsymbol{x}})$ such that (16.21) is the OLS prediction. Hint: $\hat{m}(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}$, and write $\hat{\boldsymbol{\beta}}$ in summation notation; then move everything inside the final summation (that involves Y_i).

Local polynomial kernel regression is just weighted OLS, so the linear smoother representation is similar to DQ 16.19. Let $\boldsymbol{X} = (1, X - x, (X - x)^2, \ldots, (X - x)^p)'$. The CEF estimate is $\hat{\beta}_0(x) = (1, 0, \ldots, 0)\hat{\boldsymbol{\beta}}(x)$, and $\hat{\boldsymbol{\beta}}(x)$ is linear in Y_i:

$$\hat{\boldsymbol{\beta}}(x) = \left(\sum_{i=1}^{n} K((X_i - x)/h) \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \sum_{i=1}^{n} K((X_i - x)/h) \boldsymbol{X}_i Y_i. \tag{16.22}$$

Moving the $(1, 0, \ldots, 0)$ and $(\cdots)^{-1}$ inside the last sum,

$$\begin{aligned} \hat{m}_{K,h}(x) &= \sum_{i=1}^{n} W_i(x) Y_i, \\ W_i(x) &= (1, 0, \ldots, 0) \left(\sum_{i=1}^{n} K((X_i - x)/h) \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} K((X_i - x)/h) \boldsymbol{X}_i. \end{aligned} \tag{16.23}$$

Chapter 17
Series and Sieves

Unit learning objectives for this chapter

17.1. Develop intuition about the sieve approach to nonparametric regression, and the main subcategories [TLO 2]

17.2. Qualitatively, describe how different sieve estimators work and the conditions in which they work well [TLO 1]

Constrasting the "local" approach of kernel methods (Chapter 16), the sieve approach is "global," estimating a single (flexible) function that applies everywhere. This has advantages especially in extensions like instrumenal variables. Like before, model selection (Chapter 18) is critical for good performance in practice.

Let $m(\cdot)$ denote the CEF: $m(x) = \mathrm{E}(Y \mid X = x)$, where in this chapter X is scalar.

Optional resources for this chapter

- Textbook: Hansen (2020a) Chapter 20 (nonparametric series regression)
- Textbook: Pagan and Ullah (1999)
- Textbook: Li and Racine (2007) Chapter 15
- Textbook: James et al. (2013) Chapter 7 ("Moving Beyond Linearity"), including §7.5 ("Smoothing Splines")
- Textbook: Hastie, Tibshirani, and Friedman (2009) Chapter 5 ("Basis Expansions and Regularization"), including §5.4 ("Smoothing Splines"), and Chapter 11 ("Neural Networks")
- Original sieve paper: Grenander (1981)
- Example sieve spaces: Chen (2007, §2.3)

- R: built-in package `splines` (R Core Team, 2022)
- R: built-in package `stats` (R Core Team, 2022) has functions like `smooth.spline`.

17.1 Discrete Regressor

These examples help build intuition, similar to Sections 16.1–16.3.

17.1.1 Constant Regressor

Imagine $X = 1$ is just a constant. Then $\mathrm{E}(Y \mid X = x) = \mathrm{E}(Y)$, so estimating the "CEF" is equivalent to estimating the unconditional mean. The CEF model $m(x) = \beta_0$ can be estimated by OLS: $\hat{m}(1) = \hat{\beta}_0 = \bar{Y}$, the sample average.

17.1.2 Binary Regressor

See Section 16.2.2 and Chapter 6 of Kaplan (2022b).

If $X \in \{0, 1\}$, then the true CEF can be written $m(x) = \beta_0 + \beta_1 x$. OLS is consistent: $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \xrightarrow{p} m(x)$ for both $x = 0, 1$.

Alternatively, consider the less flexible model $m(x) = \beta_0$. The OLS estimator is $\hat{m}(x) = \hat{\beta}_0$ for both $x = 0, 1$. If $m(0) \neq m(1)$, then it is biased. Put differently, the simple model $m(x) = \beta_0$ is only an approximation of the true CEF $m(x) = \beta_0 + \beta_1 x$, so there is **approximation error**. But if $m(0) = m(1) = \mathrm{E}(Y)$, then this estimator is better: it is unbiased and (usually) has lower variance because it has fewer parameters to estimate. Even if $m(0) \neq m(1)$, this estimator may still have lower MSE if the decrease in variance outweighs the increase in squared bias.

As in Section 16.2.2, there is a bias–variance tradeoff, but it is framed differently. Before, model flexibility depended on the local sample and the bandwidth. Here, model flexibility depends on the number of parameters.

17.1.3 Trinary and More

See Chapter 7 of Kaplan (2022b).

If $X \in \{1, 2, 3\}$, then the true CEF may not have the form $m(x) = \beta_0 + \beta_1 x$. That is, this linear-in-variables model may be misspecified. OLS can still estimate the BLA, but the BLA may be a poor approximation of the CEF. ("Best" does not mean "good"!) Consequently, the CEF "slopes," $m(2) - m(1)$ and $m(3) - m(2)$, may differ greatly from the BLA slope β_1.

This can be fixed by using a three-parameter CEF model. For example, there could be an intercept along with dummies $\mathbb{1}\{X = 2\}$ and $\mathbb{1}\{X = 3\}$. Or there could be an intercept along with X and X^2.

Like before, though, there is a bias–variance tradeoff: including X^2 can both reduce bias (approximation error) and increase variance.

Generalizing to J possible values of X, approximation error is completely eliminated by a regression with J coefficients like a degree $J-1$ polynomial. But, using fewer coefficients may reduce the variance enough to reduce MSE, even if the bias increases.

In the extreme, consider $J \geq n$. If $J > n$, then there are more parameters than equations, so OLS cannot even be computed. With $J = n$, a degree $n-1$ polynomial can perfectly fit all n observations in the data; this suffers from overfitting (DQ 15.2 and Figure 15.2). The approximation error may be minimized, but the variance is huge.

So with $J = n$, what is best? Maybe we should stop at the degree $n-2$ polynomial. Or maybe an even smaller degree minimizes MSE. Or maybe we should use a degree $n-1$ polynomial but only allow five non-zero coefficients. Or maybe we should not even use the polynomial structure, but another flexible structure. These rough ideas are refined in later sections.

17.2 Polynomial Series

Polynomials are not recommended in practice, but their familiarity helps intuition.

Similar to Section 17.1.3, consider a polynomial CEF approximation,

$$\sum_{j=0}^{J-1} \beta_j x^j. \tag{17.1}$$

Given J, OLS can estimate the coefficients as usual.

If the true CEF $m(\cdot)$ is continuous and X has bounded support, then the approximation error can be made arbitrarily small for large enough J (Weierstrass, 1885).

The key **smoothing parameter** that determines model flexibility is J, the number of terms. Larger J increases flexibility, decreasing bias but increasing variance. Thus, larger J is analogous to smaller bandwidth for local/kernel regression. Practical procedures for choosing smoothing parameters are discussed in Chapter 18.

The least squares minimization can be rewritten in terms of functions rather than coefficients. This is an important shift in perspective. Usually, you have seen

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{b} \in \mathbb{R}^J}{\arg\min} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{J-1} \beta_j x^j)^2, \tag{17.2}$$

minimizing over $\boldsymbol{b} \in \mathbb{R}^J$. Equivalently, consider minimization over functions within the

set of degree $J-1$ polynomials ($\mathcal{M}_J$). Then,

$$\hat{m}(\cdot) = \underset{g(\cdot)\in\mathcal{M}_J}{\arg\min} \sum_{i=1}^{n} [Y_i - g(X_i)]^2, \quad \mathcal{M}_J \equiv \{g(\cdot) : g(x) = \sum_{j=0}^{J-1} \beta_j x^j\}. \tag{17.3}$$

Here, $g(\cdot)$ is a generic function, and $\hat{m}(\cdot)$ is the estimated CEF. This is the same estimate as before, $\hat{m}(x) = \sum_{j=0}^{J-1} \hat{\beta}_j x^j$, but it emphasizes searching over a function space for the purpose of estimating a function (the CEF).

A pseudo-true parameter can be defined similar to the "best" linear approximation (e.g., Kaplan, 2022b, §7.4). Specifically, consider the function in $\mathcal{M}_J$ that's "closest" to the true $m(\cdot)$:

$$m^*(\cdot) = \underset{g(\cdot)\in\mathcal{M}_J}{\arg\min}\, \mathrm{E}[(m(X) - g(X))^2]. \tag{17.4}$$

That is, among all functions in $\mathcal{M}_J$, $m^*(\cdot)$ is the best we can hope to estimate. The difference between $m(\cdot)$ and $m^*(\cdot)$ is the approximation error.

Discussion Question 17.1 (approximate CEF)**.** Consider true CEF $m(\cdot) \in \Theta$, the space of continuous (scalar) functions. Let $Q(g(\cdot)) = \mathrm{E}[(m(X) - g(X))^2]$, with $m^*(\cdot)$ as in (17.4). Note $Q(m(\cdot)) = 0$. As in (17.4), $Q(m^*(\cdot))$ is a measure of approximation error, where $m^*(\cdot) = \arg\inf_{g(\cdot)\in\mathcal{M}_J} Q(g(\cdot))$. As in (17.3), let $\mathcal{M}_3$ be the set of quadratic functions. If the true $m(x) = \sin(x)$, then what's $Q(m^*(\cdot))$? E.g., is it infinite? close to zero? etc.

Instead of J, often the smoothing parameter is written J_n to emphasize that it can grow asymptotically. Indeed, if J were fixed asymptotically, then it's just OLS!

Discussion Question 17.2 (polynomial models)**.** Let J_n be the number of terms in the polynomial model given sample size n. Assume a fixed data-generating process.

a) For a given J, as n increases, how do the bias (or approximation error) and variance change? Why?
b) As n increases, does the MSE-optimal J_n increase, decrease, not change, or sometimes any of these? Explain.

17.3 Series Regression

The ideas in Section 17.2 generalize beyond polynomials to series regression. Generalizing (17.3),

$$\hat{m}(\cdot) = \underset{g(\cdot)\in\mathcal{M}_n}{\arg\min} \sum_{i=1}^{n} [Y_i - g(X_i)]^2, \quad \mathcal{M}_n \equiv \{g(\cdot) : g(x) = \sum_{j=1}^{J_n} \beta_j \phi_j(x)\}, \tag{17.5}$$

where $J_n \to \infty$ as $n \to \infty$. Together, the $\{\phi_j(\cdot)\}_{j=1}^{\infty}$ form a **basis** for a certain space of functions (e.g., continuous functions over a bounded interval), and each $\phi_j(\cdot)$ may be

called a **basis function**. Each $\mathcal{M}_n$ is called a **sieve space**, with the sequence of $\mathcal{M}_n$ forming a **sieve**. The general approach is known as the **method of sieves** (Grenander, 1981): minimization over a sequence of sieve spaces whose approximation error decreases to zero. This special case where the first J_n terms in a basis form the sieve space is called **series regression**.

The choice of basis remains an open problem, but it can be guided by our assumptions about the true $m(\cdot)$. Different types of functions can be approximated well by different bases. For example, see Sections 2.3.1 and 2.3.6 of Chen (2007).

17.4 Splines

See page 5571 of Chen (2007).

Splines are a popular special case of the method of sieves. Cubic splines are especially popular. They are similar to the partitioning CEF estimator (Section 16.4), but instead of being discontinuous at the partition boundaries, they are continuous and even twice continuously differentiable (with the third derivative allowed to jump discontinuously at each **knot** that separates the intervals of the partition of the support of X).

17.5 Linear vs. Nonlinear Approximation

The sieve spaces in Section 17.3 are "linear" in that if $f, g \in \mathcal{M}_n$, then so is linear combination $\lambda f + (1-\lambda)g \in \mathcal{M}_n$ for any $0 \leq \lambda \leq 1$. This is true because f and g are both defined in terms of coefficients $(\beta_1, \ldots, \beta_J)$ on the same basis functions $(\phi_1, \ldots, \phi_J)$.

More generally, nonlinear sieve spaces could be used. For example, imagine the set of polynomials with J (non-zero) terms, but allowing non-consecutive terms. If $J = 2$, this could include $\beta_0 + \beta_1 x$ but also $\beta_0 + \beta_3 x^3$ or $\beta_5 x^5 + \beta_7 x^7$. This is nonlinear because, for example, $(0.5)(\beta_0 + \beta_3 x^3) + (0.5)(\beta_5 x^5 + \beta_7 x^7)$ involves four terms, not $J = 2$ terms. There is also highly nonlinear approximation that increases flexibility by including more than just one basis, like neural networks; e.g., see White (2006).

17.6 Penalized Regression

See Section 2.3.4 of Chen (2007) and references therein. The general idea is to allow an infinite-dimensional sieve space, but add a penalty to restrict the size of the sieve space.

The most famous examples of penalized regression are ridge regression and lasso, which have generalizations like the bridge estimator and elastic net. For lasso, see James et al. (2013, §6.2.2) and Hastie, Tibshirani, and Friedman (2009, §3.4.2); for bridge and elastic net, see Hastie, Tibshirani, and Friedman (2009, §3.4.3).

First, **ridge regression** penalizes large slope coefficients according to their squared

magnitude. Given penalty parameter λ (like $\lambda = 0.1$), the ridge estimator solves

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) = \underset{b_0, b_1, \ldots, b_k}{\arg\min} \overbrace{\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_1 - \cdots - b_k X_k)^2}^{\text{SSR}} + \lambda \overbrace{\sum_{j=1}^{k} b_j^2}^{\text{penalty}}. \tag{17.6}$$

If $\lambda = 0$, then the penalty is zero, so the ridge estimator is simply OLS, minimizing the SSR. If $\lambda = \infty$, then all $\hat{\beta}_1$ through $\hat{\beta}_k$ equal zero and $\hat{\beta}_0 = \bar{Y}$: even if SSR is large, making any $\hat{\beta}_j \neq 0$ incurs an infinite penalty, so it is never worth the reduction in SSR. Thus, ridge regression is a **shrinkage estimator** that "shrinks" all the slope estimates $\hat{\beta}_j$ toward zero. When $\lambda = 0$, there is no shrinkage; as λ increases, there is more shrinkage.

Second, **lasso** (Knight and Fu, 2000; Tibshirani, 1996) replaces the b_j^2 in (17.6) with absolute values $|b_j|$:

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) = \underset{b_0, b_1, \ldots, b_k}{\arg\min} \overbrace{\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_1 - \cdots - b_k X_k)^2}^{\text{SSR}} + \lambda \overbrace{\sum_{j=1}^{k} |b_j|}^{\text{penalty}}. \tag{17.7}$$

Again, the intercept b_0 is not penalized. Again, as $\lambda \to \infty$, the slope coefficient estimates all "shrink" toward zero. However, with ridge, they never quite reach zero exactly unless $\lambda = \infty$ (which is not used in practice), whereas with lasso, usually some coefficients are shrunk all the way to $\hat{\beta}_j = 0$ exactly. This can be interpreted as lasso "selecting" only the X_j for which $\hat{\beta}_j \neq 0$. Indeed, lasso is an acronym (although it is usually written in lowercase) for "least absolute shrinkage and selection operator."

Ridge and lasso are special cases of the **bridge estimator** (Frank and Friedman, 1993; Fu, 1998; Knight and Fu, 2000), also called things like L_q lasso or L_q penalized regression. The bridge penalty replaces ridge's b_j^2 or lasso's $|b_j|$ with $|b_j|^\gamma$ for some γ. As special cases, ridge has $\gamma = 2$, and lasso has $\gamma = 1$.

Another generalization of ridge and lasso is the **elastic net**. There, both $|b_j|$ and b_j^2 are penalized, possibly in different amounts.

17.7 Linear Smoother

Some sieve CEF estimators are linear smoothers, notably series regression. Consider series regression with basis $\{\phi_j(\cdot)\}_{j=1}^{\infty}$, with sieve size J. Once J is determined, estimation is simply OLS with regressors $(\phi_1(X), \ldots, \phi_J(X))$. Thus, because OLS is a linear smoother, so is series regression.

Chapter 18
Model Selection

Unit learning objectives for this chapter

18.1. Develop intuition for different approaches to model selection, including how the capture the bias–variance tradeoff to avoid overfitting or underfitting [TLO 2]

18.2. Qualitatively, describe how different model selection methods work and when they should work well [TLO 1]

18.3. Judge whether a particular model selection procedure is appropriate in a given setting, especially with non-iid data [TLO 3]

As Box (1979, p. 2) famously wrote, "All models are wrong but some are useful."[1] This applies well to nonparametric regression: there's no pretense of finding the correct CEF model, but hopefully accounting for bias enables selection of a more useful model.

There are two main approaches to model selection. One general approach (cross-validation) uses some observations to estimate each candidate model and then tests their predictions on the rest of the observations. Another general approach (including information criteria and GCV) starts with the in-sample fit but then adds a penalty for model flexibility to avoid overfitting.

Optional resources for this chapter

- Textbook: Kaplan (2022b) Sections 8.3 (intro to model selection) and 15.2 (AIC, BIC)
- Textbooks: Konishi and Kitagawa (2008) and Claeskens and Hjort (2008)
- Textbook: Hansen (2020a) Chapter 23

[1] See https://en.wikipedia.org/wiki/All_models_are_wrong for additional discussion.

- Textbook: Hastie, Tibshirani, and Friedman (2009) Chapter 7 ("Model Assessment and Selection") and §§8.7–8.8 (model averaging)
- Bias–variance tradeoff: James et al. (2013, §2.2.2), Hastie, Tibshirani, and Friedman (2009, §§2.9,5.5.2,7.2,7.3)
- R: package `caret` (Kuhn, 2020) helps with model selection for a very wide variety of estimators
- R: some functions have (some) model selection capability built in; e.g., `smooth.spline()` in core R accepts argument `cv=TRUE` for leave-one-out cross-validation and `cv=FALSE` for generalized cross-validation (GCV).
- R: package `np` (Hayfield and Racine, 2008) has model selection functions corresponding to its kernel estimators

18.1 Purpose

The famous quote from Box (1979) begs the question: "useful" for what? Taking this question seriously has led to some innovative and practically useful model selection procedures, like that of Claeskens and Hjort (2003) or Belloni, Chernozhukov, and Hansen (2014).

For example: useful for estimation or for inference? The model that produces the best $\hat{m}(x)$ may not produce the best corresponding confidence interval.

Another example: useful for prediction or for causality? The most useful model for prediction does not necessarily produce the best structural estimates. Historically, most of the model selection literature focused on prediction. However, the literature on model selection for structural or treatment effect models is growing. For example, see Belloni, Chernozhukov, and Hansen (2014), Horowitz (2014), and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

One more example: useful for learning $m(\cdot)$ or $m(x)$ or some other summary of $m(\cdot)$?

18.2 Quantifying Flexibility of Linear Smoothers

Recall that many estimators are linear smoothers (16.21):

$$\hat{m}(x) = \sum_{i=1}^{n} W_i(x) Y_i, \tag{18.1}$$

The fitted values are thus

$$\hat{Y}_i = \hat{m}(X_i) = \sum_{j=1}^{n} W_j(X_i) Y_j, \quad i = 1, \ldots, n. \tag{18.2}$$

In matrix notation, let matrix $\underline{\boldsymbol{W}}$ have row i, column j entry $W_{ij} = W_j(X_i)$, so

$$\hat{\boldsymbol{Y}} = \underline{\boldsymbol{W}}\boldsymbol{Y}, \tag{18.3}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ and $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)'$.

To develop intuition, consider OLS. With OLS, $\underline{\boldsymbol{W}}$ is the "hat matrix" or projection matrix $\underline{\boldsymbol{X}}(\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}'$. That is, $\hat{\boldsymbol{Y}}$ is the linear projection of $\boldsymbol{Y}$ onto the column space of $\underline{\boldsymbol{X}}$, which is the $n \times k$ matrix with row i equal to $\boldsymbol{X}_i'$. The natural measure of model flexibility for OLS is the number of parameters, which is k, the number of columns in $\underline{\boldsymbol{X}}$. Further, k is the trace of $\underline{\boldsymbol{W}}$:

$$\text{tr}(\underline{\boldsymbol{W}}) = \text{tr}(\underline{\boldsymbol{X}}(\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}') = \text{tr}((\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}}) = \text{tr}(\underline{\boldsymbol{I}}_k) = k, \tag{18.4}$$

the trace of the $k \times k$ identity matrix.

Discussion Question 18.1 (OLS flexibility 1)**.** Consider OLS with k parameters (including the intercept), and assume no perfect multicollinearity. Consider when OLS perfectly fits the data, i.e., $\hat{Y}_i = Y_i$ for all $i = 1, \ldots, n$. Hint: use (18.2)–(18.4).

a) Explain how this is possible with $k = n$.
b) What is $\text{tr}(\underline{\boldsymbol{W}})$? Why?
c) What is $\underline{\boldsymbol{W}}$? In particular, what are the diagonal entries, W_{ii}?

Discussion Question 18.2 (OLS flexibility 2)**.** Consider OLS with only an intercept term. Hint: use (18.2)–(18.4).

a) What is $\hat{Y}_i$?
b) What is $\underline{\boldsymbol{W}}$? Hint: you can either use $\underline{\boldsymbol{X}} = (1, 1, \ldots, 1)'$, or use $\hat{\boldsymbol{Y}} = \underline{\boldsymbol{W}}\boldsymbol{Y}$.
c) What is k? Why?
d) What is $\text{tr}(\underline{\boldsymbol{W}})$? Why?

Equation (18.4) suggests $\text{tr}(\underline{\boldsymbol{W}})$ can quantify the flexibility of a linear smoother. This value is sometimes called the **effective number of parameters**, or **effective dimension** or **effective degrees of freedom** or **equivalent number of parameters**.

Consider a linear smoother that perfectly fits the data (extreme overfitting). That is, $\hat{Y}_i = Y_i$ for all $i = 1, \ldots, n$. Because $\hat{Y}_i = \sum_{j=1}^n W_j(X_i)Y_j$, such an estimator must have $W_i(X_i) = 1$ for $i = 1, \ldots, n$, and $W_j(X_i) = 0$ for $j \neq i$. That is, $\underline{\boldsymbol{W}}$ is the $n \times n$ identity matrix, so $\hat{\boldsymbol{Y}} = \underline{\boldsymbol{W}}\boldsymbol{Y} = \underline{\boldsymbol{I}}_n\boldsymbol{Y} = \boldsymbol{Y}$.

More generally, the main diagonal terms $W_{ii} = W_i(X_i)$ help capture overfitting. They capture the influence of Y_i on $\hat{Y}_i$:

$$\hat{Y}_i = W_{ii}Y_i + \sum_{j \neq i} W_{ij}Y_j. \tag{18.5}$$

Many model selection procedures explicitly or implicitly use $\text{tr}(\underline{\boldsymbol{W}})$ or the W_{ii}.

18.3 Bad Approaches

Don't use hypothesis tests for model selection. The question "Which model provides the best estimate?" is not answered by "I controlled the type I error rate at level α!" Another problem is sensitivity to the choice of null (vs. alternative) and choosing α. Further, recall that we don't actually want the "correct" model; we want the MSE-optimal estimate. For example, the true CEF may be a 698th-degree polynomial, but if $n = 500$, we don't want to select the true model for estimation.

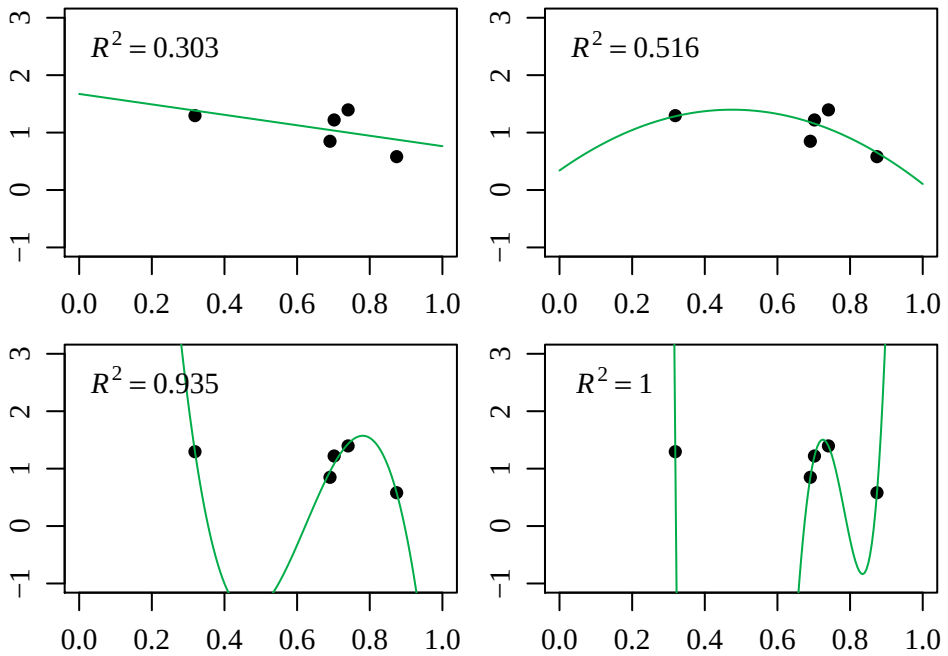


Figure 18.1: R^2 (but not accuracy) always increases with model flexibility.

Don't use R^2 or (equivalently) the sum of squared residuals (SSR). Increasing a model's flexibility always improves the in-sample fit (i.e., larger R^2, lower SSR), so R^2 or SSR would tell us to use the most flexible model possible (extreme overfitting). Figure 18.1 shows how R^2 always increases with model flexibility, even when overfitting is obvious.

Adjusted R^2 is better than R^2, but there are better approaches. Adjusted R^2 is most similar to GCV; see (18.12).

18.4 Analytic Plug-in Approach

Equation (16.20) gives the AMSE-optimal bandwidth h_* for the local linear kernel regression estimator, but h_* is **infeasible**: it depends on unknown population objects like

$f_X(x)$, $\text{Var}(Y \mid X = x)$, and most vexingly $m''(x)$. So h_* cannot directly be used in practice.

In principle, we can estimate the unknown terms and plug them into the formula, yielding a **plug-in bandwidth**. However, large estimation error may cause the plug-in bandwidth to differ significantly from the infeasible h_*.

One possibility is to iterate: once $m(\cdot)$ and $m''(\cdot)$ are estimated with an initial **pilot bandwidth**, use them to compute the plug-in bandwidth, but then use the subsequent estimates to compute yet another plug-in bandwidth, etc. However, there is no guarantee that a fixed point of this iteration (if it even exists) is the optimal bandwidth.

Because of these difficulties, the most common model selection procedures do not rely on an analytic AMSE formula.

18.5 Cross-Validation

Cross-validation is so important in statistics that the StackExchange statistics website is named Cross Validated. (Important, plus it's a good pun.) (Incidentally, it's also a very useful website.)

18.5.1 Training and Validation Paradigm

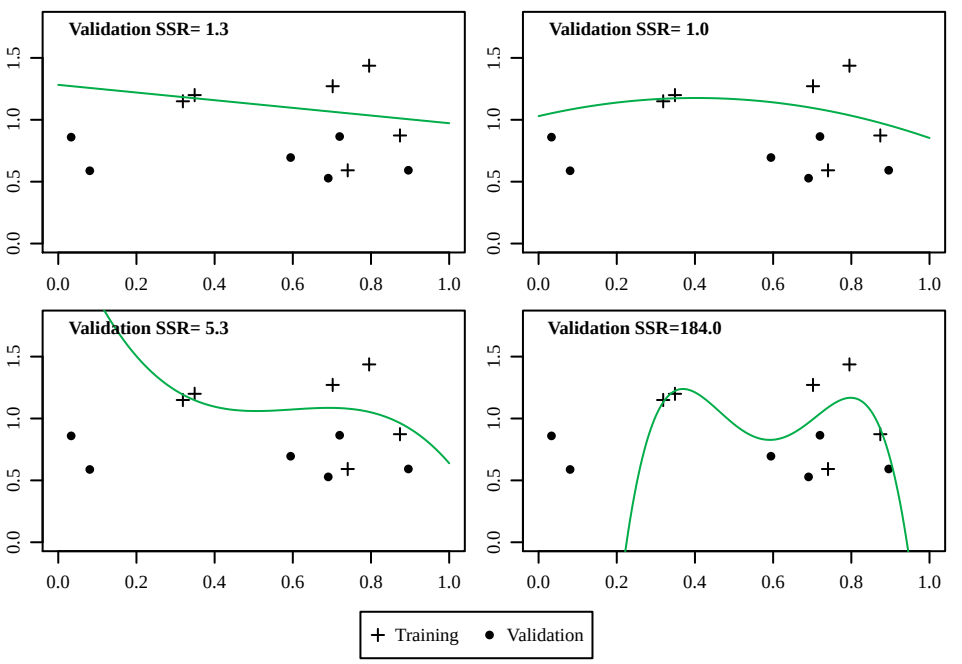


Figure 18.2: Validation data SSR for estimates from training data.

Discussion Question 18.3 (validation SSR)**.** Consider Figure 18.2. The (colored) lines are CEF estimates $\hat{m}(\cdot)$ based on (only) the "training" data shown. The "validation" data is not used for estimation. The "validation SSR" is based on the residuals $Y_i - \hat{m}(X_i)$ for validation observations i only.

a) Rank the models from least to most flexible, and explain why.
b) Rank the models from worst to best fit of the training data.
c) Rank the models from worst to best fit of the validation data.
d) Explain why the most flexible model best fits the training data.
e) Explain why the most flexible model does not best fit the validation data.
f) Recall Figure 18.1, in which SSR always decreased (R^2 always increased) with model flexibility. Why doesn't SSR always decrease with model flexibility here?

One way to describe the problem with R^2 and SSR is that they evaluate model accuracy with the same data used to estimate the model. That is, they compare the "predicted" $\hat{Y}_i$ with Y_i, but $\hat{Y}_i$ was computed using Y_i itself—this is cheating! If we're allowed to use Y_i to predict itself, then we should just predict $\hat{Y}_i = Y_i$. But we know this is (extreme) overfitting.

The **cross-validation** (CV) approach separates Y_i from the observations used to generate $\hat{Y}_i$. That is, it doesn't allow cheating. The subset of observations used to generate the predictions is the **training sample**. These predictions are then compared to the Y_i in the remaining observations, called the **validation sample**. (Sometimes the validation sample is called the testing sample, but I think technically the testing sample is something different.) That is, each model is "trained" (estimated) with one set of observations but "validated" (evaluated) using a separate set of observations.

18.5.2 LOOCV

One natural approach is to use every observation except (Y_i, X_i) to compute $\hat{Y}_i$. This is called **leave-one-out cross-validation** (LOOCV) because one observation (i) is "left out" when computing $\hat{Y}_i$. The LOOCV estimator that omits i is often denoted by a subscript $-i$ or $(-i)$. For example, $\hat{m}_{-i}(\cdot)$ is a CEF estimator based on observations $1, \ldots, i-1, i+1, \ldots, n$. The corresponding LOOCV prediction is $\hat{Y}_i = \hat{m}_{-i}(X_i)$. Adding smoothing parameter s (like bandwidth or number of series terms) to the notation, the LOOCV criterion is

$$\text{LOOCV}(s) = \sum_{i=1}^{n} [Y_i - \hat{m}_{-i}(X_i; s)]^2. \tag{18.6}$$

Given set $\mathcal{S}$ of possible candidate models, LOOCV chooses model

$$s^* = \arg\min_{s \in \mathcal{S}} \text{LOOCV}(s). \tag{18.7}$$

Note s^* also minimizes $n^{-1}\text{LOOCV}(s)$.

LOOCV can be slow to compute. With brute force, LOOCV requires computing $\text{LOOCV}(s)$ for each $s \in \mathcal{S}$, each of which requires computing $\hat{m}_{-i}(\cdot)$ for $i = 1, \ldots, n$, i.e.,

computing the estimator n different times. With large n and/or complex estimators, this could take hours or even days. Choosing values of s carefully can help a lot.

However, for linear smoothers, there is a shortcut that only requires the full-sample estimator $\hat{m}$ and not the n leave-one-out estimators $\hat{m}_{-i}$. In the notation of Section 18.2,

$$\text{LOOCV}(s) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}(X_i; s)}{1 - W_i(X_i)}\right)^2. \tag{18.8}$$

See Section 18.A for the derivation.

Given (18.8), LOOCV can be interpreted as penalized SSR. Recall from (18.5) that $W_i(X_i)$ measures the influence of Y_i on $\hat{Y}_i$, which captures the flexibility of the model. The penalty term p_i in (18.9) is larger when flexibility $W_i(X_i)$ is larger; $p_i = 1$ when $W_i(X_i) = 0$, and $p_i \to \infty$ as $W_i(X_i) \to 1$. Minimizing (18.8) is equivalent to minimizing

$$\sum_{i=1}^{n} p_i(Y_i - \hat{Y}_i)^2, \quad p_i \equiv 1/[1 - W_i(X_i)]^2, \tag{18.9}$$

where $\hat{Y}_i = \hat{m}(X_i; s)$ implicitly depends on s.

This penalty helps capture the bias–variance tradeoff. Unlike with unpenalized SSR, there is a tension in (18.8): more flexibility decreases the squared residual but increases the penalty.

18.5.3 GCV

Craven and Wahba (1978) suggest a **generalized cross-validation** (GCV) simplifying (18.8). Instead of penalizing each residual separately, they penalize the SSR by an average penalty:

$$\text{GCV} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{[1 - \text{tr}(\underline{\boldsymbol{W}})/n]^2} = n^{-1}\text{SSR}/[1 - \text{tr}(\underline{\boldsymbol{W}})/n]^2. \tag{18.10}$$

The penalty comes from

$$\frac{1}{n}\sum_{i=1}^{n}[1 - W_i(X_i)] = \frac{1}{n}(n - \sum_{i=1}^{n} W_{ii}) = 1 - \text{tr}(\underline{\boldsymbol{W}})/n. \tag{18.11}$$

Scaling by n does not change the minimizer, so GCV could also be written as

$$\text{SSR}/[n - \text{tr}(\underline{\boldsymbol{W}})]^2. \tag{18.12}$$

This is similar to the SSR adjustment in adjusted R^2, but here the penalty is squared.

Discussion Question 18.4 (model selection and the CIA)**.** You are interested in the causal effect of X_1 on Y. You have many control variables available and hope to control omitted variable bias if you choose the right ones to include as $\boldsymbol{X}_2$ in your regression. That is, you hope to choose $\boldsymbol{X}_2$ to satisfy the conditional independence assumption $U \perp\!\!\!\perp$

$X_1 \mid \boldsymbol{X}_2$. More realistically, you know it won't hold perfectly (so there will be some bias), so you want to choose the variables to include in $\boldsymbol{X}_2$ in order to minimize the MSE of $\hat{\beta}_1$ (recall MSE is variance plus squared bias).

a) What about GCV might be good for picking $\boldsymbol{X}_2$ from among a large set of available variables (and transformations of variables)?
b) What might GCV do "wrong" here?

18.5.4 Leave-d-out CV

LOOCV generalizes to **leave-*d*-out cross-validation**. It's exactly how it sounds: remove d observations from the sample, estimate the model to compute predicted $\hat{Y}_i$ for the "left out" observations, and then repeat until you have $\hat{Y}_i$ for all i. Then compute the cross-validated SSR (or other measure of fit). If the linear smoother computational shortcut of (18.8) cannot be used, then leaving out groups of d is faster: the estimator need only be computed n/d times instead of n times. LOOCV is the special case $d = 1$.

18.5.5 k-fold CV

If d is a significant fraction of n, then leave-d-out CV is called ***k*-fold cross-validation** with $k = n/d$. For whatever reason, $k = 5$ is very popular. That is, split the sample into $k = 5$ roughly equal subsamples, and rotate through: leave out one of the subsamples and compute the predicted $\hat{Y}_i$ for it based on the remaining four subsamples, and do this five times (once for each subsample). Then compute the cross-validated SSR (or other measure of fit).

As noted in Section 18.7, this has somewhat different performance than LOOCV.

18.5.6 Time Series

Although LOOCV is inappropriate for time series, the general idea of separating the training and validation data can be applied.

Basically, we can pretend to travel back to time t and generate forecasts as if we didn't know the "future" $(t + 1)$. But because we do know Y_{t+1}, we can compare the forecast $\hat{Y}_{t+1}$ to the actual Y_{t+1}. Then we can pretend we live at time $t + 1$ to generate $\hat{Y}_{t+2}$, and compare to the true Y_{t+2}, and so on, up to comparing $\hat{Y}_T$ to Y_T.

For more, see Section 3.4 of Hyndman and Athanasopoulos (2019), and the corresponding R function `tsCV()` in package `forecast` (Hyndman, Athanasopoulos, Bergmeir, Caceres, Chhay, O'Hara-Wild, Petropoulos, Razbash, Wang, and Yasmeen, 2020; Hyndman and Khandakar, 2008).

Discussion Question 18.5 (time series CV)**.** You have $T = 300$ daily observations Y_t, $t = 1, \ldots, 300$. You want to know if an AR(1) or AR(2) model gives better predictions. You deem $t = 241, \ldots, 300$ the validation data.

a) How do you compute the AR(1) and AR(2) "forecasts" for Y_{241}, the first observation in the validation data? You can use high-level descriptions like, "Estimate an AR(1) model using. . . ."
b) How do you compute the forecasts for Y_{242}?
c) How do you compute the forecasts for Y_t, $242 < t \leq 300$?
d) How do you compute the validation-sample average squared forecast error?
e) How can you decide which model produces better forecasts?

18.6 Information Criteria

Many model selection procedures use an **information criterion**. An information criterion measures "how bad" a model is. Thus, the information criterion value is computed for each candidate model, and the model with the lowest value is selected as the "best."

18.6.1 AIC and BIC

The original is the **Akaike information criterion** (AIC), proposed by Akaike (1974). Though originally formulated in the context of maximum likelihood and Kullback–Leibler divergence and written in terms of maximized likelihood $\mathcal{L}$ with k parameters, the AIC can also be written in terms of the sum of squared residuals (SSR) for linear regression with k coefficients:

$$\text{AIC} = \overbrace{-2\ln(\mathcal{L})}^{\text{fit}} + \overbrace{2k}^{\text{penalty}} \quad \text{or} \quad \overbrace{n\ln(\text{SSR})}^{\text{fit}} + \overbrace{2k}^{\text{penalty}}, \tag{18.13}$$

where n is the sample size. For both "fit" terms, smaller means better fit. More generally, k could be replaced by the effective number of parameters, like the trace of the linear smoother matrix.

Minimizing SSR alone results in overfitting, so the AIC adds a penalty for the model's flexibility. More flexibility decreases SSR but increases the penalty, so there is a tension. If the additional flexibility improves the fit greatly, then the decrease in SSR outweighs the increased penalty, resulting in lower AIC (better model). But if the fit only improves very slightly, then the penalty outweighs the reduced SSR, and the AIC says it is not worth it.

The **Bayesian information criterion** (BIC) (also SIC, SBC, or SBIC) of Schwarz (1978) has a similar form:

$$\text{BIC} = \overbrace{-2\ln(\mathcal{L})}^{\text{fit}} + \overbrace{\ln(n)k}^{\text{penalty}} \quad \text{or} \quad \overbrace{n\ln(\text{SSR})}^{\text{fit}} + \overbrace{\ln(n)k}^{\text{penalty}}. \tag{18.14}$$

BIC penalizes flexibility more than AIC. The 2 in the AIC's penalty is replaced by $\ln(n)$ in BIC. Especially with large n, $\ln(n)$ is much larger than 2, so a given increase in k corresponds to a much larger increase in BIC penalty than in AIC penalty.

Discussion Question 18.6 (AIC vs. BIC)**.** Consider four models for a particular dataset. Models A and B have $k = 2$, whereas models C and D have $k = 4$. Models A and C have $\ln(\mathcal{L}) = 2$; Model B has $\ln(\mathcal{L}) = 4$; Model D has $\ln(\mathcal{L}) = 5$. Let $n = 55$, so $\ln(n) \approx 4$. Refer to (18.13) and (18.14).

a) Compute the AIC for each model.
b) Rank the models from best to worst according to AIC.
c) Compute the BIC for each model.
d) Rank the models from best to worst according to BIC.
e) Explain which ranking seems more intuitive to you.
f) Would you guess that AIC or BIC generally tends to pick more flexible models? Why?

The AIC and BIC are often used for lag length selection in (vector) autoregression.

18.6.2 Other IC

There is GIC, FIC, (M)RIC, EAIC, NIC, and more; e.g., see Shao (1997, p. 223).

The **focused information criterion** (FIC) of Claeskens and Hjort (2003) stands out by focusing on a particular parameter rather than prediction accuracy and overall fit. DiTraglia (2016) extends this idea to GMM.

18.7 Comparison

Shao (1997) provides a unified framework for comparing many different model selection procedures. Although the setting is linear regression with homoskedastic errors (p. 224), I'd guess the qualitative results still hold for more complex models. Shao (1997) considers selecting from p_n possible regressors, where p_n may increase with n. These regressors may include nonlinear functions of an observed variable, as in Shao's Example 3; e.g., a model may include both X and X^2. Each of the p_n possible regressors can be either included or excluded in a given model, so there are 2^{p_n} possible models.

Shao (1997, p. 235) delineates three classes of model selection procedures and qualitatively compares their performance in different settings. Class 1 includes Mallows' C_p, AIC, LOOCV, and GCV. Class 2 includes BIC and delete-d CV with $d/n \to 1$. Class 3 includes delete-d CV with $d/n \to \tau \in (0, 1)$, or k-fold CV with fixed k. Shao (1997) says, "The methods in class 1 are useful in the case where there is no fixed-dimension correct model," as generally assumed with nonparametric regression. Also, "Methods in class 2 are useful in the case where there exist fixed-dimension correct models." Class 3 methods lie in between Classes 1 and 2.

For details most closely related to nonparametric regression, see Example 3 and Theorems 1(i), 3(i), 4(i,ii), and 5.

18.8 Model Averaging and Ensemble Methods

Instead of trying to pick a single best model (as in model selection), **model averaging** assigns weights to different models. The final prediction is the weighted average of all the models' predictions. Actually, model selection is a special case of model averaging where one model has weight 1 and all other models have weight 0. In many cases, model averaging produces more accurate predictions than model selection. For example, see Chapter 7 of Claeskens and Hjort (2008), who discuss both frequentist and Bayesian model averaging.

More generally, **ensemble** methods combine multiple simpler models or predictions into a more complex final prediction. Such ensemble methods include bagging (Breiman, 1996) and random forest (Breiman, 2001; Ho, 1995).

Appendix to Chapter 18

18.A LOOCV for Linear Smoothers

These are the steps to derive (18.8). Let $\hat{m}(x) = \sum_{j=1}^{n} W_j(x) Y_j$. This implies

$$\hat{m}(X_i) = \sum_{j=1}^{n} W_j(X_i) Y_j = W_i(X_i) Y_i + \sum_{j \neq i} W_j(X_i) Y_j.$$

The LOOCV estimator is

$$\hat{m}_{-i}(X_i) = \sum_{j \neq i} W_j(X_i) Y_j / \sum_{j \neq i} W_j(X_i).$$

Then, since $1 = \sum_{j=1}^{n} W_j(X_i) = W_i(X_i) + \sum_{j \neq i} W_j(X_i)$,

$$\begin{aligned}
\hat{m}_{-i}(X_i) &= \frac{\hat{m}(X_i) - W_i(X_i) Y_i}{1 - W_i(X_i)}, \\
\text{LOOCV} &= n^{-1} \sum_{i=1}^{n} [Y_i - \hat{m}_{-i}(X_i)]^2 = n^{-1} \sum_{i=1}^{n} \left[Y_i - \frac{\hat{m}(X_i) - W_i(X_i) Y_i}{1 - W_i(X_i)} \right]^2 \\
&= n^{-1} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{m}(X_i)}{1 - W_i(X_i)} \right)^2.
\end{aligned}$$

Chapter 19
Multiple Regressors

Unit learning objectives for this chapter

19.1. Develop intuition for the curse of dimensionality and how to address it [TLO 2]

19.2. Describe different functional form restrictions that help reduce the curse of dimensionality [TLO 1]

19.3. Judge which multivariate nonparametric regression model seems most appropriate in a given example [TLO 3]

Unlike with OLS, it is not trivial to simply another regressor to a nonparametric regression model. This chapters describes the difficulty and some approaches.

Optional resources for this chapter

- Textbook: Hastie, Tibshirani, and Friedman (2009) Chapter 9 and Sections 5.7 and 6.4

19.1 Curse of Dimensionality

Discussion Question 19.1 (curse of dimensionality 0)**.** Let $n = 1000$. Let $\boldsymbol{X} = (X_1, \ldots, X_6)$, where X_1, X_2, X_3, and X_4 are binary, $X_5 \in \{\text{North, South, East, West}\}$ (geographic region), and $X_6 \in \{\text{no high school, high school, college, graduate}\}$. We wish to nonparametrically estimate $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ for all possible values of $\boldsymbol{x}$, by taking the sample mean of the values Y_i for which $\boldsymbol{X}_i = \boldsymbol{x}$; i.e., taking the sample mean within each "cell." Denote subsample sizes as $N_{\boldsymbol{x}} \equiv \sum_{i=1}^{n} \mathbb{1}\{\boldsymbol{X}_i = \boldsymbol{x}\}$.

a) Among all possible samples, what's the largest possible value of $\min_{\boldsymbol{x}} N_{\boldsymbol{x}}$?

b) Is $n = 1000$ big enough for asymptotic approximations to be reasonable?

Discussion Question 19.2 (curse of dimensionality 1)**.** Consider a kernel regression estimator using a uniform kernel, first in one dimension (scalar X), then higher dimensions. Interest is in $m(\boldsymbol{x}_0) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}_0)$. Let $x_0 = 0.05$ with bandwidth $h = 0.1$. Let n denote sample size.

a) Let $X_i \sim \mathrm{Unif}(0, 1)$. What's the probability that a single X_i falls in the uniform kernel window $[x_0 - h/2, x_0 + h/2]$? That is, what's $\mathrm{P}(X_i \in [x_0 - h/2, x_0 + h/2])$?
b) Let $\boldsymbol{X}_i \sim \mathrm{Unif}([0, 1]^2)$, the uniform distribution over the unit square $[0, 1] \times [0, 1]$; i.e., the PDF of $\boldsymbol{X}_i$ is $f(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in [0, 1]^2$ and $f(\boldsymbol{x}) = 0$ elsewhere. What's the probability that a single $\boldsymbol{X}_i$ falls in the window $[x_0 - h/2, x_0 + h/2] \times [x_0 - h/2, x_0 + h/2]$?
c) What's the probability of falling in $[x_0 - h/2, x_0 + h/2]^3$ if $\boldsymbol{X}_i$ is uniformly distributed over the unit cube $[0, 1]^3$?
d) What's the probability of falling in the window $[x_0 - h/2, x_0 + h/2]^d$ if $\boldsymbol{X}_i$ is uniformly distributed over the d-dimensional hyper-cube $[0, 1]^d$?
e) For general h, if $0 \leq x_0 - h/2 < x_0 + h/2 \leq 1$ and again $\boldsymbol{X}_i$ is uniformly distributed over $[0, 1]^d$, then what's $\mathrm{P}(\boldsymbol{X}_i \in [x_0 - h/2, x_0 + h/2]^d)$?

Discussion Question 19.3 (curse of dimensionality 2)**.** Continue from DQ 19.2.

a) Let $N_0 = \sum_{i=1}^n \mathbb{1}\{\boldsymbol{X}_i \in [x_0 - h/2, x_0 + h/2]^d\}$ be the local sample size. Explain why the mean local sample size $\mathrm{E}(N_0)$ is proportional to nh^d with iid sampling. Hint: $B_i \equiv \mathbb{1}\{\boldsymbol{X}_i \in \text{window}\}$ is a Bernoulli random variable, so under iid sampling, $\sum_{i=1}^n B_i$ has a binomial distribution with parameters n (sample size) and $p = \mathrm{P}(\boldsymbol{X}_i \in \text{window})$; the mean of a Binomial$(a, b)$ rv is ab.
b) Given the same h, do you think the variance of the local constant regression estimator is smaller, larger, or the same with large d compared to $d = 1$? Why?
c) Given the same h, do you think the bias of the local constant regression estimator is smaller, larger, or the same with large d compared to $d = 1$? Why?
d) Compared to the AMSE-optimal h_* with $d = 1$, do you think the AMSE-optimal bandwidth with large d is smaller, larger, or the same? Why?
e) With the AMSE-optimal bandwidth, do you think the AMSE is smaller, larger, or the same with large d compared to $d = 1$? Why?

Generally, when there are d regressors instead of a single X, the (second-order) kernel regression estimator's convergence rate slows from $\sqrt{nh}$ to $\sqrt{nh^d}$. The latter is smaller because $h \downarrow 0$. The adverse effect of the dimension d on the convergence rate is called the **curse of dimensionality** for nonparametric estimators. Given the same bandwidth, this means a larger order of magnitude of standard errors, which are proportional to $1/\sqrt{nh^d}$. One interpretation is that now we only have nh^d relevant observations within $h/2$ of the point of interest x. Sieve estimators suffer similarly in larger dimensions.

Parametric CEFs do not suffer this curse because they impose enough structure that all observations are informative about $m(\boldsymbol{x})$, no matter how far away $\boldsymbol{X}_i$ is from $\boldsymbol{x}$.

To alleviate the curse of dimensionality, one approach is to impose more structure than a fully flexible nonparametric CEF but less than a parametric CEF. Some examples

follow in Sections 19.2–19.4. Alternatively, even with a fully flexible nonparametric CEF, a $\sqrt{n}$ convergence rate is possible for certain finite-dimensional objects of interest.

19.2 Additive Model

One way to impose structure is to exclude interaction terms between certain pairs of regressors. If $\boldsymbol{X} = (X_1, X_2)$, then instead of the fully general $m(x_1, x_2)$, excluding interactions yields $m(x_1, x_2) = g_1(x_1) + g_2(x_2)$. More generally, with k regressors in $\boldsymbol{x} = (x_1, \ldots, x_k)'$ and no interactions, the additive model is

$$m(\boldsymbol{x}) = \sum_{j=1}^{k} g_j(x_j). \tag{19.1}$$

Although this initially seems more difficult because there are now k unknown functions $g_j(\cdot)$ instead of a single unknown $m(\cdot)$, it is easier because each is a function of a scalar.

As a compromise, some interactions can be maintained, like

$$m(\boldsymbol{x}) = g_1(x_1) + g_2(x_2) + g_3(x_3, \ldots, x_k), \quad m(\boldsymbol{x}) = \sum_{j=1}^{k/2} g_j(x_{2j-1}, x_{2j}), \quad \text{etc.} \tag{19.2}$$

Discussion Question 19.4 (additive wage model)**.** Consider a CEF model of log wage in terms of education, experience, IQ score, and a dummy for being male.

a) For each of the $(4)(3)/2 = 6$ possible interactions between pairs of regressors, say why you think it is important or not.
b) Do you need to model interactions with the male dummy nonparametrically? Explain.

19.3 Partially Linear Model

Another alternative is to specify part of $m(\cdot)$ parametrically. Let $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2')'$. The **partially linear model** (PLM) has

$$m(\boldsymbol{x}) = \boldsymbol{x}_1'\boldsymbol{\beta} + g(\boldsymbol{x}_2). \tag{19.3}$$

The interpretation of (19.3) is simple if interest is only in the slopes for $\boldsymbol{x}_1$, i.e., if $\boldsymbol{x}_2$ only contains control variables. That is, the slope wrt any element of $\boldsymbol{x}_1$ is the corresponding element of $\boldsymbol{\beta}$.

The $\sqrt{n}$ convergence rate for $\hat{\boldsymbol{\beta}}$ is another benefit.

However, the usual drawback applies: the less flexible structure may be less realistic. The PLM in (19.3) excludes interactions involving any $\boldsymbol{x}_1$ variable, and it imposes linearity in certain dimensions. The vector $\boldsymbol{x}_1$ could be augmented to include certain nonlinear-in-variables terms (nonlinear functions of observed variables, and interactions between

observed variables), but it still precludes interactions with the $\boldsymbol{x}_2$ control variables, and it imposes a parametric structure of the part of the CEF involving $\boldsymbol{x}_1$.

In principle, the PLM can be combined with the additive model approach, like

$$m(\boldsymbol{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \sum_{j=3}^{k} g_j(x_j). \tag{19.4}$$

If the entire function $m(\cdot)$ is of interest, then this helps. If only $\boldsymbol{\beta}$ is of interest, then the additional additive structure within $g(\cdot)$ does not improve the convergence rate.

19.4 Single Index Model

The **single index model** keeps a nonparametric transformation but restricts $\boldsymbol{x}$ to enter through a (single) linear **index** of the form $\boldsymbol{x}'\boldsymbol{\beta}$:

$$m(\boldsymbol{x}) = h(\boldsymbol{x}'\boldsymbol{\beta}), \tag{19.5}$$

where $h(\cdot)$ is an unknown function. As in the additive model, this reduces the domain of the unknown function from $\mathbb{R}^d$ to $\mathbb{R}$, staving off the curse of dimensionality.

Single index models for binary response models were proposed by Ichimura (1993) and Klein and Spady (1993), to allow the probit model's normal CDF $\Phi(\cdot)$ or the logit's $\Lambda(\cdot)$ to be replaced by an unknown function.

19.5 Product Kernels and Bases

To estimate a nonparametric function of multiple variables, usually a **product kernel** or **tensor product basis** is used. This means multiplying together univariate kernel or basis functions. See page 5573 of Chen (2007) for examples of tensor product bases and their approximation error rates.

There is also the bother of having multiple smoothing parameters, one for each dimension. For simplicity, you could use the same smoothing parameter value everywhere, but this may not work if the dimensions are scaled very differently (kernel) or have very different smoothness.

Exercises

Exercise E19.1. Find a paper that uses nonparametric regression (including as an intermediate step to computing a finite-dimensional functional of interest), or a partially linear CEF; provide a link to the paper. The paper must be either published in a respectable economics journal[1] or be unpublished but have an author who has previously published in such a journal (like any Mizzou econ professor); or if you really want, you can use an example from an econometrics textbook. Provide a critique (not "criticism") of the paper's application, including the following.

a. Replicate one of the regression results using the paper's data and (if provided) code.

b. What is the "economic" meaning of the estimate? How/does this help address the paper's economic question?

c. How much bigger or smaller would the smoothing parameter have to be to substantially change the economic meaning of the results? ("Smoothing parameter" being the bandwidth for local/kernel regression, or the penalty and/or number of terms for sieve regression.)

d. Use an alternative model selection procedure (that wasn't used in the paper) to select the optimal smoothing parameter. How different is it from the paper's smoothing parameter? How different are the corresponding estimates?

Exercise E19.2. a. Find a paper that uses nonparametric regression (including as an intermediate step to computing a finite-dimensional functional of interest), or a partially linear CEF; provide a link to the paper. The paper must be either published in a respectable economics journal[2] or be unpublished but have an author who has previously published in such a journal (like any Mizzou econ professor); or if you really want, you can use an example from an econometrics textbook.

b. Get their data and (if available) code, and replicate one of their estimates.

c. Construct a simulation DGP based on the empirical distributions in the data. You can make small changes to simplify the DGP, but it should be plausible that the observed data came from the DGP.

d. With your DGP, run 1000 simulation replications. In each replication, draw a new dataset from the DGP, run the paper's estimator, and also run the same estimator but with a different model selection procedure (and thus different bandwidth, penalty, and/or number of terms), like LOOCV, GCV, etc.

e. Compute the simulated RMSE for both estimators. If the object of interest is a scalar, then take the square root of the simulated MSE, where MSE is variance plus squared bias. If the object is a function, then take the root integrated MSE

[1] For example, in top 500 of https://ideas.repec.org/top/top.journals.all.html

[2] For example, in top 500 of https://ideas.repec.org/top/top.journals.all.html

(RIMSE); you can just average the MSE over a grid of $\boldsymbol{x}$ values instead of actually doing an integral.

f. Report and briefly discuss the performance of the original estimator and your (slightly) modified estimator.

Exercise E19.3. Find a paper that reports a result that can be replicated with OLS; provide a link to the paper. The paper must be either published in a respectable economics journal[3] or be unpublished but have an author who has previously published in such a journal (like any Mizzou econ professor); or if you really want, you can use an example from an econometrics textbook. This includes not only OLS for a cross-sectional regression, but linear probability models (LPM) for binary choice and FE/FD estimators (which essentially are running OLS after applying the FE or FD transformation to the data), and maybe other examples.

a. Get the paper's data and (if available) code and replicate one regression result (like one column in one table).

b. Discuss one particular coefficient estimate value (like, the coefficient on education): what is the statistical interpretation, and the (hoped for) economic interpretation?

c. With the same data and regressors, use nonparametric regression. You can choose kernel or sieve, choose your favorite model selection procedure, choose an appropriate structure (additive, partially linear, fully interactive, some combination). Explain why you choose the structure you do; why is it important to allow the flexibility you allow, and why is it not important to allow even more flexibility?

d. Compute a value that can be compared to the original estimate from part (b) above. For example, if part (b) is a coefficient in a linear-in-variables model, then you should *not* just take the coefficient from that regressor in a sieve/series regression, but instead compute something like an average partial effect (APE). How does the nonparametric estimate compare to the original estimate, in terms of the economic meaning?

[3] For example, in top 500 of https://ideas.repec.org/top/top.journals.all.html

Chapter 20
Nonparametric Regression in R

Unit learning objectives for this chapter

20.1. Become familiar with some nonparametric regression estimators in R [TLO 4]

This chapter contains a few simple examples with different packages in R.

Optional resources for this chapter

- James et al. (2013) Sections 7.8 and 8.3

20.1 Splines

20.1.1 Natural Cubic Splines

The following code generates an iid dataset and fits seven natural cubic B-spline models with different degrees of freedom, using function `ns()` in the `splines` package, which is part of core R (R Core Team, 2022). The SSR and then GCV is computed for each fit. The lowest GCV corresponds to the "best" model. As the flexibility increases, initially GCV decreases (better model), but then GCV starts to increase again when the model gets "too flexible." The `caret` package (Kuhn, 2020) is also used to run 5-fold cross-validation. The same model is chosen, with the same decreasing-then-increasing pattern. The numbers are differently scaled because validation root mean squared error (RMSE) is reported for 5-fold CV instead of penalized SSR.

Figure 20.1 shows the dataset and fit.

```
library(caret); library(splines)
set.seed(112358)
```

```
CEF <- function(x)sin(1.5*x)
n <- 100; X <- rnorm(n); Y <- CEF(X)+rnorm(n)
tc <- trainControl(method="cv", number=5)
DFs <- 1:7
cvRMSEs <- GCVs <- rep(NA,length(DFs))
for (iDF in 1:length(DFs)) {
  df <- DFs[iDF]
  cvRMSEs[iDF] <-
    train(x=ns(x=X, df=df), y=Y, method="lm",
          metric="RMSE", trControl=tc)$results$RMSE
  ret <- lm(Y~ns(x=X,df=df))
  SSR <- sum(ret$residuals^2)
  GCVs[iDF] <- (n/(n-ret$rank))^2 * SSR
}
(dfstar <- DFs[which.min(cvRMSEs)])
print(GCVs)
## [1] 138 136 121 117 120 120 123
print(cvRMSEs)
## [1] 1.17 1.16 1.12 1.06 1.09 1.09 1.07
```

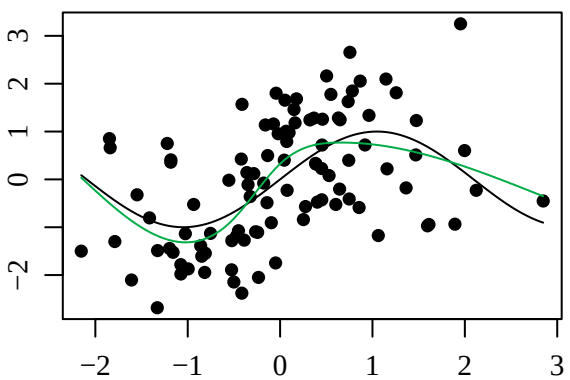


Figure 20.1: Natural cubic B-spline fit chosen by GCV (green) and true CEF (black).

20.1.2 Smoothing Spline

The following code generates the same iid data as in Section 20.1.1 but fits four cubic *smoothing* splines using the built-in function `smooth.spline()`. The smoothing spline controls flexibility by penalizing the (integrated) second derivative of the fitted function. The argument `cv=FALSE` tells it to choose the "model" (i.e., the second derivative penalty smoothing parameter) that minimizes the GCV criterion. Setting `cv=TRUE` instead uses LOOCV. The third model is too flexible (`df=n`). The fourth model is not flexible enough (`df=2`).

Figure 20.2 plots the dataset and the four fitted smoothing splines.

```
set.seed(112358)
CEF <- function(x)sin(1.5*x)
n <- 100; X <- rnorm(n); Y <- CEF(X)+rnorm(n)
df <- data.frame(X=X, Y=Y)
rets <- list()
titles <- c('GCV','LOOCV','Undersmoothed','Oversmoothed')
rets[[1]] <- smooth.spline(x=df$X, y=df$Y, cv=FALSE) #GCV
rets[[2]] <- smooth.spline(x=df$X, y=df$Y, cv=TRUE) #LOOCV
rets[[3]] <- smooth.spline(x=df$X, y=df$Y, df=n)
rets[[4]] <- smooth.spline(x=df$X, y=df$Y, df=2)
xx <- seq(from=-3, to=3, by=0.005)
mhatLOOCV <- predict(rets[[2]], x=xx) #has x and y
```

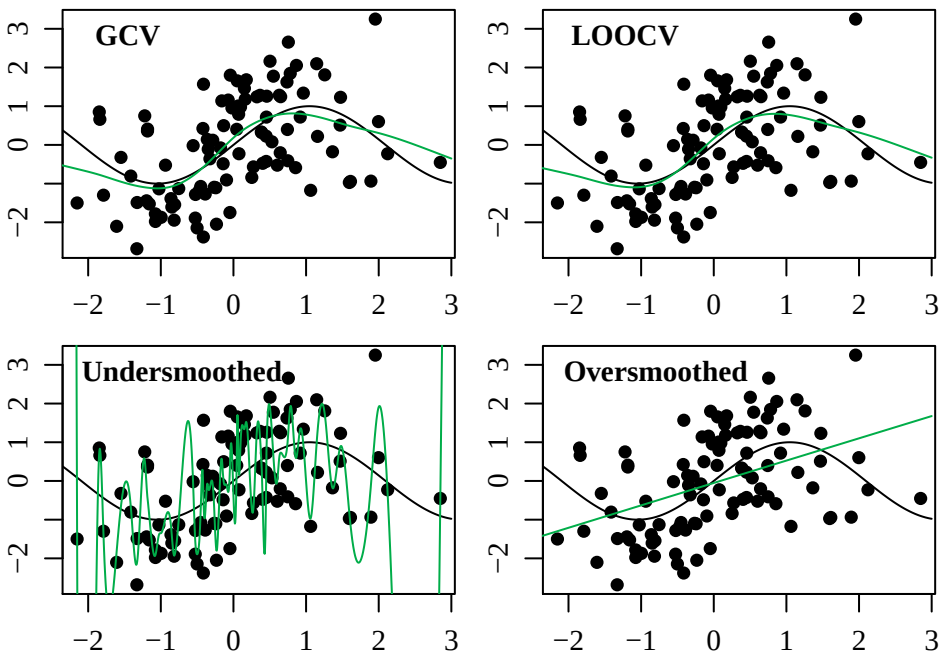


Figure 20.2: Smoothing spline estimates with same data but different penalties.

20.2 Local Polynomial Kernel Regression

The following code shows examples of local polynomial kernel regression with the same simulated data as in Sections 20.1.1 and 20.1.2. First, function `locpoly` in package `KernSmooth` (Wand, 2019) is used, with function `dpill()` for bandwidth selection. Second, function `npreg` in package `np` (Hayfield and Racine, 2008) is used, with function

npregbw for bandwidth selection. Since exdat=xx is specified when calling npreg(), the object returned from npreg() includes component mean with the estimated CEF evaluated at each point in xx. Although np is a bit more complicated, it can handle much more complex models, with many regressors, mixed data types (continuous, discrete, categorical), partially linear models, etc.

Figures 20.3 and 20.4 plot the different estimated CEFs.

```
library(KernSmooth)
set.seed(112358)
CEF <- function(x)sin(1.5*x)
n <- 100; X <- rnorm(n); Y <- CEF(X)+rnorm(n)
df <- data.frame(X=X, Y=Y)
rets <- list()
titles <- c('Local linear','Local cubic')
h <- dpill(x=df$X, y=df$Y) #for local linear only
h3 <- h*n^(4/45) #ad hoc--not recommended!
# below, degree=1 is local linear, =3 is local cubic
rets[[1]] <- locpoly(x=df$X, y=df$Y, degree=1, bandwidth=h)
rets[[2]] <- locpoly(x=df$X, y=df$Y, degree=3, bandwidth=h3)
xx <- seq(from=min(X),to=max(X),by=0.05)
for (ifig in 1:2) {
  if (ifig==2) par(mar=c(2,3,0.3,0.1))
  plot(x=df$X, y=df$Y, type='p', pch=16, cex=1, xlab='',
       ylab='', main='',cex.axis=CEXAXIS, cex.lab=CEXLAB)
  lines(x=xx, y=CEF(xx), col=1, lwd=1)
  lines(rets[[ifig]], col=ESTCOL, lwd=LWD)
  title(main=titles[ifig], line=-1, adj=0.1)
}
library(np)
set.seed(112358)
CEF <- function(x) sin(1.5*x)
n <- 100; X <- rnorm(n); Y <- CEF(X)+rnorm(n)
df <- data.frame(X=X, Y=Y)
xx <- seq(from=min(X),to=max(X),by=0.05)
bw <- npregbw(formula=Y~X, data=df, regtype='ll',
              ckertype='epanechnikov')
ret <- npreg(bws=bw, gradients=TRUE, exdat=xx)
summary(ret)  #output not shown
```

20.3 Random Forest and Neural Networks

The following example uses the same data as before but estimates the CEF with random forest or neural networks, using package caret for 5-fold CV model selection. (To clarify,

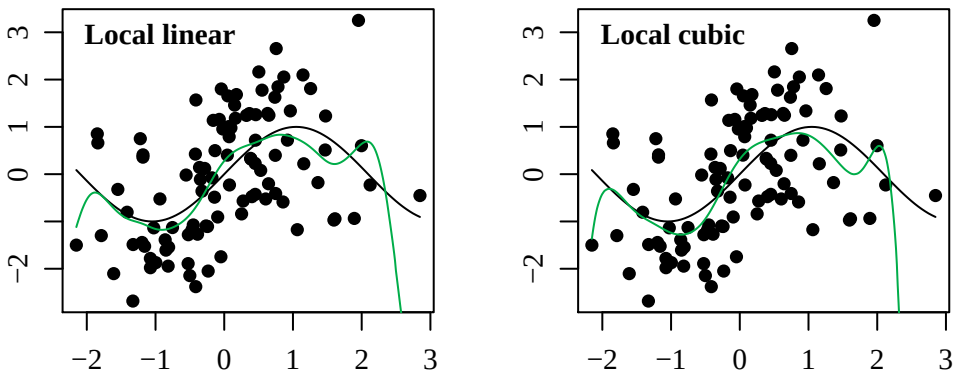


Figure 20.3: Local polynomial regression example, package KernSmooth.

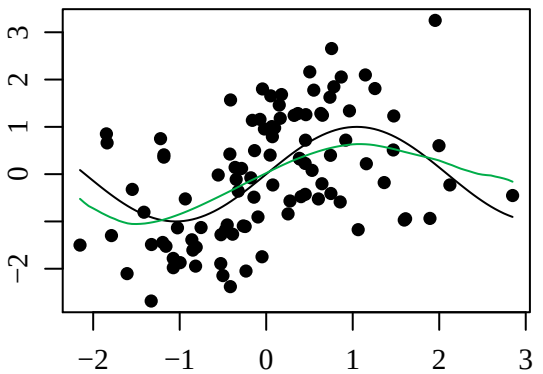


Figure 20.4: Local linear regression example, package np.

"neural network" is not a single CEF estimator, but an approach that includes many different variations with different strengths and different purposes.) With more than one X, `caret` could be used to select the tuning parameter `mtry` for `randomForest`, which is `method='rf'`.

Figure 20.5 shows the CEF estimates.

```
library(randomForest);  library(nnet);  library(caret)
set.seed(112358)
CEF <- function(x)sin(1.5*x)
n <- 100; X <- rnorm(n); Y <- CEF(X)+rnorm(n)
df <- data.frame(X=X, Y=Y)
xx <- seq(from=min(X),to=max(X),by=0.05)
rets <- list()
titles <- c('Random Forest','Neural Network')
rets[[1]] <- randomForest(x=df$X, y=df$Y, ntree=n*3)
trC <- trainControl(method='cv', number=5) # 5-fold cv
tg <- expand.grid(size=3+0:2*4, decay=10^(-2:0))
nf <- train(form=Y~X, data=df, method='nnet', maxit=100,
            tuneGrid=tg, trace=F, metric='RMSE', trControl=trC)
bt <- nf$bestTune
rets[[2]] <- nnet(formula=Y~X, data=df, linout=TRUE,
                  size=bt$size, decay=bt$decay, trace=FALSE)
mhat1 <- predict(object=rets[[1]], newdata=data.frame(X=xx))
mhat2 <- predict(object=rets[[2]], newdata=data.frame(X=xx))
```

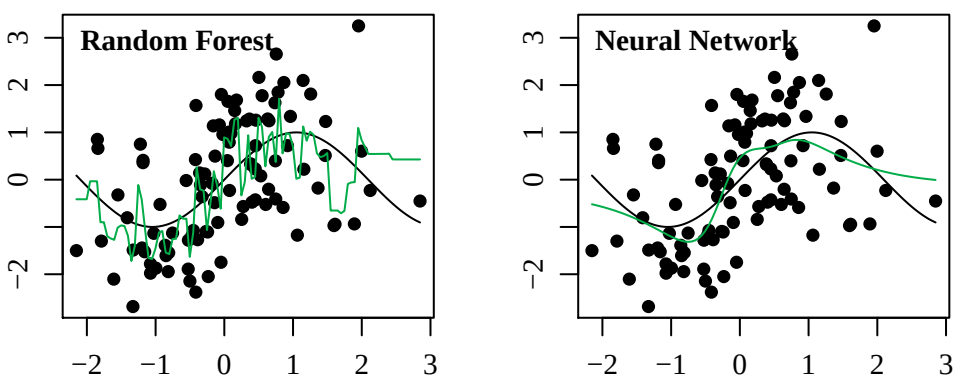


Figure 20.5: Random forest and neural network regression.

20.4 Multiple Regressors

The following code shows examples of partially linear and additive nonparametric regression. You can call `plot()` on the returned objects.

```
library(mgcv); library(splines)
set.seed(112358)
n <- 50
g1add <- function(x) { x^2 }
g1plm <- function(x) { x }
CEFadd <- function(x1,x2,x3) { g1add(x1)+x2^2+x3^2 }
CEFplm <- function(x1,x2,x3) { g1plm(x1)+(x2-x3)^2 }
df.add <- df.plm <- data.frame(X1=runif(n), X2=runif(n), X3=runif(n))
df.add$Y <- CEFadd(df.add$X1, df.add$X2, df.add$X3) + 0.1*rnorm(n)
df.plm$Y <- CEFplm(df.plm$X1, df.plm$X2, df.plm$X3) + 0.1*rnorm(n)
retadd.dfadd <- gam(Y~s(X1)+s(X2)+s(X3), data=df.add)
retadd.dfplm <- gam(Y~s(X1)+s(X2)+s(X3), data=df.plm)
retplm.dfadd <- gam(Y~X1+te(X2,X3), data=df.add)
retplm.dfplm <- gam(Y~X1+te(X2,X3), data=df.plm)
```

Part VI

Partial Identification

Introduction

This part concerns **identification**, which here means: what does the population joint distribution of observable variables tell us about our parameter of interest?

Why do we assume we know the joint distribution of all observable variables? With iid sampling, this distribution can be learned asymptotically. That is, with large enough n, we essentially "know" the joint distribution of observables. With certain non-iid types of sampling, it may not be reasonable to assume we learn this full distribution. For example, with covariance stationary time series, the (auto)covariances (but not necessarily other distributional features) can be learned asymptotically, so they are a more reasonable starting place for identification.

However, even knowing the joint population distribution may not be enough to learn θ. For example, "correlation does not imply causation": without additional assumptions, we cannot learn about causal parameters θ from regression slopes or even the full joint distribution of (Y, X).

Previously, "identification" has meant **point identification**. As in Definition 4.2, point identification means the population distribution of observables uniquely determines a single possible value (point) of the parameter θ. For example, the population median is point identified: the distribution of Y uniquely determines the median $\mathrm{Q}_{0.5}(Y)$.

If the population distribution of observables does not uniquely determine θ, but we can still learn something about θ, then it's called **partial identification** or **set identification**. (I've been convinced that "set identification" is the more proper term, but it seems "partial identification" is more popular, so I may use both.) That is, the population distribution narrows down the possible values of θ to some interval or set.

Partial identification results have been found in many econom(etr)ic fields, like game theory (IO, auctions), duration models, ordinal data models, and missing data, among others. I illustrate some basic concepts through examples.

The modern work on partial identification most directly comes from work by Manski starting in the late 1980s, although there were (much) earlier works that discussed the idea. Tamer (2010) provides a recent review.

Chapter 21

Missing Data

Unit learning objectives for this chapter

21.1. Develop intuition for when and why missing data can be problematic [TLO 2]

21.2. Describe how to construct and interpret worst-case bounds [TLO 1]

21.3. Judge whether missing data seems problematic in a particular setting [TLO 3]

Missing data is common in economics. For example, with survey data, sometimes people don't answer all the questions; somebody might report their age and education but not wage, for whatever reason. Or, in the FBI uniform crime reporting data, certain municipalities are missing certain categories of crime in certain years because they did not use the correct category definition (as Lonnie Hofmann found).

Notationally, consider Y_i and $\boldsymbol{X}_i$ for $i = 1, \ldots, n$, with $S_i = 1$ if both Y_i and $\boldsymbol{X}_i$ are observed for individual i, and $S_i = 0$ if either Y_i or $\boldsymbol{X}_i$ is missing. That is, S_i is an indicator of missing data for individual i. The magnitude of the missing data problem depends on the relationship between S_i and $(Y_i, \boldsymbol{X}_i)$.

There are at least three ways to deal with missing data. First, **complete case analysis** uses only observations i with no missing variable values, i.e., uses only observations with $S_i = 1$. Second, **imputation** tries to predict (impute) the missing values given observed values, and then computes estimates based on the "full" (imputed) data. Third, instead of a single point estimate, worst-case bounds provide a range of estimates that are collectively robust to a wide range of assumptions about the missing data. This approach goes back to Manski (1989). The benefit is that it requires less strict assumptions than complete case analysis or imputation.

Below, the focus is whether or not complete case analysis is appropriate under different mechanisms causing the missing data, and then the bounds approach is explored.

Optional resources for this chapter

- Textbook: MAR in Sections 19.4 and 19.8 of Wooldridge (2010)
- There is a fun dialog about MCAR and MAR between a fictional medical researcher and statistician here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4121561/
- Textbook: Hastie, Tibshirani, and Friedman (2009) Section 9.6 ("Missing Data")
- Textbook: Kaplan (2022b) Section 12.3.5

21.1 Best Case: MCAR

Complete case analysis works with data **missing completely at random** (MCAR), meaning that whether or not a value is missing is unrelated to either Y or $\boldsymbol{X}$. In our notation,

$$S_i \perp\!\!\!\perp Y_i, \boldsymbol{X}_i \quad \text{or equivalently} \quad \mathrm{P}(S_i = 1 \mid Y_i, \boldsymbol{X}_i) = \mathrm{P}(S_i = 1). \tag{21.1}$$

Selecting observations with $S_i = 1$ is essentially taking a random sample from within our original random sample.

Figure 21.1 shows how complete case analysis works well under MCAR. Here, Y is earnings and $X = 1$ if an individual has a college degree (and $X = 0$ otherwise). Everyone reports X_i. Some decide not to report Y_i, but the decision is unrelated to Y_i or X_i, i.e., MCAR holds as in (21.1). In the graphs, the blue shows the observed (non-missing) data and corresponding complete case OLS estimated CEF and (unconditional) complete case sample mean. The black shows the result if instead all Y_i were observed. The blue and black estimates are extremely similar, illustrating how complete case analysis is not biased by MCAR.

21.2 Fixable: MAR

Continue the example with college dummy X always observed but earnings Y sometimes unobserved, in which case $S = 0$.

Consider a weaker, conditional version of MCAR, where independence only holds within each X subpopulation (college, not). Formally,

$$S \perp\!\!\!\perp Y \mid X. \tag{21.2}$$

("Conditional on X, S and Y are independent," or "S and Y are conditionally independent given X.") That is, within the $X = 0$ subpopulation, missingness ($S = 0$) is independent of Y; and this is also true in the $X = 1$ subpopulation. This is an example of **missing**

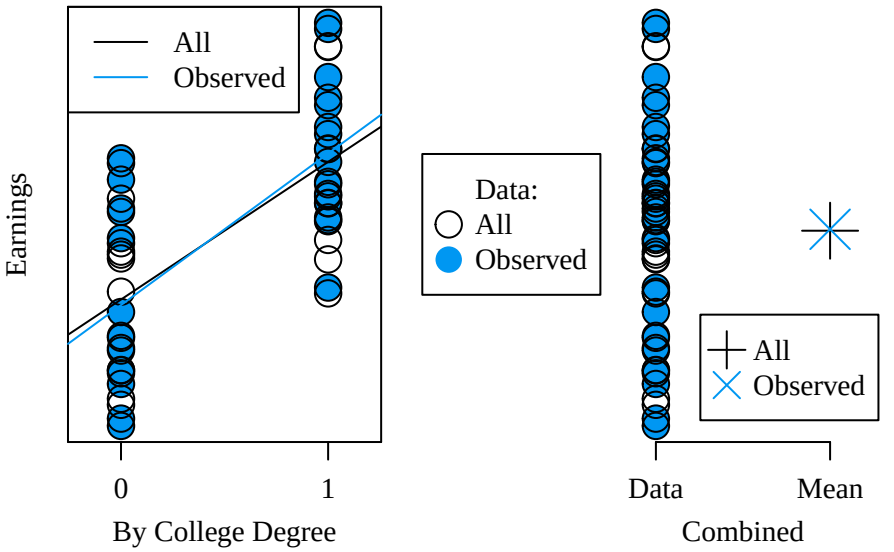


Figure 21.1: Missing completely at random: no bias of complete case analysis.

at random (MAR), where missingness ($S = 0$) is "random" (unrelated to the missing values) conditional on certain variables that are always observed.

21.2.1 Complete Case Estimation

With MAR, complete case CEF estimation is consistent, but the complete case sample mean is not. But, the CEF can be manipulated to get the unconditional mean (Section 21.2.2). Unfortunately, complete case linear projection is also biased (Section 21.2.3).

Figure 21.2 (right panel) illustrates the following example in which the complete case sample mean is biased. Everyone with $X_i = 0$ reports Y_i, but many individuals with $X_i = 1$ do not. Because individuals with a college degree tend to have higher earnings, the missing Y_i values tend to be high. Thus, if we only average observed Y_i, then our estimate of $\mathrm{E}(Y)$ has downward bias.

Figure 21.2 also shows that nonparametric CEF estimation is not biased. Because X is binary, simple OLS estimates the CEF nonparametrically, but more generally this result only holds for nonparametric regression. From (21.2), within the $X = 1$ subpopulation, missingness of Y is unrelated to the value of Y, so the complete case subpopulation mean earnings $\mathrm{E}(Y \mid X = 1, S = 1)$ equals the true subpopulation mean earnings $\mathrm{E}(Y \mid X = 1)$. Generally, $S \perp\!\!\!\perp Y \mid X$ implies that for any x subpopulation, the distribution of Y is the same for individuals who report their value ($S = 1$) and those who don't ($S = 0$), so

$$\overbrace{\mathrm{E}(Y \mid X = x, S = 1)}^{\text{CEF among "reporters"}} = \overbrace{\mathrm{E}(Y \mid X = x)}^{\text{CEF for everyone}}. \tag{21.3}$$

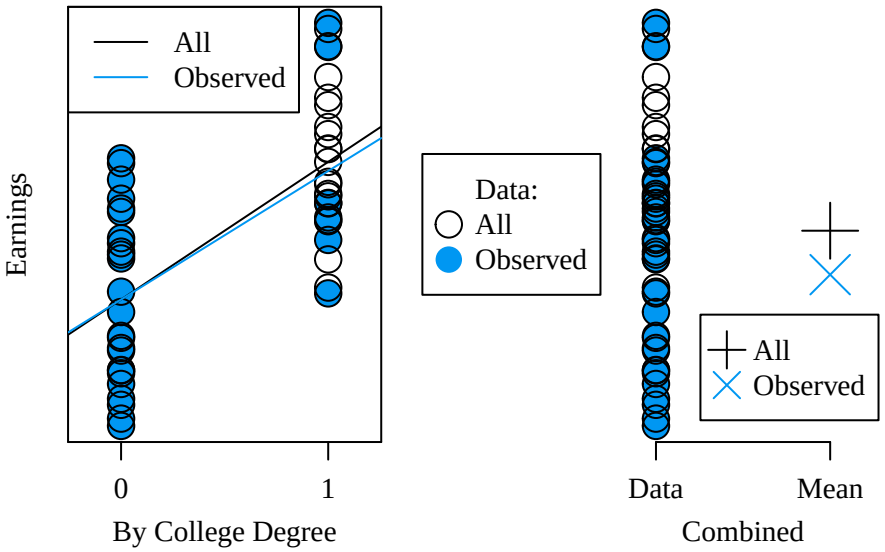


Figure 21.2: Missingness of Y based on X: example with sample mean biased, OLS not.

Fundamentally, complete case nonparametric regression consistently estimates $\mathrm{E}(Y \mid X = x, S = 1)$, and the identification result in (21.3) says this is the same as $\mathrm{E}(Y \mid X = x)$ given MAR.

21.2.2 Inverse Probability Weighting

Although the sample mean is biased, the conditional means can be combined to recover the unconditional mean. In our binary X example, given MAR,

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}(Y \mid X = 0)\,\mathrm{P}(X = 0) + \mathrm{E}(Y \mid X = 1)\,\mathrm{P}(X = 1) \\ &= \mathrm{E}(Y \mid X = 0, S = 1)\,\mathrm{P}(X = 0) + \mathrm{E}(Y \mid X = 1, S = 1)\,\mathrm{P}(X = 1). \end{aligned} \tag{21.4}$$

All four terms on the right-hand side are observable because X is always observable (and Y is always observable conditional on $S = 1$). The probability $\mathrm{P}(X = 1)$ can be estimated by the sample proportion of observations with $X_i = 1$ (regardless of whether or not Y_i is observed). Similarly, the sample proportion with $X_i = 0$ estimates $\mathrm{P}(X = 0)$. Complete case OLS can estimate $\mathrm{E}(Y \mid X = 0, S = 1)$ and $\mathrm{E}(Y \mid X = 1, S = 1)$. Then, we estimate $\mathrm{E}(Y)$ by plugging in these estimates for the four right-hand side terms in (21.4).

This approach is a special case of **inverse probability weighting** (IPW). Consider Figure 21.2. The number of sampled individuals with $X_i = 0$ equals the number with $X_i = 1$. To make the math easier, say there are 4 of each. Now, Y_i is observed for all sampled individuals with $X_i = 0$, but only for 1 individual with $X_i = 1$ (out of 4). The complete case sample mean averages 4 values of Y_i from no-college individuals with only 1 value of Y_i from college individuals, even though there are equal numbers of college

and no-college individuals in the sample. The probability weighting is a way to fix this discrepancy. The one value of Y_i essentially represents all 4 individuals with a college degree. So, instead of counting it once, we could count it 4 times. That is, if $X_i = 0$ for $i = 1, 2, 3, 4$ and $X_i = 1$ for $i = 5, 6, 7, 8$, and Y_i is observed only for $i = 1, 2, 3, 4, 5$, then our estimate could be

$$\frac{1}{8}(Y_1 + Y_2 + Y_3 + Y_4 + Y_5 + Y_5 + Y_5 + Y_5) = \frac{1}{8}(Y_1 + Y_2 + Y_3 + Y_4 + 4Y_5). \tag{21.5}$$

This is essentially a type of imputation, filling in the missing Y_i with our best guess, which is the Y_i for another individual from the same group. With MAR, this guess is justified; without MAR, it may be a bad guess.

The weight of 4 on Y_5 in (21.5) can be interpreted as an inverse probability of $S = 1$ (Y is observed) given $X = 1$. There are 4 individuals with $X_i = 1$, of whom only one has observed Y_i ($S_i = 1$), so the probability of having observed Y (i.e., having $S = 1$) among the $X = 1$ subpopulation is estimated to be 1 out of 4, or 1/4. That is,

$$\hat{\mathrm{P}}(S = 1 \mid X = 1) = 1/4, \quad \frac{1}{\hat{\mathrm{P}}(S = 1 \mid X = 1)} = 4, \tag{21.6}$$

where the inverse probability 4 is the weight that appears in (21.5). The expression $\hat{\mathrm{P}}(S = 1 \mid X = 1)$ is the estimated probability of having an observable Y value ($S = 1$) given $X = 1$.

More generally, the IPW estimator of $\mathrm{E}(Y)$ is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{S_i Y_i}{\hat{\mathrm{P}}(S = 1 \mid X = X_i)}. \tag{21.7}$$

21.2.3 Linear Projection Estimation

Unless the CEF is properly specified (like with a single binary X), complete case OLS is not consistent for the population linear projection. Even if (21.2) holds, different amounts of missing Y_i can lead to (very) different OLS slope estimates.

For example, let $X \in \{0, 1, 2\}$ with CEF values $m(0) = m(1) = 40$ and $m(2) = 60$. The population CEF slope can be anywhere between 0 and 20; if most individuals have $X = 0$ or $X = 1$, then it's close to $m(1) - m(0) = 0$, whereas if most individuals have $X = 1$ or $X + 2$, then it's closer to $m(2) - m(1) = 20$. However, (21.2) does not restrict the relationship between X and S. If everybody with $X_i = 0$ or $X_i = 1$ reports her Y_i, but nobody with $X_i = 2$ does, then OLS estimates a slope near zero. If nobody with $X_i = 0$ reports Y_i, but everybody with $X_i = 1$ or $X_i = 2$ does, then OLS estimates a slope near 20. Both examples satisfy (21.2), yet OLS $\hat{\beta}_1$ is very different. Generally, even with MAR, conditioning on $S = 1$ affects the linear projection if $\boldsymbol{X}$ has a different distribution in the $S = 1$ and $S = 0$ subpopulations, in which case complete case OLS is not consistent.

21.3 Worst Case: Non-Ignorable

The term "non-ignorable" suggests we can't ignore the missing data problem. Section 21.3.1 explains why, and Section 21.3.2 suggests one way to cope.

21.3.1 The Problem

If data are missing in a way that relates to the missing values themselves, then it is very difficult or impossible to avoid bias. This type of missing data is sometimes called **non-ignorable**.

Consider non-ignorable missing data in our example from before. Again, let $X = 1$ if somebody has a college degree and $X = 0$ otherwise, and Y is annual earnings, which is generally higher when $X = 1$ than $X = 0$. You have survey data where everyone reports X (accurately), but some people do not report Y. Specifically, people with very high earnings are less likely to report it. So, whether or not Y is missing depends on the value of Y: missingness is non-ignorable.

In this example, unlike with MCAR or MAR, even the CEF is biased. For both $X = 0$ and $X = 1$ subpopulations, the highest Y values are missing, so the observable conditional means are lower than the true conditional means. Further, most people with very high earnings (Y) have a college degree ($X = 1$), so this bias affects $\mathrm{E}(Y \mid X = 1)$ more than $\mathrm{E}(Y \mid X = 0)$. Thus, the CEF slope $\mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)$ is also downward biased.

Figure 21.3 illustrates this example. With complete case analysis, both the OLS slope and sample mean are biased downward. The OLS intercept is very slightly downward biased, too, because the very top Y_i when $X_i = 0$ are missing.

21.3.2 Worst-Case Bounds

This section uses the following notation and setup. Let scalar rv Y^* have CDF $F^*_Y(\cdot)$. Let $S = 1$ if Y^* is observed, and $S = 0$ if Y^* is missing. Individuals are randomly sampled from the population, so (Y_i^*, S_i) are iid, but Y^* is not always observed. The observable variables are S_i and Y_i, where $Y_i = Y_i^*$ if $S_i = 1$ but Y_i is missing if $S_i = 0$.

We want to learn about the population distribution of Y^* without imposing any assumption like MCAR or MAR. Unfortunately, without any such assumptions, the joint population distribution of observables (Y, S) does not uniquely determine the mean of Y^*. Instead, it determines a set of values, i.e., there is set identification. The best we can do is to learn about this identified set (the population object of interest) from the imperfect data.

Identification refers to population distributions and values, but intuition can be developed thinking about samples. (Recall samples can be thought of as discrete population distributions.) Hence, DQs 21.1 and 21.2 are not about "identification" per se but hopefully illuminate relevant concepts.

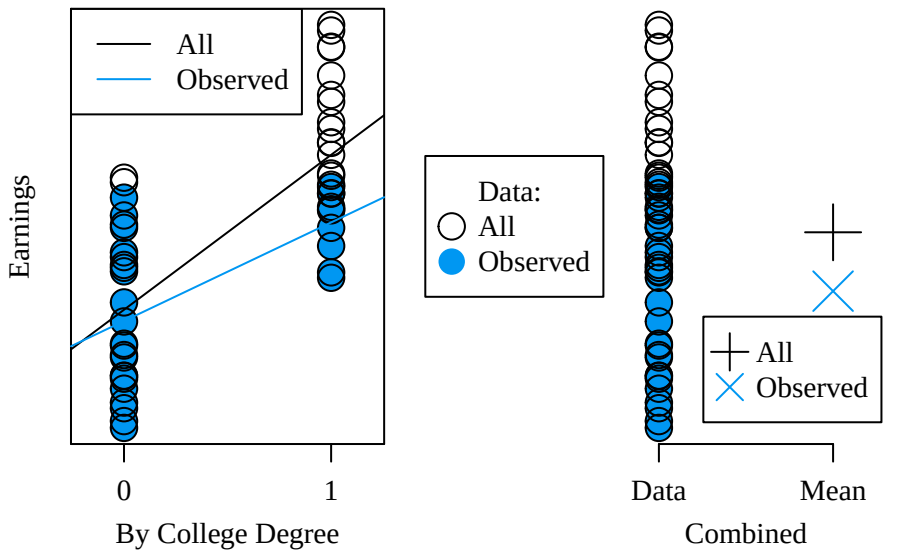


Figure 21.3: Non-ignorable missing data: bias of both OLS and sample mean.

Discussion Question 21.1 (bounds for binary sample mean)**.** Let binary unobserved $Y^* = \mathbb{1}\{\text{employed}\}$. You observe $n = 5$ values of Y_i: $(1, 1, 0, 1, \texttt{NA})$. (So the S_i are $(1, 1, 1, 1, 0)$.)

a) What are the values of Y_i^* for $i = 1, 2, 3, 4$?
b) What are the possible values of Y_5^*?
c) What's the smallest possible value of $\bar{Y}^* \equiv (1/5)(Y_1^* + Y_2^* + Y_3^* + Y_4^* + Y_5^*)$?
d) What's the largest possible value of $\bar{Y}^*$?

Discussion Question 21.2 (bounds for sample mean education)**.** Let unobserved Y^* be years of education, where $0 \leq Y^* \leq 21$. You observe $n = 5$ values of Y_i: $(12, 12, 11, 18, \texttt{NA})$. (So the S_i are $(1, 1, 1, 1, 0)$.)

a) What are the values of Y_i^* for $i = 1, 2, 3, 4$?
b) What are the possible values of Y_5^*?
c) What's the smallest possible value of $\bar{Y}^* \equiv (1/5)(Y_1^* + Y_2^* + Y_3^* + Y_4^* + Y_5^*)$?
d) What's the largest possible value of $\bar{Y}^*$?

Discussion Question 21.3 (bounds for sample median)**.** Consider the "true" sample median $\widehat{\mathrm{Q}}_{0.5}(Y^*)$. (With $n = 5$, this is the "middle" observation when sorted from low to high.)

a) In the setup of DQ 21.1, what are the smallest and largest possible values of the sample median of Y^*?
b) In the setup of DQ 21.2, what are the smallest and largest possible values of the sample median of Y^*?
c) Repeat (b) but if it's possible to get up to 120 years of education, so $0 \leq Y^* \leq 120$.

d) Imagine there are no bounds on Y^*, $-\infty < Y^* < \infty$, and $Y_i = 10$ for $i = 1, 2, 3, 4$, but $Y_5 =$ `NA` (missing). Is it still possible to get a lower and/or upper bound for the sample median? How?

The type of bounds in DQs 21.1 and 21.2 are called **worst-case bounds**, an idea from Manski. "Worst-case" suggests that they are probably conservative: they use the most extreme possible pattern of missing values, like assuming that *all* missing values were from unemployed individuals, and then alternatively assuming all were from employed individuals. But, for the same reason, they are very robust.

The worst-case bounds can be "more informative" ("tighter" bounds) in larger samples with a smaller proportion of missing values. For example, let $n = 1000$, among whom 900 are employed, 80 are not, and 20 do not answer the survey question (or do not reply at all to our request that they take the survey). Then, the bounds are more informative: $900/1000 \leq \bar{Y}^* \leq (900 + 20)/1000$, i.e., $0.90 \leq \bar{Y}^* \leq 0.92$. For comparison, the complete case sample average is $900/(900 + 80) = 0.918$. This single value is more specific, but its validity requires the very strong MCAR assumption about why data are missing. Conversely, if we think it's crazy to allow all 20 non-responses to be unemployed, then we may feel the 0.90 lower bound is too conservative.

Discussion Question 21.4 tries to use our insights from DQ 21.1 to find bounds for the population employment probability, i.e., find a partial identification result.

Discussion Question 21.4 (bounds for binary population mean)**.** Same as DQ 21.1, but in the population. We know the population distribution of (Y, S) but not binary Y^*. (Recall that this implies we also know the marginal and conditional distributions of Y and S.) What are the worst-case bounds on $\mathrm{P}(Y^* = 1) = \mathrm{E}(Y^*)$, i.e., bounds that do not assume anything about why/how data are missing? That is, find a and b such that $a \leq \mathrm{P}(Y^* = 1) \leq b$, where a and b are determined by the (Y, S) distribution. Hint: $\mathrm{E}(Y^*) = \mathrm{E}(Y^* \mid S = 1)\,\mathrm{P}(S = 1) + \mathrm{E}(Y^* \mid S = 0)\,\mathrm{P}(S = 0)$.

Discussion Question 21.5 (bounds for population mean)**.** Same setup as DQ 21.4; consider different types of Y^* variables.

a) What are bounds on $\mathrm{E}(Y^*)$ if Y^* is a binary employment indicator? Hint: recall DQs 21.1 and 21.4.
b) What are bounds on $\mathrm{E}(Y^*)$ if Y^* is years of education? Hint: recall DQ 21.2.
c) Same but Y^* is annual household consumption.
d) Same but for any variable $Y^* \in \mathbb{R}$ with unbounded support (no maximum or minimum value).

Discussion Question 21.6 (bounds for population median)**.** Like DQ 21.5 but for the median. Hint: recall DQ 21.3.

a) What are bounds on $\mathrm{Q}_{0.5}(Y^*)$ if Y^* is a binary employment indicator?
b) Same as (a) but Y^* is years of education.
c) Same as (a) but Y^* is consumption.
d) Same as (a) but for any variable Y^* with unbounded support $\mathbb{R}$.

Discussion Question 21.7 (bounds for population CDF)**.** Let $Y^* \in \mathbb{R}$. For a fixed value y, let $W^* \equiv \mathbb{1}\{Y^* \leq y\}$.

a) Bounds on $\mathrm{P}(W^* = 1)$?
b) Bounds on $\mathrm{P}(Y^* \leq y)$?
c) Bounds on $F_Y^*(\cdot)$? (A lower bound for a function is itself a function, as is the upper bound function.)

Discussion Question 21.8 (IQR bounds)**.** Assume $F_1(\cdot) \leq F^*(\cdot) \leq F_2(\cdot)$.

a) Bounds for $\mathrm{Q}_{0.75}(Y^*)$?
b) Bounds for $\mathrm{Q}_{0.25}(Y^*)$?
c) Bounds for the interquartile range, $\mathrm{Q}_{0.75}(Y^*) - \mathrm{Q}_{0.25}(Y^*)$?

Stoye (2010) gives more results like DQ 21.8.

Often, worst-case bounds are computed in addition to point estimates that require stronger assumptions. In the previous examples, worst-case bounds could be computed in addition to the complete case average (assuming MCAR) or the IPW estimator (assuming MAR). This shows the worst-case bounds that come "only from the data," and lets us see how much this changes under the stronger assumption. Of course, the stronger assumption may indeed be correct. However, we may want to think more critically about it if the assumption (rather than "just the data") is primarily driving the final result. (Similar to comparing OLS-type results with nonparametric results.)

The worst-case bounds approach can be extended to conditional distributions and regression. For example, consider binary $X_i \in \{0, 1\}$ representing low or high education. Again let $Y_i = 1$ if individual i is employed. Assume X_i is always observed but some Y_i are missing. To compute an upper bound for the OLS slope, plug in $Y_i = 0$ for all missing Y_i values when $X_i = 0$, and plug in $Y_i = 1$ for missing values when $X_i = 1$; then run OLS. To compute a lower bound, plug in $Y_i = 1$ when $X_i = 0$ and $Y_i = 0$ when $X_i = 1$; then run OLS. This approach can probably be extended to non-binary and/or multiple X, too, although I admit I don't know for sure.

21.4 R Code

Caution: by default, most commands in Stata and functions in R drop all observations (rows in your dataset) with any missing variable(s) automatically, without any error or warning message. That is, they assume you want complete case analysis. You can still figure out whether or not any observations were dropped. You can also tell R to behave differently if it encounters `NA` values. You can either do this through `options()` to change the default, or for a specific `lm` (or whatever function) call through the `na.omit` argument. See the code below.

```
n <- 5;  set.seed(112358);  options(digits=3)
Y <- rnorm(n); X <- rnorm(n)
Y[2] <- X[3] <- NA #missing values
r <- lm(Y~X) #no hint of missing/dropped obs
```

```
coef(r) #still no hint:
## (Intercept)           X
##       0.591       0.704

nrow(r$model) #aha: not n rows!
## [1] 3

#summary(r) #"(2 observations deleted due to missingness)"
options("na.action") #print current default (usually na.omit)
## $na.action
## [1] "na.omit"

predict(lm(Y~X, na.action=na.omit)) # complete case
##       1       4       5
## -0.7037 -0.0139  1.1147

predict(lm(Y~X, na.action=na.exclude)) #fill in NA
##       1       2       3       4       5
## -0.7037      NA      NA -0.0139  1.1147

lm(Y~X, na.action=na.fail) # give an error if NAs in data
## Error in na.fail.default(list(Y = c(-0.471, NA, 0.530,  :
##  missing values in object

options(na.action=na.fail) #set default to na.fail
lm(Y~X) #now gives error as default (if NA values)
## Error in na.fail.default(list(Y = c(-0.471, NA, 0.530,  :
##  missing values in object
```

Exercises

Exercise E21.1. Consider scalars Y and X, where either Y or X or both may be missing. Let Y^* and X^* be the true values that are never missing. So, either $Y = Y^*$ or Y is missing; similarly, either $X = X^*$ or X is missing. Assume iid sampling, $i = 1, \ldots, n$. It may be helpful to play around with example datasets (that you create) in R; show the data scatterplot, run `lm()` and `rq()`, change one point's value, etc. Below, "bounds" means "worst-case bounds." Section 4.6 may be helpful.

a. Imagine Y_i is missing for a single i. How/can you compute bounds for the sample average $\bar{Y}^*$? How/do the bounds depend on the assumed bounds of Y (i.e., its support)?

b. Imagine Y_i is missing for a single i. How/can you compute bounds for the sample median $\widehat{\mathrm{Q}}_{0.5}(Y^*)$? How/do the bounds depend on the assumed bounds of Y (i.e., its support)?

c. Assume binary $X \in \{0, 1\}$. Imagine a single Y_i is missing, but all X_i are observed. How/can you compute bounds for the OLS slope estimate based on the true Y_i^* and X_i^*? Hint: recall that with binary X, the OLS slope can be written as a difference of conditional means, $\hat{\beta}_{\mathrm{OLS}} = \widehat{\mathrm{E}}(Y \mid X = 1) - \widehat{\mathrm{E}}(Y \mid X = 0)$.

d. Assume binary $X \in \{0, 1\}$. Imagine a single Y_i is missing, but all X_i are observed. How/can you compute bounds for the $\tau = 0.5$ QR slope estimate based on the true Y_i^* and X_i^*? Hint: recall that with binary X, the QR slope can be written as a difference of conditional quantiles, $\hat{\beta}_\tau = \widehat{\mathrm{Q}}_\tau(Y \mid X = 1) - \widehat{\mathrm{Q}}_\tau(Y \mid X = 0)$.

e. Answer parts (c) and (d) if instead a single X_i is missing (and no Y_i are missing).

f. Imagine $n = 2$, $X_1 = Y_1 = 0$, $Y_2 = 1$ but X_2 is missing. Assume $a \leq X_2 \leq b$. What are lower and upper bounds on the OLS slope estimate based on Y_i^* and X_i^*? Explain, including why either bound does/not depend on (a, b). Hint: draw a picture.

g. Imagine $n = 2$, $X_1 = 0$ but Y_1 is missing, $Y_2 = 1$ but X_2 is missing. Assume $a \leq X_2 \leq b$ and $c \leq Y_1 \leq d$. What are lower and upper bounds on the OLS slope estimate based on Y_i^* and X_i^*? Explain, including why either bound does/not depend on (a, b, c, d). Hint: draw a picture.

h. Imagine a dataset (with $n > 500$) with a single missing X_i and all Y_i observed. Qualitatively (e.g., infinity, zero, big positive, small negative), what are the lower and upper bounds on the OLS slope estimate based on Y_i^* and X_i^*? Hint: draw a picture.

Chapter 22
Interval Data

Unit learning objectives for this chapter

22.1. Develop intuition for learning from interval-valued data, in the population and in the sample [TLO 2]

22.2. In simple examples, describe the identified set and how to estimate it, and interpret both [TLO 1]

Sometimes variables are reported as intervals instead of values. For example, instead of somebody reporting exact hourly wage, they report whether it's in the interval $[5, 10)$ or $[10, 15)$ or something. That is, instead of observing the true value Y^*, the observed variables are (Y_1, Y_2), where $Y_1 \leq Y^* < Y_2$. Assume Y^* is continuous so we won't worry about $\leq$ versus $<$.

Discussion Question 22.1 (interval-valued mean: identification)**.** The following is somewhat similar to DQ 21.4. Continute the notation where Y^* is the true but unobserved value, and the observed variables (Y_1, Y_2) satisfy $Y_1 \leq Y^* < Y_2$.

a) Is $\mathrm{E}(Y^*)$ point identified from the distribution of (Y_1, Y_2)? Why not?
b) What are bounds for $\mathrm{E}(Y^*)$ given the distribution of (Y_1, Y_2)?

Discussion Question 22.2 (interval-valued mean: estimation)**.** Continue the setup and notation from DQ 22.1.

a) Given DQ 22.1, how would you estimate the bounds? (That is, given the population "identified set" you proposed before, how would you estimate it from data?)
b) How would you interpret your proposed estimator? Is it like a confidence interval, or different?

Discussion Question 22.3 (interval regression: identification)**.** Now, scalar X is also observed. For identification, assume the joint distribution of (Y_1, Y_2, X) is known.

a) What are bounds for the CEF evaluated at a single point, i.e., for $m(x) = \mathrm{E}(Y^* \mid X = x)$ for a single x?
b) What are bounds for the CEF (as a function), $m(\cdot)$?

Discussion Question 22.4 (interval regression: estimation)**.** Continue from DQ 22.3.
a) How could you estimate your proposed bounds from DQ 22.3?
b) How do you interpret your estimated bounds? Is it like a uniform confidence band, or something else?

Discussion Question 22.5 (interval regression slope: identification)**.** Use your results from DQ 22.3 below. Let $x_2 > x_1$ be two points of evaluation. Assume $m(\cdot)$ is continuously differentiable.
a) What are bounds for $\mathrm{E}(Y^* \mid X = x_2) - \mathrm{E}(Y^* \mid X = x_1)$?
b) What are the limits of your bounds as $x_2 \downarrow x_1$?
c) What are bounds for $m'(x)$? Hint: $m'(x) = \lim_{x_2 \downarrow x_1} [m(x_2) - m(x_1)]/(x_2 - x_1)$.

Exercises

Exercise E22.1. Assume scalar $X \geq 0$. Define vector $\boldsymbol{X} = (1, X)'$. The observables are (Y_1, Y_2, X) with $Y_1 \leq Y^* \leq Y_2$ for latent Y^*. Assume iid sampling. For vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, let $\boldsymbol{a} \leq \boldsymbol{b}$ mean that the inequality holds element-wise, i.e., $a_j \leq b_j$ for each $j = 1, \ldots, \dim(\boldsymbol{a})$.

a. Consider the linear projection model in error form, $Y^* = \boldsymbol{X}'\boldsymbol{\beta} + U$ with $\mathrm{E}(\boldsymbol{X}U) = \boldsymbol{0}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Show that $\mathrm{E}(\boldsymbol{X}Y^*) = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{\beta}$.

b. Propose bounds for $\mathrm{E}(\boldsymbol{X}Y^*)$ using any feature(s) of the population joint distribution of observables. Hint: separately consider the two elements of the vector $\mathrm{E}(\boldsymbol{X}Y^*) = [\mathrm{E}(Y^*), \mathrm{E}(XY^*)]'$, and remember $X \geq 0$.

c. Denote your bounds in part (b) as $\boldsymbol{\mu}_L$ and $\boldsymbol{\mu}_U$, where $\boldsymbol{\mu}_L \leq \mathrm{E}(\boldsymbol{X}Y^*) \leq \boldsymbol{\mu}_U$. Given part (a), your bounds imply

$$\boldsymbol{\mu}_L \leq \mathrm{E}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{\beta} \leq \boldsymbol{\mu}_U. \tag{22.1}$$

Re-write (22.1) as four inequalities of the form $\beta_1 \leq a - b\beta_0$ or $\beta_1 \geq a - b\beta_0$, where the "intercept" a and "slope" b are in terms of (moments of) observable variables (i.e., not Y^*).

d. Draw an example graph of your four inequalities in part (c). That is, your graph's horizontal axis is β_0, the vertical axis is β_1; draw the four lines of the form $\beta_1 = a - b\beta_0$ (the boundaries of the inequalities), and shade/fill in the region where all four inequalities are satisfied.

e. If we are only interested in the slope β_1, can we get bounds on β_1 using this approach? You don't have to derive such bounds, just explain why or why not.

f. With $\boldsymbol{\beta} \in \mathbb{R}^2$, the bounds generate four lines in $\mathbb{R}^2$ that determine a quadrilateral subset of $\mathbb{R}^2$ containing all values of $\boldsymbol{\beta}$ consistent with the joint distribution of observables, i.e., the identified set for $\boldsymbol{\beta}$. Without necessarily solving for the bounds exactly, what is the corresponding geometry of the bounds when instead $\boldsymbol{\beta} \in \mathbb{R}^k$ for general $k > 2$? Is it still possible to get bounds for a single coefficient like β_1? Think about the structure of the generalization of (22.1) in that case; how many equations, what shape, etc. Hint: to develop intuition, you could start with $k = 3$.

Exercise E22.2. Assume binary $X \in \{0, 1\}$. The observables are (Y_1, Y_2, X, Z) with $Y_1 \leq Y^* \leq Y_2$ for latent Y^*. Assume iid sampling. Write the τ-CQF as $q_\tau(\cdot)$, where $q_\tau(x) = \mathrm{Q}_\tau(Y^* \mid X = x)$.

a. Explain why $q_\tau(0)$ and $q_\tau(1)$ are not point identified.

b. Propose bounds for $q_\tau(0)$ and $q_\tau(1)$, in terms of (features of) the population joint distribution of observables.

c. Propose estimators for your bounds in part (b).

d. Propose bounds on the quantile regression "slope" $q_\tau(1) - q_\tau(0)$.

e. Propose estimators of your bounds in part (d) based on quantile regression coefficient estimates.

Now additionally define observable binary instrument $Z \in \{0, 1\}$.

f. Consider the local average treatment effect (LATE) estimand, $[\mathrm{E}(Y^* \mid Z = 1) - \mathrm{E}(Y^* \mid Z = 0)]/[\mathrm{E}(X \mid Z = 1) - \mathrm{E}(X \mid Z = 0)]$. For simplicity, assume the denominator is strictly positive. Propose bounds for the LATE.

g. Propose estimators of your LATE bounds.

Chapter 23
Ordinal Data

Unit learning objectives for this chapter

23.1. Develop intuition about what can be learned from ordinal data about continuous latent distributions [TLO 2]

23.2. Describe which types of latent relationship are identified with ordinal data under certain assumptions, and describe the assumptions [TLO 1]

Optional resources for this chapter

- Kaplan and Zhao (2022) is the basis of this chapter
- R code: `https://kaplandm.github.io`

23.1 Latent Variable Framework

Imagine you observe an **ordinal** random variable H: its possible values are categories that are ordered (from low to high, or worst to best) but do not have a cardinal value (like 7 dollars or -90 utils). For example, H could be self-reported health status, with possible values "poor," "fair," "good," "very good," and "excellent." For convenience these may be coded as $H = 1$ through $H = 5$, but such numbers have no cardinal meaning; e.g., the fact that $4/2 = 2$ does not mean "very good" is exactly twice as good as "fair." Other ordinal variables include bond ratings, political indices (like democracy or civil rights), subjective well-being (happiness), consumer confidence, and public school ratings.

Imagine ordinal H is based on an underlying, **latent** (unobserved), continuously distributed variable H^*. There are thresholds γ_j with $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_J = \infty$ such that $H = j$ iff $\gamma_{j-1} < H^* \leq \gamma_j$.

The CDF of ordinal H is $F(\cdot)$, and the CDF of latent H^* is $F^*(\cdot)$:

$$F(j) \equiv \mathrm{P}(H \leq j) \text{ for } j \in \{1, \ldots, J\}, \quad F^*(r) \equiv \mathrm{P}(H^* \leq r) \text{ for } r \in \mathbb{R}. \tag{23.1}$$

There are $J-1$ unknown ordinal distribution parameters, $F(j)$ for $j = 1, \ldots, J-1$, because $F(J) = 1$ by definition.

Discussion Question 23.1 (ordinal: known thresholds 1)**.** Assume (unrealistically) that the γ_j are known. For "identification," imagine we fully know $F(\cdot)$ and now want to learn about $F^*(\cdot)$.

a) Explain why the events $H = 1$ and $H^* \leq \gamma_1$ are identical, i.e., either both occur or neither occurs.
b) Explain why consequently $F^*(\gamma_1) = F(1)$. Hint: write out $F(1)$ as a probability involving H, and then write out $F^*(\gamma_1)$ as a probability involving H^*.
c) Is $F^*(r)$ point identified (i.e., uniquely determined by $F(\cdot)$) for any other r? Explain.

Discussion Question 23.2 (ordinal: known thresholds 2)**.** Continue from DQ 23.1.

a) Consider any $r < \gamma_1$. Since $F^*(\cdot)$ is a CDF, we know $0 \leq F^*(r) \leq 1$ for any $r \in \mathbb{R}$. If we know $F(\cdot)$, can we get more informative (i.e., tighter, shorter interval) bounds than $F^*(r) \in [0, 1]$? Hint: we know $F^*(\gamma_1) = F(1)$, and CDFs are non-decreasing.
b) Consider $\gamma_1 < r < \gamma_2$. Propose lower and upper bounds for $F^*(r)$ that are not just $[0, 1]$ but use information from $F(1)$ and $F(2)$.
c) Most generally: find lower bound and upper bound functions $F_L^*(\cdot)$ and $F_U^*(\cdot)$ such that $F_L^*(\cdot) \leq F^*(\cdot) \leq F_U^*(\cdot)$, i.e., $F_L^*(r) \leq F^*(r) \leq F_U^*(r)$ for all $r \in \mathbb{R}$. Hint: draw it.

23.2 Inequality: Introduction

Now imagine comparing two latent population distributions represented by H^* and G^*, with CDFs $F_H^*(\cdot)$ and $F_G^*(\cdot)$. For identification, assume we know $F_H(\cdot)$ and $F_G(\cdot)$, the marginal CDFs of ordinal H and G.

Consider two types of inequality: within-group and between-group. Within-group inequality means dispersion, often quantified by interquantile ranges (differences between two quantiles). For example, to study whether "income inequality" in the U.S. has increased over time, the 90–10 interquantile range has been used to measure dispersion within the income distribution in a given year. Between-group inequality means whether one group is better off or worse off than another. For example, "racial inequality" in health means one racial group tends to have better health than another.

Sections 23.3 and 23.4 consider learning about these two types of inequality in the latent distributions, given knowledge of the ordinal distributions.

23.3 Between-Group Inequality

For between-group inequality, we want to learn if H^* is "better" than G^*.

Assume the γ_j are the same for both populations. This is critical: if one population has lower thresholds, then the ordinal values can look better even if the latent values are not. Having the same γ_j makes the ordinal distributions comparable, even if we don't know the γ_j values themselves.

23.3.1 Quantiles

Although latent means essentially cannot be compared (Bond and Lang, 2019), certain latent quantiles can be compared.

Discussion Question 23.3 (latent quantile comparison)**.** Assume the same γ_j generate G and H.

a) For any continuous CDF $F^*(\cdot)$, imagine $F^*(r) = b$. Let $c > b$; can we know if the c-quantile of the distribution is above, below, or equal to r? Explain.
b) Again $F^*(r) = b$ as in (a), but now $a < b$: what do we know about the a-quantile? Explain.
c) Let $a < b < c$ with $a = F_G(1)$ and $c = F_H(1)$. Is the b-quantile of G^* above, below, or equal to γ_1? Is the b-quantile of H^* above, below, or equal to γ_1? Explain.
d) Generalize your insights: if $F_G(j) < F_H(j)$ for some $j \in \{1, \ldots, J-1\}$, then which latent quantiles are larger for G^* than H^*? Explain.

23.3.2 Stochastic Dominance

One strong definition of "better" is first-order stochastic dominance (SD1). (Recall Chapter 10.) SD1 of G^* over H^* is written G^* SD$_1$ H^*.

There are three ways to characterize G^* SD$_1$ H^*. First, G^* has higher expected utility: $\mathrm{E}[u(G^*)] \geq \mathrm{E}[u(H^*)]$ for all (non-decreasing) utility functions $u(\cdot)$. Second, the CDF of G^* is below the CDF of H^*: $F^*_G(\cdot) \leq F^*_H(\cdot)$, i.e., $F^*_G(r) \leq F^*_H(r)$ for all $r \in \mathbb{R}$. Third, the quantiles of G^* are all higher than the corresponding quantiles of H^*: $\mathrm{Q}_\tau(G^*) \geq \mathrm{Q}_\tau(H^*)$ for all $\tau \in [0,1]$, or $Q_{G^*}(\cdot) \geq Q_{H^*}(\cdot)$.

A weaker version of SD1 is the **restricted SD1** concept of Atkinson (1987, Condition I, p. 751). If $F^*_G(r) \leq F^*_H(r)$ for all $r \in [r^-, r^+]$, then there is restricted SD1 of G^* over H^* on the interval $[r^-, r^+]$. Being weaker than SD1 makes it less helpful economically but more tractable statistically. In particular, it does not require knowledge of distributions' tails, which are especially difficult (and in this setting impossible) to learn about statistically. Atkinson (1987) originally thought about comparing income or consumption distributions: given poverty line r, G^* has lower "headcount poverty ratio" (proportion of population in poverty) if $F^*_G(r) < F^*_H(r)$. Restricted SD1 says G^* has lower poverty given any poverty line $r \in [r^-, r^+]$.

Discussion Question 23.4 (testable implication)**.** Assume the same γ_j generate G and H. Hint: draw a picture (with latent values on the horizontal axis and cumulative probabilities [CDF values] on the vertical axis).

a) Imagine $F_G(j) \leq F_H(j)$ for all j, i.e., G SD$_1$ H. Does this imply that G^* SD$_1$ H^*, i.e., that $F^*_G(r) \leq F^*_H(r)$ for all $r \in \mathbb{R}$? Why/not? Hint: focus on the range $[\gamma_1, \gamma_2]$, and which latent CDFs are consistent with the bounds implied by the ordinal CDF values $F_G(1)$, $F_G(2)$, $F_H(1)$, and $F_H(2)$.
b) Imagine $F^*_G(r) \leq F^*_H(r)$ for all $r \in \mathbb{R}$, i.e., G^* SD$_1$ H^*. Does this imply that G SD$_1$ H, i.e., that $F_G(j) \leq F_H(j)$ for all $j = 1, \ldots, J$? Why/not?
c) Imagine G does not SD1 H: there is at least one j for which $F_G(j) > F_H(j)$. Does this imply anything about latent SD1 between G^* and H^*? (In either direction, either that it does or does not hold.) What/why?

Discussion Question 23.5 (super SD1)**.** Imagine $F_G(j+1) \leq F_H(j)$ for all j. What (if anything) does this imply about restricted SD1 between G^* and H^* over the interval $[\gamma_1, \gamma_{J-1}]$? Hint: draw a picture of bounds for the latent CDFs.

23.4 Within-Group Inequality: Dispersion

Now we wish to learn whether G^* or H^* has more within-group inequality, i.e., is more dispersed. Dispersion could be measured by variance, but variance is sensitive to the extreme tails, which we cannot learn about here. Dispersion can also be measured by interquantile ranges, which are not sensitive to the tails. This is similar to the reason for using the 0.9–0.1 interquantile range as a measure of income inequality when income data are top-coded; see Section 4.5.

To start, assume G and H share the same γ_j thresholds. For simplicity, assume both $F^*_G(\cdot)$ and $F^*_H(\cdot)$ are strictly increasing, so the quantile function is the inverse CDF.

Discussion Question 23.6 (latent quantiles)**.** Section 4.2 may be helpful here; e.g., how to find a quantile value on a CDF graph.

a) Explain why γ_1 is the $F_G(1)$-quantile of G^*.
b) Explain why γ_2 is the $F_G(2)$-quantile of G^*.
c) Imagine you know the τ_2-quantile of continuous random variable Y^*, and you know its τ_1-quantile, but nothing else. Given $\tau_1 \leq \tau \leq \tau_2$, explain why $\mathrm{Q}_{\tau_1}(Y^*) \leq \mathrm{Q}_{\tau}(Y^*) \leq \mathrm{Q}_{\tau_2}(Y^*)$.
d) Let $F_G(1) < \tau < F_G(2)$. Using (a)–(c), provide bounds for the τ-quantile of G^*.

The following are less directly related to DQ 23.7, but helpful for intuition.

e) What is the $F_H(j)$-quantile of H^*? Why?
f) If $F^*_G(r) \leq F^*_H(r)$ for all $r \in \mathbb{R}$, which distribution has the bigger τ-quantile? Why?

Discussion Question 23.7 (single crossing)**.** Your thoughts from DQs 23.3 and 23.6 will help greatly here; pictures also help. Imagine the ordinal CDFs cross: $F_G(1) < F_H(1)$ but $F_G(2) > F_H(2)$.

a) What is the $F_H(1)$-quantile of H^*?
b) Derive (and explain) bounds for the $F_H(1)$-quantile of G^*.
c) Explain why G^* has a larger $F_H(1)$-quantile than H^*.

d) What is the $F_H(2)$-quantile of H^*?
e) Derive (and explain) bounds for the $F_H(2)$-quantile of G^*.
f) Explain why H^* has a larger $F_H(2)$-quantile than G^*.
g) Use your previous results to argue that there is evidence of H^* being "more dispersed" than G^*. Specifically, explain why H^* has a larger $F_H(2)$–$F_H(1)$ interquantile range. (Recall $\mathrm{Q}_b(Y^*) - \mathrm{Q}_a(Y^*)$ is the b–a interquantile range of Y^*.) That is, show that $\mathrm{Q}_{F_H(2)}(H^*) - \mathrm{Q}_{F_H(1)}(H^*)$ is greater than $\mathrm{Q}_{F_H(2)}(G^*) - \mathrm{Q}_{F_H(1)}(G^*)$.

Discussion Question 23.8 (single crossing with shift)**.** Now let G and H have different thresholds. The thresholds for G are γ_j^G, and the H thresholds are $\gamma_j^H = \gamma_j^G + \Delta$, where possibly $\Delta \neq 0$ but Δ does not depend on j. Do the results from DQ 23.7 still hold? Why/not? Hint: if you shift a distribution left or right, does that change its dispersion? Or if you add c to all quantiles, does that change its interquantile ranges?

Kaplan and Zhao (2022) also discuss frequentist and Bayesian inference (using Kaplan and Zhuo, 2021) and provide some R code. "Regression" is also discussed but not detailed.

23.5 Parametric Approach

Alternatively, the latent distributions could be specified parametrically, in which case maximum likelihood can be run. This is similar to a probit model with latent Y^* and observed $Y = \mathbb{1}\{Y^* > 0\}$, but now there are multiple categories. This yields an "ordered probit" model. The ordered probit approach can be used for regression or simply unconditionally. As with the probit, the scale parameter is not identified; changing the latent variables scale (standard deviation) is observationally equivalent to scaling the γ_j by an equivalent amount.

Bond and Lang (2019) point out why ordered probit is not a great approach for happiness (and health, etc.). As you guessed, results are sensitive to the parametric (mis)specification. (They show how simply adding skewness to the specified parametric distribution can reverse the $+/-$ sign of estimates from several prominent published papers on happiness.) And since latent variables are by definition unobserved, it's impossible to learn the true shape of their distributions.

23.6 Inequality Indices

There is a literature about computing an **inequality index** from an observed ordinal distribution. This provides a single number that (supposedly) measures how much inequality there is in the distribution, which is convenient. Thus, you can definitively compare any two ordinal distributions. However, there are many such indices, and sometimes "one" index requires you to choose the value of some parameter(s), so really there is not a single definitive number after all. And if any justification is given, it often presumes the latent variable has only J possible different values (i.e., not continuous), which does not seem realistic.

There is a Stata .ado program `ineqord` available from SSC (`ssc install ineqord`) that calculates a slew of ordinal inequality indices, as described nicely by Jenkins (2020).

Exercises

Exercise E23.1. a. Find a published paper that at some point compares two (or more) ordinal distributions, like for self-reported health status or happiness.

b. Replicate one of their comparisons (ideally using their provided code and data). If they only have "regression" results (like ordered probit), then you could try to discretize the $\boldsymbol{X}$ and/or drop certain regressors to see if you can get a qualitatively similar result. For example: if an ordered probit has ordinal happiness as Y, and $\boldsymbol{X}$ includes an individual's education, sex, and height, you could do something like drop height and make a dummy for high (vs. low) education, which leaves four Y distributions to compare, i.e., when the new simplified "$\boldsymbol{X}$" is $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$.

c. Compare the two ordinal distributions with some of the methods from Kaplan and Zhao (2022), optionally using the R or Stata code at `https://kaplandm.github.io/`. Most importantly, verbally interpret your results clearly in terms of the underlying latent distributions. The empirical illustrations from Kaplan and Zhao (2022) may be helpful examples for reference.

d. Discuss similarities and differences between the original results and your new results.

Bibliography

Abadie, Alberto and Guido W. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76 (6):1537–1557. URL https://www.jstor.org/stable/40056514. [140]

Akaike, Hirotugu. 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control* 19 (6):716–723. URL https://doi.org/10.1109/TAC.1974.1100705. [203]

Anderson, T. W. and D. A. Darling. 1952. "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes." *Annals of Mathematical Statistics* 23 (2):193–212. URL https://projecteuclid.org/euclid.aoms/1177729437. [99]

Anderson, Theodore Wilbur and Donald A. Darling. 1954. "A test of goodness of fit." *Journal of the American Statistical Association* 49 (268):765–769. [99]

Andrews, Donald W. K. and Moshe Buchinsky. 2000. "A Three-Step Method for Choosing the Number of Bootstrap Repetitions." *Econometrica* 68 (1):23–51. URL https://www.jstor.org/stable/2999474. [145]

———. 2001. "Evaluation of a three-step method for choosing the number of bootstrap repetitions." *Journal of Econometrics* 103 (1):345–386. URL https://doi.org/10.1016/S0304-4076(01)00047-1. [145]

———. 2002. "On the Number of Bootstrap Repetitions for BC_a Confidence Intervals." *Econometric Theory* 18 (4):962–984. URL https://www.jstor.org/stable/3533421. [145]

Andrews, Isaiah and Anna Mikusheva. 2016. "Conditional Inference With a Functional Nuisance Parameter." *Econometrica* 84 (4):1571–1612. URL https://doi.org/10.3982/ECTA12868. [90]

Angrist, Joshua, Victor Chernozhukov, and Iván Fernández-Val. 2006. "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure." *Econometrica* 74 (2):539–563. URL https://doi.org/10.1111/j.1468-0262.2006.00671.x. [69, 73, 74, 75]

Arellano, Manuel and Stéphane Bonhomme. 2016. "Nonlinear panel data estimation via quantile regressions." *Econometrics Journal* 19 (3):C61–C94. URL https://doi.org/10.1111/ectj.12062. [92]

Atkinson, A. B. 1987. "On the Measurement of Poverty." *Econometrica* 55 (4):749–764.

URL https://www.jstor.org/stable/1911028. [112, 245]

Bååth, Rasmus. 2018. *bayesboot: An Implementation of Rubin's (1981) Bayesian Bootstrap*. URL https://CRAN.R-project.org/package=bayesboot. R package version 0.2.2. [147]

Banks, James, Richard Blundell, and Arthur Lewbel. 1997. "Quadratic Engel curves and consumer demand." *Review of Economics and Statistics* 79 (4):527–539. URL https://doi.org/10.1162/003465397557015. [25]

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *Review of Economic Studies* 81 (2):608–650. URL https://doi.org/10.1093/restud/rdt044. [196]

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society: Series B* 57 (1):289–300. URL https://www.jstor.org/stable/2346101. [118]

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 2nd ed. URL https://doi.org/10.1007/978-1-4757-4286-2. [147]

Blundell, Richard and James L. Powell. 2003. "Endogeneity in Nonparametric and Semiparametric Regression Models." In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, *Econometric Society Monographs*, vol. 2, edited by Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, chap. 8. Cambridge: Cambridge University Press, 312–357. URL https://doi.org/10.1017/CBO9780511610257. [167]

Bond, Timothy N. and Kevin Lang. 2019. "The Sad Truth About Happiness Scales." *Journal of Political Economy* 127 (4):1629–1640. URL https://doi.org/10.1086/701679. [245, 247]

Box, G. E. P. 1979. "Robustness in the Strategy of Scientific Model Building." Tech. Rep. 1954, Mathematics Research Center, University of Wisconsin–Madison. URL http://www.dtic.mil/docs/citations/ADA070213. [195, 196]

Breiman, Leo. 1996. "Bagging predictors." *Machine Learning* 24 (2):123–140. URL https://doi.org/10.1007/BF00058655. [205]

———. 2001. "Random Forests." *Machine Learning* 45 (1):5–32. URL https://doi.org/10.1023/A:1010933404324. [205]

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90 (3):414–427. URL https://doi.org/10.1162/rest.90.3.414. [142]

Canty, Angelo and B. D. Ripley. 2019. *boot: Bootstrap R (S-Plus) Functions*. URL https://cran.r-project.org/web/packages/boot. R package version 1.3-23. [125, 137]

Card, David. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, edited by Louis N. Christophides, E. Kenneth Grant, and Robert

Swidinsky. University of Toronto Press, 201–222. [89]

Cattaneo, Matias D. and Max H. Farrell. 2013. "Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators." *Journal of Econometrics* 174 (2):127–143. URL https://doi.org/10.1016/j.jeconom.2013.02.002. [178]

Chamberlain, Gary and Guido W. Imbens. 2003. "Nonparametric Applications of Bayesian Inference." *Journal of Business & Economic Statistics* 21 (1):12–18. URL https://www.jstor.org/stable/1392346. [147, 154, 156, 160]

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." In *Handbook of Econometrics*, vol. 6B, edited by James J. Heckman and Edward E. Leamer, chap. 76. Elsevier, 5549–5632. URL https://doi.org/10.1016/S1573-4412(07)06076-X. [171, 189, 193, 212]

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *Econometrics Journal* 21 (1):C1–C68. URL https://doi.org/10.1111/ectj.12097. [196]

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. "Inference on Counterfactual Distributions." *Econometrica* 81 (6):2205–2268. URL https://www.jstor.org/stable/23524318. [84, 131]

Chernozhukov, Victor and Christian Hansen. 2005. "An IV Model of Quantile Treatment Effects." *Econometrica* 73 (1):245–261. URL https://www.jstor.org/stable/3598944. [87, 88]

———. 2008. "Instrumental Variable Quantile Regression: A Robust Inference Approach." *Journal of Econometrics* 142 (1):379–398. URL https://doi.org/10.1016/j.jeconom.2007.06.005. [90]

Chernozhukov, Victor, Christian Hansen, and Michael Jansson. 2009. "Finite sample inference for quantile regression models." *Journal of Econometrics* 152 (2):93–103. URL https://doi.org/10.1016/j.jeconom.2009.01.004. [75, 90]

Chernozhukov, Victor, Christian Hansen, and Kaspar Wüthrich. 2017. "Instrumental Variable Quantile Regression." In *Handbook of Quantile Regression*, edited by Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng, chap. 9. CRC/Chapman-Hall, 119–144. URL https://www.routledgehandbooks.com/doi/10.1201/9781315120256. [87, 90]

Claeskens, Gerda and Nils Lid Hjort. 2003. "The Focused Information Criterion." *Journal of the American Statistical Association* 98 (464):900–916. URL https://www.jstor.org/stable/30045340. [196, 204]

———. 2008. *Model Selection and Model Averaging*. Cambridge: Cambridge University Press. URL https://doi.org/10.1017/CBO9780511790485. [195, 205]

Craven, Peter and Grace Wahba. 1978. "Smoothing noisy data with spline functions." *Numerische Mathematik* 31 (4):377–403. URL https://doi.org/10.1007/BF01404567. [201]

Davidson, Russell and Jean-Yves Duclos. 2000. "Statistical Inference for Stochastic Dom-

inance and for the Measurement of Poverty and Inequality." *Econometrica* 68 (6):1435–1464. URL https://www.jstor.org/stable/3003995. [112]

———. 2013. "Testing for Restricted Stochastic Dominance." *Econometric Reviews* 32 (1):84–125. URL https://doi.org/10.1080/07474938.2012.690332. [109, 112]

Davidson, Russell and James G. MacKinnon. 2000. "Bootstrap tests: how many bootstraps?" *Econometric Reviews* 19 (1):55–68. URL https://doi.org/10.1080/07474930008800459. [145]

Davison, A. C. and D. V. Hinkley. 1997. *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press. URL http://statwww.epfl.ch/davison/BMA. [125, 137]

de Castro, Luciano, Antonio F. Galvao, David M. Kaplan, and Xin Liu. 2019. "Smoothed GMM for quantile models." *Journal of Econometrics* 213 (1):121–144. URL https://doi.org/10.1016/j.jeconom.2019.04.008. [83, 89]

DiTraglia, Francis J. 2016. "Using invalid instruments on purpose: Focused moment selection and averaging for GMM." *Journal of Econometrics* 195 (2):187–208. URL https://doi.org/10.1016/j.jeconom.2016.07.006. [204]

Dong, Qi, Michael R. Elliott, and Trivellore E. Raghunathan. 2014. "A nonparametric method to generate synthetic populations to adjust for complex sampling design features." *Survey Methodology* 40 (1):29. URL https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X201400114003. [147]

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7 (1):1–26. URL https://projecteuclid.org/euclid.aos/1176344552. [126]

Efron, Bradley and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*, *Monographs on Statistics and Applied Probability*, vol. 57. Chapman & Hall/CRC. [125, 129, 131, 136, 139, 145]

Engel, Ernst. 1857. "Die Productions- und Consumtionsverhältnisse des Königreichs Sachsen." *Zeitschrift des Statistischen Bureaus des Königlich Sächsischen, Ministerium des Inneren* 8–9:1–54. [25]

Falk, Michael and Edgar Kaufmann. 1991. "Coverage Probabilities of Bootstrap-Confidence Intervals for Quantiles." *Annals of Statistics* 19 (1):485–495. URL https://projecteuclid.org/euclid.aos/1176347995. [136]

Fan, Yanqin and Sang Soo Park. 2010. "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference." *Econometric Theory* 26 (3):931–951. URL https://doi.org/10.1017/S0266466609990168. [79]

———. 2012. "Confidence intervals for the quantile of treatment effects in randomized experiments." *Journal of Econometrics* 167 (2):330–344. [79]

Ferguson, Thomas S. 1973. "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics* 1 (2):209–230. URL https://projecteuclid.org/euclid.aos/1176342360. [154]

———. 1974. "Prior Distributions on Spaces of Probability Measures." *Annals of Statistics* 2 (4):615–629. URL https://projecteuclid.org/euclid.aos/1176342752. [154]

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2009. "Unconditional Quantile

Regression." *Econometrica* 77 (3):953–973. URL https://www.jstor.org/stable/40263848. [84]

Frank, Ildiko E. and Jerome H. Friedman. 1993. "A Statistical View of Some Chemometrics Regression Tools." *Technometrics* 35 (2):109–135. URL https://www.jstor.org/stable/1269656. [194]

Fu, Wenjiang J. 1998. "Penalized Regressions: The Bridge Versus the Lasso." *Journal of Computational and Graphical Statistics* 7 (3):397–416. URL https://www.jstor.org/stable/1390712. [194]

Gneezy, Uri and John A. List. 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." *Econometrica* 74 (5):1365–1384. URL https://doi.org/10.1111/j.1468-0262.2006.00707.x. [106, 120]

Goldman, Matt and David M. Kaplan. 2017. "Fractional order statistic approximation for nonparametric conditional quantile inference." *Journal of Econometrics* 196 (2):331–346. URL https://doi.org/10.1016/j.jeconom.2016.09.015. [67]

———. 2018a. "Comparing distributions by multiple testing across quantiles or CDF values." *Journal of Econometrics* 206 (1):143–166. URL https://doi.org/10.1016/j.jeconom.2018.04.003. [99, 109, 114, 115, 122]

———. 2018b. "Non-parametric inference on conditional quantile differences and linear combinations, using L-statistics." *Econometrics Journal* 21 (2):136–169. URL https://doi.org/10.1111/ectj.12095. [67]

Grenander, Ulf. 1981. *Abstract Inference*. New York: Wiley. URL https://www.worldcat.org/oclc/6708583. [189, 193]

Hahn, Jinyong. 1997. "Bayesian Bootstrap of the Quantile Regression Estimator: A Large Sample Study." *International Economic Review* 38 (4):795–808. URL https://www.jstor.org/stable/2527216. [75]

Hanck, Christoph, Martin Arnold, Alexander Gerber, and Martin Schmelzer. 2018. "Introduction to Econometrics in R." URL https://www.econometrics-with-r.org. Department of Business Administration and Economics, University of Duisburg-Essen. [30]

Hansen, Bruce E. 2020a. "Econometrics." URL https://www.ssc.wisc.edu/~bhansen/econometrics. Textbook draft. [iv, xvii, 3, 53, 61, 65, 69, 72, 73, 78, 100, 167, 168, 173, 189, 195]

———. 2020b. "Introduction to Econometrics." URL https://www.ssc.wisc.edu/~bhansen/probability. Textbook draft. [xvii, 3]

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd ed. URL https://web.stanford.edu/~hastie/ElemStatLearn. Corrected 12th printing, January 13, 2017. [173, 189, 193, 196, 209, 228]

Hayfield, Tristen and Jeffrey S. Racine. 2008. "Nonparametric Econometrics: The np Package." *Journal of Statistical Software* 27 (5):1–32. URL https://doi.org/10.18637/jss.v027.i05. [174, 196, 217]

Heckman, James J. and Edward Vytlacil. 2001. "Policy-Relevant Treatment Effects."

American Economic Review (Papers & Proceedings) 91 (2):107–111. URL https://www.jstor.org/stable/2677742. [80]

Heckman, James J. and Edward J. Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics*, vol. 6B, edited by James J. Heckman and Edward E. Leamer, chap. 70. Elsevier, 4779–4874. URL https://doi.org/10.1016/S1573-4412(07)06070-9. [80]

Heiss, Florian. 2016. *Using R for Introductory Econometrics*. CreateSpace. URL http://www.urfie.net/read.html. [30]

Ho, Tin Kam. 1995. "Random decision forests." In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 278–282. URL https://doi.org/10.1109/ICDAR.1995.598994. [205]

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2):65–70. URL https://www.jstor.org/stable/4615733. [118]

Horowitz, Joel L. 2014. "Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter." *Journal of Econometrics* 180 (2):158–173. URL https://doi.org/10.1016/j.jeconom.2014.03.006. [196]

Hutson, Alan D. 2007. "An 'exact' two-group median test with an extension to censored data." *Journal of Nonparametric Statistics* 19 (2):103–112. URL https://doi.org/10.1080/10485250701464657. [128]

Hyndman, Rob, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. 2020. *forecast: Forecasting functions for time series and linear models*. URL https://pkg.robjhyndman.com/forecast. R package version 8.11. [202]

Hyndman, Rob J. and George Athanasopoulos. 2019. *Forecasting: Principles and Practice*. OTexts. URL https://otexts.com/fpp2. [202]

Hyndman, Rob J. and Yeasmin Khandakar. 2008. "Automatic time series forecasting: the forecast package for R." *Journal of Statistical Software* 26 (3):1–22. URL https://www.jstatsoft.org/article/view/v027i03. [202]

Ichimura, Hidehiko. 1993. "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models." *Journal of Econometrics* 58 (1):71–120. URL https://doi.org/10.1016/0304-4076(93)90114-K. [212]

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2):467–475. URL https://www.jstor.org/stable/2951620. [90]

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer, 1st ed. URL https://faculty.marshall.usc.edu/gareth-james/ISL/. Corrected 8th printing, 2017. [29, 173, 189, 193, 196, 215]

Jenkins, Stephen P. 2020. "Comparing distributions of ordinal data." *Stata Journal* 20 (3):505–531. URL https://doi.org/10.1177/1536867X20953565. [248]

Jun, Sung Jae. 2008. "Weak Identification Robust Tests in an Instrumental Quantile Model." *Journal of Econometrics* 144 (1):118–138. URL https://doi.org/10.1016/j.jeconom.2007.12.006. [90]

Kaplan, David M. 2015. "Improved quantile inference via fixed-smoothing asymptotics and Edgeworth expansion." *Journal of Econometrics* 185 (1):20–32. URL https://doi.org/10.1016/j.jeconom.2014.08.011. [25, 67]

———. 2019. "distcomp: Comparing distributions." *Stata Journal* 19 (4):832–848. URL https://doi.org/10.1177/1536867x19893626. [99, 105, 109]

———. 2022a. "Inference on Consensus Ranking of Distributions." Working paper available at https://kaplandm.github.io/. [113, 114, 115, 117]

———. 2022b. *Introductory Econometrics: Description, Prediction, and Causality*. Columbia, MO: Mizzou Publishing, 3rd ed. URL https://www.themizzoustore.com/p-236916-introductory-econometrics-description-prediction-and-causality.aspx. [xvii, 29, 51, 77, 128, 147, 167, 168, 190, 192, 195, 228]

———. 2022c. "Smoothed instrumental variables quantile regression." *Stata Journal* 22 (2):379–403. URL https://doi.org/10.1177/1536867X221106404. [89]

Kaplan, David M. and David M. Blei. 2007. "A computational approach to style in American poetry." In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007)*. Omaha, NE: IEEE Computer Society, 553–558. URL https://doi.org/10.1109/ICDM.2007.76. [25]

Kaplan, David M. and Lonnie Hofmann. 2020. "High-order coverage of smoothed Bayesian bootstrap intervals for population quantiles." Working paper available at https://kaplandm.github.io/. [67, 136, 156]

Kaplan, David M. and Yixiao Sun. 2017. "Smoothed Estimating Equations for Instrumental Variables Quantile Regression." *Econometric Theory* 33 (1):105–157. URL https://doi.org/10.1017/S0266466615000407. [89]

Kaplan, David M. and Wei Zhao. 2022. "Comparing Latent Inequality with Ordinal Data." Working paper available at https://kaplandm.github.io/. [243, 247, 249]

Kaplan, David M. and Longhao Zhuo. 2021. "Frequentist properties of Bayesian inequality tests." *Journal of Econometrics* 221 (1):312–336. URL https://doi.org/10.1016/j.jeconom.2020.05.015. [247]

Kiefer, J. and J. Wolfowitz. 1956. "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters." *Annals of Mathematical Statistics* 27 (4):887–906. URL https://projecteuclid.org/euclid.aoms/1177728066. [148]

Kleiber, Christian and Achim Zeileis. 2008. *Applied Econometrics with R*. New York: Springer. URL https://eeecon.uibk.ac.at/~zeileis/teaching/AER. [30]

Klein, Roger W. and Richard H. Spady. 1993. "An Efficient Semiparametric Estimator for Binary Response Models." *Econometrica* 61 (2):387–421. URL https://www.jstor.org/stable/2951556. [212]

Knight, Keith and Wenjiang Fu. 2000. "Asymptotics for Lasso-Type Estimators." *Annals of Statistics* 28 (5):1356–1378. URL https://www.jstor.org/stable/2674097. [194]

Koenker, Roger. 2004. "Quantile regression for longitudinal data." *Journal of Multivariate Analysis* 91 (1):74–89. URL https://doi.org/10.1016/j.jmva.2004.05.006. [91]

———. 2005. *Quantile Regression*, *Econometric Society Monographs*, vol. 38. Cambridge University Press. URL https://doi.org/10.1017/CBO9780511754098. [59, 66, 69]

———. 2019. *quantreg: Quantile Regression*. URL https://CRAN.R-project.org/package=quantreg. R package version 5.51. [70, 73]

Kolmogorov, Andrey Nikolaevich. 1933. "Sulla determinazione empirica di una legge di distribuzione." *Giornale dell'Istituto Italiano degli Attuari* 4 (1):83–91. [99]

Konishi, Sadanori and Genshiro Kitagawa. 2008. *Information Criteria and Statistical Modeling*. New York: Springer. URL https://doi.org/10.1007/978-0-387-71887-3. [195]

Kuhn, Max. 2020. *caret: Classification and Regression Training*. URL https://CRAN.R-project.org/package=caret. R package version 6.0-85. [174, 196, 215]

Lehmann, E. L. and Joseph P. Romano. 2005a. "Generalizations of the Familywise Error Rate." *Annals of Statistics* 33 (3):1138–1154. URL https://projecteuclid.org/euclid.aos/1120224098. [118]

———. 2005b. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 3rd ed. URL https://books.google.com/books?id=Y7vSVW3ebSwC. [116, 118]

Li, Qi and Jeffrey Scott Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press. [173, 189]

Liu, Xin. 2019. "Averaging estimation for instrumental variables quantile regression." Working paper available at https://ideas.repec.org/p/umc/wpaper/1907.html. [145]

———. 2020. "Panel Quantile Regression with Time-Invariant Rank." Working paper available at https://liuxinecon.weebly.com/research.html. [92]

Machado, José A. F. and J. M. C. Santos Silva. 2019. "Quantiles via moments." *Journal of Econometrics* 213 (1):145–173. URL https://doi.org/10.1016/j.jeconom.2019.04.009. [90]

MacKinnon, James G. 2002. "Bootstrap inference in econometrics." *Canadian Journal of Economics / Revue canadienne d'Economique* 35 (4):615–645. URL https://www.jstor.org/stable/3131829. [125]

———. 2006. "Bootstrap Methods in Econometrics." *Economic Record* 82 (s1):S2–S18. URL https://doi.org/10.1111/j.1475-4932.2006.00328.x. [125]

Manski, Charles F. 1989. "Anatomy of the Selection Problem." *Journal of Human Resources* 24 (3):343–360. URL https://www.jstor.org/stable/145818. [227]

Melly, Blaise and Kaspar Wüthrich. 2017. "Local Quantile Treatment Effects." In *Handbook of Quantile Regression*, edited by Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng, chap. 10. CRC/Chapman-Hall, 145–164. URL https://www.routledgehandbooks.com/doi/10.1201/9781315120256. [87, 90, 91]

Newey, Whitney K. and Kenneth D. West. 1987. "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (3):703–708. URL https://www.jstor.org/stable/1913610. [186]

Owen, Art B. 1988. "Empirical likelihood ratio confidence intervals for a single functional." *Biometrika* 75 (2):237–249. URL https://doi.org/10.1093/biomet/75.2.237. [148]

———. 2001. *Empirical Likelihood*. Chapman & Hall/CRC. [148]

Pagan, Adrian and Aman Ullah. 1999. *Nonparametric Econometrics*. Cambridge University Press. URL https://doi.org/10.1017/CBO9780511612503. [173, 187, 189]

Politis, Dimitris N. and Joseph P. Romano. 1992. "A Circular Block-Resampling Procedure for Stationary Data." In *Exploring the Limits of Bootstrap*, edited by Raoul LePage and Lynne Billard. Wiley, 263–270. URL https://books.google.com/books?id=ZJzIpNZNVLgC. [144]

———. 1994a. "Large sample confidence regions based on subsamples under minimal assumptions." *Annals of Statistics* 22 (4):2031–2050. [140, 141]

———. 1994b. "The Stationary Bootstrap." *Journal of the American Statistical Association* 89 (428):1303–1313. URL https://doi.org/10.1080/01621459.1994.10476870. [144]

Powell, David. 2020. "Quantile Regression with Nonadditive Fixed Effects." Working paper available at https://sites.google.com/site/davidmatthewpowell/quantile-regression-with-nonadditive-fixed-effects. [92]

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org. [173, 190, 215]

Racine, Jeffrey S. 2008. *Nonparametric Econometrics: A Primer*. now publishers. URL https://socialsciences.mcmaster.ca/racinej/ECO0301.pdf. [173]

Rothe, Christoph. 2010. "Nonparametric estimation of distributional policy effects." *Journal of Econometrics* 155 (1):56–70. URL https://doi.org/10.1016/j.jeconom.2009.09.001. [84]

Rousseeuw, Peter J. 1984. "Least Median of Squares Regression." *Journal of the American Statistical Association* 79 (388):871–880. URL https://www.jstor.org/stable/2288718. [66]

Rubin, Donald B. 1981. "The Bayesian Bootstrap." *Annals of Statistics* 9 (1):130–134. URL https://projecteuclid.org/euclid.aos/1176345338. [155]

Sasaki, Yuya, Takuya Ura, and Yichong Zhang. 2020. "Unconditional Quantile Regression with High-Dimensional Data." Working paper available at https://arxiv.org/abs/2007.13659. [84]

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6 (2):461–464. URL https://projecteuclid.org/euclid.aos/1176344136. [203]

Shao, Jun. 1997. "An Asymptotic Theory for Linear Model Selection." *Statistica Sinica* 7 (2):221–242. URL https://www.jstor.org/stable/24306073. [204]

Shao, Jun and Dongsheng Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer-Verlag. URL https://link.springer.com/978-1-4612-0795-5. [125, 137, 145]

Smirnov, N. 1948. "Table for estimating the goodness of fit of empirical distributions." *Annals of Mathematical Statistics* 19 (2):279–281. URL https://www.jstor.org/stable/2236278. [99]

Stoye, Jörg. 2010. "Partial identification of spread parameters." *Quantitative Economics* 1 (2):323–357. URL https://doi.org/10.3982/QE24. [235]

Tamer, Elie. 2010. "Partial Identification in Econometrics." *Annual Review of Economics* 2:167–195. URL https://doi.org/10.1146/annurev.economics.050708.143401. [225]

Tan, Li. 2021. "Imputing Top-Coded Income Data in Longitudinal Surveys." *Oxford Bulletin of Economics and Statistics* 83 (1):66–87. URL https://doi.org/10.1111/obes.12400. [64]

Tibshirani, Robert J. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B* 58 (1):267–288. URL www.jstor.org/stable/2346178. [194]

van der Vaart, Aad W. and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. New York: Springer. URL https://doi.org/10.1007/978-1-4757-2545-2. [138]

Wand, Matt. 2019. *KernSmooth: Functions for Kernel Smoothing Supporting Wand and Jones (1995)*. URL https://CRAN.R-project.org/package=KernSmooth. R package version 2.23-16. [174, 217]

Weierstrass, Karl. 1885. "Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen." *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin* 2:633–639. [191]

White, Halbert. 2006. "Approximate Nonlinear Forecasting Methods." In *Handbook of Economic Forecasting*, vol. 1, edited by G. Elliott, C. W. J. Granger, and A. Timmermann, chap. 9. Elsevier, 459–512. URL https://doi.org/10.1016/S1574-0706(05)01009-8. [193]

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd ed. URL https://www.worldcat.org/oclc/831625495. [xvii, 50, 228]

Wu, C. F. J. 1986. "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis." *Annals of Statistics* 14 (4):1261–1295. URL https://projecteuclid.org/euclid.aos/1176350142. [139]

Index