Distributional and Nonparametric Econometrics

First edition

David M. Kaplan
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 viii
List of Discussion Questions xvii
Preface xxii
Notation 1
Statistical Software Overview 5

I Writing, Coding, and Logic 9

Introduction 11
1 Writing and Typesetting 13
2 R: Some Basics 27
3 Logic 49

II Quantile Methods 53

Introduction 55
4 Quantiles: Description and Prediction 57
5 Quantile Regression: Description and Prediction 67
BRIEF CONTENTS

6 Quantile Regression: Causality 77

7 Quantile Regression with Endogeneity 85

III Distributional Methods 93

Introduction 95

8 Distributional Inference: One-Sample, Two-Sided 97

9 Distributional Inference: Two-Sample, Two-Sided 107

10 Distributional Inference: Stochastic Dominance 111

11 Distributional Inference: Multiple Testing 117

IV Bootstrap and Friends 125

Introduction 127

12 Bootstrap: Basics 129

13 Bootstrap Extensions and Subsampling 143

14 Bayesian Bootstrap 155

V Nonparametric Regression 173

Introduction 175

15 Nonparametric Methods: Preliminaries 177

16 Local (Kernel) Regression 185

17 Series and Sieves 205

18 Model Selection 213
BRIEF CONTENTS

19 Multiple Regressors 227
20 Nonparametric Regression in R 233
21 Regression Discontinuity 239

VI Partial Identification 253

Introduction 255
22 Missing Data 257
23 Interval Data 269
24 Ordinal Data 273

Bibliography 279

Index 291
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contents</td>
<td>viii</td>
</tr>
<tr>
<td>List of Discussion Questions</td>
<td>xvii</td>
</tr>
<tr>
<td>Preface</td>
<td>xxi</td>
</tr>
<tr>
<td>Notation</td>
<td>1</td>
</tr>
<tr>
<td>Statistical Software Overview</td>
<td>5</td>
</tr>
<tr>
<td><strong>1 Writing, Coding, and Logic</strong></td>
<td>9</td>
</tr>
<tr>
<td>Introduction</td>
<td>11</td>
</tr>
<tr>
<td><strong>1 Writing and Typesetting</strong></td>
<td>13</td>
</tr>
<tr>
<td>1.1 <strong>LaTeX</strong></td>
<td>13</td>
</tr>
<tr>
<td>1.2 Writing Advice</td>
<td>14</td>
</tr>
<tr>
<td>1.2.1 Striving</td>
<td>14</td>
</tr>
<tr>
<td>1.2.2 Suppositions</td>
<td>15</td>
</tr>
<tr>
<td>1.2.3 Structure</td>
<td>16</td>
</tr>
<tr>
<td>1.2.4 Simplicity</td>
<td>18</td>
</tr>
<tr>
<td>1.2.5 Segues (and Sentence Structure)</td>
<td>19</td>
</tr>
<tr>
<td>1.2.6 Summary</td>
<td>21</td>
</tr>
<tr>
<td>1.3 Common Minor Mistakes</td>
<td>23</td>
</tr>
<tr>
<td>Exercises</td>
<td>26</td>
</tr>
<tr>
<td><strong>2 R: Some Basics</strong></td>
<td>27</td>
</tr>
<tr>
<td>2.1 Getting Help</td>
<td>27</td>
</tr>
<tr>
<td>2.2 Getting Started</td>
<td>28</td>
</tr>
<tr>
<td>2.2.1 Running R</td>
<td>28</td>
</tr>
</tbody>
</table>
CONTENTS

2.2.2 Packages .......................................................... 28
2.2.3 RStudio Interface .................................................. 29
2.2.4 Readability ......................................................... 29
2.3 Data Types ............................................................. 30
2.4 Basic Data Manipulation ............................................ 32
  2.4.1 Numerical Operations .......................................... 32
  2.4.2 Combining Data .................................................. 32
  2.4.3 String Manipulation ............................................. 32
2.5 Functions .............................................................. 35
2.6 Data File Input ....................................................... 36
2.7 Basic Statistics ....................................................... 37
2.8 Basic Plotting (Graphs) ............................................ 38
2.9 Saving Text Output .................................................. 39
2.10 Probability Distributions and Random Numbers .......... 39
  2.11 Control Flow: If, Loops, Errors .............................. 40
    2.11.1 If-Else Statements ........................................ 40
    2.11.2 For and While Loops ....................................... 41
    2.11.3 Try-Catch, Warnings, Errors ............................. 42
2.12 Time and Timing ................................................... 43
2.13 Parallel Computing (On Your Laptop) ...................... 43
2.14 Simulation: Example #1 .......................................... 44
2.15 Simulation: Example #2 .......................................... 45
Exercises ................................................................. 48

3 Logic ........................................................................... 49
  3.1 Terminology ......................................................... 49
  3.2 Assumptions .......................................................... 51
  3.3 Theorems ............................................................. 51

II Quantile Methods .................................................. 53

  Introduction ................................................................ 55

4 Quantiles: Description and Prediction .................... 57
  4.1 Description .......................................................... 57
  4.2 Formal Definitions .................................................. 57
  4.3 Prediction ............................................................. 58
  4.4 Estimation and Sample Quantiles ............................ 61
  4.5 Censoring ............................................................. 62
  4.6 Robustness and Efficiency ....................................... 64
CONTENTS

4.7 Inference .................................................. 65

5 Quantile Regression: Description and Prediction 67
5.1 Description ................................................. 68
  5.1.1 Conditional Quantile Function ..................... 68
  5.1.2 CQF Models .................................. 69
  5.1.3 Monotonicity .................................. 69
5.2 Prediction .................................................. 70
5.3 QR with Misspecification .................................. 71
  5.3.1 “Best” Linear Predictor ......................... 71
  5.3.2 “Best” Linear Approximation .................... 71
5.4 Estimation .................................................. 72
5.5 Asymptotic Properties ................................... 73
5.6 Inference ................................................... 74
5.7 Censoring .................................................. 75

6 Quantile Regression: Causality 77
6.1 Background: Potential Outcomes and ATE ............. 77
6.2 Quantile Treatment Effects ............................. 78
6.3 Background: Random Coefficients ..................... 80
6.4 Another Random Coefficients Model .................. 81
  6.4.1 Heteroskedasticity ......................... 83
6.5 Unconditional Quantile Regression .................... 83

7 Quantile Regression with Endogeneity 85
7.1 Instrumental Variables Quantile Regression .......... 85
  7.1.1 Reminder: Usual IV Regression .............. 86
  7.1.2 IVQR Identification ......................... 86
  7.1.3 IVQR Estimation ......................... 87
  7.1.4 IVQR Inference ......................... 89
7.2 Other Approaches to Endogeneity ..................... 89
  7.2.1 Triangular Model ......................... 89
  7.2.2 Local Quantile Treatment Effect .......... 89
7.3 Panel Data with Fixed Effects ......................... 90
  Exercises .................................. 92

III Distributional Methods 93

Introduction .................................................... 95

8 Distributional Inference: One-Sample, Two-Sided 97
# CONTENTS

8.1 Discrete and Categorical Distributions .................................................. 97
8.2 Preliminary Results for Continuous Distributions ..................................... 98
8.3 Goodness-of-Fit Testing .............................................................................. 98
8.4 Kolmogorov–Smirnov Test ........................................................................ 99
8.5 Uniform Confidence Band .......................................................................... 100
  8.5.1 Test Inversion: Scalar ........................................................................... 100
  8.5.2 Test Inversion: Vectors and Functions ................................................ 100
  8.5.3 Uniform Confidence Band ....................................................................... 101
8.6 Dirichlet Approach .................................................................................... 102
  8.6.1 Dirichlet Uniform Confidence Band .................................................... 102
  8.6.2 Dirichlet GOF Test ................................................................................ 103
8.A ECDF: Asymptotic Properties .................................................................. 105

9 Distributional Inference: Two-Sample, Two-Sided ..................................... 107
  9.1 Setup .......................................................................................................... 107
  9.2 Exact Finite-Sample Testing ....................................................................... 108
  9.3 Asymptotic KS ........................................................................................... 109
  9.4 Dirichlet Test ............................................................................................. 110

10 Distributional Inference: Stochastic Dominance ....................................... 111
  10.1 First-Order Stochastic Dominance ........................................................... 111
  10.2 Null of Dominance .................................................................................. 112
    10.2.1 KS Test .............................................................................................. 113
    10.2.2 Dirichlet Test ...................................................................................... 113
  10.3 Null of Non-Dominance ........................................................................... 113
    Exercises ........................................................................................................ 116

11 Distributional Inference: Multiple Testing .............................................. 117
  11.1 Multiple Testing: Concepts and Terms ..................................................... 117
    11.1.1 Familywise Error Rate ...................................................................... 118
    11.1.2 Alternatives to FWER ....................................................................... 118
    11.1.3 Other Ways to Improve Power .......................................................... 119
  11.2 One-Sample, Two-Sided ......................................................................... 119
    11.2.1 KS ...................................................................................................... 120
    11.2.2 Dirichlet ............................................................................................. 120
  11.3 Two-Sample and/or One-Sided ............................................................... 120
    Exercises ........................................................................................................ 123

IV Bootstrap and Friends ............................................................................... 125

  Introduction .................................................................................................... 127
CONTENTS

12 Bootstrap: Basics 129
  12.1 Introduction .............................................. 129
  12.2 Preliminaries: The Plug-in Principle .................. 130
    12.2.1 Example: Mean ...................................... 130
    12.2.2 Example: OLS ....................................... 131
    12.2.3 Other Types of Parameters ........................... 131
  12.3 The Real World and the Bootstrap World ............... 132
    12.3.1 The Real World ..................................... 133
    12.3.2 The Bootstrap World ................................ 133
  12.4 Empirical Bootstrap ....................................... 134
  12.5 Standard Errors ........................................... 135
  12.6 Confidence Intervals ...................................... 136
    12.6.1 CI Properties ....................................... 136
    12.6.2 Normal CI, Bootstrapped SE ........................... 137
    12.6.3 Root Method .......................................... 138
    12.6.4 Percentile Bootstrap CI ............................... 140
    12.6.5 Studentized Bootstrap CI ................................ 141

13 Bootstrap Extensions and Subsampling 143
  13.1 Exchangeable Weights Bootstrap ........................ 143
  13.2 Other Bootstraps ......................................... 144
    13.2.1 Parametric Bootstrap ................................. 144
    13.2.2 Residuals Bootstrap .................................. 144
    13.2.3 Bias-Corrected Bootstrap ............................ 145
    13.2.4 Wild Bootstrap ....................................... 145
    13.2.5 Smoothed and Iterated Bootstraps ...................... 145
  13.3 Bootstrap Failure .......................................... 145
  13.4 Subsampling .............................................. 146
    13.4.1 Subsampling Consistency ............................. 146
    13.4.2 Standard Errors ...................................... 147
  13.5 Clustered Data ............................................ 148
  13.6 Time Series Data .......................................... 149
    13.6.1 Moving Blocks Bootstrap ............................ 149
    13.6.2 Circular Block Bootstrap ............................ 150
    13.6.3 Stationary Bootstrap ................................ 150
  13.7 Bias Correction ........................................... 151
  13.8 Choice of B .............................................. 152
  13.9 Bootstrap-Based Model Selection ......................... 153
    Exercises ..................................................... 154

14 Bayesian Bootstrap 155
  14.1 Bayesian Basics ........................................... 155
CONTENTS

14.1 Beliefs, Data, and Model ........................................ 155
  14.1.1 Beliefs, Data, and Model ........................................ 155
  14.1.2 The Likelihood .................................................. 156
  14.1.3 Bayes' Theorem .................................................. 156
  14.1.4 Strengths and Weaknesses ................................... 157
14.2 Beta–Binomial Model ............................................. 158
  14.2.1 Likelihood, Prior, and Posterior .............................. 158
  14.2.2 Technical Details ............................................... 161
14.3 Conjugacy .......................................................... 162
14.4 Dirichlet–Multinomial Model .................................... 163
14.5 Improper Priors .................................................... 164
14.6 Nonparametric Bayes ............................................. 165
14.7 Bayesian Bootstrap ................................................ 166
  14.7.1 Population Mean ................................................. 166
  14.7.2 Other Population Features .................................... 168
  14.7.3 Population CDF .................................................. 168
Exercises ................................................................ 171

V Nonparametric Regression .......................................... 173

Introduction ................................................................ 175

15 Nonparametric Methods: Preliminaries .......................... 177
  15.1 Motivation ............................................................. 177
  15.2 Simple Examples for Intuition .................................. 178
  15.3 Terminology .......................................................... 182

16 Local (Kernel) Regression ......................................... 185
  16.1 Constant “Regressor” ............................................... 186
  16.2 Binary Regressor ..................................................... 186
    16.2.1 Estimation .......................................................... 186
    16.2.2 Consistency ........................................................ 186
    16.2.3 Bias–Variance Tradeoff ...................................... 187
    16.2.4 Binary Regressor: Small Probability .................... 188
  16.3 Discrete Regressor ................................................... 188
    16.3.1 Estimation .......................................................... 188
    16.3.2 Local Sample Size .............................................. 189
    16.3.3 Bias–Variance Tradeoff ...................................... 190
  16.4 Continuous Regressor: Introduction ......................... 191
    16.4.1 Estimation with Discretization ......................... 191
    16.4.2 Bias–Variance Tradeoff ...................................... 192
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.4.3</td>
<td>Types of “Discretization”</td>
<td>192</td>
</tr>
<tr>
<td>16.5</td>
<td>Local Constant Regression</td>
<td>193</td>
</tr>
<tr>
<td>16.6</td>
<td>Local Linear Regression</td>
<td>198</td>
</tr>
<tr>
<td>16.7</td>
<td>Local Polynomial Regression</td>
<td>200</td>
</tr>
<tr>
<td>16.8</td>
<td>Kernel Regression</td>
<td>200</td>
</tr>
<tr>
<td>16.8.1</td>
<td>Intuition</td>
<td>200</td>
</tr>
<tr>
<td>16.8.2</td>
<td>Local Linear Regression: Uniform Kernel</td>
<td>201</td>
</tr>
<tr>
<td>16.8.3</td>
<td>Other Second-Order Kernels</td>
<td>201</td>
</tr>
<tr>
<td>16.8.4</td>
<td>Effect on AMSE</td>
<td>202</td>
</tr>
<tr>
<td>16.8.5</td>
<td>Higher-Order Kernels</td>
<td>203</td>
</tr>
<tr>
<td>16.9</td>
<td>Linear Smoother</td>
<td>203</td>
</tr>
</tbody>
</table>

### 17 Series and Sieves

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.1</td>
<td>Discrete Regressor</td>
<td>205</td>
</tr>
<tr>
<td>17.1.1</td>
<td>Constant Regressor</td>
<td>206</td>
</tr>
<tr>
<td>17.1.2</td>
<td>Binary Regressor</td>
<td>206</td>
</tr>
<tr>
<td>17.1.3</td>
<td>Trinary and More</td>
<td>206</td>
</tr>
<tr>
<td>17.2</td>
<td>Polynomial Series</td>
<td>207</td>
</tr>
<tr>
<td>17.3</td>
<td>Series Regression</td>
<td>208</td>
</tr>
<tr>
<td>17.4</td>
<td>Splines</td>
<td>209</td>
</tr>
<tr>
<td>17.5</td>
<td>Linear vs. Nonlinear Approximation</td>
<td>209</td>
</tr>
<tr>
<td>17.6</td>
<td>Penalized Regression</td>
<td>209</td>
</tr>
<tr>
<td>17.7</td>
<td>Linear Smoother</td>
<td>211</td>
</tr>
</tbody>
</table>

### 18 Model Selection

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.1</td>
<td>Purpose</td>
<td>214</td>
</tr>
<tr>
<td>18.2</td>
<td>Quantifying Flexibility of Linear Smoothers</td>
<td>214</td>
</tr>
<tr>
<td>18.3</td>
<td>Bad Approaches</td>
<td>215</td>
</tr>
<tr>
<td>18.4</td>
<td>Analytic Plug-in Approach</td>
<td>216</td>
</tr>
<tr>
<td>18.5</td>
<td>Cross-Validation</td>
<td>217</td>
</tr>
<tr>
<td>18.5.1</td>
<td>Training and Validation Paradigm</td>
<td>217</td>
</tr>
<tr>
<td>18.5.2</td>
<td>LOOCV</td>
<td>218</td>
</tr>
<tr>
<td>18.5.3</td>
<td>GCV</td>
<td>219</td>
</tr>
<tr>
<td>18.5.4</td>
<td>Leave-d-out CV</td>
<td>220</td>
</tr>
<tr>
<td>18.5.5</td>
<td>k-fold CV</td>
<td>220</td>
</tr>
<tr>
<td>18.5.6</td>
<td>Time Series</td>
<td>221</td>
</tr>
<tr>
<td>18.6</td>
<td>Information Criteria</td>
<td>221</td>
</tr>
<tr>
<td>18.6.1</td>
<td>AIC and BIC</td>
<td>221</td>
</tr>
<tr>
<td>18.6.2</td>
<td>Other IC</td>
<td>222</td>
</tr>
<tr>
<td>18.7</td>
<td>Comparison</td>
<td>222</td>
</tr>
<tr>
<td>18.8</td>
<td>Model Averaging and Ensemble Methods</td>
<td>223</td>
</tr>
<tr>
<td>18.A</td>
<td>LOOCV for Linear Smoothers</td>
<td>225</td>
</tr>
</tbody>
</table>
CONTENTS

22.3.2 Worst-Case Bounds .................................................. 263
22.4 Summary ................................................................. 266
22.5 R Code ................................................................. 266
   Exercises ............................................................. 268

23 Interval Data ............................................................ 269
   Exercises ............................................................. 271

24 Ordinal Data ............................................................. 273
   24.1 Latent Variable Framework ....................................... 273
   24.2 Between-Group Inequality: Stochastic Dominance........... 274
   24.3 Within-Group Inequality: Dispersion ......................... 275
   24.4 Parametric Approach ............................................. 276
   24.5 Inequality Indexes ............................................... 277
   Exercises ............................................................. 278

Bibliography ................................................................. 279

Index ................................................................. 291
List of Discussion Questions

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>writing: striving</td>
<td>15</td>
</tr>
<tr>
<td>1.2</td>
<td>writing: structure</td>
<td>17</td>
</tr>
<tr>
<td>1.3</td>
<td>writing: simplicity</td>
<td>19</td>
</tr>
<tr>
<td>1.4</td>
<td>writing: segues</td>
<td>21</td>
</tr>
<tr>
<td>1.5</td>
<td>writing: sentence structure</td>
<td>21</td>
</tr>
<tr>
<td>1.6</td>
<td>writing: summary</td>
<td>22</td>
</tr>
<tr>
<td>2.1</td>
<td>R: simulation example 1</td>
<td>45</td>
</tr>
<tr>
<td>2.2</td>
<td>R: simulation example 2</td>
<td>47</td>
</tr>
<tr>
<td>3.1</td>
<td>logic</td>
<td>51</td>
</tr>
<tr>
<td>3.2</td>
<td>median theorem logic</td>
<td>52</td>
</tr>
<tr>
<td>3.3</td>
<td>mean theorem logic</td>
<td>52</td>
</tr>
<tr>
<td>3.4</td>
<td>logic with feathers</td>
<td>52</td>
</tr>
<tr>
<td>4.1</td>
<td>quantiles from CDF</td>
<td>58</td>
</tr>
<tr>
<td>4.2</td>
<td>population minimization: quadratic and absolute loss</td>
<td>58</td>
</tr>
<tr>
<td>4.3</td>
<td>check function</td>
<td>60</td>
</tr>
<tr>
<td>4.4</td>
<td>sample minimization: quadratic and absolute loss</td>
<td>61</td>
</tr>
<tr>
<td>4.5</td>
<td>inequality measures</td>
<td>63</td>
</tr>
<tr>
<td>4.6</td>
<td>identification with top-coding: mean</td>
<td>63</td>
</tr>
<tr>
<td>4.7</td>
<td>identification with top-coding: median</td>
<td>63</td>
</tr>
<tr>
<td>4.8</td>
<td>robustness to outliers: population</td>
<td>64</td>
</tr>
<tr>
<td>5.1</td>
<td>context for QR</td>
<td>68</td>
</tr>
<tr>
<td>5.2</td>
<td>motivation for quantile regression</td>
<td>68</td>
</tr>
<tr>
<td>5.3</td>
<td>QR monotonicity</td>
<td>70</td>
</tr>
<tr>
<td>5.4</td>
<td>quantile crossing problem</td>
<td>73</td>
</tr>
<tr>
<td>6.1</td>
<td>ATE, QTE</td>
<td>78</td>
</tr>
<tr>
<td>6.2</td>
<td>effect heterogeneity</td>
<td>79</td>
</tr>
<tr>
<td>6.3</td>
<td>random coefficient exogeneity: wage</td>
<td>80</td>
</tr>
<tr>
<td>6.4</td>
<td>structural random coefficient wage model</td>
<td>81</td>
</tr>
<tr>
<td>6.5</td>
<td>random coefficient model conditional quantiles</td>
<td>82</td>
</tr>
</tbody>
</table>
LIST OF DISCUSSION QUESTIONS

6.6 QR monotonicity 1 .................................................... 82
6.7 QR monotonicity 2 .................................................... 83
6.8 random coefficient model: heteroskedasticity ................. 83

7.1 MSE of difference .................................................. 88
7.2 panel FE QR: airfare .............................................. 91

8.1 KS tail power 1 ....................................................... 99
8.2 KS tail power 2 ....................................................... 100
8.3 pointwise vs. joint .................................................. 101
8.4 KS band ............................................................... 102
8.5 probability integral transform: discrete? ...................... 103
8.6 Dirichlet tail power ................................................ 104

9.1 two-sample GOF .................................................... 108
9.2 two-sample KS vs. Dirichlet .................................... 110

10.1 SD1 and nonSD1 .................................................... 112
10.2 null of SD1 .......................................................... 112
10.3 null vs. alternative ................................................ 113
10.4 null of non-SD1 ..................................................... 113

11.1 FWER ................................................................... 117
11.2 k-FWER ................................................................ 119
11.3 GOF vs. MTP ......................................................... 121
11.4 MTP for SD1 ........................................................ 121
11.5 MTP for distributional treatment effects ................... 122

12.1 frequentist bias ...................................................... 132
12.2 frequentist SE and CI ............................................. 132
12.3 bootstrap world ...................................................... 133
12.4 empirical bootstrap 1 .............................................. 135
12.5 empirical bootstrap 2 .............................................. 135
12.6 CI properties ........................................................ 137
12.7 deriving CI from sampling distribution ..................... 139

13.1 bootstrap max ........................................................ 145
13.2 bootstrap world correlation: pairs .......................... 148
13.3 bootstrap world correlation: time ............................ 148
13.4 time series bootstrap .............................................. 149
13.5 bootstrap block length ............................................ 150

14.1 are you sick? ........................................................ 157
14.2 beta–binomial vs. frequentist 1 ............................... 160
LIST OF DISCUSSION QUESTIONS

14.3 beta–binomial vs. frequentist
14.4 posterior mean consistency
14.5 Dirichlet posterior mean
14.6 Bayesian bootstrap mean
14.7 Bayesian bootstrap CDF
14.8 Bayesian bootstrap CDF
14.9 Bayesian bootstrap CDF
15.1 best fit: scatter
15.2 best fit: comparison
15.3 curve fitting
15.4 curve fitting
15.5 bias–variance tradeoff
16.1 CEF with binary regressor: tradeoff
16.2 small probability of conditioning event
16.3 discrete kernel regression
16.4 small probability of conditioning event
16.5 local sample size rate
16.6 discrete CEF bias and variance
16.7 discrete CEF MSE
16.8 discrete CEF tradeoff
16.9 local constant local sample size
16.10 local constant flexibility
16.11 local constant boundary
16.12 local constant smoothness
16.13 local constant smoothness
16.14 local constant bandwidth
16.15 local constant bandwidth
16.16 local constant bandwidth
16.17 local constant CI
16.18 local linear boundary bias
16.19 Y as a linear smoother
16.20 OLS as a linear smoother
17.1 approximate CEF
17.2 polynomial models
18.1 OLS flexibility
18.2 OLS flexibility
18.3 validation SSR
18.4 model selection and the CIA
18.5 time series CV
LIST OF DISCUSSION QUESTIONS

18.6 AIC vs. BIC .......................................................... 222
19.1 curse of dimensionality 0 .......................................... 227
19.2 curse of dimensionality 1 .......................................... 227
19.3 curse of dimensionality 2 .......................................... 228
19.4 curse of dimensionality 3 .......................................... 228
19.5 additive wage model .............................................. 229
21.1 RD: scholarship estimand ......................................... 240
21.2 RD: scholarship OLS ............................................. 240
21.3 RD: scholarship heterogeneity .................................... 241
21.4 RD: sharp vs. fuzzy ................................................ 243
21.5 RD: local L/ATE .................................................... 245
21.6 RD: manipulation .................................................. 246
21.7 RD: bandwidth ..................................................... 247
22.1 point ID w/ missing data: mean .................................. 263
22.2 point ID w/ missing data: median ................................ 263
22.3 bounds for binary sample mean .................................. 263
22.4 bounds for sample mean education ............................... 263
22.5 bounds for sample median ........................................ 264
22.6 bounds for binary population mean .............................. 264
22.7 bounds for population mean ....................................... 265
22.8 bounds for population median .................................... 265
22.9 bounds for population CDF ....................................... 265
22.10 IQR bounds ........................................................ 265
23.1 interval-valued mean: identification ............................ 269
23.2 interval-valued mean: estimation .................................. 269
23.3 interval regression: identification ................................. 269
23.4 interval regression: estimation ..................................... 269
23.5 interval regression slope: identification ........................ 269
24.1 ordinal: known thresholds 1 ....................................... 273
24.2 ordinal: known thresholds 2 ....................................... 274
24.3 testable implication ............................................... 275
24.4 super SD1 ............................................................ 275
24.5 latent quantiles ..................................................... 275
24.6 single crossing ..................................................... 276
24.7 single crossing with shift ......................................... 276
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 2020b), 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 or from my website.

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

https://www.overleaf.com/read/bbmwhsvfwgfc
http://faculty.missouri.edu/kaplandm/teach.html
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, ..., $T$, 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

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

and a $k$-dimensional non-random vector is

\[
z = \begin{pmatrix}
z_1 \\
z_2 \\
\vdots \\
z_k
\end{pmatrix}
\]

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

\[
z' = (z_1, z_2, \ldots, z_k)
\]

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

Greek letters like $\beta$ and $\theta$ generally denote fixed population parameters. I sometimes make exceptions to match convention. For example, $\epsilon$ 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 $\theta$ 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 $\theta$: $\hat{\theta}$, $\tilde{\theta}$, 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.

$\begin{align*}
  y & \quad \text{scalar fixed (non-random) value} \\
  Y & \quad \text{scalar random variable} \\
  \theta & \quad \text{scalar non-random value} \\
  \hat{\theta} & \quad \text{scalar random variable} \\
  x & \quad \text{non-random column vector} \\
  x' & \quad \text{transpose of } w \\
  X & \quad \text{random column vector} \\
  \beta & \quad \text{non-random column vector} \\
  \hat{\beta} & \quad \text{random column vector} \\
  w & \quad \text{non-random matrix} \\
  w' & \quad \text{transpose of } w \\
  W & \quad \text{random matrix} \\
  \Omega & \quad \text{non-random matrix} \\
  \hat{\Omega} & \quad \text{random matrix}
\end{align*}$

**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.

\[ \implies \quad \text{implies; see Chapter 3} \]
\[ \iff \quad \text{is implied by; see Chapter 3} \]
\[ \iff \quad \text{if and only if; see Chapter 3} \]
\[ \lim_{n \to \infty} \quad \text{limit} \]
\[ \text{plim}_{n \to \infty} \quad \text{probability limit} \]
\[ \to \quad \text{converges to (deterministic)} \]
\[ \overset{p}{\to} \quad \text{converges in probability to; see \text{Hansen} (2020b, §7.3)} \]
\[ \overset{a.s.}{\to} \quad \text{converges almost surely to; see \text{Hansen} (2020b, §7.14)} \]
\[ \overset{d}{\to} \quad \text{converges in distribution to; see \text{Hansen} (2020b, §8.2)} \]
NOTATION

\[ \nearrow \] converges weakly to
\[ \equiv \] is defined as
\[ \approx \] approximately equals
\[ \cong \] equals when ignoring smaller-order terms
\[ \sim \] is distributed as
\[ \overset{\sim}{\sim} \] is distributed approximately (or asymptotically) as
\( X \perp Y \) \( X \) and \( Y \) are statistically independent
\( N(\mu, \sigma^2) \) normal distribution with mean \( \mu \) and variance \( \sigma^2 \)
\( N(0,1) \) standard normal distribution
\( \Phi(\cdot) \) cumulative distribution function (CDF) of \( N(0,1) \)
\( \phi(\cdot) \) probability density function (PDF) of \( N(0,1) \)
\( F_Y(\cdot) \) cumulative distribution function (CDF) of \( Y \)
\( Q_\tau(Y) \) \( \tau \)-quantile function of \( Y \)
\( Q_\tau(Y | X = x) \) conditional \( \tau \)-quantile function (\( \tau \)-CQF)
\( f_Y(\cdot) \) probability density function (PDF) of \( Y \) (or PMF if discrete)
\( \mathbb{1}\{\cdot\} \) indicator function: \( \mathbb{1}\{A\} = 1 \) if event \( A \) occurs, else \( \mathbb{1}\{A\} = 0 \)
\( P(A) \) probability of event \( A \)
\( P(A | B) \) conditional probability of \( A \) given \( B \)
\( E(Y) \) expected value of \( Y \)
\( \hat{E}(Y) \) expectation for sample distribution; same as \( \frac{1}{n} \sum_{i=1}^n Y_i \)
\( E(Y | X = x) \) CEF (function of \( x \)); see Hansen (2020a, §2.5)
\( E(Y | X) \) expected value of \( Y \) given \( X \); this is a random variable
\( Q_\tau(Y) \) \( \tau \)-quantile of \( Y \); see Section 4.2
\( Q_\tau(Y | X = x) \) conditional \( \tau \)-quantile function (\( \tau \)-CQF); see Section 5.1
\( \text{Var}(Y) \) variance of \( Y \)
\( \text{Var}(Y | X = x) \) conditional variance (a non-random value)
\( \text{Var}(Y | X) \) conditional variance (a random variable)
\( \text{Cov}(Y,X) \) covariance
\( \text{Corr}(Y,X) \) correlation
\( b \in \{a,b,c\} \) \( b \) is in the set containing \( a \), \( b \), and \( c \)
\( S_1 \cup S_2 \) the union of sets \( S_1 \) and \( S_2 \)
\( \bigcup_{j=1}^J S_j \) the union of \( S_1, \ldots, S_J \)
\( \bigcap_{j=1}^J S_j \) the intersection of \( S_1, \ldots, S_J \)
\( N \) 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
\( \mathbb{R} \) the extended real numbers, \( \mathbb{R} \cup \{-\infty, \infty\} \)
\( \mathbb{R}^k \) \( k \)-dimensional Euclidean space
\( Z \) the set of integers, \( \{\ldots,-2,-1,0,1,2,\ldots\} \)
NOTATION

\begin{align*}
Z_{\geq 0}, Z_{> 0} & \quad \text{analogous to } \mathbb{R}_{\geq 0} \text{ and } \mathbb{R}_{> 0} \\
\text{SE}(\hat{\theta}) & \quad \text{standard error of estimator } \hat{\theta} \\
\arg \min_g f(g) & \quad \text{the value of } g \text{ that minimizes } f(g) \\
I_k & \quad k \times k \text{ identity matrix (ones on main diagonal, zeros elsewhere)} \\
\| \cdot \| & \quad \text{norm (Euclidean unless otherwise defined)} \\
\text{tr}(\mathbf{v}) & \quad \text{trace of matrix } \mathbf{v} \\
\mathbf{v}' & \quad \text{transpose of matrix } \mathbf{v} \\
\mathbf{v}^{-1} & \quad \text{inverse of matrix } \mathbf{v} \\
\mathbf{v} > 0 & \quad \text{matrix } \mathbf{v} \text{ is positive definite} \\
\mathbf{v} \geq 0 & \quad \text{matrix } \mathbf{v} \text{ is positive semi-definite}
\end{align*}
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. 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.

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.).
5. The graphics look the nicest to me.
6. It’s flexible (easy to create new functions, etc.).

https://doit.missouri.edu/services/software/software-anywhere
https://doit.missouri.edu/services/computing-sites/sites-software
STATISTICAL SOFTWARE OVERVIEW

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 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 ⇒ 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.
   The first Google hit is currently a Princeton professor’s tutorial; I haven’t looked through it, but it’s probably pretty good, right?

https://stats.idre.ucla.edu/stata
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](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....
STATISTICAL SOFTWARE OVERVIEW
Part I

Writing, Coding, and Logic
Introduction

This first part relates to econometrics more tangentially. I provide my opinions about effective written communication, some basics for coding in R, and basic logic concepts and terminology.
Chapter 1

Writing and Typesetting

Optional resources for this chapter

- Overleaf registration: https://www.overleaf.com/register
- My \LaTeX\ templates, including job market candidate (JMC) templates: https://www.overleaf.com/read/gtfzpkrzhhw
- 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. 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 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/register
CHAPTER 1. WRITING AND TYPESETTING

- https://www.overleaf.com/read/mvkpywqyxhvt (paper and talks/slides)
- 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, which have sometimes popped up on Google for me and seem to be helpful. The StackExchange site is also helpful (and is usually the first Google result).

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.”

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. 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?

https://www.overleaf.com/learn
https://tex.stackexchange.com
http://stat.wharton.upenn.edu/~buja/sci.pdf
1.2. WRITING ADVICE

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.
CHAPTER 1. WRITING AND TYPESETTING

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.

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...
1.2. WRITING ADVICE

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 es-
timates 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 discriminatino 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 exam-
ple, 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 appro-
priate. Instead, stylistic features like sound devices (rhyme, alliteration, etc.) and
part-of-speech frequencies take prominence.
CHAPTER 1. WRITING AND TYPESETTING

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).
1.2. WRITING ADVICE

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.”
20

CHAPTER 1. WRITING AND TYPESETTING

• “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 $\tau$. 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 \( \frac{\partial m(x)}{\partial x_k} = \beta_k \) 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 everything 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
1.3. COMMON MINOR MISTAKES

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 $H_0: \beta = 0$ 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*.

1.3 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).

2. It is simply “University of Missouri” and not “University of Missouri–Columbia” as stated in the official MU style guide.

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 \texttt{Dr. K} to \texttt{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.

[https://styleguide.missouri.edu/term/university-identification](https://styleguide.missouri.edu/term/university-identification)
CHAPTER 1. WRITING AND TYPESETTING

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 \textup or \textup. Example: instead of $FWER = \alpha$, write $FWER = \alpha$; 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).

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 =)

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,...,k,...,n.

13. In math mode, I’d suggest just using \ldots instead of guessing whether to use \ldots or \cdots, since it usually gets it right automatically; using only \ldots: $1,\ldots,n$ 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 $P(\cdot)$ instead of $P(\cdot)$, since upright $P$ looks like an operator while slanted $P$ looks like a variable. Similarly for $E(\cdot)$ and $Q(\cdot)$. But CDFs and quantile functions are functions, not operators on random variables, hence $F_Y(\cdot)$ and $Q_Y(\cdot)$.

\[http://blog.apastyle.org/apastyle/2012/01/that-versus-which.html\]
1.3. COMMON MINOR MISTAKES

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 ×.

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.
CHAPTER 1. WRITING AND TYPESETTING

Exercises

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

Exercise E1.2. At https://www.overleaf.com/read/gtfzpkrzhhw get the _paper.tex template. 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/gtfzpkrzhhw get the _talk.tex template. 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

See also Chapter 1 of Kaplan (2020b).

My goals for this section are to give you the basic tools to be able to
1. download, run, and maintain R software,
2. write/run/save new data analysis with .csv or .dta data,
3. write/run/save new Monte Carlo simulations, and
4. learn new things on your own.

There are 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.

2.1 Getting Help

At first, it may help to have some quick reference “cheat sheets.” 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)
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
6. CRAN

https://cran.r-project.org/doc/contrib/Short-refcard.pdf
http://www.cyclismo.org/tutorial/R
http://cran.r-project.org/doc/manuals/r-release/R-intro.html
CHAPTER 2. R: SOME BASICS

7. No longer free beyond first chapter: courses at datacamp.com like Introduction to R.

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 (Spring 2020) the best option seems to be Microsoft Azure.
A free account is required to sign in.
Download R for Windows: Google “r windows” and try the first result.
Download R for Mac: Google “r mac” and try the first result to find the newest .pkg.
Download RStudio (free, nicer interface): Google “rstudio download” and try the first result.

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 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. GETTING STARTED

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 since 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 (these are hyperlinks):

- Cyclismo: data types
- CRAN: commands, case sensitivity
- CRAN: simple numerical manipulations
- CRAN: arrays and matrices
- CRAN: lists and data frames
- Cyclismo: vector indexing

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:

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

---

[1](http://www.cyclismo.org/tutorial/R/types.html)
[2](http://cran.r-project.org/doc/manuals/r-release/R-intro.html#R-commands_/zero.alt3/zero.alt33b-case-sensitivity-etc)
[3](http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Simple-manipulations-numbers-and-vectors)
[4](http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Arrays-and-matrices)
[5](http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Lists-and-data-frames)
[6](http://www.cyclismo.org/tutorial/R/vectorIndexing.html)
2.3. DATA TYPES

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 readability 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).[16] 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
```

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., \(x[c(\text{FALSE}, \text{TRUE}, \text{TRUE})]\) returns the 2nd and 3rd elements of \(x\). You can extract multiple named columns from a data frame with something like \(d[,c(\text{'age'}, \text{'edu'}, \text{'wage'})]\).

### 2.4 Basic Data Manipulation

See also: [http://www.cyclismo.org/tutorial/R/basicOps.html](http://www.cyclismo.org/tutorial/R/basicOps.html)

#### 2.4.1 Numerical Operations

Most of the numerical operators are relatively intuitive. For example, \(2+2\), \(4/2\), \(2*2\), \(2^2\), \(\sqrt{4}\), \(\exp(1)\), \(\log(2.71)\), \(\log10(100)\), \(\text{abs}(-2)\), \(\text{floor}(2.9)\), \(\text{ceiling}(1.1)\), \(\text{round}(2.49)\). A few less intuitive things: \(\%\%\) for modulo/remainder; \(\text{round}(2.5)\) 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 \(\text{seq}(\text{from}=1, \text{to}=3, \text{by}=1)\). You can repeat values/sequences like \(\text{rep}(1:3, \text{each}=2)\) or \(\text{rep}(1:3, \text{times}=2)\) (these results differ; try them). You can fill a matrix like \(\text{matrix}(1:6, \text{nrow}=3)\) or equivalently \(\text{matrix}(1:6, \text{ncol}=2)\).

#### 2.4.2 Combining Data

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

#### 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 \(\text{paste0()}\):

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

Substring:
2.4. BASIC DATA MANIPULATION

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

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= 5.2/zero.alt3"

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

You can also insert strings:

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 "
TABLE 2.1: Table generated from R output.

\begin{tabular}{lSSSlSSS}
\toprule & \multicolumn{3}{c}{Bias} && \multicolumn{3}{c}{Variance} \\
Method & DGP1 & DGP2 & DGP3 & DGP1 & DGP2 & DGP3 \\
\midrule
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 \\
\bottomrule
\end{tabular}

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

\begin{verbatim}
> sprintf("Hello, %s",c("world","Dave"))
[1] "Hello, world" "Hello, Dave"
\end{verbatim}

These can in turn be combined into a single text string with \texttt{paste0}:

\begin{verbatim}
> paste0(sprintf("Hello, %s",c("world","Dave")),collapse="; ")
[1] "Hello, world; Hello, Dave"
\end{verbatim}

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

\begin{verbatim}
> set.seed(112358)
> head <- \"\begin{tabular}{lSSSlSSS}\n\toprule\n& \multicolumn{3}{c}{Bias} && \multicolumn{3}{c}{Variance} \\
Method & \rot{DGP1} & \rot{DGP2} & \rot{DGP3} & DGP1 & DGP2 & DGP3 \\
\midrule
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 \\
\bottomrule\end{tabular}\n\end{verbatim}
2.5 Functions

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

```r
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:

```r
> 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:

```r
> 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](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:

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

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

```r
> 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 [http://faculty.missouri.edu/kaplandm/data/OXY_daily_data_no_holidays.csv](http://faculty.missouri.edu/kaplandm/data/OXY_daily_data_no_holidays.csv) into the working directory. Then,

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

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

```r
> 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
```
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 \texttt{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:

\begin{verbatim}
> 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
\end{verbatim}

\section*{2.7 Basic Statistics}

See also: Cyclismo tutorials on basic operations\footnote{http://economics.mit.edu/faculty/angrist/data1/data/angchefer06} and OLS\footnote{http://www.cyclismo.org/tutorial/R/linearLeastSquares.html}.

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 \texttt{apply()} below.

For example, there is \texttt{mean()}, \texttt{median()}, \texttt{sd()}, \texttt{quantile()}, \texttt{sum()}, \texttt{min()}, \texttt{max}(), etc. Note if you want a “parallel” (element-wise) min or max, then use \texttt{pmin()} or \texttt{pmax()}:

\begin{verbatim}
> min(matrix(1:6,ncol=2))
[1] 1
> pmin(1:3,4:2)
[1] 1 2 2
\end{verbatim}

See also \texttt{which.min()} and \texttt{which.max()}, which return the index of the minimum and maximum. There’s also a \texttt{which()} that returns the indices of \texttt{TRUE} elements in a logical vector.

\footnote{http://economics.mit.edu/faculty/angrist/data1/data/angchefer06} \footnote{http://www.cyclismo.org/tutorial/R/basicOps.html} \footnote{http://www.cyclismo.org/tutorial/R/linearLeastSquares.html}
CHAPTER 2. R: SOME BASICS

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 22. But, sometimes you just want to remove all the NA. Compare:

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

See Section 22.5.5 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`):

```r
> (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[^20] and CRAN[^21].

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

```r
> plot(x=1:1/zero.alt3, y=11:2/zero.alt3)
```

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)L{$\LaTeX$} 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.

[^21]: http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Graphics
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:

```r
> (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[^22]

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 \( \mu \), and the standard deviation \( \sigma \), 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., 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,...,\( n \). For example,

```r
> 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:

[^22]: http://www.cyclismo.org/tutorial/R/probability.html
> 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\textsuperscript{23} and CRAN\textsuperscript{24}

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 \texttt{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 \texttt{NULL}
or \texttt{NA}. The condition needs to be inside parentheses (unlike Matlab, etc.). You can also
insert any number of \texttt{else if} blocks. Some examples:

```r
> 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")
```

\textsuperscript{23}http://www.cyclismo.org/tutorial/R/scripting.html
\textsuperscript{24}http://cran.r-project.org/doc/manuals/r-release/R-intro.html#Loops-and-conditional-execution
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. 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:

```r
> 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
> 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:

Footnote:

25 See also [http://www.cyclismo.org/tutorial/R/types.html#logical](http://www.cyclismo.org/tutorial/R/types.html#logical)
> 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:

```r
> for (i in 1:4) {
+   if (i%%2 == /zero.alt3) { 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
2.12 TIME AND TIMING

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

```r
(x <- Sys.time())
> 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:

```r
set.seed(112358); x <- rbinom(n=1e8, size=1, 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
```

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:

```r
library(doRNG); library(doParallel)
library(foreach); library(parallel)
workers <- makeCluster(detectCores()) #use everything available
```

CHAPTER 2. R: SOME BASICS

> 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,] 0.4143905 -0.49700764 1.2869777 -1.29213296  0.5669822
[2,] 0.4143905 -0.49700764 1.2869777 -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).

```r
# 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(paste("NREPLIC=%d, N=%d, MU=%g, ","SIGMA=%g"),
    NREP, N, MU, SIGMA),
```
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).
CHAPTER 2. R: SOME BASICS

# 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
  betas <- 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%*%betas + 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='
'), '
',
       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
2.15. SIMULATION: EXAMPLE #2

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 \texttt{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 \texttt{set.seed} were moved above the first \texttt{for}? Why/not?

d) Try changing \texttt{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 \texttt{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 \texttt{lm} include an intercept term? How do you know?

h) What does the code \texttt{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. Let $n = 100$. 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 $\rho$. So, $\rho$ controls the degree of endogeneity (with $\rho = 0$ for exogenous $X$), and $\gamma_1$ controls the degree of relevance of the instrument, or the “strength” of the instrument (with $\gamma_1$ implying the instrument is not relevant and thus not valid). Let $\rho = 0.5$. Try different values of $\gamma_1$, 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 $\hat{\beta} = (Z'X)^{-1}Z'Y$, where $Z$ contains a column of ones and a column of $Z_i$, and $X$ contains a column of ones and a column of $X_i$, and $Y = (Y_1, \ldots, Y_n)'$. Compute the estimated covariance matrix like in (5.34) of Wooldridge (2010), $(\hat{X}'\hat{X})^{-1}(\sum_{i=1}^{n} \hat{U}_i^2 \hat{X}_i \hat{X}_i') (\hat{X}'\hat{X})^{-1}$ where $\hat{X}_i = (1, \hat{X}_i)$ with $\hat{X}_i = \hat{\gamma}_0 + \hat{\gamma}_1 Z_i$ (estimated linear projection); the standard error for $\beta_1$ is the square root of the $(2, 2)$ entry in that matrix. Compute a “95%” CI for $\beta_1$ as $\hat{\beta}_1 \pm 1.96 \text{SE}(\hat{\beta}_1)$. 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 $\beta_1$. Also compute the proportion of replications in which the true $\beta_1$ lay in 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

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.

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. $B \iff A$
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):

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>A $\implies$ B</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

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 $A \iff B$)
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):

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>A $\iff$ B</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

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.)
3.2 ASSUMPTIONS

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 assumptions $A$ and conclusion $B$, let $A$ and $B$ denote different assumptions. For example, let $A$ be $E(Y^4) < \infty$, and let $B$ be $E(Y^2) < \infty$. Any random variable $Y$ with finite $E(Y^4)$ also has finite $E(Y^2)$, but some have finite $E(Y^2)$ and infinite $E(Y^4)$. That is, $A \implies B$. Thus, people say “$E(Y^4) < \infty$ is a stronger assumption than $E(Y^2) < \infty$,“ or equivalently, “$E(Y^2) < \infty$ is weaker than $E(Y^4)$.”

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 $E(U \mid X) = 0$, and let $B$ be $E(U) = 0$ and 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, it is nicer to have results based on weaker assumptions. In the previous example, the weaker assumption $E(Y^2) < \infty$ is true for more distributions of $Y$ than is the stronger assumption $E(Y^4) < \infty$. If $Y$ has a distribution with $E(Y^2) < \infty$ but $E(Y^4) = \infty$, then only results based on the weaker assumption can be applied. For this reason, results with weaker assumptions are often called “more general” than results based on stronger assumptions: they apply to more situations.

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.
CHAPTER 3. LOGIC

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

Quantiles can be useful for both description and prediction. By “description,” I mean it can help us learn about features of the population distribution. For prediction (guessing $Y$), although the mean is optimal for quadratic loss, quantiles are optimal for other loss functions that in particular allow asymmetry. This chapter considers the unconditional distribution of $Y$ to introduce concepts. Chapter 5 extends this to the conditional (on $X = x$) distribution of $Y$.

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 about the full distribution. Some common features are the mean and variance, $E(Y) = \int_{\mathbb{R}} y \, dF_Y(y)$ and $\text{Var}(Y) = E[(Y - E(Y))^2] = \int_{\mathbb{R}} [y - E(Y)]^2 \, dF_Y(y)$. If $Y$ is normally distributed, then this is sufficient. However, most variables are not normally distributed.

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 $\tau$-quantile is the same as the $100\tau$th percentile: the value for which $\tau$ proportion of the population has a smaller value. (There are some caveats; see below.) This tells us the opposite of the CDF $F_Y(\cdot)$: $F_Y(y) = P(Y \leq y)$.

The $\tau$-quantile’s formal definition and notation follow. Let $\tau \in [0, 1]$ denote the quantile index (or quantile level). Let $Q_{\tau}(Y)$ denote the $\tau$-quantile of random variable
Y. The intuition from above suggests \( Q_\tau(Y) \) satisfies \( P(Y \leq Q_\tau(Y)) = \tau \), or equivalently \( F_Y(Q_\tau(Y)) = \tau \). This further suggests \( Q_\tau(Y) = F_Y^{-1}(\tau) \) if the CDF \( F_Y(\cdot) \) is invertible. More generally,

\[
Q_\tau(Y) \equiv \inf\{y : F_Y(y) \geq \tau\}. \tag{4.1}
\]

The quantile function \( Q_Y(\cdot) \) instead expresses the quantiles of \( Y \) as a function of \( \tau \). 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).

The reliance of the definitions in (4.1) and (4.2) on the CDF emphasizes that quantiles are features that help describe the distribution of \( Y \).

**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 \( Q_{0.5}(Y) \).

c) Visually locate \( Q_{0.8}(Y) \), iv) sketch \( Q_Y(\cdot) \).

**4.3 Prediction**

Like the mean, quantiles have both descriptive and predictive interpretations.

**Discussion Question 4.2** (population minimization: quadratic and absolute loss). Let \( Y \) be a discrete rv with \( P(Y = 1) = P(Y = 2) = P(Y = 99) = 1/3 \).

a) Compute \( \theta_1 = \arg \min_{t \in \mathbb{R}} \mathbb{E}[(Y - t)^2] \).

b) Compute \( \theta_2 = \arg \min_{t \in \mathbb{R}} \mathbb{E}(|Y - t|) \).
4.3. PREDICTION

c) What are the common names for $\theta_1$ and $\theta_2$?

Some vocabulary will be 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 long-run average loss given fixed $g$ is thus the **expected loss** (also called **risk**), where the expectation is wrt the distribution of $Y$: $E[L(Y, g)]$. 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 $y$. In the frequentist framework, given loss function $L$, the optimal predictor minimizes risk (expected loss):

$$g^*_L \equiv \arg\min_g E[L(Y, g)].$$

(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

$$L_2(y, g) = \rho_2(y - g) = (y - g)^2.$$  

(4.4)

Then, the mean is optimal in that

$$E(Y) = \arg\min_g E[L_2(Y, g)].$$

(4.5)

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

$$0 = \frac{d}{dg} E[L_2(Y, g)] \bigg|_{g=\hat{g}_2} = \frac{d}{dg} E[(Y - g)^2] \bigg|_{g=\hat{g}_2} = 2 E[Y - \hat{g}_2^2],$$

so $\hat{g}_2^2 = E(Y)$.

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 minimizes $E[L_1(Y, g)]$:

$$Q_{0.5}(Y) = \arg\min_g E[L_1(Y, g)].$$

(4.6)

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

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

(4.7)
CHAPTER 4. QUANTILES: DESCRIPTION AND PREDICTION

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

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

More generally,

$$Q_{\tau}(Y) = \arg\min_g E[\rho_{\tau}(Y - g)]. \quad (4.8)$$

That is, given Definition 4.1, the $\tau$-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 $\tau$?

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 $Q_{0.95}(Y)$ is a better predictor than $Q_{0.5}(Y)$. 
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 \( \hat{\theta}_2 \)?

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 (unconditional, conditional, panel, IV, censored, etc.).

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, let the CDF estimator be

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

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. The population mean is

\[
E(Y) = \int_{\mathbb{R}} y \, dF_Y(y).
\]

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

\[
\hat{E}(Y) = \int_{\mathbb{R}} y \, d\hat{F}_Y(y) = \sum_{i=1}^{n} Y_i (1/n) = \bar{Y}_n,
\]

the familiar sample mean. For \( Q_\tau(Y) \), we can replace \( F_Y(\cdot) \) in (4.1) to get

\[
\hat{Q}_\tau(Y) = \inf\{y : \hat{F}_Y(y) \geq \tau\},
\]

often called the sample \( \tau \)-quantile.

Second, related to prediction: we can solve the sample version of the population minimization characterizing the “best” predictor. For the mean, we had

\[
E(Y) = \arg \min_g \mathbb{E}[(Y - g)^2]
\]

in the population. Replacing the population expectation \( \mathbb{E}[\cdot] \) with the sample expectation (sample average) \( \hat{\mathbb{E}}[\cdot] \), although I suppose it’s somewhat circular in this case,

\[
\hat{E}(Y) = \arg \min_g \hat{\mathbb{E}}[(Y - g)^2] = \arg \min_g \frac{1}{n} \sum_{i=1}^{n} (Y_i - g)^2.
\]
CHAPTER 4. QUANTILES: DESCRIPTION AND PREDICTION

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

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

Although the minimization approach was used first for QR, the CDF-based approach has also been used, especially for nonparametric estimation or distributions that may have (some) discrete mass point(s).

Yet a third approach is quasi-maximum likelihood estimation. Recall that using a $N(\mu, \sigma^2)$ likelihood leads to $\hat{\mu} = \bar{Y}_n$, a consistent estimator of $E(Y)$ even if the true distribution is not normal. Instead using an asymmetric Laplace distribution likelihood yields a consistent estimator of $Q_\tau(Y)$ even if the true distribution is not asymmetric Laplace.

Variations of these approaches have been suggested to improve finite-sample (or higher-order asymptotic) properties. For example, Nadaraya (1964b) proposes using a “smoothed” $\hat{F}_Y(\cdot)$ in the minimization approach. My personal favorite is a weighted average of certain order statistics (i.e., sample values sorted into increasing order) that you can get in R’s quantile() function with argument type=6.

4.5 Censoring

Quantiles (and QR) can be 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 common example of censoring is top-coding of earnings data. This is seen in the Survey of Income and Program Participation (SIPP), the National Longitudinal Survey(s) of Youth (NLSY), and Current Population Survey (CPS). The motivation for this type of censoring is (I think) to protect privacy for individuals with notably higher earnings. The simplest version of top-coding replaces earnings values above some threshold (like $150,000/yr) with the threshold value. Other versions replace the true income with an average over a certain demographic group.

One approach to top-coding is to try to impute values, i.e., guess what the true values are. After imputation, the resulting dataset is treated like any other dataset. MU econ PhD Li Tan has 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 (2017).

Another approach to top-coding is to ask economic questions that don’t rely on the very upper tail. Consider income inequality. We could quantify “inequality” as variance, but variance depends crucially on the very upper tail that we don’t observe. (Similarly, the sample variance can become arbitrarily large if just a single observation is made...
4.5. CENSORING

arbitrarily large.) Alternatively, we could quantify inequality as the difference between the 0.9-quantile and the 0.1-quantile. This is also called the 0.9–0.1 interquartile range. (The 0.75–0.25 interquartile range is called the interquartile range; a bit confusing.)

**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)](Hansen2020) §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 \( F \) be a set of possible joint distributions of observable variables. Parameter \( \theta \in \mathbb{R} \) is **identified** on \( F \) if \( F \) uniquely determines \( \theta \) for all \( F \in F \).

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

\[
Y = \begin{cases} 
Y^* & \text{if } Y^* \leq c \\
 0 & \text{if } Y^* > c.
\end{cases}
\]  

Since \( P(Y = c) = P(Y^* \geq c) \), 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

\[
F(y) = \begin{cases} 
F^*(y) & \text{if } y \leq c \\
 1 & \text{if } y > c
\end{cases}
\]  

**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 \( E(Y^*) \).

**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 $F$ be restricted to ensure they always have the same median?

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

This idea can be extended to quantile regression, too.

### 4.6 Robustness and Efficiency

You may hear that the median is more “robust” than the mean. Any time you hear the word “robust,” you should first ask: robust to what? Here, people would say, “robust to outliers.” But that begs the question: what’s an “outlier”? I’ll try to give more objective descriptions below.

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. There are other particularly “fat-tailed” distributions for which the mean is not defined, but the median always is.

Even if the mean is defined, fat tails may make the sample mean’s variance much larger than the sample median’s variance. That is, the median could be preferred because of better estimation efficiency. In the extreme, the sample mean’s variance could be infinite if the population has infinite variance, like with a $t$-distribution with two degrees of freedom. For regression, efficiency depends on the conditional distribution of $Y$.

Also, the median is less sensitive to very large but unlikely values. This is true both in the population and in the sample. This was seen (somewhat) in DQs 4.2 and 4.4. For example, imagine a discrete population distribution with

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

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

**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 mean capture that the median doesn’t?

The median is also less sensitive to single large values in the sample. We can 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 $\bar{Y}_n = Y_n \to \infty$ 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 somewhat more robust to outliers than OLS. Specifically, it is relatively robust to outliers in $Y_i$, but the slope estimate can
be made arbitrarily large (or small) by changing just a single point \((Y_i, X_i)\), i.e., making \(X_i\) arbitrarily large (and either setting \(Y_i = \bar{Y}\) to make the slope near zero or taking \(Y_i \to \infty\) to make the slope arbitrarily large). Quantile regression is thus not particularly robust to measurement error in \(X_i\).

There are more robust regression methods if that is your primary concern. The **breakdown point** is the smallest fraction of the sample that if contaminated could cause the estimate to deviate arbitrarily far from its value with the uncontaminated sample. For OLS and QR, a single observation is enough to arbitrarily change the estimate, even as \(n \to \infty\), so the breakdown point is \(1/n \to 0\) asymptotically. In contrast, the “least median of squares” (Rousseeuw, 1984) regression estimator has a much higher breakdown point, asymptotically \(1/2\). (This is as good as possible without becoming nonsense: the estimate should change if, say, 90% of the sample changes.) The least median of squares estimator is defined as the set of parameters that minimizes the median of \((Y_i - X_i'\beta)^2\) rather than the average (as OLS does). Koenker (2005, §8.5) discusses proposals for higher-breakdown quantile regression methods.

### 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, \(\sqrt{n}(\bar{Y}_n - E(Y)) \overset{d}{\to} N(0, \sigma^2)\), 2) estimate the unknown \(\sigma^2\) 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, \(\sqrt{n}(\hat{Q}_\tau(Y) - Q_\tau(Y)) \overset{d}{\to} N(0, \sigma^2)\), \(\sigma^2 = \tau(1 - \tau)/[f_Y(Q_\tau(Y))]^2\), (4.19) where \(f_Y(\cdot)\) is the PDF of \(Y\). Thus, given \(\hat{\sigma} \overset{p}{\to} \sigma\), the CI \(\hat{Q}_\tau(Y) \pm 1.96\hat{\sigma}/\sqrt{n}\) has coverage probability approaching 95% as \(n \to \infty\).

However, \(\sigma\) is “more difficult” to estimate with quantiles. Assuming a parametric assumption (like “\(Y\) is normal”) is unreasonable, \(f_Y(Q_\tau(Y))\) must be estimated nonparametrically. Computationally, this is still easy; e.g., there are many nonparametric “kernel density estimators” available in R. However, the estimation error tends to be larger, which is captured theoretically by the convergence rate. For the mean, \(\sqrt{n}(\hat{\sigma} - \sigma)\) converges to a normal distribution, whereas here, the \(\sqrt{n}\) is replaced by \(n^{2/5}\).

Theoretically, there is no problem since \(\hat{\sigma}\) is still consistent: \(\hat{\sigma} \overset{p}{\to} \sigma\). There is no requirement on the rate of convergence for the CI to have asymptotically correct coverage probability. However, in practice (with finite \(n\)), naturally the CI accuracy depends on the accuracy of \(\hat{\sigma}\). That is, for a given \(n\), the standard CI for \(E(Y)\) is probably more accurate than the corresponding CI for \(Q_\tau(Y)\).

Consequently, 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 (2019). 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, which is impossible for the mean. Let $F_Y(\cdot)$ be the continuous CDF of $Y$. Then, $F_Y(Y_{n:k}) \sim \text{Beta}(k, n + 1 - k)$; e.g., see Wilks (1962, pp. 236–238). The notation $Y_{n:k}$ is the $k$th order statistic: if we order the sample $Y_i$ values as $Y_{n:1} < Y_{n:2} < \cdots < Y_{n:k} < \cdots < Y_{n:n}$, then this is the $k$th-smallest value in the sample. Then the coverage probability of the one-sided CI $[Y_{n:k}, \infty)$ for $Q_\tau(Y)$ is

$$P(Y_{n:k} \leq Q_\tau(Y)) = P(F_Y(Y_{n:k}) \leq F_Y(Q_\tau(Y))) = P(\text{Beta}(k, n + 1 - k) \leq \tau), \quad (4.20)$$

which can be computed by \texttt{pbeta(tau,k,n+1-k)} in R. That is, for any $k$, the finite-sample coverage probability can be computed exactly.
Chapter 5

Quantile Regression: Description and Prediction

Like unconditional quantiles, quantile regression (QR) helps with description and prediction. For description, QR captures more of the conditional distribution of $Y$ given $X = x$ than just the mean. For prediction (guessing $Y$ given $X = x$), 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 some of those in Chapter 2 of Hansen (2020a) for mean regression.

Historically, while the idea behind QR goes back to Laplace, development of the modern QR literature is largely due to Roger Koenker. Koenker and Bassett (1978) introduced QR to the econometrics literature, and Koenker (2005) explains the different
CHAPTER 5. QUANTILE REGRESSION: DESCRIPTION AND PREDICTION

directions and developments in the intervening decades. The quantreg package in R, also Koenker’s work, has made many of these methods easy to use. More recently, many other different econometricians have contributed new models and methodology to the quantile literature.

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 (i.e., describing features of the joint distribution of observable variables)
b) Prediction (guessing $Y$ based on $X$);
c) Causality

5.1 Description

Consider the conditional distribution of $Y$ given $X = x$. That is, within the overall population, there is a subpopulation with $X = x$, 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.

Parallel to how “regression” refers to the conditional mean as a function of $x$, “quantile regression” refers to the conditional quantiles. Parallel to how the word “regression” is used with multiple meanings, the phrase “quantile regression” is also used with multiple meanings, so beware.

5.1.1 Conditional Quantile Function

Previously, you have learned about the conditional expectation function (CEF), $E(Y \mid X = x)$. The CEF tells us the mean of the conditional distribution of $Y$ given $X = x$, for every $x$. This lets us see how the mean of $Y$ varies across subpopulations defined by $X = x$.

Similar to the unconditional setting, there are other features of the conditional distribution of $Y$ that are interesting. The conditional variance $\text{Var}(Y \mid X = x)$ is one measure of how dispersion varies across subpopulations. Of course, we will focus on conditional quantiles.

The conditional $\tau$-quantile of $Y$ given $X = x$ is

$$Q_\tau(Y \mid X = x) \equiv \inf\{y : F_{Y \mid X}(y \mid X = x) \geq \tau\}.$$  (5.1)
5.1. DESCRIPTION

This parallels (4.1), just conditioning on \( x \). Parallel to the term conditional expectation function (which is also a function of \( x \)), \( Q_\tau(Y \mid X = x) \) is called a **conditional quantile function** (CQF). When specifying \( \tau \) explicitly is important, I write \( \tau \)-CQF, although of course it should be \( C\{\tau\text{-Q}\}F \) or something. Also, parallel to (4.2), the quantile function of \( Y \) conditional on \( X = x \) is

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

Like the CEF, the CQFs describe how the distribution of \( Y \) varies with \( x \) without overwhelming us with the full conditional CDF. The conditional CDF is like a function-valued function of \( x \), i.e., a different function at each possible \( x \). The CEF and CQFs are functions of \( x \), but at least scalar-valued; for scalar \( x \), they are easily plotted. 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 \( q_\tau(x) \) by any of these:

\[
q_\tau(x) = Q_\tau(Y \mid X = x),
\] (5.3)

\[
Y = q_\tau(X) + V, \quad Q_\tau(V \mid X) = 0,
\] (5.4)

\[
\tau = P(Y \leq q_\tau(X) \mid X).
\] (5.5)

If a linear model is (optimistically) specified, then

\[
q_\tau(x) = Q_\tau(Y \mid X = x) = x'\beta(\tau),
\] (5.6)

\[
Y = X'\beta(\tau) + V, \quad Q_\tau(V \mid X) = 0,
\] (5.7)

\[
\tau = P(Y \leq X'\beta(\tau) \mid X).
\] (5.8)

5.1.3 Monotonicity

Without further assumptions, the CQFs could have any shape with respect to \( x \), but they must obey a certain monotonicity in \( \tau \). Let \( 0 < s < t < 1 \). In the unconditional case, by definition, \( Q_s(Y) \leq Q_t(Y) \). This remains true conditional on any \( X = x \):

\[
Q_s(Y \mid X = x) \leq Q_t(Y \mid X = x).
\]

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

Alternatively, monotonicity can be written in terms of (conditional) quantile functions. Again let \( 0 < s < t < 1 \). Unconditionally, by definition, \( Q_Y(s) \leq Q_Y(t) \). Conditionally, \( Q_{Y|X}(s \mid X = x) \leq Q_{Y|X}(t \mid X = x) \) for any \( x \).
This is analogous to CDF monotonicity. That is, for \( c < d \), \( F_Y(c) \leq F_Y(d) \) unconditionally, and \( F_{Y|X}(c \mid X = x) \leq F_{Y|X}(d \mid X = x) \) conditional on any \( x \). It makes sense that (conditional) quantile functions and CDFs share some properties since the quantile function is the (generalized) inverse of the CDF.

Monotonicity plays a large role in Chapter 6.

Discussion Question 5.3 (QR monotonicity). Consider a simple quantile regression model with intercept \( \alpha(\tau) \) and slope \( \beta(\tau) \).

a) Does quantile monotonicity imply that we (should) have \( \beta(0.25) < \beta(0.5) \)?

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 CEF is the “best” predictor of \( Y \) given \( X \) under quadratic loss. If we imagine repeated draws of \( Y \) given a single fixed \( X = x \), then we are essentially in the unconditional setting: we simply treat the subpopulation with \( X = x \) as the population and apply (4.5). If instead we imagine repeated draws of \( (Y, X') \) from the joint population distribution, then the optimality of the CEF is less trivial but still true. As shown by Hansen (2020a, §2.11),

\[
E(Y \mid X) = \arg \min_{g(X)} E[\rho_2(Y - g(X))] = \arg \min_{g(X)} E[\rho_2(Y - g(X))] = \arg \min_{g(X)} E[(Y - g(X))^2].
\]

(5.9)

Here, \( E(Y \mid X) \) is a random variable rather than a function of \( x \): it conditions on the random variable \( X \), not on a particular value \( X = x \). In \( g(X) \), \( g(\cdot) \) could be any possible scalar-valued function of vectors the dimension of \( X \). If we observe \( X = x \), then we guess (predict) \( y = g(x) \); if we observe \( X = \tilde{x} \), then we guess \( g(\tilde{x}) \); etc. Thus, the probability distribution of \( X \) induces a probability distribution of \( g(X) \), a random variable. The “best” predictor \( g(\cdot) \) gets the random variable \( g(X) \) “closest” to \( Y \) in the stochastic sense of minimizing mean squared prediction error, \( E[(Y - g(X))^2] \). Defining “closest” differently results in a different “best” predictor.

Replacing the quadratic loss function in (5.9) can make the optimal predictor not the CEF but rather the CQF. This extends the result in (4.8) for unconditional quantiles as optimal predictors. Now considering predictors that use information in \( X \),

\[
Q_{\tau}(Y \mid X) = \arg \min_{g(\cdot)} E[\rho_{\tau}(Y - g(X))]. \tag{5.10}
\]

The check function allows over-prediction to be worse than under-prediction (with the same absolute error magnitude), or vice-versa. Like before, when \( \tau \) 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 $\tau$ is near 0, over-prediction is very bad, so low conditional quantiles are better since they more often avoid over-prediction.

Section 5.3.1 discusses how we get a “best linear predictor” if the functional form of $g(\cdot)$ is restricted to $g(X) = X'b$.

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 $X$; the linear projection is both the “best” linear approximation of the CEF as well as the “best” linear predictor of $Y$ given $X$. QR has similar properties.

5.3.1 “Best” Linear Predictor

Even if the CEF is not of the form $x'\beta$, OLS still consistently estimates the “best” linear predictor ([Hansen 2020a], §2.18). If the CEF actually is of that form, then OLS estimates the CEF, which is the “best” predictor ([Hansen 2020a], §2.11). I put “best” in quotes because it is defined as minimizing the expectation of a particular loss function, specifically quadratic loss. Under other loss functions, the CEF is not the best predictor, and OLS does not estimate the best linear predictor. In particular, for check function loss, QR estimates the best linear predictor, which is also the best predictor if the CQF is properly specified.

From Theorem 5.1, even if the CQF is not linear in $x$, QR is consistent for the population vector

$$\beta(\tau) = \arg \min_b E[\rho(\tau(Y - X'b))].$$

That is, among predictors of the form $X'b$, the predictor $X'\beta(\tau)$ minimizes risk (expected loss) under the loss function $L(y,g) = \rho(\tau(y - g))$. Thus, $X'\beta(\tau)$ is the “best” linear predictor if we define “best” according to $\rho(\cdot)$, 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 linear projection of $Y$ on $X$ is the “best” linear approximation of the CEF. OLS estimates the linear projection (coefficient). Thus, even if the CEF is not linear in $x$, the linear projection $x'\beta$ is the “best” possible approximation of that form, so OLS essentially estimates an approximate CEF. Further, if the CEF is linear in $x$, then OLS estimates the CEF.

Formally, the “best linear approximation” of the CEF is written as follows. Let $m(x) \equiv E(Y \mid X = x)$ be the CEF. Consider all functions $g(\cdot)$ of the form $g(x) = x'b$. Then, the
linear projection coefficient $\beta = [E(XX')^{-1}E(XY)]$ satisfies
\[ \beta = \arg \min_{b} E[(X'b - m(X))^2], \] (5.11)
where the expectation is wrt the distribution of random vector $X$. In this mean squared error sense, $X'b$ is the “best” linear approximation of the CEF $m(x)$.

Angrist, Chernozhukov, and Fernández-Val (2006) provide a similar result for QR. Given the QR plim $\beta(\tau)$ from Theorem 5.1, the function $x'b$ has a descriptive interpretation in addition to the “best linear predictor” property in Section 5.3. Specifically, $x'b$ is the “best” linear approximation of the true CQF in terms of a weighted mean squared error criterion. Skipping the (complicated) definition of the weight,
\[ \beta(\tau) = \arg \min_{b} E\{\text{weight}(\tau, X, b) \times [Q_{\tau}(Y | X) - X'b]^2}\}. \] (5.12)
This is not as pleasing or easy to interpret as an unweighted mean squared error, but it’s something. For details, see Theorem 1 of Angrist, Chernozhukov, and Fernández-Val (2006).

5.4 Estimation

The same approaches from Section 4.4 can be used for quantile regression. Originally, Koenker and Bassett (1978) proposed the minimization approach for linear QR.

R function rq() in package quantreg (Koenker, 2019) is essentially the quantile analog of lm().

Similar to OLS being consistent for the LP/BLP/BLA rather than the CEF, the QR estimator is generally consistent for the population objects in Section 5.3 rather than the CQF. If magically the CQF’s true functional form has been specified, then the CQF is also consistently estimated. Otherwise, nonparametric QR should be used to estimate the CQF (rather than the objects in Section 5.3). See Chapters 16 and 17 for general nonparametric estimation approaches and R packages (e.g., np has function npqreg for nonparametric QR).

As in Section 4.4, smoothing has the potential to improve estimation precision here. Fernandes, Guerre, and Horta (2020) describe this and derive many theoretical results.

Analogous to (4.15) for unconditional quantile estimation, the QR estimator is
\[ \hat{\beta}(\tau) = \arg \min_{b} \hat{E}[\rho_{\tau}(Y - X'b)] = \arg \min_{b} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau}(Y_i - X'_i b). \] (5.13)
This is the same as the OLS estimator with $\rho_{\tau}(\cdot)$ replacing the function $\rho_2(u) = u^2$. Instead of minimizing the sum (or average) of squared residuals, (5.13) minimizes the sum (average) of “checked” residuals.

Computationally, (5.13) is more difficult than the OLS minimization. For OLS, we can take a first-order conditional and analytically derive a closed-form expression for $\beta$. This
is not possible for QR: \( \rho_\tau(\cdot) \) is not differentiable at zero. Nonetheless, Koenker figured out how to write the minimization as a linear program that could be solved quickly, and today’s algorithms are very fast. So, unless you are a computer scientist, you should probably not spend too much time thinking about QR computation.

**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.

### 5.5 Asymptotic Properties

Angrist, Chernozhukov, and Fernández-Val (2006, Thm. 3, p. 549) establish consistency and asymptotic normality of \( \hat{\beta}(\tau) \) (for the corresponding population minimizer), uniformly over a continuum of \( \tau \), 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.

Angrist, Chernozhukov, and Fernández-Val (2006, p. 548) define \( \mathcal{T} \) as (a closed subset of) the interval \( [\epsilon, 1 - \epsilon] \) for some \( \epsilon > 0 \). That is, \( \mathcal{T} \) excludes zero and one (and values within \( \epsilon \) of either). We cannot learn about the extreme tails.

Condition (i) in Theorem 3 of Angrist, Chernozhukov, and Fernández-Val (2006) is iid sampling. This is sufficient, but not necessary. There are asymptotic results on QR with time series data, for example.

Condition (ii) is about the conditional PDF of \( Y \), \( f_Y(y \mid X = x) \). The first assumption is that this PDF exists. In particular, this excludes discrete \( Y \). There are also some smoothness assumptions. If only a single \( \tau \) is of interest, only a neighborhood of \( y = Q_\tau(Y \mid X = x) \) is involved, but their results consider a continuum of \( \tau \) uniformly.

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

Condition (iv) requires finite variance (slightly stronger) for \( X \), 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 \( \beta(\tau) = \arg\min_b E[\rho_\tau(Y - X'b)] \). Under conditions (i)–(iv) in Theorem 3 of Angrist, Chernozhukov, and Fernández-Val (2006),

\[
\hat{\beta}(\tau) \overset{p}{\to} \arg\min_b E[\rho_\tau(Y - X'b)].
\]

(5.14)

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

\[
\sup_{\tau \in \mathcal{T}} \| \hat{\beta}(\tau) - \beta(\tau) \|_p \overset{p}{\to} 0.
\]

(5.15)
CHAPTER 5. QUANTILE REGRESSION: DESCRIPTION AND PREDICTION

To get very basic intuition for uniform consistency, recall that $W_n \overset{p}{\to} w$ is equivalent to $W_n - w \overset{p}{\to} 0$. Similarly, $\hat{\beta}(\tau) \overset{p}{\to} \beta(\tau)$ is equivalent to $\|\hat{\beta}(\tau) - \beta(\tau)\| \overset{p}{\to} 0$. So Theorem 5.1 just extends this by taking a supremum over $\tau \in T$.

As mentioned earlier, if the $\tau$-CQF has the form $x'\beta(\tau)$, then it is consistently estimated by QR. Angrist, Chernozhukov, and Fernández-Val (2006, Thm. 3) also establish asymptotic normality of the QR estimator under misspecification. This includes “pointwise” asymptotic normality, for a single $\tau$. They also show that the random function $\hat{\beta}(\cdot)$ over $\tau \in T$ is “asymptotically normal,” i.e., when centered and scaled converges to a Gaussian process. A 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 $\beta(\tau)$ for different $\tau$. For example, we could construct a uniform confidence band that includes the true function $\beta(\cdot)$ with $1 - \alpha$ probability (asymptotically). Or, we could run hypothesis tests involving multiple $\tau$.

Theorem 5.2 (Theorem 3 of Angrist, Chernozhukov, and Fernández-Val (2006) continued). Define

$$J(\tau) \equiv \mathbb{E}[f_Y | X (X'\beta(\tau) | X)XX'],$$
$$\Sigma(\tau, \tau') \equiv \mathbb{E}[(\tau - 1\{Y < X'\beta(\tau)\}) (\tau' - 1\{Y < X'\beta(\tau')\})XX'].$$  \hspace{1cm} (5.16) \hspace{1cm} (5.17)

Under conditions (i)–(iv) in Theorem 3 of Angrist, Chernozhukov, and Fernández-Val (2006), given some $\tau \in (0, 1)$,

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \overset{d}{\to} N(0, \Sigma_\tau), \quad \Sigma_\tau \equiv [J(\tau)]^{-1} \Sigma(\tau, \tau)[J(\tau)]^{-1}. \hspace{1cm} (5.18)$$

More strongly, $J(\cdot)\sqrt{n}(\hat{\beta}(\cdot) - \beta(\cdot))$ converges in distribution to $Z(\cdot)$, a zero-mean Gaussian process with covariance function $\mathbb{E}[Z(\tau_1)Z(\tau_2)'] = \Sigma(\tau_1, \tau_2)$ in $\overset{\sim}{\equiv}$ (5.17).

5.6 Inference

One option for inference (confidence intervals, hypothesis testing) is to use the asymptotic normality results from Theorem 5.2. In principle, the asymptotic covariance matrix/function could be estimated and plugged into the usual formulas. However, as in Section 4.7, the asymptotic variance is “difficult” to estimate due to the (now conditional) PDF term.

There are many different approaches to QR in the literature, although many of them historically have assumed homoskedasticity, which economists usually try to avoid. Angrist, Chernozhukov, and Fernández-Val (2006) suggest subsampling for inference on the function $\beta(\cdot)$; see their Section 3 (and see my Section 13.4 for a basic introduction).

For pointwise (single $\tau$ at a time) inference, Bayesian bootstrap is one possibility; see Chapter 14 and Hahn (1997), for example.
Chernozhukov, Hansen, and Jansson (2009) offer a clever approach that’s exact even in finite samples, similar to the exact approach (based on order statistics) in Section 4.7. However, it relies on having a properly specified conditional quantile function. The general idea is: if $q_\tau(x) = Q_\tau(Y | X = x)$, then $P(Y \leq q_\tau(X)) = \tau$, so (with iid sampling) $1\{Y \leq q_\tau(X)\}$ are iid Bernoulli($\tau$).

### 5.7 Censoring

The ideas in Section 4.5 extend to QR. In Stata, try the `cqiv` command available through SSC.
CHAPTER 5. QUANTILE REGRESSION: DESCRIPTION AND PREDICTION
Chapter 6

Quantile Regression: Causality

Optional resources for this chapter

- *Handbook of Quantile Regression*, [http://merlin.lib.umsystem.edu/record=b12097203-s1](http://merlin.lib.umsystem.edu/record=b12097203-s1)
- R and Stata code (various): [https://sites.google.com/site/blaisemelly/home/computer-programs](https://sites.google.com/site/blaisemelly/home/computer-programs)

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 coefficient model under certain assumptions. Both approaches can be extended to allow endogeneity as in Chapter 7.

6.1 Background: Potential Outcomes and ATE

The following is a reminder (not even enough details to be called a “review”); e.g., see Section 4.3 of [Kaplan (2020b)](https://example.com) for more 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. 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.
CHAPTER 6. QUANTILE REGRESSION: CAUSALITY

Table 6.1: Potential outcomes example for DQ 6.1

<table>
<thead>
<tr>
<th>( Y_0 )</th>
<th>( Y_1 )</th>
<th>( Y_1 - Y_0 )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>-3</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Different population objects can be formed from these potential outcomes \((Y_1, Y_0)\). The treatment effect for an individual is defined as \( C = Y_1 - Y_0 \). (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: \( \text{ATE} = E(C) = E(Y_1 - Y_0) \). By linearity of expectation, \( E(Y_1 - Y_0) = E(Y_1) - E(Y_0) \), the treatment effect on the mean. That is, \( E(Y_1) \) is the mean outcome in the parallel universe where everyone is treated, \( E(Y_0) \) is the mean outcome in the parallel universe where nobody is treated, and \( E(Y_1) - E(Y_0) \) shows how the mean outcome changes (the effect on the mean) when we move from the untreated universe to the treated universe. Even though it has a different interpretation, this is often just called the ATE since it is mathematically equivalent. However, we cannot estimate \( E(Y_1 - Y_0) \) by \( \hat{E}(Y_1 - Y_0) \). However, we can get more traction on the formulation \( E(Y_1) - E(Y_0) \): we can at least take the sample average outcome of treated individuals and subtract the sample average outcome of untreated individuals.

Assumptions are required for the ATE to be identified. 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 \( E(Y_0) \) and \( E(Y_1 - Y_0) \); e.g., see Theorem 2.11 in Hansen (2020a, §2.29).

Alternatively, if \( Y_0, Y_1 \perp X \), then

\[
E(Y \mid X = 1) - E(Y \mid X = 0) = E(Y_1 \mid X = 1) - E(Y_0 \mid X = 0) = E(Y_1) - E(Y_0),
\]

(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 \( Y_1 \)). Thus, the ATE can be estimated by \( \hat{E}(Y \mid X = 1) - \hat{E}(Y \mid X = 0) \), the difference of the subsample averages.

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
6.2. QUANTILE TREATMENT EFFECTS

outcome pair.

a) Compute $E(Y_0)$.
b) Compute $E(Y_1)$.
c) Compute $E(Y_1) - E(Y_0)$.
d) Compute $E(Y_1 - Y_0)$.
e) Compute $Q_{0.4}(Y_1) - Q_{0.4}(Y_0)$.
f) Compute $Q_{0.4}(Y_1 - Y_0)$.

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

As DQ 6.1 illustrates, $Q_\tau(Y_1 - Y_0) \neq Q_\tau(Y_1) - Q_\tau(Y_0)$. Unlike the expectation operator, the quantile operator is nonlinear. Thus, the $\tau$-quantile of the population distribution of individual treatment effects $C = Y_1 - Y_0$ differs from the treatment effect on the $\tau$-quantile of the outcome distribution.

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

This is one of two reasons to focus on $Q_\tau(Y_1) - Q_\tau(Y_0)$, called the quantile treatment effect (QTE), or more specifically the $\tau$-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 $Y_0, Y_1 \perp \perp X$:

$$Q_\tau(Y \mid X = 1) - Q_\tau(Y \mid X = 0) = Q_\tau(Y_1 \mid X = 1) - Q_\tau(Y_0 \mid X = 0) = Q_\tau(Y_1) - Q_\tau(Y_0) = \tau\text{-QTE.} \tag{6.2}$$

Thus, the $\tau$-QTE can be estimated by $\hat{Q}_\tau(Y \mid X = 1) - \hat{Q}_\tau(Y \mid X = 0)$, the difference of subsample $\tau$-quantiles.

More generally, independence identifies the full marginal distributions of $Y_1$ and $Y_0$, so any summary of these distributions is also identified. This includes how the treatment affects the mean and quantiles (as shown), as well as how treatment affects the standard deviation, interquantile ranges, the upper tail, the 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 for 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
CHAPTER 6. QUANTILE REGRESSION: CAUSALITY

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

In the usual structural model $Y = X' \beta + V$, the uppercase letters denote random variables, whereas the coefficient vector $\beta$ is non-random. “Random” just means that each individual has their own $(Y, X, V)$, and that vector has some probability distribution in the population. In contrast, $\beta$ is a constant; it does not depend on the individual.

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. Instead of “random coefficient,” it could be called “individual-specific coefficient.” To model this, the constant $\beta$ can be replaced with random vector $B$.

The resulting structural model is $Y = X' B$. Assuming $X$ includes a constant, the additive error $V$ would be redundant; it can just be absorbed into the random intercept term. For example, if $Y = B_0 + B_1 X + V$, then we can equivalently write $Y = (B_0 + V) + B_1 X$ and just redefine the random intercept as $B_0 + V$. The population is now the joint distribution of $(Y, X, B)$.

Exogeneity here means $B$ is unrelated to $X$, like $B \perp X$. The weaker mean independence condition $E(B | X) = E(B)$ also suffices below. At a high level, the idea is the same as usual: the regressors are unrelated to the unobserved determinants of $Y$. The difference is that before the “unobserved determinants” were aggregated into the additive structural error term $V$, whereas now they are all in $B$.

Given exogeneity, a certain feature of the structural model is identified by the CEF. To see this, plug the structural model into the CEF:

$$E(Y | X = x) = E(X'B | X = x) = x'E(B | X = x) = x'E(B).$$

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

Discussion Question 6.3 (random coefficient exogeneity: wage). At some point, you’d 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, $\beta_0$ and $\beta_1$ are fixed constant parameters,
6.4. ANOTHER RANDOM COEFFICIENTS MODEL

and \( U \) is other determinants of \( Y \). Now, keep the same \( Y \) and \( X \), but consider the structural random coefficients model \( Y = B_0 + B_1X \).

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 \( B \) besides its mean?

6.4 Another Random Coefficients Model

To link the structural random coefficient model to conditional quantiles, some additional restrictions are imposed beyond Section 6.3. To be concrete, imagine \( Y \) is log wage and \( X = 1 \) for “high education” and \( X = 0 \) for “low education.” A general structural random coefficient model is \( Y = B_0 + B_1X \), 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 as well as a different return to schooling.

Continue to assume individuals may have different intercepts and slopes, but now assume this heterogeneity can be represented by a scalar random variable \( U \). That is, each individual has their own \((Y, X, U)\). Then, 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 coefficient model is

\[
Y = \beta_0(U) + \beta_1(U)X.
\]  

(6.3)

This model in (6.3) is more restrictive than the general random coefficient model \( Y = B_0 + B_1X \), but less restrictive than the model \( Y = \beta_0 + \beta_1X + U \). Just as (6.3) is a special case of \( Y = B_0 + B_1X \), \( Y = \beta_0 + \beta_1X + U \) is a special case of (6.3): set \( \beta_1(U) = \beta_1 \) and \( \beta_0(U) = \beta_0 + U \).

It is usually assumed (without loss of generality, given continuous \( Y \)) that \( U \sim \text{Unif}(0, 1) \). Imagine the model \( Y = \tilde{\beta}_0(V) + \tilde{\beta}_1(V) \) where \( V \) has continuous CDF \( F_V(\cdot) \). By the probability integral transform, \( U \equiv F_V(V) \sim \text{Unif}(0, 1) \). Let \( \tilde{\beta}_0(u) = \tilde{\beta}_0(F_V^{-1}(u)) \) and \( \tilde{\beta}_1(u) = \tilde{\beta}_1(F_V^{-1}(u)) \). Then, \( Y = \beta_0(U) + \beta_1(U)X \). Hence, the original model with non-uniform \( V \) is equivalent to a model with uniform \( U \), by adjusting the coefficient functions accordingly.

Discussion Question 6.4 (structural random coefficient wage model). Consider (6.3), where \( Y \) is log wage and \( X = 1 \) if high education (\( X = 0 \) if low). Let \( u_2 > u_1 \). Based on economic theory and/or your intuition, what do you think is the relationship between the objects in each of the following pairs?

a) Between \( \beta_0(u_2) \) and \( \beta_0(u_1) \)?
82

CHAPTER 6. QUANTILE REGRESSION: CAUSALITY

b) Between $\beta_1(u_2)$ and $\beta_1(u_1)$?

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

Although there is no quantile analog of $E(X'B | X) = X'E(B | X)$, 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 coefficient model $Y = X'\beta(U)$, $Y$ is strictly increasing in $U$ given any $X = x$. That is, given any $x$ in the support of $X$, the function $x'\beta(u)$ is strictly increasing in $u$ over $0 \leq u \leq 1$.

To get you a head start on DQ 6.5, consider the following proof that $Q_{0.5}(Y | X = 0) = \beta_0(0.5)$ given (6.3) and [A6.1] and independence ($U \perp X$). By (6.3), 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 $Q_{r}(f(W)) = f(Q_{r}(W))$ 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,

$$Q_{0.5}(Y | X = 0) = Q_{0.5}(\beta_0(U) | X = 0) = Q_{0.5}(\beta_0(U)) = \beta_0(Q_{0.5}(U)) = \beta_0(0.5). \quad (6.4)$$

**Discussion Question 6.5** (random coefficient model conditional quantiles). Consider (6.3) with $X \perp U$, Assumption [A6.1], and the normalization $U \sim \text{Unif}(0,1)$. Similar to (6.4), 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, $Q_{0.5}(Y | X = 1)$.

b) The difference between the two prior objects, $Q_{0.5}(Y | X = 1) - Q_{0.5}(Y | X = 0)$.

c) Other quantiles of log wage in the low-education subpopulation: $Q_{0.25}(Y | X = 0)$ and $Q_{0.75}(Y | X = 0)$.

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 $X = 1$. 

6.5. UNCONDITIONAL QUANTILE REGRESSION

   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 $X \perp U$ affect any of your answers?

6.4.1 Heteroskedasticity

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

As DQ 6.8 shows, there can be heteroskedasticity even if $U \perp X$. This is different than with an additively separable error, like $Y = X'\beta + U$. In that case, if $U \perp X$, then

$$\text{Var}(Y | X = x) = \text{Var}(X'\beta + U | X = x) = \text{Var}(U | X = x) = \text{Var}(U),$$

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

This is important to remember for Section 7.1. Unlike in the standard IV regression model, having independence between the unobserved variable and the instrument vector does not preclude heteroskedasticity.

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 some 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 $w(Y)$ summarizes “how good” that distribution of $Y$ is into a single number. 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 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 $X$ 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 \( \beta(u) \) for various \( 0 < u < 1 \). The “change” is a change in the marginal distribution of \( X \). The “effect” assumes the conditional distributions of \( Y \) given any \( X = x \) 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. For more discussion of when UQR can be used to estimate policy effects in practice, see Kaplan (2019b).
Chapter 7

Quantile Regression with Endogeneity

Optional resources for this chapter

- R (and eventually Stata) code for IVQR: [http://faculty.missouri.edu/kaplanmd](http://faculty.missouri.edu/kaplanmd)
- R and Stata code (various): [https://sites.google.com/site/blaisemelly/home/computer-programs](https://sites.google.com/site/blaisemelly/home/computer-programs)
- Stata: *cqiv* can be installed from SSC for control function estimation (with or without censoring)

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

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. There are other quantile models that identify causal objects under endogeneity; see Sections 7.2.1 and 7.2.2
CHAPTER 7. QUANTILE REGRESSION WITH ENDOGENEITY

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 satisfies certain moment conditions. Relevance ensures that the moment conditions are uniquely satisfied by the true parameter vector.

For example, consider structural model \( Y = X'\beta + U \) with endogeneity so \( E(XU) \neq 0 \) but the full instrument vector \( Z \) satisfies \( E(ZU) = 0 \), where \( Z \) includes both exogenous regressors and excluded instruments (that do not appear in the structural model). Then, using the structural model to replace \( U = Y - X'\beta \), the moment condition \( E[Z(Y - X'b)] = 0 \) is solved by \( b = \beta \).

The relevance condition ensures that there are no other values of \( b \) that solve the moment conditions. For example, with exact identification, if \( E(ZX') \) is invertible, then the moment conditions can be solved uniquely: \( \beta = [E(ZX')]^{-1}E(ZY) \).

Given the moment conditions and identification, GMM consistently estimates the structural parameters. (GMM with weighting matrix \( \hat{E}(ZZ') \) is 2SLS.)

7.1.2 IVQR Identification

Following a special case of Chernozhukov and Hansen (2005), it is shown that the structural parameters in a random coefficient model like 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 more general functional form, but the core intuition is the same as below.

As in Section 6.4, consider structural model

\[
Y = X'\beta(U), \quad U \sim \text{Unif}(0, 1).
\]

(7.1)

Let \( Z \) be the full vector of instruments, including both exogenous regressors and excluded instruments. Here, “exogeneity” means \( Z \perp U \). Recall from Section 6.4.1 that there can still be heteroskedasticity through \( \beta(\cdot) \), unlike in a fixed-coefficients model with independent error term. Monotonicity (Assumption A6.1) is again assumed.

The following derivation is similar to (6.4) and DQ 6.5. Independence and monotonicity are both crucial. Independence implies that probabilities involving (only) \( U \) are unaffected by conditioning on \( Z \). Monotonicity implies that, given \( X = x \), the relative value of \( Y \) depends on \( U \). The conclusion is that the structural parameter \( \beta(\tau) \) for some \( 0 < \tau < 1 \) solves a particular conditional probability that involves only observable variables. Subsequently, this conditional probability is rewritten as a (conditional) moment condition, which also implies unconditional moments.
7.1. INSTRUMENTAL VARIABLES QUANTILE REGRESSION

Formalizing the above verbal arguments,

\[ P(Y \leq X'\beta(\tau) \mid Z) = P(X'\beta(U) \leq X'\beta(\tau) \mid Z) \]

by (7.1) \hspace{1cm} (7.2)

\[ = P(U \leq \tau \mid Z) \]

by A6.1 \hspace{1cm} (7.3)

\[ = P(U \leq \tau) \]

by \( U \perp \perp Z \) \hspace{1cm} (7.4)

\[ = \tau \]

by (7.1), \( U \sim \text{Unif}(0, 1) \). \hspace{1cm} (7.5)

That is, the structural \( \beta(\tau) \) satisfies a particular conditional quantile restriction. However, it is different than 2SLS in that it does not correspond to a regular QR where the regressors are fitted values from a first-stage regression.

The conditional quantile restriction can be written as a conditional moment condition, which further implies unconditional moments. Recall for event \( A \) that \( P(A) = E[I\{A\}] \). Thus, \( P(Y \leq X'\beta(\tau) \mid Z) = \tau \) becomes

\[ E[I\{Y \leq X'\beta(\tau)\} \mid Z] = \tau, \]

\[ E[I\{Y \leq X'\beta(\tau)\} - \tau \mid Z] = 0. \] \hspace{1cm} (7.6)

Unconditional moments can be generated by multiplying the main part by any function of \( Z \). Although not “optimal,” a common choice is just \( Z \) itself. This is similar to using the moments \( E(ZU) = 0 \) given the conditional moment \( E(U \mid Z) = 0 \). Here,

\[ 0 = E[Z(I\{Y \leq X'\beta(\tau)\} - \tau)]. \] \hspace{1cm} (7.7)

7.1.3 IVQR Estimation

Since we have moment conditions from (7.7), GMM estimation seems natural. However, there are computational challenges. The parameter is “stuck” inside the indicator function \( I\{\cdot\} \). 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. Standard numerical solvers may fail to find the global minimizer of the GMM criterion function.

Because of this, there have been many different approaches to estimation. Since I’m writing this, I’ll focus on the approach of Kaplan and Sun (2017), which seems fast, reliable, scalable, and has code on my website.

The basic idea is to replace \( I\{\cdot\} \) with a smoothed (differentiable) version. The amount of smoothing is determined by bandwidth \( h \), where the smoothed version approaches \( I\{\cdot\} \) in the limit as \( h \downarrow 0 \). The sample moment conditions still can’t be solved analytically for \( \hat{\beta}(\tau) \), but standard gradient-based numerical solvers can be used, like \texttt{newtonsyst} in the \texttt{pracma} package (Borchers, 2019). With overidentification, it’s impossible to know if you found the global minimizer of the GMM objective function or just a local minimum. Consequently, to produce an exactly identified system, the regressor vector is linearly projected onto the full instrument vector to get a vector of instruments the same length as the regressor vector. Then, once the numerical solver thinks it finds \( \hat{\beta}(\tau) \), it
can be plugged into the sample moments to check that they are all zero (plus or minus some numerical tolerance). If not, the amount of smoothing \( h \) is increased to make the numerical problem easier, until a solution is found. (The bandwidth \( h \) can usually then be reduced again, because using the large-\( h \) estimate as the initial value helps the numerical solver.)

There is a tradeoff between large and small \( h \). As \( h \downarrow 0 \), the moment conditions approach the original IVQR moments, but the numerical computation becomes intractable. As \( h \) increases, computation is easier, and also the estimator’s variance decreases. In fact, there is always some \( h > 0 \) that minimizes the estimator’s mean squared error (MSE): a small amount of bias is introduced, but the variance decreases rapidly enough that MSE decreases, too. Similar MSE improvements from smoothing in (non-IV) QR are detailed by Fernandes, Guerre, and Horta (2020).

The first-order results of Kaplan and Sun (2017) are extended to more general quantile restriction models and non-iid data by de Castro et al. (2019), who also propose a smoothed two-step GMM estimator.

One possible argument against using the pointwise MSE-optimal bandwidth in Kaplan and Sun (2017) is that MSE fails to penalize bias appropriately. With (IV)QR, often the hope is to learn about heterogeneity across \( \tau \). If the 2SLS \( \hat{\beta} \) has a much lower variance than the IVQR \( \hat{\beta}(\tau) \) for all \( \tau \in (0, 1) \), then even if there is heterogeneity so that 2SLS is biased, it is still a “better” estimator according to (pointwise in \( \tau \)) MSE. However, using the same 2SLS estimate for all \( \tau \) makes it look like there is zero heterogeneity. Here, the problem is less with MSE than with the object of interest: do we care only about a single \( \beta(\tau) \), or do we care about differences like \( \beta(\tau_2) - \beta(\tau_1) \)? A separate, more general objection is that MSE under-values bias in the big-picture context of developing a scientific body of evidence; see Kaplan (2019c).

Discussion Question 7.1 (MSE of difference). Consider parameters \( \theta_1 \) and \( \theta_2 \). Consider respective estimators \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \), as well as the pair of estimators \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \). Let \( \hat{\theta}_1 \sim N(\theta_1, \sigma^2) \), \( \hat{\theta}_2 \sim N(\theta_2, \sigma^2) \), with \( \sigma^2 = 100 \) and \( \hat{\theta}_1 \perp \!\!\!\perp \hat{\theta}_2 \). Let \( \hat{\theta}_1 = \theta_1 - 9 \) and \( \hat{\theta}_2 = \theta_2 + 9 \). Recall MSE equals variance plus squared bias.

a) Compute the bias and variance of \( \hat{\theta}_1 \) as an estimator of \( \theta_1 \), to show its MSE is 100.

b) Compute the bias and variance of \( \hat{\theta}_2 \) as an estimator of \( \theta_2 \), to show its MSE is 100.

c) Compute the bias and variance of \( \hat{\theta}_2 \) as an estimator of \( \theta_1 \), to show its MSE is 81 (better than \( \hat{\theta}_1 \)).

d) Compute the bias and variance of \( \hat{\theta}_2 \) as an estimator of \( \theta_2 \), to show its MSE is 81 (better than \( \hat{\theta}_2 \)).

e) Compute the bias and variance of \( \hat{\theta}_2 - \hat{\theta}_1 \) as an estimator of \( \theta_2 - \theta_1 \), to show its MSE is 200.

f) Compute the bias and variance of \( \hat{\theta}_2 - \hat{\theta}_1 \) as an estimator of \( \theta_2 - \theta_1 \), to show its MSE is 324 (worse than \( \hat{\theta}_2 - \hat{\theta}_1 \)).

g) Discuss what point you think I was trying to illustrate through the above calculations. (Bonus if you can connect it to IVQR.)
7.2. OTHER APPROACHES TO ENDOGENEITY

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 should work fine. Unfortunately, the justification for the first-stage $F > 10$ rule-of-thumb for 2SLS does not apply to IVQR; currently, I’m unaware of any well-justified formal assessment of instrument strength for IVQR. The exact finite-sample approach of Chernozhukov, Hansen, and Jansson (2009) 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 $Y = g(D, \epsilon)$, where scalar continuous $D$ is endogenous and modeled by $D = h(Z, \eta)$, in which one of the instruments in $Z$ must also be continuous. (Other exogenous regressors can be added, too.) Identification follows from monotonicity of $h(z, \eta)$ in $\eta$ as well as the exogeneity condition $Z \perp \perp (\epsilon, \eta)$.

Generally, the triangular model restricts the selection equation (for $D$) more than IVQR, but restricts the structural (outcome) equation less than IVQR. For example, $\epsilon$ 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 and 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 only receive treatment
(X = 1) 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 discuss results comparing LQTE and IVQR. Like the triangular model, the LQTE model restricts the selection equation more than the outcome equation (compared to IVQR). Under LQTE assumptions, even if some IVQR assumptions fail, the IVQR estimand is closely related to the LQTE, just at shifted \( \tau \) values. So, if the IVQR assumptions hold, IVQR estimates the QTE (by implicitly extrapolating from the compliers), and if not, IVQR still has a well-defined interpretation related to LQTE.

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.

### 7.3 Panel Data with Fixed Effects

The first take on QR 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., \( Q_\tau(Y_2 - Y_1) \neq Q_\tau(Y_2) - Q_\tau(Y_1) \). 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: \( Y = X' \beta(U) \). 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

\[
Y_t = X' \beta(U + V_t).
\]

(7.8)

In contrast, the original panel QR models had only \( V_i \) as the rank variable, and added individual heterogeneity through a dummy regressor. Adding the \( i \) subscript explicitly,

\[
Y_{it} = X'_{it} \beta(V_{it}) + \eta_i \gamma(V_{it}),
\]

(7.9)

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

\[
Y_t = X' \beta(V_i) + U,
\]

(7.10)

\[
Y_t = X' \beta(U) + V_i,
\]

(7.11)
etc. Most (not all) of these have the standard FE model \( Y_t = X_t'\beta + U + V_t \) 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 from Powell (2016) and Liu (2020).

**Discussion Question 7.2** (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?
CHAPTER 7. QUANTILE REGRESSION WITH ENDOGENEITY

Exercises

Exercise E7.1.  a. Find a published paper from a respectable economics journal that runs a cross-sectional IV regression (and makes its data publicly available); provide a link to the paper. 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 $\tau$ (0.5, 0.25, etc.), using the R code available at [http://faculty.missouri.edu/kaplanmd/code](http://faculty.missouri.edu/kaplanmd/code) (file ivqr See.R has the main ivqr see() function; gmmq.R contains helper functions). (Or: try my new Stata code to help me make sure it works; have me email you the sivqr.ado file and help file.) Provide the code you write/run. Try both the plug-in bandwidth (i.e., don’t specify argument $h$) as well as a very small bandwidth (for which you can just specify $h=0$), and other $h$ values as desired. Note the warnings in the comments at the top of ivqr see.R.

d. Discuss any similarities and differences across quantile levels ($\tau$) 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.

---

1 E.g., in top 500 of [https://ideas.repec.org/top/top.journals.all.html](https://ideas.repec.org/top/top.journals.all.html)
2 E.g., if nobody’s claimed it yet, [https://doi.org/10.1016/j.jhealeco.2016.08.002](https://doi.org/10.1016/j.jhealeco.2016.08.002)
Part III

Distributional Methods
Introduction

This part concerns learning about entire distributions.
Chapter 8

Distributional Inference: One-Sample, Two-Sided

Here, the goal 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). Perhaps most importantly, this chapter introduces ideas useful in one-sided and/or two-sample extensions that are usually more economically interesting.

8.1 Discrete and Categorical Distributions

The focus of this chapter (and subsequent chapters) is continuous CDFs, but this Section 8.1 has some notes on discrete and categorical distributions.

Discrete and categorical distributions can be characterized by their probability mass function (PMF). That is, if the possible values or categories are written as $v_1, \ldots, v_J$, then the vector $p = (p_1, \ldots, p_J)$ with $p_j \equiv P(Y = v_j)$ fully describes the distribution of $Y$.

Thus, standard results for finite-dimensional parameters apply. Assuming $J$ is not too big compared to $n$, the estimators $\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} I\{Y_i = v_j\}$ work well and are jointly asymptotically normal. Since the joint distribution is (asymptotically) known, joint hypothesis tests involving multiple (even all) $p_j$ are possible. Presumably such $\chi^2$ tests can be refined to improve power or higher-order properties, but the basic approach is relatively straightforward.
8.2 Preliminary Results for Continuous Distributions

Many approaches are based on the empirical CDF (ECDF), also known as the empirical distribution function (EDF). Given iid sampling, the ECDF is

$$\hat{F}(y) \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y\}, \quad y \in \mathbb{R}. \quad (8.1)$$

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. Even some finite-sample properties are known. Some details are in Section 8.A.

8.3 Goodness-of-Fit Testing

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

$$H_0: F(\cdot) = F_0(\cdot), \quad (8.2)$$

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 $H_0: F(\cdot) \leq F_0(\cdot)$, or $\geq$, but intuition is easier with (8.2).

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 the null $F = F_0$. There are different ways to measure the distance, corresponding to different tests. The Kolmogorov–Smirnov (KS) approach (Section 8.4) uses

$$D_n \equiv \sup_{r \in \mathbb{R}} |\hat{F}(r) - F_0(r)|, \quad (8.3)$$

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

$$W_n^2 \equiv n \int_{\mathbb{R}} [\hat{F}(r) - F_0(r)]^2 \, dF_0(r). \quad (8.4)$$

Both (8.3) and (8.4) can be simplified in terms of $F_0(\cdot)$ evaluated at the sample order statistics. The Anderson–Darling AD test (Anderson and Darling, 1952, 1954) usually...
8.4 Kolmogorov–Smirnov Test

To control size, the distribution of $D_n$ in (8.3) 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

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

where again $B(\cdot)$ is the standard Brownian bridge.

The KS test then rejects $H_0$ when $\sqrt{n}D_n > K_{1-\alpha}$, the $(1-\alpha)$-quantile of the Kolmogorov distribution in (8.6). Size is controlled asymptotically since if $H_0$ is true then $P(\sqrt{n}D_n > K_{1-\alpha}) \to 1 - \alpha$.

In fact, size can be controlled exactly in finite samples. Such finite-sample critical values are available in ks.test() in R, for example. Assuming (still) that $F(\cdot)$ is continuous, then $F(Y) \sim \text{Unif}(0,1)$. Thus, $H_0: F(\cdot) = F_0(\cdot)$ is equivalent to $H_0: F_0(Y) \sim \text{Unif}(0,1)$. That is, we can use the transformed data $Z_i = F_0(Y_i)$; under the null, $Z_i \simiid \text{Unif}(0,1)$, so we can compute $D_n$ for comparing the ECDF of $Z_i$ to the $\text{Unif}(0,1)$ CDF. Under the null, we know the distribution of $D_n$. We can easily simulate it: just take many random samples of $Z_i \simiid \text{Unif}(0,1)$, and compute the KS statistic $D_n$ for each sample. Then the critical value $K_{1-\alpha}/\sqrt{n}$ should be the $(1-\alpha)$-quantile among the simulated $D_n$ values.

Despite all these nice properties, there is one significant deficiency seen in DQs 8.1 and 8.2, which motivates Section 8.6.

Discussion Question 8.1 (KS tail power 1). Let $H_0: Y_i \simiid \text{Unif}(0,1)$. Let $n = 20$. Use critical value $K_{1-\alpha}/\sqrt{n} = 0.26$, for $\alpha \approx 10\%$, so the KS rejects $H_0$ when $D_n > 0.26$. Imagine a dataset with $Y_i = i/20$ for $i = 1, \ldots, k$, and $Y_i = i + 1000$ for $i > k$. 

refers to a normalized version of CvM,

$$A_n^2 \equiv n \int_{\mathbb{R}} \frac{[\hat{F}(r) - F_0(r)]^2}{F_0(r)[1 - F_0(r)]} dF_0(r),$$

but it can also mean a similarly normalized KS statistic. And, fun fact: the two-sample “CvM” test is actually due to Anderson (1962), too.

Which is best? They can all control size. One difference is which alternatives they have the most power against. For example, if $F(\cdot)$ and $F_0(\cdot)$ are the same except at a certain point where there’s a big difference, the KS should have better power; conversely, if there is a small but constant difference across the whole distribution, then CvM should do better. The normalization (for either test) increases power in the tails, at the expense of decreasing it in the middle. Another difference is that the KS test can be inverted into a uniform confidence band. It can also be used for a multiple testing procedure to detect where the two CDFs differ. See Section 8.5 and Chapter 11 for details.
a) Let $k = 20$ and graph $F_0(\cdot)$, the 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 “distance” between $\hat{F}(\cdot)$ and $F_0(\cdot)$) 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?

You can also try with(data=list(k=20), expr=ks.test(c(1:k/2/zero.alt3,1/zero.alt3/zero.alt3/(k+1):2/zero.alt3),punif,exact=T)) in R, replacing $k=20$ with other values.

## 8.5 Uniform Confidence Band

### 8.5.1 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 $\theta$, a hypothesis test can be “inverted” into a confidence interval (CI). Specifically, for any value $t \in \mathbb{R}$, define null hypothesis $H_t: \theta = t$. The CI from test inversion with confidence level $1 - \alpha$ is

$$\hat{\text{CI}}_{1-\alpha} \equiv \{t : H_t \ not \ rejected \ at \ level \ \alpha\}. \quad (8.7)$$

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 $\alpha$ means that $H_\theta$ is rejected with probability $\alpha$ (or less), where $\theta$ is the true value. Simplifying slightly more, assume $P(H_\theta$ rejected at level $\alpha) = \alpha$ exactly. Then,

$$P(\hat{\text{CI}}_{1-\alpha} \ni \theta) = P(H_\theta \ not \ rejected \ at \ level \ \alpha) = 1 - P(H_\theta \ rejected) = 1 - \alpha. \quad (8.8)$$

### 8.5.2 Test Inversion: Vectors and Functions

Nothing in the argument of (8.8) is special to the dimension of $\theta$. The argument goes through with vector $\theta$. It also works if the parameter is a function (“infinite-dimensional”), like the population CDF.
8.5. **UNIFORM CONFIDENCE BAND**

Without trying to picture it yet, imagine we can define a set \( \hat{F}_{1-\alpha} \) of all CDFs that cannot be rejected by the KS test at level \( \alpha \). That is,

\[
\hat{F}_{1-\alpha} = \{ F_0(\cdot) : \text{KS does not reject } H_0: F(\cdot) = F_0(\cdot) \text{ at level } \alpha \}.
\]  

(8.9)

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

\[
P(\hat{F}_{1-\alpha} \ni F(\cdot)) = P(\text{true } F(\cdot) \text{ not rejected by KS at level } \alpha) \\
= 1 - P(\text{KS rejects true CDF at level } \alpha) \\
= 1 - \alpha.
\]

That is, there is exactly \( 1 - \alpha \) probability that this random (data-dependent) set of functions contains the true \( F(\cdot) \). This \( \hat{F} \) can be called a confidence set for \( F(\cdot) \). As seen below, it can also be called a **uniform confidence band**.

### 8.5.3 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.” That is, \( \hat{L}(r) \leq \hat{U}(r) \) for all \( r \in \mathbb{R} \), also written \( \hat{L}(\cdot) \leq \hat{U}(\cdot) \). Let \( 1 - \alpha \) be the confidence level. Let \( F(\cdot) \) be the true function. Then, at least asymptotically,

\[
1 - \alpha = P(\hat{L}(r) \leq F(r) \leq \hat{U}(r)) = P(\hat{L}(r) \leq \hat{U}(r) \leq F(r) \text{ for all } r \in \mathbb{R}). \tag{8.10}
\]

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 \( P(\hat{L}(r) \leq F(r) \leq \hat{U}(r)) = 1 - \alpha \) (or \( \rightarrow 1 - \alpha \)) 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 \( \theta_1 \) and \( \theta_2 \). (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 \( \hat{\theta}_1 \sim N(\theta_1, SE^2_1) \) and \( \hat{\theta}_2 \sim N(\theta_2, SE^2_2) \).

a) Briefly explain why \( \hat{C}I_1 = \hat{\theta}_1 \pm 1.64SE_1 \) is a 90% pointwise CI for \( \theta_1 \), and similarly \( \hat{C}I_2 = \hat{\theta}_2 \pm 1.64SE_2 \) for \( \theta_2 \).

b) Assuming \( \theta_1 \perp \theta_2 \), what is the joint coverage probability of the pointwise CIs, i.e., \( P(\theta_1 \in \hat{C}I_1 \text{ and } \theta_2 \in \hat{C}I_2) \)?

c) Without the independence assumption, is \( P(\theta_1 \in \hat{C}I_1 \text{ and } \theta_2 \in \hat{C}I_2) = 90\% \) 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 \( \alpha \), as in (8.9). In DQ 8.4 we’ll try to picture this band more concretely.

**Discussion Question 8.4 (KS band).** Consider \( \hat{F}_{1-\alpha} \) from (8.9). Recall that the KS rejects \( H_0 : F(\cdot) = F_0(\cdot) \) at level \( \alpha \) when \( D_n > c_{n,\alpha} \), with \( D_n \) defined in (8.3) and \( c_{n,\alpha} \equiv K_{1-\alpha}/\sqrt{n} \). The particular value of \( \alpha \) and the critical value is 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 \( F_0(\cdot) \in \hat{F}_{1-\alpha} \) satisfies \( \hat{L}(r) \leq F_0(r) \leq \hat{U}(r) \) 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 \( F_0(\cdot) \in \hat{F}_{1-\alpha} \) satisfies \( \hat{L}(\cdot) \leq F_0(\cdot) \leq \hat{U}(\cdot) \).

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 \( \mathbb{P}(\hat{L}(\cdot) \leq F(\cdot) \leq \hat{U}(\cdot)) \)? Why? Hint: recall \( \mathbb{P}(\hat{F}_{1-\alpha} \ni F(\cdot)) = 1 - \alpha \).

### 8.6 Dirichlet Approach

See Goldman and Kaplan (2018a) and references therein for details, and Kaplan (2019a) for an overview of the methodology and Stata `distcomp` command.

The problem with the KS is that it uses the same vertical distance threshold everywhere in the distribution, and symmetrically around \( \hat{F}(\cdot) \). Large values of \(|\hat{F}(r) - F(r)|\) are more likely in the middle of the distribution than in the tails. Further, the distribution of \( \hat{F}(r) - F(r) \) is not symmetric about zero when \( r \) is in the lower or upper tail. The downside of this is very low power in the tails, as seen in DQs 8.1 and 8.2. Equivalently, the uniform confidence band is “too wide” in the tails. You may have noticed in DQ 8.4 that \( \hat{F}(r) - c_{n,\alpha} < 0 \) when \( r \) is in the lower tail, and \( \hat{F}(r) + c_{n,\alpha} > 1 \) for \( r \) in the upper tail. Of course, \( \hat{L}(\cdot) \) and \( \hat{U}(\cdot) \) can be redefined to always be between 0 and 1, but this alludes to a deeper problem.

Improvement is possible by allowing the relationship between \( \hat{L}(r), \hat{F}(r), \) and \( \hat{U}(r) \) to differ appropriately across the distribution.

#### 8.6.1 Dirichlet Uniform Confidence Band

The probability integral transform says if \( F(\cdot) \) is continuous, then \( F(Y) \sim \text{Unif}(0, 1) \). If \( Y_i \id \sim F(\cdot) \), then \( F(Y_i) \id \sim \text{Unif}(0, 1) \). Of course, we do not know the true \( F(\cdot) \) in practice,
8.6. DIRICHLET APPROACH

but this is still useful. Further, consider the order statistics

\[ Y_{n:1} < Y_{n:2} < \cdots < Y_{n:k} < \cdots < Y_{n:n}, \tag{8.11} \]

formed by ordering the \( Y_i \) into increasing order, with \( Y_{n:1} \) the sample minimum and \( Y_{n:n} \) the sample maximum. Sometimes these are denoted \( Y_{(k)} \) instead of \( Y_{n:k} \) for the \( k \)th order statistic, i.e., the \( k \)th-smallest sample value. Given \( U_i \overset{\text{iid}}{\sim} \text{Unif}(0,1) \), define the uniform order statistics as

\[ U_{n:1} < \cdots U_{n:k} < \cdots < U_{n:n}. \tag{8.12} \]

Then, \( F(Y_{n:k}) \overset{d}{=} U_{n:k} \). Further, the full joint finite-sample distribution of the transformed \( F(Y_{n:k}) \) is known:

\[ (F(Y_{n:1}), F(Y_{n:2}) - F(Y_{n:1}), \ldots, F(Y_{n:n}) - F(Y_{n:n-1}), 1 - F(Y_{n:n})) \overset{d}{=} (U_{n:1}, U_{n:2} - U_{n:1}, \ldots, U_{n:n} - U_{n:n-1}, 1 - U_{n:n}) \sim \text{Dir}(1,1,\ldots,1). \tag{8.13} \]

The marginals are beta distributions:

\[ F(Y_{n:k}) \overset{d}{=} U_{n:k} \sim \text{Beta}(k, n + 1 - k). \tag{8.14} \]

These results are in Wilks (1962, pp. 236–238), for example.

**Discussion Question 8.5** (probability integral transform: discrete?). Find a discrete \( F(\cdot) \) such that \( F(Y) \) is not \( \text{Unif}(0,1) \). That is, choose \( F(\cdot) \), derive the distribution of \( F(Y) \), and check that it’s not uniform. Hint: keep it as simple as possible.

The Dirichlet-based uniform confidence band is constructed in three steps. First, consider a pointwise equal-tailed \( 1 - \alpha \) CI for the unknown \( F(\cdot) \) at each order statistic \( Y_{n:k} \), \( k = 1, \ldots, n \). From (8.14), for any \( k \), \( F(Y_{n:k}) \sim \text{Beta}(k, n + 1 - k) \) exactly. Thus, the \( \tilde{\alpha}/2 \) and \( 1 - \tilde{\alpha}/2 \) quantiles of the \( \text{Beta}(k, n + 1 - k) \) distribution provide respective lower and upper bounds for an exact finite-sample pointwise CI. Second, to ensure joint \( 1 - \alpha \) coverage, we can simulate the Dirichlet distribution (or, simulate uniform random numbers) from (8.13) to find the \( \tilde{\alpha} \) that yields the desired joint coverage. Third, the \( n \) joint CIs can be connected by a stair-step pattern that retains the same uniform coverage probability since \( F(\cdot) \) must be non-decreasing (since it’s a CDF). That is, for \( a < c \), if \( L_a \leq F(a) \leq U_a \) and \( L_c \leq F(c) \leq U_c \), then for any \( b \in (a, c) \), \( L_a \leq F(b) \leq U_c \).

### 8.6.2 Dirichlet GOF Test

Instead of inverting the KS test into a uniform confidence band, the Dirichlet band can be used for a GOF test. That is, any \( F_0(\cdot) \) not completely contained within the band is rejected (meaning \( H_0: F(\cdot) = F_0(\cdot) \) is rejected). The testing analog of the even pointwise coverage probabilities is even type I error rates across \( k = 1, \ldots, n \) when \( F(\cdot) = F_0(\cdot) \).
Discussion Question 8.6 (Dirichlet tail power). Recall DQs [8.1] and [8.2]. Recall the dataset is \( Y_i = \frac{i}{20} \) for \( i = 1, \ldots, k \), and \( Y_i = i + 1000 \) for \( i > k \), with \( n = 20 \), and the null hypothesis is \( H_0: F(\cdot) = F_0(\cdot) \) where \( F_0(\cdot) \) is the \( \text{Unif}(0, 1) \) CDF. Let \( k = 19 \).

a) What is \( F_0(Y_{n:n}) \)? \( (Y_{n:n} = 1020) \).

b) Can the Dirichlet uniform confidence band ever have \( \hat{U}(Y_{n:n}) = 1 \) exactly? Or is it always \( \hat{U}(Y_{n:n}) < 1 \)? Hint: recall \( \hat{U}(Y_{n:n}) \) is the \( 1 - \tilde{\alpha}/2 \) quantile of the \( \text{Beta}(n, 1) \) distribution.

c) Does \( \hat{L}(Y_{n:n}) \leq F_0(Y_{n:n}) \leq \hat{U}(Y_{n:n}) \) here, or not? Does it depend on \( \alpha \)? Explain.

d) Is \( F_0(\cdot) \) contained within the Dirichlet uniform confidence band? Why/not?

e) Explain why this is more reasonable than the result of DQ 8.2.
Appendix to Chapter 8

8.A ECDF: Asymptotic Properties

Like many estimators, the ECDF is consistent and asymptotically normal, but in the sense of a function. Specifically, it is uniformly consistent:

$$\sup_{r \in \mathbb{R}} |\hat{F}(r) - F(r)| \xrightarrow{p} 0.$$  \hspace{1cm} (8.15)

The result also holds almost-surely, with $\xrightarrow{a.s.}$ replacing $\xrightarrow{p}$. This result is also called the Glivenko–Cantelli Theorem.

The asymptotic normality result is an example of a functional central limit theorem (FCLT). It is also called Donsker’s Theorem and says

$$\sqrt{n}(\hat{F}(\cdot) - F(\cdot)) \xrightarrow{d} B(F(\cdot)),$$  \hspace{1cm} (8.16)

where $B(\cdot)$ is a standard Brownian bridge on the unit interval $[0, 1]$, which is a particular type of Gaussian process. That is, $B(\cdot)$ is a random function with a particular distribution. Just as scalar random variable $W$ could have realization $W = 1.53$ or $W = -2.2$, the realization of a random function $B(\cdot)$ is a particular function on $[0, 1]$, called a sample path. Just as some (ranges of) values of $W$ are more likely than others according to its probability distribution, similarly some realizations of $B(\cdot)$ are “more likely” than others according to its distribution. Here, for any $0 \leq t_1 < \cdots < t_k \leq 1$, the finite vector $(B(t_1), \ldots, B(t_k))$ has a multivariate normal distribution with mean $0$ and a particular covariance matrix: for any $s < t$, $\text{Cov}(B(s), B(t)) = s(1 - t)$. Instead of a covariance matrix, the Gaussian process has a covariance function, which describes the covariance between any two possible points. Figure 8.1 shows some sample paths from $B(\cdot)$, 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

Optional resources for this chapter

- R: ks.test()
- R: code for “Comparing distributions by multiple testing across quantiles or CDF values” at [http://faculty.missouri.edu/kaplandm](http://faculty.missouri.edu/kaplandm)
- Stata: command distcomp described by [Kaplan (2019a)](http://faculty.missouri.edu/kaplandm)
- Stata: ksmirnov

The “one-sample” methods, in which a single CDF is unknown, can be extended to “two-sample” methods where two CDFs are unknown. Here, there is one sample from one population, and another (independent) sample from another population. The assumptions of iid sampling and continuous CDFs are continued throughout.

9.1 Setup

The setup for goodness-of-fit (GOF) testing follows. The two-sample, two-sided GOF null hypothesis is

\[ H_0: F_1(\cdot) = F_2(\cdot), \]  

(9.1)

where \( F_1(\cdot) \) and \( F_2(\cdot) \) are both unknown continuous CDFs. We have iid samples of size \( m \) and \( n \), respectively, from which the ECDFs \( \hat{F}_1(\cdot) \) and \( \hat{F}_2(\cdot) \) are computed. The samples are also independent of each other. A level-\( \alpha \) GOF test promises not to reject with probability above \( \alpha \) when in fact \( F_1(\cdot) = F_2(\cdot) \), i.e., the null is true. Ideally, it also has good power, i.e., high rejection probability when \( F_1(\cdot) \neq F_2(\cdot) \).
Alternative, define
\[ \Delta(\cdot) \equiv F_1(\cdot) - F_2(\cdot). \] (9.2)

The GOF null is equivalent to \( H_0: \Delta(r) = 0 \) for all \( r \in \mathbb{R} \). We may also want a uniform confidence band for \( \Delta(\cdot) \), i.e., \( \hat{L}(\cdot) \) and \( \hat{U}(\cdot) \) such that \( P(\hat{L}(\cdot) \leq \Delta(\cdot) \leq \hat{U}(\cdot)) = 1 - \alpha \).

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
\[ D_{m,n} \equiv \sup_{r \in \mathbb{R}} |\hat{F}_1(r) - \hat{F}_2(r)|. \] (9.3)

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 treatment group was surprised with double their anticipated hourly wage; i.e., they got paid $20/hr, whereas the control group got paid $10/hr as anticipated. You’re curious if there is any evidence of “gift exchange,” where productivity increases in response to the “gift” of higher wage. You test the GOF null hypothesis \( H_{0a} \) that the treatment and control productivity distributions are identical. You also test the null hypothesis \( H_{0b} \) that the treatment and control population means are identical.

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

9.2 Exact Finite-Sample Testing

The approach from Section 8.4 does not work here because \( H_0 \) does not determine either true CDF. Before, the null specified a single \( F_0(\cdot) \), so we could apply \( F_0(Y_i) \) in the data. Now, the two-sample null specifies \( F_1(\cdot) = F_2(\cdot) \), but they could be anything, so we cannot take \( F_1(Y_i) \).

However, finite-sample size control is possible using a randomization test or permutation test with the KS statistic. Consider when \( H_0: F_1(\cdot) = F_2(\cdot) \) holds, and let \( F(\cdot) = F_1(\cdot) = F_2(\cdot) \) be the shared distribution. Then, the \( Y_i \) in both samples are iid from the same distribution, \( F(\cdot) \). More formally, imagine labeling observations from both samples together, \( Y_i \) 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 \( i = 1, \ldots, m + n \). 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, even if we have
9.3. ASYMPTOTIC KS

\[ i = 1, \ldots, n \text{ from the second sample and } i = n + 1, \ldots, m + n \text{ from the first sample, we still have } Y_i \overset{iid}{\sim} F(\cdot). \]

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.

There is actually a stronger but similar property specific to the KS test that can be leveraged. Imagine ordering all the \( Y_i \) from both samples combined, and then labeling them “1” or “2” (which sample). So, we summarize the dataset with something like 11121221. The statistic \( D_{m,n} \) only depends on this ordering; the actual \( Y_i \) values are irrelevant. This is because, given any \( r \in \mathbb{R} \), \( \hat{F}_1(r) \) is the proportion of \( Y_i \) below \( r \) in sample 1, and \( \hat{F}_2(r) \) is the proportion of \( Y_i \) below \( r \) in sample 2. Further, \( D_{m,n} \) only depends on the data through the maximum \( |\hat{F}_1(r) - \hat{F}_2(r)| \), which only depends on the ordering. Under \( H_0 \), all orderings are equally likely, irrespective of \( F(\cdot) \). So, the null distribution of \( D_{m,n} \) can be simulated by simulating repeated datasets of size \( m \) and \( n \) from any distribution, say \( \text{Unif}(0,1) \), and computing \( D_{m,n} \) for each. One advantage of this approach over a permutation test is that the distribution (given \( m \) and \( n \)) can be simulated ahead of time, rather than requiring permutations of the observed data. However, the permutation test can alternatively use the CvM or AD statistic, whereas this approach cannot (since they depend on more than the ordering).

In sum: it is straightforward to compute exact, finite-sample \( p \)-values for two-sample GOF testing. For example, just specify the argument `exact=TRUE` in R’s `ks.test()`.

9.3 Asymptotic KS

The asymptotic argument for one-sample KS extends readily to the two-sample case. Since the two samples are independent, it is straightforward to take a difference of ECDFs. Assuming \( n/m \to \lambda \in (0, \infty) \), 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 approach allows for having non-zero \( \Delta(\cdot) \) from (9.2) under \( H_0 \). E.g., it allows something like \( \Delta(r) = 0.1 \mathbb{1}_{\{0 \leq r \leq 1\}} r(1-r) \). Let \( \Delta(\cdot) \) continue to be the
true \( F_1(\cdot) - F_2(\cdot) \), and \( \hat{\Delta}(\cdot) = \hat{F}_1(\cdot) - \hat{F}_2(\cdot) \). Then,

\[
\sqrt{n}(\hat{\Delta}(\cdot) - \Delta(\cdot)) = \sqrt{n}[\hat{F}_1(\cdot) - \hat{F}_2(\cdot) - (F_1(\cdot) - F_2(\cdot))]
\]

\[
= \sqrt{n}[\hat{F}_1(\cdot) - F_1(\cdot)] - \sqrt{n}[\hat{F}_2(\cdot) - F_2(\cdot)]
\]

\[
\rightarrow \frac{\lambda}{\sqrt{\frac{n}{m}}} \text{ Gaussian limit } \frac{\lambda}{\sqrt{\frac{m}{n}}} \text{ Gaussian limit }
\]

\[
\rightarrow \lambda \sqrt{\frac{n}{m}} \sqrt{\frac{m}{n}} [\hat{F}_1(\cdot) - F_1(\cdot)] - \frac{\lambda}{\sqrt{\frac{m}{n}}} [\hat{F}_2(\cdot) - F_2(\cdot)].
\]

This should also allow a uniform confidence band for \( \Delta(\cdot) \) by test inversion. (At least, I think this should all work out; I haven’t actually seen anybody use it.)

### 9.4 Dirichlet Test

The Dirichlet approach can also be applied to a two-sample GOF test. At least to me, it is not immediately obvious how to do this, or why the proposed method works, but see Goldman and Kaplan (2018a) for details (and code on my website).

(I could not figure out how to construct a uniform confidence band for \( \Delta(\cdot) \) using this approach, but that doesn’t mean it’s impossible.)

**Discussion Question 9.2** (two-sample KS vs. Dirichlet). Consider testing \( H_0: F_1(\cdot) = F_2(\cdot) \) by both the finite-sample KS and Dirichlet tests.

a) If the Dirichlet rejects but KS does not, then how do you interpret the results?

b) If the KS rejects but Dirichlet does not, then how do you interpret the results?
Chapter 10

Distributional Inference: Stochastic Dominance

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 (2019a)

In addition to learning whether or not two distribution are equal (as in GOF testing), we may also want to know which distribution is “better.” We could just compare the distributions’ means, or certain quantiles; here, stochastic dominance is considered.

Here, the variables are assumed to have cardinal meaning (like dollars). If they are ordinal, then see Chapter 24.

10.1 First-Order Stochastic Dominance

The concept of first-order stochastic dominance (SD1) has been used in economics as an unequivocal assessment of which distribution is “better.” Consider $Y_1 \sim F_1(\cdot)$ and $Y_2 \sim F_2(\cdot)$. Then, SD1 can be characterized as

$$Y_1 \ SD_1 \ Y_2 \iff E[u(Y_1)] \geq E[u(Y_2)] \text{ for all } u(\cdot) \in U,$$

where $U$ is the set of all (non-decreasing) utility functions. (Given the eventual statistical focus, the difference between “weak” and “strong” SD1 is ignored here.) 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.
CHAPTER 10. DISTRIBUTIONAL INFERENCE: STOCHASTIC DOMINANCE

Notationally, let $Y_1 \text{ nonSD}_1 Y_2$ mean “$Y_1$ does not first-order stochastically dominate $Y_2$.” Thus, for any $(Y_1, Y_2)$, either $Y_1 \text{ SD}_1 Y_2$ or $Y_1 \text{ nonSD}_1 Y_2$, but not both (and not neither).

SD1 has been considered in various settings: $Y$ could represent returns from a financial portfolio, possible yield of an agricultural area, consumption of an individual in a particular country, bids in an auction, etc.

SD1 is very useful because it provides such a strong (unequivocal) assessment, but the stringency of SD1 is also a drawback. For some pairs of $(Y_1, Y_2)$, there is not SD1 in either direction; neither can be said “better” or “worse.” Because of this, SD1 is an example of a partial ordering (rather than a complete ordering). It is also 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,

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

It’s often confusing since the “better” CDF is lower, which usually we interpret as “worse.” It’s perhaps more intuitive (though mathematically equivalent) to express SD1 in terms of quantile functions,

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

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

**Discussion Question 10.1** (SD1 and nonSD1). For each of the following, draw a picture of $F_1(\cdot)$ and $F_2(\cdot)$, or explain why it’s impossible.

a) $Y_1 \text{ SD}_1 Y_2$, $Y_2 \text{ nonSD}_1 Y_1$

b) $Y_1 \text{ nonSD}_1 Y_2$, $Y_2 \text{ SD}_1 Y_1$

c) $Y_1 \text{ SD}_1 Y_2$, $Y_2 \text{ SD}_1 Y_1$

d) $Y_1 \text{ nonSD}_1 Y_2$, $Y_2 \text{ nonSD}_1 Y_1$

10.2 Null of Dominance

Most of the literature (with continuous $Y_1$ and $Y_2$) considers

$$H_0: Y_1 \text{ SD}_1 Y_2 \text{ or equivalently } H_0: F_1(\cdot) \leq F_2(\cdot). \quad (10.4)$$

Implicitly, the alternative hypothesis is $Y_1 \text{ nonSD}_1 Y_2$.

**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.

a) Reject $H_0$: $Y_1 \text{ SD}_1 Y_2$; do not reject $H_0$: $Y_2 \text{ SD}_1 Y_1$

b) Do not reject $H_0$: $Y_1 \text{ SD}_1 Y_2$; reject $H_0$: $Y_2 \text{ SD}_1 Y_1$

c) Reject both $H_0$: $Y_1 \text{ SD}_1 Y_2$ and $H_0$: $Y_2 \text{ SD}_1 Y_1$

d) Do not reject either $H_0$: $Y_1 \text{ SD}_1 Y_2$ or $H_0$: $Y_2 \text{ SD}_1 Y_1$
10.3. NULL OF NON-DOMINANCE

10.2.1 KS Test

The KS test extends to one-sided testing (either one-sample or two-sample) by only considering deviations against $H_0$. For example, instead of $|\hat{F}_1(r) - \hat{F}_2(r)|$ for the two-sided statistic, 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 $r \in \mathbb{R}$. Having a large $\hat{F}_1(r) - \hat{F}_2(r)$ provides evidence against $H_0: F_1(\cdot) \leq F_2(\cdot)$, whereas a large $\hat{F}_2(r) - \hat{F}_1(r)$ provides evidence against $H_0: F_1(\cdot) \geq F_2(\cdot)$.

In R, just 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 are a little confusing (I always forget which means which), so it’s good to sanity-check your results.

10.2.2 Dirichlet Test

The Dirichlet approach also extends to one-sided testing, for both one-sample and two-sample. As with two-sided testing, it spreads power more evenly across the distribution (whereas the KS has very 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 the population parameter $\theta$, like the ATE (or LATE or ATT or whatever; that’s not the point here). You estimate $\hat{\theta} > 0$ but know that might be due to random sampling variation even if really $\theta = 0$. 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) Why do people test this $H_0$ instead of $H_0$: $\theta > 0$, when they’re trying to show evidence of $\theta > 0$?

Some authors like [Davidson and Duclos (2013)] have argued that instead of taking SD1 as the null as in (10.4), it should be the alternative. Equivalently, this leaves non-SD1 as the null:

$$H_0: Y_1 \text{ nonSD}_1 Y_2 \text{ or equivalently } H_0: F_1(r) > F_2(r) \text{ for some } r \in \mathbb{R}.$$  (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 $H_0: Y_1 \text{ nonSD}_1 Y_2$; do not reject $H_0: Y_2 \text{ nonSD}_1 Y_1$

b) Do not reject $H_0: Y_1 \text{ nonSD}_1 Y_2$; reject $H_0: Y_2 \text{ nonSD}_1 Y_1$

c) Reject both $H_0: Y_1 \text{ nonSD}_1 Y_2$ and $H_0: Y_2 \text{ nonSD}_1 Y_1$

d) Do not reject either $H_0: Y_1 \text{ nonSD}_1 Y_2$ or $H_0: Y_2 \text{ nonSD}_1 Y_1$
Unfortunately, with continuous $Y$, it is impossible to truly test (10.5) and control size at level $\alpha$. This is essentially because $Y_1$ SD $Y_2$ can be violated by even a very small probability on a very high value of $Y_1$. For example, imagine it is almost true that $Y_1$ SD $Y_2$, except that there is some $M$ larger than any possible value of $Y_1$ with $P(Y_2 = M) = c/n$, for some $c > 0$. If a dataset contains even a single observation of $Y_2$ equal to $M$, it will be clear that $Y_1$ nonSD $Y_2$ is true. However, if the dataset does not contain any such $M$ observations, then it looks like $Y_1$ nonSD $Y_2$ should be rejected.

What’s the probability of each type of dataset? The probability of a single draw of $Y_2$ being below $M$ is $P(Y_2 < M) = 1 - P(Y_2 = M) = 1 - c/n$. Given $n$ iid draws of $Y_2$, then the probability that none of the draws equals $M$ is $(1 - c/n)^n$. As $n \to \infty$, the limit is

$$\lim_{n \to \infty} P(\text{all below } M) = \lim_{n \to \infty} (1 - c/n)^n = e^{-c}. \quad (10.6)$$

This becomes arbitrarily close to zero as $c \uparrow \infty$. We can tolerate probability $\alpha$ of incorrectly concluding $Y_1$ SD $Y_2$, but probability $e^{-c} < \alpha$ means it’s impossible to control size at level $\alpha$.

There are two responses, neither of which is fully satisfactory. First, we could give up and go back to the null of dominance. But then we can never get very convincing positive evidence of $Y_1$ SD $Y_2$; 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 $H_0: \beta = 0$, my conclusion is that there is zero effect”?)

Second, we could weaken SD1 to something that’s statistically tractable when used as the alternative hypothesis. Davidson and Duclos (2000, 2013) suggest using “restricted” SD1 as introduced in Condition 1 of Atkinson (1987, p. 751). This version of restricted SD1 weakens the CDF characterization of (10.2). Specifically, it “restricts” the CDF ordering to interval $[a, b] \subset \mathbb{R}$. That is, there is restricted SD1 of $Y_1$ over $Y_2$ when $F_1(r) \leq F_2(r)$ for all $r \in [a, b]$, rather than all $r \in \mathbb{R}$. 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. See also Chapter 11, which can apply to non-SD1 testing.

Another option is to weaken the expected utility characterization in (10.1). Instead of defining $U$ as the set of all utility functions, a restricted subset is used, $\tilde{U} \subset U$ analogous to $[a, b] \subset \mathbb{R}$. For example, if only concave (risk-averse) utility functions are considered, then SD1 is weakened to second-order stochastic dominance (SD2). Further restricting the set leads to third-order and higher-order stochastic dominance. (Bootstrap tests with a null of any order stochastic dominance are proposed by Barrett and Donald (2003).) Or, we could restrict attention to a certain parameteric class of utility functions (like CRRA). Again, the tradeoff is that we are more likely to find positive statistical evidence of “dominance” for weaker (less definitive) 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 (2020) for details and examples. The CDF-based “restriction” is easier to approach statistically, but more difficult to interpret economically.
10.3. NULL OF NON-DOMINANCE

Kaplan (2020a) 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 \( E[u(Y_1)] \geq E[u(Y_2)] \) for all \( u(\cdot) \in \mathcal{D} \). Then, Kaplan (2020a) constructs \( \mathcal{D}_1 \) and \( \mathcal{D}_2 \), respectively called “inner” and “outer” confidence sets for \( \mathcal{D} \), such that

\[
P(\mathcal{D}_1 \subseteq \mathcal{D}) \to 1 - \alpha, \quad P(\mathcal{D}_2 \supseteq \mathcal{D}) \to 1 - \alpha.
\]  

(10.7)
Exercises

Exercise E10.1. Apply the methodology from Goldman and Kaplan (2018a) and/or Kaplan (2020a) for assessing first-order stochastic dominance. Code in R and Stata is available at http://faculty.missouri.edu/kaplandm.

a. Find a paper involving stochastic dominance testing that was published in a respectable economics journal\(^1\) and includes publicly available data and ideally code, too; provide a link to the paper.

b. Replicate one of the original paper’s stochastic dominance tests.

c. Run the new methodology to assess stochastic dominance on the same data.

d. Compare your new results to the original results, both statistically and economically.

\(^1\)E.g., in top 500 of https://ideas.repec.org/top/top.journals.all.html
Chapter 11

Distributional Inference: Multiple Testing

Optional resources for this chapter

- Goldman and Kaplan (2018a): CDF comparison across values
- Kaplan (2020a): expected utility comparison across utility functions

Tests for the various null hypotheses in Chapters 8–10 can only report one of two results: reject, or don’t reject. Contrasting the complexity of the population parameters and concepts, the information such a test provides is exceedingly simple. “Simple” is always a tradeoff: easy to communicate, but possibly missing important information. For example, the NBER time series of whether or not the US economy is in a recession is often very helpful, but reducing the entire economy to only two possible states may lose important information for certain applications.

In this chapter, we’ll think about ways to learn more about distributions, still taking statistical uncertainty into account.

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 $f = P(\text{reject } H_{0a} \text{ and/or } H_{0b} | \text{ both true})$. 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 $\alpha$ so that $f \leq 0.1$? (This conservative adjustment is called a Bonferroni adjustment or Bonferroni correction.)

Imagine testing multiple hypotheses simultaneously. Instead of a single $H_0$, maybe we have $H_{0h}$ 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 $H_0: \beta_1 = \beta_2 = 0$. 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.

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 $T \equiv \{h : H_{0h} \text{ is true}\}$ be the set of indexes of true hypotheses. The familywise error rate (FWER) is

$$\text{FWER} \equiv P(\text{reject any } H_{0h} \text{ with } h \in T).$$

**Definition 11.2** (weak and strong control of FWER). Given Definition 11.1, **weak control of FWER** at level $\alpha$ requires FWER $\leq \alpha$ if each $H_{0h}$ is true; **strong control of FWER** requires FWER $\leq \alpha$ for any $T$.

Weak control of FWER is not particularly useful since nothing is guaranteed if even a single $H_{0h}$ is false. It can be a helpful intermediate step theoretically, but in practice, you should be skeptical of an MTP that cannot guarantee strong control of FWER.

11.1.2 Alternatives to FWER

There are alternative false positive rate measures besides FWER. The 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.

Most related to FWER, the $k$-FWER is the probability of making at least $k$ familywise errors:

$$k\text{-FWER} \equiv P(\text{reject at least } k \text{ null hypotheses } H_{0h} \text{ with } h \in T)$$

$$= P(k \text{ or more false positives}). \quad (11.1)$$
11.2. ONE-SAMPLE, TWO-SIDED

The original FWER is the special case $k = 1$. This concept and related procedures were introduced by [Lehmann and Romano (2005a)].

Discussion Question 11.2 ($k$-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 $\gamma$ with high probability $1 - \alpha$, i.e., $P(\text{FDP} \leq \gamma) \geq 1 - \alpha$ or equivalently $P(\text{FDP} > \gamma) \leq \alpha$. See [Lehmann and Romano (2005a)] for details.

Alternatively, the false discovery rate (FDR) is the expected value of the FDP: $\text{FDR} = \mathbb{E}(\text{FDP})$. An MTP could be designed to control $\text{FDP} \leq \alpha$. See [Benjamini and Hochberg (1995)].

11.1.3 Other Ways to Improve Power

Even with strong control of FWER, there are other ways to improve power. In particular, “stepdown” and “pre-test” procedures may help.

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 usually improve power without sacrificing any FWER control, although they do result in more false positives (but only in cases that do not affect FWER or $k$-FWER). Stepdown procedures seem to trace back to [Holm (1979)]; see also [Lehmann and Romano (2005b), Ch. 9].

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 $\alpha$, to add negligible impact to the FWER.

11.2 One-Sample, Two-Sided

The family of null hypotheses is

$$H_{0r}: F(r) = F_0(r), \quad r \in \mathbb{R}.$$  (11.2)
CHAPTER 11. DISTRIBUTIONAL INFERENCE: MULTIPLE TESTING

Any MTP must decide whether or not to reject each of the uncountably infinite number of $H_0$. This sounds daunting, but the monotonicity of $F(\cdot)$ and the discreteness of the sample data combine to simplify the structure.

In contrast, the GOF $H_0: F(\cdot) = F_0(\cdot)$ was only a single hypothesis that warranted a single decision. The MTP provides more detailed information, at least in the case when the GOF rejects.

11.2.1 KS

Recall that the KS approach can generate a uniform confidence band. One natural MTP is to reject any $H_0_r$ for which $F_0(r)$ lies outside the band. It turns out this MTP has strong control of FWER.

There is a general equivalence between uniform confidence bands and MTPs that control FWER. I’ll use CDF notation, but imagine $F(\cdot)$ to be any function over any domain. Assume there exists a uniform confidence band $[\hat{L}(\cdot), \hat{U}(\cdot)]$ such that

$$P(\hat{L}(\cdot) \leq F(\cdot) \leq \hat{U}(\cdot)) = 1 - \alpha. \quad (11.3)$$

Consider the MTP that rejects $H_0_r: F(r) = F_0(r)$ when $F_0(r)$ lies outside the band. The true hypotheses are $H_0_r$ for $r \in \mathcal{T}$. Then,

$$\text{FWER} = P(\text{reject at least one true } H_0_r)$$
$$= P(F_0(r) \notin [\hat{L}(r), \hat{U}(r)] \text{ for some } r \in \mathcal{T})$$
$$= 1 - P(\hat{L}(r) \leq F_0(r) \leq \hat{U}(r) \text{ for all } r \in \mathcal{T})$$
$$= 1 - P(\hat{L}(r) \leq F(r) \leq \hat{U}(r) \text{ for all } r \in \mathcal{T})$$
$$\leq 1 - P(\hat{L}(r) \leq F(r) \leq \hat{U}(r) \text{ for all } r \in \mathbb{R})$$
$$= 1 - (1 - \alpha) = \alpha.$$

Thus, the KS MTP has strong control (i.e., for any $\mathcal{T}$) of FWER at level $\alpha$.

11.2.2 Dirichlet

Since the Dirichlet approach can also construct a uniform confidence band, the same arguments apply. This relatively simple argument does not apply to the two-sample case where the Dirichlet approach cannot provide a uniform confidence band, so alternative arguments are developed by Goldman and Kaplan (2018a).

11.3 Two-Sample and/or One-Sided

Similar approaches can be used for two-sample and/or one-sided MTPs.

As before, the two-sample case is much more common in economics, where we have two unknown population distributions. Here, instead of simply saying “they’re different”
or “I can’t reject that they’re identical,” the MTP can reject \( H_{0r} : F_1(r) = F_2(r) \) for each \( r \in \mathbb{R} \). You can also think of this as generating an “inner” confidence set (CS) for the true set \( F = \{ r : H_{0r} \text{ is false} \} \) of false hypotheses. That is, the set of rejected hypotheses \( \hat{F} = \{ r : H_{0r} \text{ is rejected} \} \) is a conservative “estimate” of the true \( F \) in the sense that

\[
P(\hat{F} \subseteq F) \geq 1 - \alpha.
\]

That is, if \( r \in \hat{F} \), then we can feel relatively confident that \( H_{0r} \) is indeed false. This would not be true if only the pointwise type I error rates were controlled, \( P(\text{reject } H_{0r} \mid H_{0r} \text{ true}) \).

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), as in Chapter 10. In particular, it relates to the idea of restricted SD1. Let \( H_{0r} : F_1(r) \geq F_2(r) \), so rejecting \( H_{0r} \) contributes evidence toward restricted SD1 of \( F_1(\cdot) \) over \( F_2(\cdot) \). Then \( F \), the set of false hypotheses, is the set over which there is restricted SD1 since \( F = \{ r : F_1(r) < F_2(r) \} \). Thus, the MTP with strong FWER control at level \( \alpha \) also generates an inner \( 1 - \alpha \) CS for \( F \) as described above. That is, the set of rejected hypotheses \( \hat{F} \) is the set of points over which we are confident that \( F_1(\cdot) \) is below \( F_2(\cdot) \), in the sense that \( P(\hat{F} \subseteq F) \geq 1 - \alpha \). So, the MTP is able to provide stronger, more positive evidence in favor of restricted SD1 than when SD1 is the null, and it is able to do so without pre-specifying a range over which to test restricted SD1. Further, there may be cases where there are small “holes” in \( \hat{F} \), e.g., \([10, 20] \cup [21, 30] \cup [31, 40]\). Generating this \( \hat{F} \) with a single MTP run is much simpler than trying to test many different ranges for restricted SD1 until \([10, 20], [21, 30], [31, 40] \) are separately found (and having been found separately, it’s unclear that they can simply be combined).

**Discussion Question 11.3** (GOF vs. MTP). Consider \( H_{0r} : F_1(r) = F_2(r) \) for \( r \in \mathbb{R} \). Consider an MTP with strong control of FWER at level \( \alpha \). Consider the GOF test that rejects \( H_0 : F_1(\cdot) = F_2(\cdot) \) if and only if at least one \( H_{0r} \) is rejected by the MTP.

a) Prove that this GOF test controls size at level \( \alpha \).

b) Prove that the MTP is more informative in the sense 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 11.4** (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 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 whether the treatment distribution first-order stochastically dominates the control distribution. All tests/MTPs use level \( \alpha = 0.05 \).

a) A global test does not reject \( H_0 : F_T(\cdot) \leq F_C(\cdot) \). What do you learn about SD1?
CHAPTER 11. DISTRIBUTIONAL INFERENCE: MULTIPLE TESTING

b) A global test rejects $H_0: F_T(\cdot) \geq F_C(\cdot)$. What do you learn about SD1?

c) An MTP rejects $H_{0r}: F_T(r) \geq F_C(r)$ for all $r$ between the sample 5th and 95th percentiles of the control distribution. What do you learn about SD1?

d) An MTP rejects $H_{0r}: F_T(r) \geq F_C(r)$ for all $r$ between the sample 5th and 65th percentiles of the control distribution, and it rejects $H_{0r}: F_T(r) \leq F_C(r)$ 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 11.5** (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 $H_{0r}: F_T(r) = F_C(r)$ 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 $H_{0r}: F_T(r) = F_C(r)$ 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 $H_0: F_T(\cdot) = F_C(\cdot)$? Why/not?

d) An MTP for $H_{0r}: F_T(r) \geq F_C(r)$ rejects (only) for all $r$ between the 5th and 25th sample percentiles of the control distribution, and an MTP for $H_{0r}: F_T(r) \leq F_C(r)$ 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 $H_0: F_T(\cdot) \geq F_C(\cdot)$ and $H_0: F_T(\cdot) \leq F_C(\cdot)$? 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.

a. Find a paper published in a respectable economics journal that includes (publicly available) data from a randomized experiment; provide a link to the paper.

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.).

\[http://faculty.missouri.edu/kaplanm\]

\[E.g.,\ in\ top\ 500\ of\ https://ideas.repec.org/top/top.journals.all.html\]
CHAPTER 11. DISTRIBUTIONAL INFERENCE: MULTIPLE TESTING
Part IV

Bootstrap and Friends
Introduction

This part concerns bootstrap methods and related approaches like subsampling and non-parametric Bayesian inference.
Chapter 12

Bootstrap: Basics

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](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](http://merlin.lib.umsystem.edu/record=b2717057-S1)
- Textbook: Davison and Hinkley (1997)? MU: [http://merlin.lib.umsystem.edu/record=b3774612-S1](http://merlin.lib.umsystem.edu/record=b3774612-S1)
- R: try package boot ([Canty and Ripley](http://rlang.rstudio.com), [Davison and Hinkley](http://statlab.cam.ac.uk/ftp/old/tech-reports/reports/dh250.pdf) (1997))

I use scalar notation in most of this section because it’s easier, but almost everything immediately applies to vectors, too.

My bootstrap and subsampling education is partly due to the class by Dimitris Politis, but of course all errors are mine.

12.1 Introduction

Bootstrap techniques have gained enormous popularity in applied economics. Although people reference “the bootstrap,” there are actually many different types of bootstrap, and multiple ways to construct a confidence interval for each type. The same bootstrap approach can be used for a wide variety of estimators and models. However, it is not
magic, and in some settings it fails. In such settings, sometimes the related technique
of subsampling can work. However, subsampling is not strictly better than bootstrap:
subsampling involves an additional smoothing parameter and often has lower power.

This chapter introduces the main ideas of bootstrap and subsampling and discusses
some tradeoffs. For more details, see the references herein, or try Google.

“The” bootstrap was introduced by Efron (1979). It improved upon the Quenouille–
Tukey jackknife in certain ways, though jackknife methods are still used in some cases.
As computation power has grown since 1979, bootstrap methods have become more con-
venient in practice, although with complex estimators and/or large datasets they may
still be computationally demanding.

I focus on confidence intervals, but hypothesis testing may also be done via bootstrap.
For example, if a 95% CI exists, then the test rejecting any hypothesis outside the CI is
a valid level-5% test.

The bootstrap is primarily a frequentist tool, but see Chapter [14] for an objective
nonparametric Bayesian interpretation of a particular bootstrap.

12.2 Preliminaries: The Plug-in Principle

To understand the core bootstrap idea in Section [12.3], it helps to review the plug-in
principle, also called the analogy principle.

Notationally, let $F(\cdot)$ denote the population joint distribution of all observable vari-
ables. Let $\theta = \theta(F(\cdot))$ be the population parameter of interest. That is, $\theta$ 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 [8.2].

The plug-in principle suggests replacing $F(\cdot)$ with $\hat{F}(\cdot)$ to estimate $\theta$. That is, it
suggests

$$\hat{\theta} = \theta(\hat{F}(\cdot)) \quad \text{estimates} \quad \theta = \theta(F(\cdot)). \quad (12.1)$$

The same summary function $\theta(\cdot)$ is applied to empirical distribution $\hat{F}(\cdot)$.

12.2.1 Example: Mean

Consider the mean of a scalar population distribution. In terms of $F(\cdot)$, the population
mean is

$$\theta(F(\cdot)) = \int_{\mathbb{R}} y \, dF(y). \quad (12.2)$$

---

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 extremely similar when applied
to linear statistics/estimators like the mean, but the jackknife can be much worse for nonlinear statistics.
12.2. PRELIMINARIES: THE PLUG-IN PRINCIPLE

Here, \( \theta(\cdot) \) is the function that computes the mean corresponding to a CDF. 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 \( \int_{\mathbb{R}} x \, d\Phi(x) = \int_{\mathbb{R}} x \phi(x) \, dx = 0 \). Assume the EDF \( \hat{F}(\cdot) \) is computed from an iid sample of \( Y_i \overset{iid}{\sim} F(\cdot) \), \( i = 1, \ldots, n \). Analogous to (12.2), we could use \( \theta(\cdot) \) but plug in \( \hat{F}(\cdot) \) instead of \( F(\cdot) \):

\[
\theta(\hat{F}(\cdot)) = \int_{\mathbb{R}} y \, d\hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} Y_i (1/n) = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y},
\]

(12.3)

the usual sample mean. That is, the “sample mean” can be interpreted as taking the mean of the “sample distribution” or 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, i.e., the usual formula for the mean of a discrete distribution.

12.2.2 Example: OLS

As another example, consider OLS with iid data. Let \( F(\cdot) \) be the population joint CDF of \( (Y, X) \). Let \( \hat{F}(\cdot) \) be the empirical CDF. The population parameter is the linear projection coefficient. The expectations in the usual formula can be written in terms of \( F(\cdot) \) for emphasis:

\[
\beta(F(\cdot)) = \left[ \int x x' \, dF_X(x) \right]^{-1} \int x y \, dF_{X,Y}(x, y) = \left[ \text{E}(X_i X'_i) \right]^{-1} \text{E}(X_i Y_i),
\]

(12.4)

for simplicity omitting the limits of integration on the definite integrals. Applying the same function \( \beta(\cdot) \) to \( \hat{F}(\cdot) \) instead of \( F(\cdot) \),

\[
\beta(\hat{F}(\cdot)) = \left[ \int x x' \, d\hat{F}_X(x) \right]^{-1} \int x y \, d\hat{F}_{X,Y}(x, y)
\]

\[
= \left[ \frac{1}{n} \sum_{i=1}^{n} (X_i X'_i)(1/n) \right]^{-1} \frac{1}{n} \sum_{i=1}^{n} (X_i Y_i)(1/n)
\]

\[
= \left[ \frac{1}{n} \sum_{i=1}^{n} X_i X'_i \right]^{-1} \frac{1}{n} \sum_{i=1}^{n} X_i Y_i,
\]

the OLS estimator \( \hat{\beta} \). This is the same as replacing the \( \text{E} \) (population expectation operator) with \( \hat{\text{E}} \) (sample expectation operator) and skipping the more complicated-looking (but equivalent) integrals.

12.2.3 Other Types of Parameters

Some parameters are not defined as functions of \( F(\cdot) \). Notably, causal parameters are not defined in terms of \( F(\cdot) \). However, if they are identified, then they can be. Like
most statistical inference approaches, the bootstrap helps us quantify uncertainty about
the feature of \( F(\cdot) \) that has a causal interpretation if identification holds, but it does not
quantify uncertainty about identification. Thus, if identification fails, the bootstrap may
give us a fine confidence interval for the statistical parameter \( \theta(F(\cdot)) \), but the interval
could be way off from the causal parameter of interest.

12.3 The Real World and the Bootstrap World

**Discussion Question 12.1** (frequentist bias). In the frequentist framework, if you ac-
tually knew the true population \( F(\cdot) \), how could you figure out the bias of an estimator,
\( \text{Bias}(\hat{\theta}) \)? 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). In the frequentist framework, imagine
you actually knew the true population \( F(\cdot) \). Hint for both parts: what are the defi-
nitions/properties of SE and CI?

a) How could you figure out the standard error \( \text{SE}(\hat{\theta}) \) of an estimator \( \hat{\theta} \)? Explain.
b) How could you figure out the coverage probability of a proposed confidence interval
for \( \theta \)? Explain.

Although bootstraps usually involve Monte Carlo simulation, the bootstrap idea itself
does not intrinsically involve computation. This is parallel to how concepts like bias,
standard error, and confidence interval are not intrinsically defined in terms of Monte
Carlo simulation, although you could propose such an approach in DQs [12.1] and [12.2].
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.)

The goal of bootstrapping is to learn about the sampling distribution of an estimator,
say \( \hat{\theta} \). If we knew the true population distribution \( F(\cdot) \), then we could learn about the
sampling distribution and its features using the approaches you proposed in DQs [12.1]
and [12.2]. Of course, if we knew \( F(\cdot) \), then we could simply compute \( \theta(F(\cdot)) \) directly and
not worry about statistical inference. Since we don’t know \( F(\cdot) \), the bootstrap proposes
other ways to mimic the sampling distribution of \( \hat{\theta} \).

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.

Following Section [12.2], what if the plug-in principle were extended from estimation
to inference? That is, what if we “plug in” \( \hat{F}(\cdot) \) for \( F(\cdot) \) when considering the sampling
distribution of \( \hat{\theta} \)? From Section [8.2], \( \hat{F}(\cdot) \) \( \xrightarrow{a.s.} \) \( F(\cdot) \) uniformly; it is a “good” estimator in
that sense. Hopefully the corresponding estimator of the sampling distribution is also
“good.” Computationally, we can repeatedly draw random samples from \( \hat{F}(\cdot) \), and see
how \( \hat{\theta} \) varies across these samples to estimate its frequentist properties.
12.3. THE REAL WORLD AND THE BOOTSTRAP WORLD

12.3.1 The Real World

The idea of sampling from $\hat{F}(\cdot)$ mimics the bigger picture view of the frequentist framework. Imagine $X$ is a person’s height (in meters). There is some population distribution $F(\cdot)$. There is iid sampling, $X_i \overset{iid}{\sim} F(\cdot)$. After we take the 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 instead, 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.

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 EDF $\hat{F}(\cdot)$ very different from $F(\cdot)$. If we could just keep drawing random samples from $F(\cdot)$, then this would be straightforward. But in practice, we only have one sample from $F(\cdot)$.

Operators like expectation $\mathbb{E}(\cdot)$ are usually implicitly defined wrt $F(\cdot)$. For example, consider the expected value of estimator $\hat{\theta} = \theta(X)$, where $X = (X_1, \ldots, X_n)$ is the full dataset, and $X_i \overset{iid}{\sim} F(\cdot)$. The distribution $\hat{F}(\cdot)$ of $X$ is determined by $F(\cdot)$. The expectation of $\hat{\theta}$ means the weighted average over all possible samples we could have drawn from $\hat{F}(\cdot)$:

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}[\theta(X)] = \int \theta(x) d\hat{F}(x).$$

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 $X$ or equivalently $\hat{F}(\cdot)$. The real-world estimator is computed from the real-world sample: $\hat{\theta} = \theta(X)$, or $\hat{\theta} = \theta(\hat{F}(\cdot))$.

12.3.2 The Bootstrap World

In parallel to the real world is the bootstrap world. 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 $X^*$ or equivalently $\hat{F}^*(\cdot)$. The bootstrap-world estimator is computed from the bootstrap-world sample: $\hat{\theta}^* = \theta(X^*)$, or $\hat{\theta}^* = \theta(\hat{F}^*(\cdot))$. That is, we treat the EDF as if it were the population, and define other objects accordingly. See Table 12.1.

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 $F(\cdot)$.

---

$^3$David Freedman’s term, according to Efron and Tibshirani (1993) p. 86; and see their Figure 8.3 on page 87.
CHAPTER 12. BOOTSTRAP: BASICS

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 $\hat{F}(\cdot)$ in the bootstrap world.

Table 12.1: The Real World and the parallel Bootstrap World.

<table>
<thead>
<tr>
<th>Object</th>
<th>Real World</th>
<th>Bootstrap World</th>
</tr>
</thead>
<tbody>
<tr>
<td>pop. distribution</td>
<td>$F(\cdot)$</td>
<td>$\hat{F}(\cdot)$</td>
</tr>
<tr>
<td>pop. parameter</td>
<td>$\theta(F(\cdot))$</td>
<td>$\theta(\hat{F}(\cdot))$</td>
</tr>
<tr>
<td>sample data</td>
<td>$X_i \sim F(\cdot)$, $i = 1, \ldots, n$</td>
<td>$X^*_i \sim \hat{F}(\cdot)$, $i = 1, \ldots, n$</td>
</tr>
<tr>
<td>sample dist/EDF</td>
<td>$F(\cdot)$</td>
<td>$\hat{F}^*(\cdot)$</td>
</tr>
<tr>
<td>estimator</td>
<td>$\hat{\theta} = \theta(\hat{F}(\cdot))$</td>
<td>$\hat{\theta}^* = \theta(\hat{F}^*(\cdot))$</td>
</tr>
<tr>
<td>root</td>
<td>$\hat{\theta} - \theta$</td>
<td>$\hat{\theta}^* - \hat{\theta}$</td>
</tr>
</tbody>
</table>

12.4 Empirical Bootstrap

This section discusses one specific bootstrap approach to estimating the sampling distribution of an estimator, assuming iid sampling. Other bootstraps (in later sections) could also be used to estimate the sampling distribution. Section 12.6 discusses how to use such an estimated sampling distribution 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** since it samples “pairs” $(Y_i, X_i)$ from the empirical distribution, instead of keeping the $X_i$ fixed and only resampling residuals like in Section 13.2.2.

Method 12.1 describes how to estimate the sampling distribution of $\hat{\theta}$ using the empirical bootstrap. As shown later, this can be used subsequently to compute standard errors, confidence intervals, etc.

**Method 12.1** (empirical bootstrap). Assume (in the real world) $W_i \overset{iid}{\sim} F(\cdot)$; e.g., perhaps $W_i = (Y_i, X_i^\prime)^\prime$. Let $n$ denote the sample size (in both the real world and bootstrap world). Let $B$ denote the number of bootstrap replications. Let $W^*_i$ 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 $W^*_i$ randomly from the original $W_i$; i.e., $W^*_i = W_S$ with $P(S = j) = 1/n$ for $j = 1, \ldots, n$. Do this independently and with replacement, i.e., it is fine to get the same $S$ twice for
12.5. STANDARD ERRORS

A given \( b \) and thus have \( W_{i}^{*b} = W_{j}^{*b} \) for \( i \neq j \). This could also be written as \( W_{i}^{*b} \simiid \hat{F}(\cdot) \), where \( \hat{F}(\cdot) \) is the empirical distribution.

Let \( \theta \) denote the population parameter. Let \( \hat{\theta} \) denote an estimator computed from the original \( W_{i}, i = 1, \ldots, n \). Let \( \overline{\theta}^{*b} \) denote the same estimator but computed from bootstrap sample \( b \) of observations \( W_{i}^{*b}, i = 1, \ldots, n \). The empirical bootstrap estimates the real-world sampling distribution of \( \hat{\theta} - \theta \) by the bootstrap-world distribution of \( \overline{\theta}^{*b} - \theta \). More specifically, the above procedure provides \( B \) random draws of \( \overline{\theta}^{*b} - \theta \) from its probability distribution (conditional on the original sample).

Although Method \text{[12.1]} consistently estimates the sampling distribution for a wide variety of estimators, it can fail in some cases. See Sections \text{[13.3][13.6]} for examples of failure and possible solutions.

**Discussion Question 12.4** (empirical bootstrap 1). Let \( n = 2 \) with \( Y_{1} = 0 \) and \( Y_{2} = 1 \). Let \( \theta = E(Y) \).

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 (a)? (Sanity check: should sum to 100% probability.)

**Discussion Question 12.5** (empirical bootstrap 2). Continue from DQ \text{[12.4]}

a) What are the possible values of the bootstrap-world estimator \( \overline{\theta}^{*} = (Y_{1}^{*} + Y_{2}^{*})/2 \)?

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 \( \overline{\theta}^{*b} \) in each sample \( (b = 1, \ldots, B) \), then what’s the bootstrap approximation of the bootstrap-world probability that \( \overline{\theta}^{*} = 0 \)?

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 \( \overline{\theta}^{*b} - \hat{\theta} \). These could be produced by the empirical bootstrap in Method \text{[12.1]} or by any other bootstrap.

The standard error of \( \hat{\theta} \) in the real world is simply the standard deviation of its sampling distribution. Since \( \theta \) is a constant, the standard error also equals the standard deviation of the sampling distribution of \( \hat{\theta} - \theta \).

Method \text{[12.2]} is essentially Algorithm 6.1 of \text{Efron and Tibshirani (1993)}.
CHAPTER 12. BOOTSTRAP: BASICS

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

\[
\hat{SE}(\hat{\theta}) = \sqrt{(B - 1)^{-1} \sum_{b=1}^{B} (\hat{\theta}^b - \bar{\theta}^*)^2}, \quad \bar{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^b.
\]

If the sampling distribution is asymptotically normal, then an alternative based on the interquartile range can be used, as suggested by Chernozhukov, Fernández-Val, and Melly (2013, p. 2222–2223). This may be less sensitive than Method 12.2 to “outlier” draws of \( \hat{\theta}^b \). The idea is that if \( W \sim N(0, \sigma^2) \), then its CDF is \( F_W(w) = \Phi(w/\sigma) \), so \( Q_\tau(W) \) solves \( \Phi(Q_\tau(W)/\sigma) = \tau \), which means \( Q_\tau(W)/\sigma \) is the \( \tau \)-quantile of a standard normal distribution, say \( z_\tau \). Since \( Q_\tau(W)/\sigma = z_\tau \), then \( Q_\tau(W) = \sigma z_\tau \), so

\[
Q_{0.75}(W) - Q_{0.25}(W) = \sigma z_{0.75} - \sigma z_{0.25} = \sigma (z_{0.75} - z_{0.25}) \Rightarrow \sigma = \frac{Q_{0.75}(W) - Q_{0.25}(W)}{z_{0.75} - z_{0.25}}.
\]

(12.5)

That is, if we know \( W \) is normal, then its standard deviation can be written as the interquartile range of \( W \) divided by the standard normal interquartile range. This is the approach of Method 12.3. In principle, there is nothing special about 0.25 and 0.75, but the quantile indices should be symmetric about 0.5 and probably not too close together nor too far into the tails, so 0.25 and 0.75 seem reasonable.

Method 12.3 (empirical bootstrap: SE for normal). First compute \( B \) values of \( \hat{\theta}^b \) using Method 12.1 or another bootstrap. Let \( q^*_\tau \) denote the sample \( \tau \)-quantile among \( \hat{\theta}^b \) over \( b = 1, \ldots, B \). Let \( z_\tau \) denote the \( \tau \)-quantile of a \( N(0,1) \) distribution. Then the estimated standard error is

\[
\hat{SE}(\hat{\theta}) = \frac{q^*_{0.75} - q^*_{0.25}}{z_{0.75} - z_{0.25}}.
\]

In either case, as \( B \to \infty \), the bootstrapped standard error estimator approaches the bootstrap world’s “population” standard error of \( \hat{\theta} \). Again, the hope is that the bootstrap-world standard error is close to the real-world standard error.

12.6 Confidence Intervals

This section describes multiple ways to construct a confidence interval (CI) for \( \theta \) based on \( B \) draws of \( \hat{\theta}^b - \hat{\theta} \). These draws could be produced by the empirical bootstrap in Method 12.1 or by any other bootstrap.

12.6.1 CI Properties

The focus is on two-sided (asymptotic) equal-tailed CIs. Write the CI as \([\hat{L}, \hat{U}]\). If the CI has 1 – \( \alpha \) coverage probability (CP), then \( P(\hat{L} \leq \theta \leq \hat{U}) = 1 - \alpha \). The equal-tailed
12.6. CONFIDENCE INTERVALS

property further specifies
\[ P(\hat{L} > \theta) = P(\hat{U} < \theta), \] (12.6)
i.e., there is equal probability of the CI being “too low” or “too high.” To satisfy the overall CP, this implies \( P(\hat{L} > \theta) = P(\hat{U} < \theta) = \alpha / 2 \). This further implies
\[ P(\hat{L} \leq \theta) = 1 - \alpha / 2, \quad P(\theta \leq \hat{U}) = 1 - \alpha / 2, \] (12.7)
i.e., \([\hat{L}, \infty)\) and \((-\infty, \hat{U})\) are one-sided \( 1 - \alpha / 2 \) CIs. Thus, generally, a two-sided equal-tailed \( 1 - \alpha \) CI can be constructed as the intersection of two one-sided \( 1 - \alpha / 2 \) CIs. In practice, these probabilities are often (though not always) asymptotic, e.g., \( P(\hat{L} \leq \theta \leq \hat{U}) \to 1 - \alpha \), \( P(\hat{L} \leq \theta) \to 1 - \alpha / 2 \), \( P(\theta \leq \hat{U}) = 1 - \alpha / 2 \).

If the (asymptotic) distribution of an estimator is normal, then the equal-tailed CI is also (asymptotically) symmetric, due to the normal distribution’s symmetry. A symmetric CI is “symmetric” about the estimator \( \hat{\theta} \) in the sense that \( \hat{U} - \hat{\theta} = \hat{\theta} - \hat{L} \); it can be written in the form \( \hat{\theta} \pm \hat{c} \), like \( \hat{c} = 1.96 \hat{SE}(\hat{\theta}) \). There are details below, but it is essentially due to properties of the standard normal CDF \( \Phi(\cdot) \), like \( \Phi(-q) = 1 - \Phi(q) \) and thus \( \Phi(q) - \Phi(-q) = 2\Phi(q) - 1 \), for \( q > 0 \).

Discussion Question 12.6 (CI properties). Consider the degenerate CI \([\hat{L}, \hat{U}]\) with \( P(\hat{L} = \hat{U} = 3.14159) = \alpha \) and \( P(\hat{L} = -\infty, \hat{U} = \infty) = 1 - \alpha \).

a) Starting from the definition of coverage probability, compute the coverage probability.

b) Is the CI “equal-tailed”? 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,
\[ \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2). \] (12.8)
If a consistent estimator \( \hat{\sigma}^2 \xrightarrow{p} \sigma^2 \) exists, then defining \( \hat{SE}(\hat{\theta}) = \hat{\sigma} / \sqrt{n} \),
\[ \hat{Z}_n \equiv \frac{\hat{\theta} - \theta}{\hat{SE}(\hat{\theta})} \xrightarrow{d} Z \sim N(0, 1). \] (12.9)

The left-hand side has been Studentized since the estimator \( \hat{\theta} \) was “centered” (at the true \( \theta \)) and “scaled” by 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. We have seen this to be useful earlier in Section 8.6, which used the fact that if \( X_i \iid F(\cdot) \) then \( F(X_i) \iid \text{Unif}(0, 1) \) if \( F(\cdot) \) is continuous, and further \( F(X_{n,k}) \sim \text{Beta}(k, n + 1 - k) \); these are finite-sample pivotal distributions, as opposed to asymptotic pivotal distributions. A statistic with an (asymptotic) pivotal distribution can be called an (asymptotic) pivot.
Consider a two-sided equal-tailed 95% CI for \( \theta \). With \( Z \) in \((12.9)\), \( P(-1.96 < Z < 1.96) = 0.95 \). Consequently, with \( \hat{\theta} \) meaning we drop asymptotically negligible terms,

\[
0.95 \doteq P(-1.96 < \frac{\hat{\theta} - \theta}{\text{SE}(\hat{\theta})} < 1.96)
\]

\[
= P(-1.96 \hat{\text{SE}}(\hat{\theta}) - \hat{\theta} < -\theta < 1.96 \hat{\text{SE}}(\hat{\theta}) - \hat{\theta})
\]

\[
= P(\hat{\theta} - 1.96 \hat{\text{SE}}(\hat{\theta}) < \theta < \hat{\theta} + 1.96 \hat{\text{SE}}(\hat{\theta}))
\]

which is our familiar CI \( \hat{\theta} \pm 1.96 \hat{\text{SE}}(\hat{\theta}) \).

More generally, consider a two-sided equal-tailed \( 1 - \alpha \) CI, and let \( Z \) follow any known pivotal continuous distribution (with standard normal as a special case). Let \( z_\tau \) denote the \( \tau \)-quantile of \( Z \), so \( P(Z \leq z_\tau) = \tau \) for \( 0 < \tau < 1 \). (If \( Z \) were allowed to be discrete or have mass points, then the \( \pm \) becomes \( \leq \) and \( \geq \) and similar modifications must be made below, but the intuition is the same.) Since \( Z \) is continuous, \( P(Z \leq z) = P(Z < z) \), so I just use \( < \) for simplicity. For any \( \alpha \), \( P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha \). Thus,

\[
1 - \alpha = P(z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\text{SE}(\hat{\theta})} < z_{1-\alpha/2})
\]

\[
= P(z_{\alpha/2} \hat{\text{SE}}(\hat{\theta}) - \hat{\theta} < -\theta < z_{1-\alpha/2} \hat{\text{SE}}(\hat{\theta}) - \hat{\theta})
\]

\[
= P((-1)(z_{\alpha/2} \hat{\text{SE}}(\hat{\theta}) - \hat{\theta}) > (-1)(-\theta) > (-1)(z_{1-\alpha/2} \hat{\text{SE}}(\hat{\theta}) - \hat{\theta}))
\]

\[
= P(\hat{\theta} - z_{1-\alpha/2} \hat{\text{SE}}(\hat{\theta}) < \theta < \hat{\theta} - z_{\alpha/2} \hat{\text{SE}}(\hat{\theta}))
\]

\((12.10)\)

the CP of the CI \( [\hat{\theta} - z_{1-\alpha/2} \hat{\text{SE}}(\hat{\theta}), \hat{\theta} - z_{\alpha/2} \hat{\text{SE}}(\hat{\theta})] \). This looks unfamiliar and initially wrong. But if \( Z \sim N(0, 1) \), then \( z_{\alpha/2} = -z_{1-\alpha/2} \) and \((12.10)\) becomes \( \hat{\theta} \pm z_{1-\alpha/2} \hat{\text{SE}}(\hat{\theta}) \), the usual normality-based CI.

**Method 12.4** (bootstrap CI: SE). Let \( \hat{\text{SE}}(\hat{\theta}) \) be a bootstrap standard error estimator as in Section \(12.5\). A two-sided \( 1 - \alpha \) CI is

\[
[\hat{\theta} - z_{1-\alpha/2} \hat{\text{SE}}(\hat{\theta}), \hat{\theta} - z_{\alpha/2} \hat{\text{SE}}(\hat{\theta})]
\]

where \( z_\tau \) is the \( \tau \)-quantile of the \( Z \) in \((12.9)\). For example, if \( Z \sim N(0, 1) \), 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, \( \hat{\theta} - \theta \). (This is somewhat obscured in the notation of \((12.13)\).) 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 getting to the method itself, some motivation is provided.
12.6. CONFIDENCE INTERVALS

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 \(-z_\alpha\) by \(+z_{1-\alpha}\).

Besides the math in (12.10), a simple example may help intuition. Imagine \(\hat{\theta} \sim N(\theta, 1)\), so \(P(\theta - 1.96 < \hat{\theta} < \theta + 1.96) = 0.95\). If we happen to sample a dataset with \(\hat{\theta} = \theta + 1.96\), then \(\theta = \hat{\theta} - 1.96\), i.e., the true \(\theta\) is below our estimate by 1.96. This is our lower endpoint: we’ll include values up to 1.96 below our estimate, but exclude values even farther below our estimate. The true value \(\theta\) is farthest below our estimated \(\hat{\theta}\) when \(\hat{\theta}\) is drawn from the upper quantiles of its sampling distribution. Similarly, if we sample a dataset with \(\hat{\theta} = \theta - 1.96\), then \(\theta = \hat{\theta} + 1.96\), i.e., the true \(\theta\) 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: \(\hat{\theta} - \theta \sim \text{Exp}(1)\). The CDF is \(F(x) = 1 - e^{-x}\). The \(\tau\)-quantile is \(F^{-1}(\tau) = -\ln(1-\tau)\). Imagine a two-sided 90% CI for \(\theta\). With \(\alpha = 0.1\), (12.10) suggests the two-sided equal-tailed CI \([\hat{\theta} - F^{-1}(0.95), \hat{\theta} - F^{-1}(0.05)]\).

Here, you’ll consider why other variations wouldn’t work well. Draw an example with each of these; try drawing the PDF of \(\hat{\theta}\) on a graph first, with the true \(\theta\) 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 \([\hat{\theta} - F^{-1}(0.95), \hat{\theta} + F^{-1}(0.05)]\)?
- c) What’s the CP of \([\hat{\theta} + F^{-1}(0.05), \hat{\theta} + F^{-1}(0.95)]\)?
CHAPTER 12. BOOTSTRAP: BASICS

Then,

\[ 1 - \alpha = P(r_{\alpha/2} < \hat{\theta} - \theta < r_{1-\alpha/2}) \]

\[ = P(r_{\alpha/2} - \hat{\theta} < -\theta < r_{1-\alpha/2} - \hat{\theta}) \]

\[ = P((-1)(r_{\alpha/2} - \hat{\theta}) > (-1)(-\theta) > (-1)(r_{1-\alpha/2} - \hat{\theta})) \]

\[ = P(\hat{\theta} - r_{1-\alpha/2} < \theta < \hat{\theta} - r_{\alpha/2}), \quad (12.11) \]

the coverage probability of the CI \([\hat{\theta} - r_{1-\alpha/2}, \hat{\theta} - r_{\alpha/2}]\).

The \(r_\tau\) in (12.11) can be replaced by bootstrap estimates. That is, \(r_\tau\) can be replaced by \(r^*_\tau\), the \(\tau\)-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 \(\tau\)-quantile of \(\hat{\theta}^*\) first and then subtracting \(\hat{\theta}\). That is, if \(q^*_\tau\) is the bootstrap \(\tau\)-quantile of \(\hat{\theta}^*\), then \(r^*_\tau = q^*_\tau - \hat{\theta}\). Thus, the CI \([\hat{\theta} - r^*_1, \hat{\theta} - r^*_\alpha]\) can be estimated by the bootstrap version \([\hat{\theta} - r^*_1, \hat{\theta} - r^*_\alpha]\) or equivalently

\[ [\hat{\theta} - (q^*_1 - \hat{\theta}), \hat{\theta} - (q^*_{\alpha/2} - \hat{\theta})] = [2\hat{\theta} - q^*_1, 2\hat{\theta} - q^*_{\alpha/2}]. \quad (12.12) \]

This is summarized in Method [12.5]

**Method 12.5** (bootstrap CI: basic/standard/root method). Assume we have computed a set of \(B\) bootstrapped estimators \(\hat{\theta}^b\) for \(b = 1, \ldots, B\). Then, a two-sided equal-tailed \(1 - \alpha\) confidence interval for \(\theta\) is

\[ [2\hat{\theta} - q^*_1, 2\hat{\theta} - q^*_{\alpha/2}], \quad (12.13) \]

where \(q^*_\tau\) denotes the sample \(\tau\)-quantiles of \(\hat{\theta}^b\). This is equivalent to the CI \([\hat{\theta} - r^*_1, \hat{\theta} - r^*_\alpha]\), where \(r^*_\tau\) is the sample \(\tau\)-quantile of the bootstrapped roots, \(\hat{\theta}^b - \hat{\theta}\). \(\Box\)

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 two-sided equal-tailed percentile bootstrap \(1 - \alpha\) confidence interval for \(\theta\) is

\[ [q^*_{\alpha/2}, q^*_{1-\alpha/2}], \quad (12.14) \]

where \(q^*_\tau\) is the sample \(\tau\)-quantile of \(\hat{\theta}^b\). \(\Box\)
12.6. CONFIDENCE INTERVALS

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 $\hat{\theta}^{sb}$. I have even heard it criticized for being “backward,” which is fine if the sampling distribution is symmetric but not with any asymmetry. At one point, I planned to mention such criticisms in a referee report of a paper that focused exclusively on the percentile bootstrap. But, I wanted better evidence than just rumors, so I ran some simulations. I tried to use examples where the finite-sample sampling distribution is very skewed, like with sample quantiles for $\tau$ closer to 0 or 1. Despite this, the percentile CIs performed very well, even better than other methods.

The percentile CI is the most closely related to Bayesian approaches; see 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 $(\hat{\theta} - \theta) / \hat{\text{SE}}(\hat{\theta})$. Although bootstrapping the estimator or root is equivalent (since they differ by only a constant), bootstrapping the Studentized estimator is not equivalent since the denominator is not a constant.

With a Studentized estimator, both the estimator and the standard error estimator must be recomputed in each bootstrap sample. Notationally, let $\hat{Z} = (\hat{\theta} - \theta) / \hat{\text{SE}}$ in the real world and $\hat{Z}^{sb} = (\hat{\theta}^{sb} - \theta) / \hat{\text{SE}}^{sb}$ in bootstrap sample $b$, $b = 1, \ldots, B$. The standard error estimator $\hat{\text{SE}}^{sb}$ could use a formula (like for heteroskedasticity-robust OLS standard errors). Or, it may itself be 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, in theory, as long as the estimator is “smooth” enough (e.g., excluding quantile regression), Studentization always increases accuracy. 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.

The CI in Method 12.7 is similar to that in Method 12.4 but the latter uses quantiles $z_\tau$ from the asymptotic distribution of the Studentized estimator, whereas Method 12.7 uses the bootstrap to estimate that distribution’s quantiles as $\hat{z}_\tau$.

Method 12.7 (bootstrap-t/percentile-t/Studentized bootstrap CI). Assume we have computed a set of $B$ bootstrapped Studentized estimators $\hat{Z}^{sb} = (\hat{\theta}^{sb} - \theta) / \hat{\text{SE}}^{sb}$, $b = 1, \ldots, B$. Then, a two-sided equal-tailed $1 - \alpha$ confidence interval for $\theta$ is

$$[\hat{\theta} - \hat{z}_{1-\alpha/2} \hat{\text{SE}}, \hat{\theta} - \hat{z}_{\alpha/2} \hat{\text{SE}}],$$

where $\hat{z}_\tau$ is the sample $\tau$-quantile of the $B$ values of $\hat{Z}^{sb}$.

\[\square\]
Chapter 13

Bootstrap Extensions and Subsampling

This chapter describes alternatives to the empirical bootstrap. Some apply to the same settings, whereas others can work even when 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.

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 $W$ be a random vector of weights, independent of the data, with

$$W = (W_1, \ldots, W_n) \sim \text{Multinomial}(n; 1/n, \ldots, 1/n),$$

(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 $W$-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 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 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 since \( W_i = n - \sum_{i=1}^{n-1} W_i \), since the weights sum to n. Instead, 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 \( W_i / \sum_{i=1}^{n} W_i \) weight on observation i, since these probabilities sum to 1 by construction.

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 \( E_i \sim \text{Exp}(1) \) and \( W_i = E_i / \sum_{i=1}^{n} E_i \) makes W \( \sim \text{Dir}(1, 1, \ldots, 1) \), called the Bayesian bootstrap, although from this perspective it is a valid frequentist bootstrap; see Chapter 14.

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 \( N(\hat{\mu}, \hat{\sigma}^2) \). As you might guess, this is very rarely (if ever) 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 \( Y_i = X_i'\beta + \epsilon_i \). The X_i remain the same in the bootstrap world, but \( Y_i^* = X_i'\hat{\beta} + \hat{\epsilon}_i^* \), 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 \( \epsilon_i \), 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. There is also a bias-corrected and accelerated bootstrap, BCₐ, 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, \( \hat{\epsilon}_i^* = \hat{V}_i^* \hat{\epsilon}_i \), where \( \hat{V}_i^* \) is drawn randomly in each bootstrap replication. For example, one could use \( \hat{V}_i^* \overset{iid}{\sim} \mathcal{N}(0,1) \), or a discrete weight distribution with \( P(\hat{V}_i^* = -1) = P(\hat{V}_i^* = 1) = 1/2 \). 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 \( X_i \overset{iid}{\sim} \text{Unif}(0, \theta) \) for some \( \theta > 0 \) that represents the upper bound of the distribution, like \( \theta = \text{Q}_1(X) \). Let \( X_{n:n} \) denote the original sample maximum, and let \( X_{n:n}^{*b} \) denote the \( b \)th bootstrap sample maximum.

a) What’s \( P(X_{n:n} < \theta) \)?

b) What’s \( P(X_{n:n} = \theta) \)?

c) What’s \( P(X_{n:n}^{*b} < X_{n:n}) \)? Hint: all the \( X_i \) are unique wp1, and the \( X_i^* \) are resampled with replacement and independently, so you can use the fact that \( P(A \text{ and } B) = P(A) P(B) \) if \( A \perp B \).

d) What’s \( P(X_{n:n}^{*b} = X_{n:n}) \)? How does this bootstrap world probability compare with the real world probability in (b)?

e) Does the discrepancy disappear as \( n \to \infty \)? Hint: \( \lim_{n \to \infty} (1 + c n^{-1})^n = e^c \).

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.”
CHAPTER 13. BOOTSTRAP EXTENSIONS AND SUBSAMPLING

13.4 Subsampling

Much of the bootstrap’s appeal is the ability to compute a confidence interval even when we have no analytic expression for one. However, there are cases where the bootstrap does not perform as desired, even asymptotically.

Subsampling 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 behind subsampling is essentially the same as that behind bootstrap. That is, we can repeatedly sample from \( \hat{F}(\cdot) \), compute our estimator/statistic, and look at the resulting distribution. The difference is that instead of bootstrap samples of size \( n \) drawn with replacement, we take subsamples of size \( m < n \) without replacement.\(^1\)

13.4.1 Subsampling Consistency

**Assumption A13.1** (subsampling limit distribution). For some convergence rate \( \tau_n = n^\alpha \) for \( \alpha > 0 \), with iid sampling,

\[
J_n(\cdot) \equiv P\{\tau_n(\hat{\theta}_n - \theta) \leq \cdot\} \to J(\cdot)
\]

for some CDF \( J(\cdot) \), where population parameter \( \theta \) is estimated by \( \hat{\theta}_n \).

**Assumption A13.2** (subsampling subsample size). As \( n \to \infty, m \to \infty \) and \( m/n \to 0 \) (so \( \tau_m/\tau_n = (m/n)^\alpha \to 0 \)).

**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 \( X_{m,i} \) be the \( i \)th subsample of size \( m \), and let \( \hat{\theta}_{m,i} \) be the estimator calculated from \( X_{m,i} \). To estimate \( J_n(\cdot) \), the subsampling distribution is

\[
L_{m,n}(\cdot) \equiv \frac{1}{|Q_m|} \sum_{i=1}^{|Q_m|} 1\{\tau_m(\hat{\theta}_{m,i} - \hat{\theta}_n) \leq \cdot\}.
\]

In practice, if \( |Q_m| \) is prohibitively large, some number of subsamples may be selected at random in a Monte Carlo approach similar to that for bootstrap. Alternatively, certain subsamples may be selected intentionally, e.g., if there is time series data and consecutive blocks are drawn similar to the moving blocks bootstrap (Method 13.2).\(^2\)

\(^1\)There is also an \( m \)-out-of-\( n \) bootstrap that takes samples of size \( m < n \), with replacement, but obviously not a method with subsamples of size \( n \) without replacement.
13.4. SUBSAMPLING

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, assuming \( d \) is fixed as \( n \to \infty \) (i.e., not \( d \to \infty \)), both of these violate A13.2. That is, \( (n-1)/n \to 1 \) and \( (n-d)/n \to 1 \), violating the assumption that \( m/n \to 0 \).

**Theorem 13.1** (subsampling). Under A13.1 and A13.2, for all \( x \) that are points of continuity of \( J(\cdot) \), \( L_{m,n}(x) \overset{p}{\to} J(x) \).

**Proof.** The following gives the structure and some intermediate steps in the proof. Let

\[
U_{m,n}(x) = \frac{1}{|Q_m|} \sum_{i=1}^{|Q_m|} \mathbf{1}\{\tau_m(\hat{\theta}_{m,i} - \theta) \leq x\},
\]

\[
L_{m,n}(x) = \frac{1}{|Q_m|} \sum_{i=1}^{|Q_m|} \mathbf{1}\{\tau_m(\hat{\theta}_{m,i} - \theta) + \tau_m(\theta - \hat{\theta}_n) \leq x\}.
\]

From A13.1, \( \hat{\theta}_n - \theta = O_p(1/\tau_n) \), so

\[
\tau_m(\theta - \hat{\theta}_n) = O_p(\tau_m/\tau_n) = o_p(1) \overset{p}{\to} 0.
\]

Consequently, \( L_{m,n}(x) \approx U_{m,n}(x) \). Taking expectation, and noting that \( E[\mathbb{1}\{\cdot\}] = \mathbb{P}(\cdot) \),

\[
E[U_{m,n}(x)] = \frac{1}{|Q_m|} \sum_{i=1}^{|Q_m|} \mathbb{P}\{\tau_m(\hat{\theta}_{m,i} - \theta) \leq x\} = \mathbb{P}\{\tau_m(\hat{\theta}_{m,i} - \theta) \leq x\}
\]

since the \( \hat{\theta}_{m,i} \) have the same distribution for all \( i \) (by the iid assumption). By A13.1, simply replacing \( n \) with \( m \),

\[
\mathbb{P}\{\tau_m(\hat{\theta}_m - \theta) \leq x\} \to J(x)
\]

as \( m \to \infty \), which it does by A13.2. To complete the proof, it can be shown that \( \text{Var}(U_{m,n}(x)) \to 0 \) as \( m/n \to 0 \) (which it does by A13.2) using Hoeffding’s Inequality for \( U \)-statistics.

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 \( \tau_n \neq \tau_m \). We want the standard error of the original sample’s estimator, \( \hat{\theta}_n \).

For simplicity, assume \( \tau_n = \sqrt{n} \), so \( \tau_m = \sqrt{m} \). Since \( \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} W \), defining \( W \) as a random variable with \( W \sim J(\cdot) \), then

\[
\text{Var}(\sqrt{n}(\hat{\theta}_n - \theta)) = \text{Var}(W) \implies \text{Var}(\hat{\theta}_n) = \frac{\text{Var}(W)}{n}.
\]

Now, \( \sqrt{m}(\hat{\theta}_m - \theta) \) has the same limit \( W \), which means

\[
\text{Var}(\sqrt{m}(\hat{\theta}_m - \theta)) = \text{Var}(W) \implies \text{Var}(\hat{\theta}_m) = \text{Var}(W)/m = \frac{\text{Var}(W)/n}{n/m}.
\]

Thus, if we can estimate \( \text{Var}(\hat{\theta}_m) \) from subsampling, the standard error estimate for \( \hat{\theta}_n \) should be

\[
\text{SE}(\hat{\theta}_n) = \sqrt{\text{Var}(\hat{\theta}_n)} = \sqrt{(m/n) \text{Var}(\hat{\theta}_m)} = \sqrt{m/n} \text{SE}(\hat{\theta}_m).
\]

That is, take the standard deviation of your subsampled \( \hat{\theta}_{m,i} \) (i.e., your estimated \( \text{SE}(\hat{\theta}_m) \)) 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 \( Y_i \) 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?
13.6. TIME SERIES 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.,

\[
(Y_1^*, \ldots, Y_T^*) = (Y_4, Y_5, \ldots, Y_{10}, Y_{35}, Y_{36}, \ldots, Y_{41}, Y_{20}, Y_{21}, \ldots, Y_{26}, Y_{21}, Y_{22}, \ldots, Y_{27}, Y_2, Y_3, \ldots, Y_8, Y_{11}, Y_{12}, \ldots, Y_{17}).
\]

(13.2)

**Method 13.2** (moving blocks bootstrap). Let \(T\) denote sample size. Given block length \(\ell\), the moving blocks bootstrap samples \(k = T/\ell\) blocks of length \(\ell\) 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

\[
(X_{I_1}, X_{I_1+1}, \ldots, X_{I_1+\ell-1}, X_{I_2}, \ldots, X_{I_2+\ell-1}, \ldots, X_{I_k}, \ldots, X_{I_k+\ell-1}).
\]

**Discussion Question 13.5** (bootstrap block length). Consider the choice of \(\ell\) for Method 13.2.

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 \(\ell\) 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 \(\ell\) is not too small or too big. The block length \(\ell\) 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. Please let me know of any good references to optimal selection of \(\ell\).

**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 1994]. The main difference is that block length is no longer a constant \(\ell\) but randomly drawn (from
13.7. BIAS CORRECTION

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 (1994, p. 1304) with tuning parameter value $p = n^{-1/3}$ (that determines the distribution of block length), the rate for $p$ that they suggest on page 1306.

13.7 Bias Correction

The bootstrap may also be used to estimate (and correct for) the bias of an estimator. When models get more complicated, bias may be unavoidable. For example, in nonlinear dynamic panel fixed effects models with fixed $T$, the incidental parameters problem of having $N$ fixed effects $\alpha_i$ (that cannot be consistently estimated with only $T$ observations per $i$) causes $\hat{\theta}$ to be biased.

Method [12,1] and other bootstraps estimate the full sampling distribution of $\hat{\theta} - \theta$, so it can be used to estimate the bias. That is, if we can estimate the full distribution of $\hat{\theta} - \theta$, then we can look specifically at its mean, $E(\hat{\theta} - \theta)$, which is the bias of $\hat{\theta}$.

Notationally, now let $R$ denote the number of bootstrap replications, so we can use $B$ for the bias, $B \equiv \text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$. The estimated bias is $\hat{B}$.

Computationally, a bootstrap gives us $R$ draws of $\hat{\theta}^* - \hat{\theta}$ for $r = 1, \ldots, R$. Thus, the mean is approximated by

$$\hat{B} = \text{Bias}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^* - \hat{\theta}).$$

(13.3)

By the weak law of large numbers (WLLN), as $R \to \infty$, the sample average of $R$ iid random variables converges to the random variables’ mean. That is, as $R \to \infty$, the bias estimator converges to the bootstrap-world mean. The bootstrap-world mean generally differs from the real-world mean. However, the hope is that the bootstrap world and real world are very similar with large $n$, enough that this difference is small.

The bias-corrected estimator simply subtracts the estimated bias from the original estimator. The bias-corrected estimator is

$$\tilde{\theta} = \hat{\theta} - \hat{B}.$$ 

(13.4)

Lower bias sounds good, but there is usually a price to pay: larger variance. Consider estimator $\hat{\theta}$ with true bias $B = E(\hat{\theta}) - \theta$ and bias estimator $\hat{B}$. (The following is not specific to bootstrap bias estimators.) Assume $\hat{B}$ is itself unbiased for $B$, i.e., $E(\hat{B}) = B$. Then the bias-corrected estimator $\tilde{\theta} = \hat{\theta} - \hat{B}$ is unbiased:

$$E(\tilde{\theta}) = E(\hat{\theta} - \hat{B}) = E(\hat{\theta}) - E(\hat{B}) = (\theta + B) - B = \theta.$$ 

(13.5)

However, the variance is usually larger. For simplicity, let $\hat{\theta} \perp \perp \hat{B}$. Then

$$\text{Var}(\tilde{\theta}) = \text{Var}(\hat{\theta} - \hat{B}) = \text{Var}(\hat{\theta}) + \text{Var}(\hat{B}).$$

(13.6)
CHAPTER 13. BOOTSTRAP EXTENSIONS AND SUBSAMPLING

The MSE of the bias-corrected estimator is thus
\[
\text{MSE}(\tilde{\theta}) = [\text{Bias}(\tilde{\theta})]^2 + \text{Var}(\tilde{\theta}) = [0]^2 + \text{Var}(\hat{\theta}) + \text{Var}(\hat{B}).
\] (13.7)

The original estimator’s MSE is \(\text{MSE}(\hat{\theta}) = B^2 + \text{Var}(\hat{\theta})\), so the MSE difference is
\[
\text{MSE}(\tilde{\theta}) - \text{MSE}(\hat{\theta}) = [\text{Var}(\tilde{\theta}) + \text{Var}(\hat{B})] - [B^2 + \text{Var}(\hat{\theta})] = \text{Var}(\hat{B}) - B^2.
\] (13.8)

If \(B^2 > \text{Var}(\hat{B})\), then the bias-corrected estimator is better (smaller MSE). But if \(B^2 < \text{Var}(\hat{B})\), then the bias-corrected estimator is worse (larger MSE). Thus, in general, the effect of bias correction on MSE is ambiguous.

However, confidence intervals are “more sensitive” to bias than point estimators, so sometimes bias correction is useful for confidence intervals even if not for estimation. (Although it seems usually bias correction is not optimal in such settings; see papers by Tim Armstrong.)

For an example of bootstrap bias correction, see [Kim and Sun (2016), “Bootstrap and \(k\)-step bootstrap bias corrections for the fixed effects estimator in nonlinear panel data models.”]

13.8 Choice of \(B\)

In one sense, the number of bootstrap replications \(B\) is like a smoothing parameter: we must choose it, and its value affects performance. However, unlike with smoothing parameters, larger \(B\) is always better theoretically. The goal of computationally approximating the sampling distribution under the EDF is achieved as \(B \to \infty\). The tradeoff is not bias vs. variance, but rather accuracy vs. computation time.

How big must \(B\) be for “good” accuracy? It depends on the data, estimator, definition of “good,” and what we want. For example, if we want a two-sided 99% CI using the root method bootstrap, then \(B = 20\) is not big enough; imagine taking the sample 0.005-quantile with only \(B = 20\) observations.

In statistical software, the default value for \(B\) is sometimes too small. For example, this was (and maybe still is?) the case with quantile regression in Stata for a long time \((B = 20)\).

For a long-term research project, I suggest the following general approach. At first, while still exploring the data, use a relatively small value of \(B\) to save yourself time. Or if it’s fast to compute even with large \(B\), use a large \(B\). Later, for your final results, use a very large value of \(B\). You may need to run the bootstrap overnight, but if this is only for the very final result, then it is not too inconvenient.

There are some papers dealing with the optimal number of bootstrap replications. See especially [Andrews and Buchinsky (2000) [2001] [2002], and Davidson and MacKinnon (2000)]. There are also other suggestions coming from other fields (e.g., computer science), some in the form of optimal stopping criteria: keep drawing bootstrap samples until some condition is met.
13.9 Bootstrap-Based Model Selection

For an example of bootstrap-based model selection for IV quantile regression estimation, see Xin Liu’s 2019 working paper, “Averaging estimation for instrumental variables quantile regression.”

There exist model selection procedures based on the bootstrap. If I recall, Shao and Tu (1995) discuss some of these, and Jun Shao has written papers on model selection (including some with bootstrap). One motivation for cross-validation is “testing” an estimator using observations not in the “training” data, and bootstrap resampling lets us sample entire new datasets from the EDF. (Well, ok, I don’t know much about bootstrap model selection, but I’m guessing the idea is similar?) See also Efron and Tibshirani (1993, §17.6–18).
Chapter 13. Bootstrap Extensions and Subsampling

Exercises

Exercise E13.1. a. Find a published paper with replication materials (i.e., data and code) available.

b. Replicate one standard error estimate 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. Make sure they are appropriate to 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.

d. Qualitatively discuss your results (compared to each other and to the original paper’s result).

Exercise E13.2. a. Find a paper published in a respectable economics journal\footnote{E.g., in top 500 of \url{https://ideas.repec.org/top/top.journals.all.html}} that uses a bootstrap to get a CI for their main parameter of interest; provide a link to the paper. 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.
Chapter 14
Bayesian Bootstrap

This chapter is the long answer to, “What’s the Bayesian bootstrap?” Although the Bayesian bootstrap has a frequentist interpretation (as a special case of “exchangeable weights” bootstrap), I focus on the Bayesian interpretation.

Sampling is assumed iid throughout (unless otherwise noted).

14.1 Bayesian Basics

Generally, the Bayesian approach helps us update our beliefs based on observed data, using our model. The prior is our beliefs about parameters before seeing the data. The likelihood is the model of how data is generated depending on the parameters. The posterior is 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, Data, and Model

The Bayesian approach concerns our beliefs about unknown parameters. A belief about a parameter’s value is quantified as a probability distribution. Let $\beta$ be the parameter of interest. We think there is some true value of $\beta$, but we don’t know what it is. Let random variable $B$ represent our beliefs about the fixed parameter $\beta$. Maybe we believe
there’s a 50% chance that \( \beta \) is negative, and a 75% chance that \( \beta \) is below 10. These beliefs can be written as \( P(B \leq 0) = 0.5 \) and \( P(B \leq 10) = 0.75 \). In principle, we could keep asking ourselves what we believe to be the chance that \( \beta \) is below \( b \), for more and more \( b \in \mathbb{R} \). This would give us \( P(B \leq b) \), i.e., the CDF of \( B \) evaluated at \( b \), for many \( b \).

If we did this for all \( b \in \mathbb{R} \), then we’d fully describe the CDF of \( B \), i.e., we’d have fully quantified our belief about \( \beta \). Or, we could take a shortcut and say \( B \sim N(\mu, \sigma^2) \) and just pick \((\mu, \sigma^2)\) to best match our beliefs. This process of quantifying real-world beliefs is called prior elicitation.

To confuse you, instead of writing \( B \sim N(\mu, \sigma^2) \) to describe our beliefs, we write \( \beta \sim N(\mu, \sigma^2) \). This looks like we don’t believe in a single “true” value of \( \beta \) (which indeed is one less-common interpretation). However, it is just a notational difference: in Bayesian analysis, the “parameter” actually represents our beliefs about the parameter, which are naturally expressed as random variables.

Another main difference with the frequentist approach that is not merely notational is how the data are treated. In the frequentist framework, observations like \( Y_1 \) 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 \( Y_i \).

However, the Bayesian framework still considers how the data is generated from the population distribution. There is some model of the data-generating process and how observations are generated depending on the population parameters.

### 14.1.2 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”):

\[
P(B \mid A) = \frac{P(B) P(A \mid B)}{P(A)}. \tag{14.1}
\]

There’s a version with PDFs that we’ll see later. Abstracting somewhat, think of \( A \) like data, and \( B \) the parameter (or, the parameter being in some range of values). The LHS is like the posterior: what do we believe about the parameter conditional on the data we
14.1. BAYESIAN BASICS

saw? The RHS says this equals our prior $P(B)$ times the likelihood $P(A \mid B)$, normalized by something that doesn’t depend on the parameter, $P(A)$.

**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 $P(X = 1 \mid \theta = 0) = \alpha = 5\%$. 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 $P(\theta = 1)$ matter here? Hint: make a table of the joint probability distribution of $(X, \theta)$, and see the effect of changing various values.

The cookie example at [https://stats.stackexchange.com/a/2287](https://stats.stackexchange.com/a/2287) is also very insightful. In the future, I’ll try to explain it here and add another discussion.

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 “objective”?
  - · But the frequentist estimate is equivalent to a Bayesian estimate with a certain prior...
  - 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.
CHAPTER 14. BAYESIAN BOOTSTRAP

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, 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 are subject to misuse (whether intentional or not).

14.2 Beta–Binomial Model

The first step toward understanding the Bayesian bootstrap as nonparametric Bayesian inference is to understand the beta–binomial model.

The notation in this section is not all conventional but hopefully helpful for getting started. Here, the true population parameter is the non-random \( \theta \), while beliefs about \( \theta \) are represented by random variable \( P \). Further, \( p \) is a dummy variable (in the calculus sense, not the econometrics sense), representing any possible value of \( \theta \) 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, \( \theta \) might be used for \( \theta, P, \) 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 \( \theta \) (or whatever variable) refers to the true \( \theta \), 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 \( Y \in \{0, 1\}, P(Y = 1) = \theta, \) so \( P(Y = 0) = 1 - \theta, \) where \( \theta \) is the unknown (non-random) population parameter of interest. That is, \( Y \) has a **Bernoulli distribution** with parameter \( \theta \).

If the likelihood is Bernoulli, why is it called “binomial”? Define

\[
N_1 \equiv \sum_{i=1}^{n} 1\{Y_i = 1\}, \quad N_0 \equiv \sum_{i=1}^{n} 1\{Y_i = 0\} = n - N_1.
\]  

(14.2)
14.2. BETA–BINOMIAL MODEL

Given iid Bernoulli $Y_i$, the sampling distribution of $N_1$ given $p$ is $N_1 \sim \text{Binomial}(n, \theta)$, a binomial distribution with parameters $n$ and $\theta$ (i.e., how many “successes” out of $n$, where the “success” probability is $\theta$). 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:

$$n_1 \equiv \sum_{i=1}^{n} 1\{y_i = 1\}, \quad n_0 \equiv \sum_{i=1}^{n} 1\{y_i = 0\} = n - n_1.$$  \hfill (14.3)

Prior

Second, the prior: how can we quantify our beliefs about $\theta$? We know $0 \leq \theta \leq 1$, so we should use a distribution with support on $[0, 1]$. One such option (that turns out to be very convenient) is the beta distribution. Write the prior as

$$P \sim \text{Beta}(a, b).$$  \hfill (14.4)

The beta distribution is restrictive. For example, it cannot put positive probability mass on any particular value (except in degenerate cases where 0 or 1 has 100% probability). Nor can we have a bimodal belief with modes at 0.25 and 0.75. But, it is often a reasonable approximation of our belief.

Posterior

Third, the posterior: given the prior and likelihood, what do we believe about $p$ after seeing the data? Magically, the posterior is also a beta distribution. This property of the posterior being in the same distributional family as the prior is called conjugacy. Specifically, the posterior is

$$P \mid y \sim \text{Beta}(a + n_1, b + n_0),$$  \hfill (14.5)

where $y = (y_1, \ldots, y_n)'$ is the observed data, $(a, b)$ are from the prior in \hfill (14.4), and $(n_1, n_0)$ are from \hfill (14.3). That is, given our prior belief $p \sim \text{Beta}(a, b)$ and the iid Bernoulli sampling (likelihood), our belief about $p$ after seeing the data $y$ 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 $P \sim \text{Beta}(a, b)$ could be thought of as having previously seen data with $a$ observations of $y_i = 1$ and $b$ observations of $y_i = 0$.

The more observations we have seen previously, the “stronger” our prior belief is (more concentrated). This is indeed true of a beta distribution as $(a, b)$ are increased. Similarly, the data have less effect on the posterior when the prior is stronger, since $(n_0, n_1)$ are small compared with $(a, b)$.

Conversely, if $a = b = 0$, then the posterior seems entirely driven by the data. This makes such a prior a candidate for being an “objective” prior. However, there are other candidates and considerations (that we will not discuss).
CHAPTER 14. BAYESIAN BOOTSTRAP

Posterior Mean

The **posterior mean** is an important feature of the posterior. It is often reported as the “point estimate.” Assuming quadratic loss, the posterior mean minimizes posterior expected loss. That is, with quadratic loss function $L^2(\cdot)$,

$E(P \mid y) = \arg \min_{g \in [0,1]} E[(P - g)^2 \mid y]$, \hspace{1cm} (14.6)

where the (conditional) expectation is wrt the posterior of $P$ (given data $y$). Thus, in a Bayesian sense, the posterior mean is our “best guess” of $\theta$ under quadratic loss. Under other loss functions, it may be optimal to report the median or other quantiles of the posterior, or the posterior mode.

The mean of the posterior in (14.5) is easily computed. The mean of a Beta($\alpha, \beta$) distribution is $\alpha/\alpha + \beta$. Thus, the posterior mean is

$E(P \mid y) = \frac{a + n_1}{a + b + n}$. \hspace{1cm} (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 (we have a strong prior) relative to $n$, then the posterior mean will not differ much from 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 $\theta = P(Y = 1)$ is $\hat{\theta} = N_1/n$.

a) What is the frequentist justification of $\hat{\theta}$ as a “good” estimator? (Hint: recall $P(Y = 1) = E[I\{Y = 1\}]$.)

b) Is there any prior (i.e., any $a$ and $b$) that makes the posterior mean equal the frequentist estimator, i.e., $E(P \mid y) = \hat{\theta}$?

**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 $E(P \mid y)$ above, below, or equal to $\hat{\theta}$?

b) If $a/(a + b) \in [0, 1]$ is fixed but $a$ and $b$ increase, then does $E(P \mid y)$ get closer to or farther from $\hat{\theta}$, or does it not change? Why?

c) Let $N_1 = 750$ and $n = 1000$, so $\theta = 0.75$. Is there any $(a, b)$ that makes $E(P \mid y) = 0.5$? What values/why?

**Discussion Question 14.4** (posterior mean consistency). Continue to consider the beta–binomial model with iid data. Consider the posterior mean as an estimator of the true population parameter $p$. Is it consistent? That is, given a fixed prior Beta($a, b$), as $n \to \infty$, does $E(P \mid Y) \overset{p}{\to} \theta$? (Note the uppercase $Y$ here; alternatively, one could ask about convergence for almost-all sequences $y_1, y_2, \ldots$, but I think this way will seem more familiar to you.)
14.2. BETA–BINOMIAL MODEL

14.2.2 Technical Details

This section shows the technical details for deriving the beta posterior in (14.5).

First, the prior PDF is \( \pi(\cdot) \). Sometimes \( p(\cdot) \) is used for the prior, but \( p \) below is the dummy variable for \( P \) or \( \theta \), so \( p(p) \) would be confusing. The PDF of the Beta\((a,b)\) prior distribution is

\[
\pi(p) = f_{a,b}(p) = \text{constant} \times p^{a-1}(1-p)^{b-1}.
\]  

(14.8)

The constant does not involve \( p \) and will not be necessary to compute the posterior.

Second, consider the likelihood function. For a single \( Y_i \), \( P(Y_i = 1) = \theta \), so the likelihood may be written

\[
\ell(y_i \mid \theta = p) = p^{1\{y_i=1\}}(1-p)^{1\{y_i=0\}}.
\]  

(14.9)

Although for maximum likelihood it is common to write likelihoods with reverse notation \( \ell(p \mid y) \), I write \( \ell(y \mid p) \) to emphasize the use of Bayes’ theorem, parallel to (14.1). With iid sampling, the likelihood for the full vector \( y = (y_1,\ldots,y_n)' \) is the product of the individual likelihoods,

\[
\ell(y \mid p) = \prod_{i=1}^{n} p^{1\{y_i=1\}}(1-p)^{1\{y_i=0\}} = p^{n_1}(1-p)^{n_0},
\]  

(14.10)

using \( n_1 \) and \( n_0 \) defined in (14.3).

Bayes’ theorem in (14.1) extends to PDFs. Generally, consider data \( W \) and parameter \( \theta \), with \( f_W(\cdot) \) the marginal PDF of \( W \), \( \pi(\cdot) \) the prior on \( \theta \), \( \ell(W \mid \theta = t) \) the likelihood, and \( \pi(\cdot \mid W) \) the posterior for \( \theta \). Then, letting \( t \) be some possible value of \( \theta \), and \( w \) a value of \( W \),

\[
\pi(t \mid w) = \frac{\pi(t)\ell(w \mid t)}{f_W(w)}.
\]  

(14.11)

Further, any PDF must integrate to 1. Thus, integrating the posterior (over \( t \)) must equal 1. Thus, we can ignore any constant terms (not depending on \( t \)) 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,

\[
\pi(t \mid w) \propto \pi(t)\ell(w \mid t).
\]  

(14.12)

In our beta–binomial example,

\[
\pi(p \mid y) = \frac{\pi(p)\ell(y \mid p)}{f_Y(y)} \propto \pi(p)\ell(y \mid p).
\]  

(14.13)

Determining the constant is straightforward. Let \( \tilde{f}(\cdot) \) be the unscaled PDF (without the proper constant). The constant must be \( 1/\int_{\mathbb{R}} \tilde{f}(t) \, dt \). This ensures

\[
\int_{\mathbb{R}} C\tilde{f}(t) \, dt = C\int_{\mathbb{R}} \tilde{f}(t) \, dt = 1.
\]  

(14.14)
This is true whether $\tilde{f}(\cdot)$ represents the posterior PDF or any PDF.

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 $\theta$, it integrates out $\theta$ according to the prior. That is, it integrates the numerator over $p$, as in (14.14). Thus, it does not contain any “new” information; it’s just some constant that makes the right-hand side of Bayes’ theorem integrate to 1, irrespective of the parameter of interest.

Other terms that do not depend on the parameter can also be removed to simplify further. The part of the PDF that depends on the parameter ($p$ in our example) is called the kernel. (This differs from the “kernel” for nonparametric smoothing.) From (14.11), generally, the kernel of the posterior equals the kernel of the prior times the kernel of the likelihood. Thus, the posterior is proportional to the prior kernel times the likelihood kernel (up to a constant).

In the beta–binomial model, the kernel approach is used as follows. From (14.8), the kernel of the beta prior is $p^{a-1}(1-p)^{b-1}$. From (14.10), 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

$$
\pi(p \mid y) \propto p^{a-1}(1-p)^{b-1} \times p^{n_1}(1-p)^{n_0} = p^{a+n_1-1}(1-p)^{b+n_0-1}.
$$

(14.15)

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 Beta$(a + n_1, b + n_0)$, as stated (without proof) in (14.5).

### 14.3 Conjugacy

The nice property in (14.15) 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 $\mathbf{P} = (P_1, \ldots, P_J)'$ follows a Dirichlet distribution, then the marginal distribution of each $P_j$ is a beta distribution. Details are on Wikipedia, for example.

14.4 Dirichlet–Multinomial Model

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.\footnote{https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions}

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 \( P(Y = 1) = \theta \) and \( P(Y = 0) = 1 - \theta \), let \( \theta_1 = P(Y = 0) \) and \( \theta_2 = P(Y = 1) \). More generally, \( Y = 0 \) and \( Y = 1 \) could be replaced by \( Y = v_1 \) and \( Y = v_2 \). Clearly \( \theta_1 + \theta_2 = 1 \) and \( \theta_1, \theta_2 \geq 0 \). Let vector \( \mathbf{P} = (P_1, P_2) \) describe our belief about \( \theta = (\theta_1, \theta_2) \). The prior can be written

\[
\mathbf{P} \sim \text{Dir}(a_1, a_2). \tag{14.16}
\]

This is the same as \( P_2 \sim \text{Beta}(a_2, a_1) \) and \( P_1 = 1 - P_2 \); yes, the parameter order is reversed, \( (a_2, a_1) \). The posterior is

\[
\mathbf{P} \mid \mathbf{y} \sim \text{Dir}(a_1 + n_0, a_2 + n + 1). \tag{14.17}
\]

From here, it is more clear how to generalize to \( J \) possible values of \( Y \). That is, let \( Y \in \{v_1, \ldots, v_J\} \). Let

\[
\theta = (\theta_1, \ldots, \theta_J)', \quad \theta_j \equiv P(Y = v_j), \ j = 1, \ldots, J, \tag{14.18}
\]

and continue to assume iid sampling. Let \( \mathbf{P} = (P_1, \ldots, P_J)' \) represent the belief about \( \theta \). The Dirichlet prior is

\[
\mathbf{P} \sim \text{Dir}(a_1, \ldots, a_J). \tag{14.19}
\]

In the observed data, define

\[
n_j \equiv \sum_{i=1}^{n} \mathbb{1}\{y_i = v_j\}. \tag{14.20}
\]

Then, the posterior is

\[
\mathbf{P} \mid \mathbf{y} \sim \text{Dir}(a_1 + n_1, \ldots, a_J + n_J). \tag{14.21}
\]

As for the beta–binomial model, the posterior can be derived formally using the prior kernel and likelihood. The Dirichlet prior’s kernel is

\[
\prod_{j=1}^{J} p_j^{a_j-1} = p_1^{a_1-1} \times \cdots \times p_J^{a_J-1}. \tag{14.22}
\]
CHAPTER 14. BAYESIAN BOOTSTRAP

The likelihood is
\[
\prod_{j=1}^{J} p_j^{n_j} = p_1^{n_1} \times \cdots \times p_J^{n_J}.
\] (14.23)

Thus,
\[
\pi(p \mid y) \propto \left( \prod_{j=1}^{J} p_j^{a_j-1} \right) \left( \prod_{j=1}^{J} p_j^{n_j} \right) = \prod_{j=1}^{J} p_j^{a_j+n_j-1},
\] (14.24)

which is another Dirichlet PDF kernel. Specifically, it corresponds to the posterior already stated in (14.21).

For the posterior mean, one can apply the formula for the mean of a Dirichlet distribution. Specifically,
\[
P \sim \text{Dir}(\alpha_1, \ldots, \alpha_J) \implies E(P) = (\alpha_1, \ldots, \alpha_J)/\sum_{j=1}^{J} \alpha_j.
\] (14.25)

Discussion Question 14.5 (Dirichlet posterior mean). Consider the notation, prior, posterior, and properties in (14.19)–(14.21) and (14.25).

a) For any \(j\), what’s the posterior mean of \(P_j\)? That is, what’s the \(j\)th component of (14.25) when you plug in the Dirichlet parameters from the posterior in (14.21)?

b) Which prior makes the posterior means all equal the corresponding frequentist estimators? That is, which \(a = (a_1, \ldots, a_J)\)' makes \(E(P_j \mid y) = n_j/n\) 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 Beta\((a, b)\) “prior” had \(a = b = 0\), i.e., a Beta\((0, 0)\) “distribution.” Even Wikipedia knows that beta distributions need \(a > 0\) and \(b > 0\); there is no such thing as a Beta\((0, 0)\) distribution. Similarly, Dir\((0, 0, \ldots, 0)\) is not a real distribution.

A technical interpretation of improper priors thus requires some nuance. An improper prior can be interpreted as the limit of a sequence of proper priors. In the beta case, consider using prior Beta\((a, a)\) with \(a \downarrow 0\). There is a corresponding sequence of Beta\((a + n_1, a + n_0)\) posteriors. As \(a \downarrow 0\), the posterior limit is Beta\((n_1, n_0)\). Similarly, the improper Dirichlet prior is the limit when taking a sequence of Dir\((a, a, \ldots, a)\) priors as \(a \downarrow 0\); the limit of the corresponding sequence of posteriors is Dir\((n_1, n_2, \ldots, n_J)\).
14.6. NONPARAMETRIC BAYES

There are other examples of improper priors. If \( P \sim N(\mu, \tau^2) \), then one could take \( \tau \to \infty \). If \( P \sim \text{Unif}(-a,a) \), then one could take \( a \to \infty \). Also, the Bayesian bootstrap can be seen as a particular nonparametric Bayesian approach with an improper prior.

14.6 Nonparametric Bayes

The most popular “nonparametric” (in the sense of not assuming that the distribution is known up to a finite-dimensional parameter vector) approach comes from [Ferguson (1973), 1974] and [Rubin (1981)]. 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 \( \theta \), this approach uses a Dirichlet process to describe a probability distribution over possible values of the population CDF. (A distribution of distributions.)

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 \( \mu = (\mu_1, \ldots, \mu_J) \) 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 \( \alpha = (a_1, \ldots, a_J) \) 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 \( \theta_j = P(Y = v_j) \) being high relative to the other \( \theta_j \), \( a(\cdot) \) 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

\[
P(\cdot) \sim \text{DP}(a(\cdot)),
\]

then for any partition with any \( J < \infty \), the finite-dimensional vector

\[
(P(B_1), \ldots, P(B_J)) \sim \text{Dir}(a(B_1), \ldots, a(B_J)).
\]

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 \( \delta_v(x) = 1\{x = v\} \), the posterior is

\[
P(\cdot) \mid y \sim \text{DP}(a(\cdot) + \sum_{i=1}^{n} \delta_{y_i}(\cdot)).
\]

It is especially easy to use an improper prior. This generalizes the improper Dirichlet prior that took the limit of priors \( \text{Dir}(a, a, \ldots, a) \) 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 (14.28), the posterior becomes $\text{DP}(\delta_{y_1}(\cdot) + \cdots + \delta_{y_n}(\cdot))$. Using (14.28), 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 $\theta = (\theta_1, \ldots, \theta_n)$ with $\theta_j = P(Y = y_j)$, and the posterior is

$$P \mid y \sim \text{Dir}(1, \ldots, 1).$$

(14.29)

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).

Also, a possibly unsettling consequence of this improper prior is that the posterior only has at most $n$ points of support, $\{y_i\}_{i=1}^n$. Even though in principle our belief $P(\cdot)$ can include continuous probability distributions (measures), the posterior only includes discrete distributions. That is, taken literally, it says we have 100% belief that the true distribution of $Y$ is discrete. 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. Not only are many distributions actually discrete (e.g., years of education), but some are discretely discrete. For example, variables like income, wage, and price come in discrete units of $\$0.01$. Technically, storing numbers in a computer makes everything discrete. The question is more "how discrete" than "discrete vs. continuous." Nonetheless, even if a variable is technically discrete, it’s possible that $n$ is much smaller than the number of points of support, and in some cases that may cause problems.

### 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 takes the posterior in (14.29). Below, properties of this posterior are explored, with particular attention to comparison with frequentist properties.

#### 14.7.1 Population Mean

Consider the posterior for the population mean, $\mu = E(Y)$. Let random variable $M$ represent our belief about the non-random value $\mu$. The posterior of $M$ follows from the posterior of $P$ in (14.29) because a specific value of $P = p$ uniquely determines the value $M = m$. This can be seen from the expected value formula for a discrete distribution. Specifically, recall that the posterior only includes discrete distributions on
values \(y_1, \ldots, y_n\), with respective probabilities \(P(Y = y_j) = P_j\). Thus,

\[
M = \mathbb{E}(Y \mid \theta = \mathbf{P}) = \sum_{i=1}^{n} y_i P_i = \mathbf{y}' \mathbf{P}. \tag{14.30}
\]

In practice, the distribution of \(M\) can be approximated by repeatedly drawing \(\mathbf{P}\) from its posterior and computing the corresponding value of \(M\).

The expression for \(M\) in (14.30) can be used to compute the posterior mean, \(\mathbb{E}(M \mid \mathbf{y})\). This can be interpreted as a Bayesian point estimate of the population mean \(\mu\). Using (14.30),

\[
\mathbb{E}(M \mid \mathbf{y}) = \mathbb{E}\left(\sum_{i=1}^{n} y_i P_i\right) = \sum_{i=1}^{n} y_i \mathbb{E}(P_i). \tag{14.31}
\]

The marginal distribution of each \(P_i\) is beta, which follows from the joint Dirichlet distribution. Specifically,

\[
(P_1, \ldots, P_J) \mid \mathbf{y} \sim \text{Dir}(1, \ldots, 1) \implies P_j \mid \mathbf{y} \sim \text{Beta}(1, n - 1), \ j = 1, \ldots, J. \tag{14.32}
\]

The mean of a Beta\((a, b)\) distribution is \(a/(a + b)\), so the mean of Beta\((1, n - 1)\) is \(1/n\). Thus, for all \(i = 1, \ldots, n\),

\[
\mathbb{E}(P_i \mid \mathbf{y}) = 1/n. \tag{14.33}
\]

Plugging this into (14.31),

\[
\mathbb{E}(M \mid \mathbf{y}) = \sum_{i=1}^{n} y_i (1/n) = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}, \tag{14.34}
\]

the sample mean. 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. (I find this reassuring, but you are free to feel disappointed.)

However, the Bayesian approach provides more than just a point estimate. It provides an entire posterior distribution describing our belief about \(\mu\), which is more useful than the asymptotic normal sampling distribution. For example, imagine we need to make a decision whose consequences depend on \(\mu\). 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 \(\mathbf{P} = (P_1, \ldots, P_J)'\) refer to the probabilities \(P(Y = v_j)\). Then,

\[
\mathbf{P} \mid \mathbf{y} \sim \text{Dir}(f_1, \ldots, f_J). \tag{14.35}
\]
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 (14.35) to (14.29).

**Discussion Question 14.6** (Bayesian bootstrap mean). Imagine we have a \( Y \) with a discrete distribution. You observe \( n = 8 \) values, \( y = (1, 1, 2, 2, 2, 3, 3, 4)' \). To approximate the posterior, we take random draws of \( P = (P_1, P_2, P_3, P_4)' \) from the appropriate \( \text{Dir}(2, 3, 2, 1) \) distribution. The marginals are
\[
\begin{align*}
P_1 | y & \sim \text{Beta}(2, n - 2), \\
P_2 | y & \sim \text{Beta}(3, n - 3), \\
P_3 | y & \sim \text{Beta}(2, n - 2), \\
P_4 | y & \sim \text{Beta}(1, n - 1).
\end{align*}
\]
(14.36)

a) What’s the mean of the posterior of \( M \), our belief about \( \mu \)?
b) Does the result in (14.34) showing \( \text{E}(M | y) = \bar{y} \) depend on \( Y \) having a continuous distribution and/or \( y \) having unique elements? If so, is there any (easy) way to “correct” this discrepancy?

### 14.7.2 Other Population Features

Not only \( \mu \), 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 2019](#) show higher-order frequentist accuracy of BB confidence intervals for unconditional population quantiles.

#### 14.7.3 Population CDF

**Discussion Question 14.7** (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 \( P(Y = 1) \), \( P(Y = 2) \), and \( P(Y = 3) \); we assume \( Y \) is discrete with only those possible values. The Bayesian bootstrap posterior here is \( P | y \sim \text{Dir}(1, 1, 1) \), as usual, with marginal distributions \( P_j | y \sim \text{Beta}(1, 2) \).

a) Why is \( P_j \) a random variable? E.g., where does the “randomness” come from / represent?
b) If we sampled a different dataset with different \( y_i \), how/would the meaning of \( P \) differ?

c) Draw the CDF of \( Y \) evaluated at point \( y \), \( F_Y(y) \), and label important values in terms of \( P(Y = 1) \), \( P(Y = 2) \), and \( P(Y = 3) \) (which sum to 1).
d) What is the CDF corresponding to a particular value of \( P = p \)?

**Discussion Question 14.8** (Bayesian bootstrap CDF 2). Continue from DQ 14.7

a) What’s the mean of our posterior belief about \( F_Y(y_1) \)? Hint: recall the mean of \( \text{Beta}(a, b) \) is \( a/(a + b) \).
b) What’s the mean of our posterior belief about \( F_Y(y_2) \)? Hint: see previous hint, and \( \text{E}(A + B) = \text{E}(A) + \text{E}(B) \).
14.7. BAYESIAN BOOTSTRAP

c) What’s the mean of our posterior belief about $F_Y(\cdot)$?

Discussion Question 14.9 (Bayesian bootstrap CDF 3). Continue from DQ 14.6 in particular (14.36).

a) What is the CDF corresponding to a particular value of $P$?

b) What’s the mean (over the posterior of $P$) of the CDF in (a)?

There are some equivalence results for Bayesian bootstrap and frequentist inference on the population CDF itself.

First, consider the posterior mean. Let $y_{n:1} \leq y_{n:2} \leq \cdots \leq y_{n:n}$ denote the sample order statistics. The cumulative probability at $y_{n:k}$ is $P_1 + \cdots + P_k$. The posterior mean is

$$E(P_1 + \cdots + P_k | y) = E(P_1 | y) + \cdots + E(P_k | y) = \sum_{j=1}^{k} E(P_j | y) = (k)(1/n) = k/n. \tag{14.37}$$

Moreover, for any point of evaluation $r \in \mathbb{R}$, the posterior mean of the cumulative probability is

$$\sum_{i=1}^{n} \frac{1}{n} \mathbbm{1}\{y_i \leq r\} = \frac{1}{n} \sum_{i=1}^{n} \mathbbm{1}\{y_i \leq r\} = \frac{1}{n} \sum_{i=1}^{n} \mathbbm{1}\{y_i \leq r\}. \tag{14.38}$$

This is the usual (frequentist) empirical CDF, $\hat{F}(r)$, as in Section 8.2.

Second, consider the posterior probability of the Dirichlet-based (frequentist) uniform confidence band in Section 8.5. That method was based on the joint frequentist sampling distribution of $\{F(Y_{n,k}) - F(Y_{n,k-1})\}_{i=1}^{n+1}$ (with $F(Y_{n,0}) \equiv 0$, $F(Y_{n,n+1}) \equiv 1$), which was also a Dirichlet distribution. However, that distribution had $n+1$ instead of $n$ (or, $n-1$ instead of $n-1$). For example, marginals of “spacings” are $[F(Y_{n,k}) - F(Y_{n,k-1})] \sim$ Beta($1, n$) rather than $P_k | y \sim$ Beta($1, n-1$). The biggest difference is manifest at the sample maximum: in the Bayesian bootstrap posterior, $F(y_{n:n}) = 1$ wp 1, whereas (of course) there is uncertainty about $F(Y_{n:n})$ in the frequentist sampling distribution.

The posterior and sampling distributions for $F(Y_{n:1}), \ldots, F(Y_{n:n})$ do match exactly when applying a “continuity correction” (smoothing) to the Bayesian bootstrap posterior, as done by Banks (1988). The interpretation is still different, but the $1 - \alpha$ uniform confidence band from Section 8.5 is also a valid $1 - \alpha$ uniform credible band, i.e., there is $1 - \alpha$ posterior probability of the CDF lying in the band. Unfortunately, the Banks (1988) correction makes it less of a true Bayesian method.

Even when there is a single band with the same level of uniform confidence (frequentist coverage probability) and credibility (Bayesian posterior probability), there may be very different conclusions in hypothesis testing. In the case of the CDF, perhaps the most common economic hypothesis is about stochastic dominance relationships. Kaplan and Zhuo (XXX) show how frequentist and Bayesian evaluations of stochastic dominance hypotheses may be very different. Zhuo (2017, Ch. 1) provides nonparametric Bayesian inference
on first-order and higher-order stochastic dominance, allowing for sampling weights and clustering.
EXERCISES

171

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 posterior’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.
CHAPTER 14. BAYESIAN BOOTSTRAP
Part V

Nonparametric Regression
Introduction

This part concerns nonparametric regression. Both kernel and sieve approaches are discussed. Model selection is a main focus. Regression discontinuity is discussed as a popular application of nonparametric regression.

At a high level, there are three steps for nonparametric regression, meaning flexible CEF estimation. 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”). 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.

I focus mostly on the case of a single regressor, and some on multiple regressors, but there are of course extensions. In particular, it is straightforward to compute 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.

Thanks to Yixiao Sun for his excellent graduate class on which some of this material is based; of course, all errors are mine.
Chapter 15

Nonparametric Methods: Preliminaries

Optional resources for this chapter

- Textbook: Kaplan (2020b, §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 $m(x) = E(Y | X = x)$. For description, the CEF describes a statistical relationship between $Y$ and $X$; i.e., it summarizes the joint distribution of $(Y, X)$. For prediction, the CEF provides the “best” (under quadratic loss) predictor of $Y$ given $X = x$. For 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 Chapter 6 and Section 10.6 of [Kaplan (2020b)] and Sections 2.5, 2.11, and 2.30 of [Hansen (2020a)].

Since those references don’t actually mention the average structural function (ASF), here’s a bit of detail. Consider structural function $Y = h(X, U)$, where $(Y, X)$ is observed and $U$ is unobserved. Similar to the (conditional) average structural effect, the ASF averages over the unobserved components. Specifically, the ASF first plugs in $X = x$ and then averages over the unconditional distribution of $U$. That is,

$$\text{ASF}(x) = E[h(x, U)].$$

(15.1)

In the special case of a scalar additive error, $h(X, U) = m(X) + U$. Then

$$E[h(x, U)] = E[m(x) + U] = m(x) + E(U).$$
Thus, the ASF is \( m(x) \), possibly shifted up/down if \( E(U) \neq 0 \). If \( E(U \mid X) = 0 \), then the ASF is identified: it equals the CEF. That is, the ASF is \( m(x) \) since \( E(U) = E[E(U \mid X)] = 0 \), and the CEF is

\[
E(Y \mid X = x) = E(m(x) + U \mid X = x) = m(x) + E(U \mid X = x) = m(x).
\]

For more on the ASF, see Blundell and Powell (2003).

The CEF is useful, but previously we only estimated an approximation of it. For example, Chapter 7 of Kaplan (2020b) explains how to interpret what we’re actually estimating with OLS: a linear projection, or “best” linear approximation of the CEF, or “best” linear predictor. 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) \( U \perp X \), so \( m(\cdot) \) is the CEF. Nonetheless, if we estimate the model \( Y = \beta_0 + \beta_1 X + U \), then our estimates may be very biased.

Here, instead of learning how to properly interpret estimates under misspecification, we’ll actually try 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

**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) an \( n \)th-degree polynomial (\( n \) is sample size).
15.2. SIMPLE EXAMPLES FOR INTUITION

Figure 15.2: An example of overfitting.

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 with \( \hat{\beta} = 2.4 \). In contrast, a ninth-degree polynomial fits all points exactly, yet looks unreasonable. Including the intercept, it has 10 parameters, and there are \( n = 10 \) observations, so it can be “solved” to go through each observation exactly, i.e., solved for \( (\beta_0, \ldots, \beta_9) \) to set \( Y = \sum_{k=0}^{9} X_i \beta_k \) for all \( i = 1, \ldots, n \). Despite such precision, the ninth-degree polynomial produces a totally unreasonable function, ranging far outside the plot window in many places. Even if the true CEF were not a perfectly straight line (perhaps even a ninth-degree polynomial), I can’t think of any economic example where it would be as crazy as the polynomial estimate. The ninth-degree polynomial here is “too flexible” and suffers from overfitting.

If we can’t even fit a ninth-degree polynomial in the case of Figure [15.2] then what is a “nonparametric model” really? Indeed, it’s not magic; we can’t actually estimate the true CEF perfectly in finite samples. It’s more about taking seriously the inevitable misspecification bias, and trying to balance this bias against overfitting. Trying to find such a “balanced” model is called model selection and is a critical component of all nonparametric methods.

How do we know if a nonparametric method works well? As usual, there are two types of evidence: theoretical and simulation. Theory can tell us how fast the estimation error shrinks to zero as \( n \to \infty \) (i.e., the convergence rate), given certain assumptions. If we have a “small” sample, though, this may not be reassuring, and it may be difficult to know what “small” means. A simulation can show us exactly how the method performs for a given sample size and given true model, but it can take a long time just to look at one very specific model and sample size, let alone all (infinite) possible true models and
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”).

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 nonparametric regression and model selection.
15.3 Terminology

What does “nonparametric” mean? What’s the precise definition of “parametric”? Currently, Wikipedia agrees with my understanding, but also agrees that these terms are not always used the same way by everyone, so always feel free to ask for clarification. I basically go by the definitions in Footnote 1 of Chen (2007).

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 if it is specified up to a finite number of parameters of interest, but at least one infinite-dimensional nuisance parameter. 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, consider the CEF \( m(x) = x'\beta \). Since \( \beta \) is a finite-dimensional vector, this is a parametric CEF. But consider the same model in error form: \( Y = X'\beta + U, \text{E}(U \mid X) = 0 \). Here, the conditional (on \( X \)) distribution of \( U \) is restricted to have mean zero, but otherwise it is unrestricted. If the conditional CDF \( F_{U\mid X}(\cdot) \) is

\[ \text{http://en.wikipedia.org/wiki/Parametric_model} \]
15.3. TERMINOLOGY

considered a nuisance parameter in the model, then the model is semiparametric since $F_{U|X}(\cdot)$ is a function (infinite-dimensional), but (presumably) only finite-dimensional $\beta$ is of interest. This “semiparametric regression model” differs from a parametric regression model with a stronger assumption like $U | X \sim N(0, \sigma_U^2)$. (Even then, the marginal distribution of $X$ is technically left unspecified, unless the $X_i$ are treated as fixed values.) So, depending on the setting, people may refer to specifying $x'\beta$ 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 $X$, 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 $m(x_1,x_2) = x'_1\beta + g(x_2)$, where only finite-dimensional $\beta$ 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 15. NONPARAMETRIC METHODS: PRELIMINARIES
Chapter 16

Local (Kernel) Regression

Optional resources for this chapter

- Textbook: Hansen (2020a), Chapter 19 (kernel/local nonparametric regression) and Chapter 23 (model selection).
- Textbook: Li and Racine (2007)
- 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, 2019) 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.
- 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.

This chapter introduces one of the two main approaches to nonparametric regression. Here, $X$ is scalar; for vector $X$, 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 $m(x) \equiv E(Y \mid X = x)$. 

185
16.1 Constant “Regressor”

To build intuition, consider a silly example: \( X = 1 \). There is no variation in the regressor \( X \); it is a constant.

Thus, the CEF is a single point: \( m(1) \). Moreover, \( m(1) = E(Y) \), the unconditional mean.

The unconditional mean can be estimated “nonparametrically” by the sample mean, \( \bar{Y} \). (I think this is actually “semiparametrically”: there is only one unknown parameter of interest, \( m(1) \); although the distribution of \( Y \) is unrestricted, i.e., there is an infinite-dimensional nuisance parameter, the parameter of interest is clearly finite-dimensional.) That is, \( \hat{m}(1) = \bar{Y} \). To match with later notation, this can also be written as

\[
\hat{m}(1) = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} 1 \{X_i = 1\}},
\]

(16.1)

where the denominator equals \( n \) since \( X_i = 1 \) for \( i = 1, \ldots, n \).

16.2 Binary Regressor

Let \( X \in \{0, 1\} \). The CEF consists of \( m(0) \) and \( m(1) \). Here, \( m(0) = E(Y \mid X = 0) \) is the mean \( Y \) value for the \( X = 0 \) subpopulation, and \( m(1) = E(Y \mid X = 1) \) is the mean \( Y \) in the \( X = 1 \) subpopulation.

16.2.1 Estimation

The subpopulation means are naturally estimated by the corresponding subsample averages. That is, to estimate the mean \( Y \) in the subpopulation with \( X = 0 \), take the average \( Y_i \) in the subsample with \( X_i = 0 \). Similarly, to estimate the mean \( Y \) in the subpopulation with \( X_i = 1 \), take the average \( Y_i \) in the subsample with \( X_i = 1 \). Extending (16.1), these subsample averages are

\[
\hat{m}(0) = \frac{\sum_{i=1}^{n} Y_i 1 \{X_i = 0\}}{\sum_{i=1}^{n} 1 \{X_i = 0\}}, \quad \hat{m}(1) = \frac{\sum_{i=1}^{n} Y_i 1 \{X_i = 1\}}{\sum_{i=1}^{n} 1 \{X_i = 1\}},
\]

(16.2)

The estimators in (16.2) are equivalent to regressing \( Y_i \) on \((1, X_i)\) and taking \( \hat{m}(0) = \hat{\beta}_0 \), \( \hat{m}(1) = \hat{\beta}_0 + \hat{\beta}_1 \). For example, see Section 6.3.4 (“Linear CEF”) of [Kaplan (2020b)](#208).

16.2.2 Consistency

To check consistency, first imagine stratified sampling with \( n/2 \) observations having \( X_i = 0 \) and \( n/2 \) having \( X_i = 1 \). That is, we have \( n/2 \) iid draws of \( Y_i \) from the subpopulation with \( X = 0 \), and another \( n/2 \) iid draws of \( Y_i \) from the subpopulation with \( X = 1 \). Thus, each estimator in (16.2) is the sample average of \( n/2 \) iid random variables, which
converges in probability to the corresponding expected value as \( n/2 \to \infty \) by the weak law of large numbers (WLLN). That is, as \( n \to \infty \), \( \hat{m}(0) \overset{P}{\to} m(0) \) and \( \hat{m}(1) \overset{P}{\to} m(1) \).

Consistency can be shown with (non-stratified) iid sampling of \( (Y_i, X_i) \), too. Dividing numerator and denominator in (16.2) by \( n \), then applying the WLLN to numerator and denominator separately, and combining with the continuous mapping theorem (CMT) yields the result. The denominator (divided by \( n \)) converges in probability to \( P(X = 0) \) or \( P(X = 1) \). The numerator is trickier than it looks but converges in probability to \( m(0) P(X = 0) \) or \( m(1) P(X = 1) \). The CMT applies as long as \( P(X = 0) > 0 \) and \( P(X = 1) > 0 \).

### 16.2.3 Bias–Variance Tradeoff

The following appears frivolous now, but it plants an important seed.

Imagine you compute \( \hat{m}(0) \) and \( \hat{m}(1) \) in (16.2), but you are unhappy with how big the standard errors are. Lacking ideas to improve the estimators themselves, better precision (lower standard errors) seems to require simply more data. You think: Aha! I am only using \( \sum_{i=1}^{n} 1 \{X_i = 0\} \) observations to estimate \( \hat{m}(0) \), even though I have \( n \) observations. Why waste observations? Let’s use all \( n \) to estimate \( \hat{m}(0) \), and use all \( n \) again to estimate \( \hat{m}(1) \).

This is possible if you assume \( m(0) = m(1) \). This assumption implies \( m(0) = m(1) = E(Y) \). Thus, you can use \( \hat{m}(0) = \hat{m}(1) = \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \).

As usual, there is a tradeoff. Your estimators’ standard errors will indeed be lower. However, if your assumption is wrong, then your estimators are biased even asymptotically (i.e., not consistent). Clearly \( \operatorname{plim}_{n \to \infty} \hat{m}(0) = E(Y) \), but if \( E(Y) \neq E(Y \mid X = 0) \), then the estimator is not consistent. Depending how much the variance decreases and how much the squared bias increases, mean squared error (MSE) could either decrease or increase. But it’s tough to guess how much bias there is since the bias depends on \( E(Y \mid X = 0) \), which is the unknown parameter we’re trying to estimate in the first place.

The tradeoff can also be framed in terms of model flexibility. If we use the more flexible model that allows \( m(0) \) and \( m(1) \) to differ arbitrarily, then the bias is lower but the variance is higher. If we use the less flexible model that imposes the structure \( m(0) = m(1) \) on \( m(\cdot) \), then variance is lower but bias is higher.

This tension is called the **bias–variance tradeoff**. By changing the model’s flexibility, we can trade off higher bias for lower variance, or vice-versa.

**Discussion Question 16.1** (CEF with binary regressor: tradeoff). Consider stratified sampling where the \( Y_i \) are sampled iid from the \( X = 0 \) subpopulation for \( i = 1, \ldots, 10 \) and iid from the \( X = 1 \) subpopulation for \( i = 11, \ldots, 20 \). Recall the variance of \( (W_1 + \cdots + W_m)/m \) is \( \operatorname{Var}(W)/m \), where the \( W_j \) are iid realizations of \( W \). Write \( m(x) = E(Y \mid X = x) \) for \( x = 0, 1 \). Assume the bias is \( m(1) - E(Y) = E(Y) - m(0) = b \geq 0 \). Assume \( \operatorname{Var}(Y \mid X = x) = \sigma^2 \) for both \( x = 0, 1 \). Consider the estimators in (16.2) along with \( \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \).

a) Explain why \( \operatorname{Var}(\hat{m}(0)) = \sigma^2/10 \).
b) Assuming (only for this part, for simplicity) that instead the \( n = 20 \) observations were all iid (not stratified), explain why \( \text{Var}(\bar{Y}) = \sigma^2/20 \). (Stratification would make the variance even smaller than this.)

c) Explain why \( \hat{m}(0) \) is unbiased, i.e., \( E[\hat{m}(0)] = m(0) \).

d) Using your results from the previous parts, what is the value of \( b \) such that \( \hat{m}(0) \) and \( \bar{Y} \) have the same MSE when taken as estimators of \( m(0) \)? Hint: recall MSE equals variance plus squared bias, so set the variance of \( \hat{m}(0) \) (whose bias is zero) equal to the variance of \( \bar{Y} \) plus its squared bias \( (b^2) \), and solve for \( b \) in terms of \( \sigma \).

e) Is \( \hat{m}(0) \) preferred to (i.e., has lower MSE than) \( \bar{Y} \) when \( b \) is smaller or larger than this indifference point? Why?

f) If \( n \) were much larger, would the indifference point increase, decrease, or stay the same? Why?

16.2.4 Binary Regressor: Small Probability

**Discussion Question 16.2** (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 \( N_1 \equiv \sum_{i=1}^n 1\{X_i = 1\} \).

Let \( p_x \equiv P(X = x) > 0 \) 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 \( E(N_1) \)?

e) If \( p_1 \) is fixed as \( n \to \infty \), then explain why it’s impossible to have \( E(N_1) \to 1 \) 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 \( E(N_1) \to 1 \) as \( n \to \infty \).

In DQ 16.2, \( N_1 \) 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 rest are completely ignored. This is essentially the tradeoff in Section 16.2.3: having a local (effective) sample size smaller than \( n \) can avoid large bias but at the expense of larger 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 \( P(X = x) > 0 \). Then,

\[
\hat{m}(x) = \frac{n^{-1} \sum_{i=1}^n Y_i 1\{X_i = x\}}{n^{-1} \sum_{i=1}^n 1\{X_i = x\}} \xrightarrow{p} \frac{m(x) P(X = x)}{P(X = x)} = m(x)
\] (16.3)
16.3. DISCRETE REGRESSOR

by the WLLN and CMT. (Again, the numerator is a little tricky, but I’m not sure the details are worth the opportunity cost.)

Again, (16.3) is just a subsample average. The “subsample” is the subset of the overall sample observations for which \( X_i = x \). The corresponding \( Y_i \) values are then averaged.

**Discussion Question 16.3** (discrete kernel regression). You observe \( n = 100 \) observations of wage and age, from a survey of 18- to 65-year-olds. Is

\[
\hat{m}(a) = \frac{\sum_{i=1}^{n} Y_i \mathbb{I}\{\text{age}_i = a\}}{\sum_{i=1}^{n} \mathbb{I}\{\text{age}_i = a\}}
\]

a good estimator? Why/not?

16.3.2 Local Sample Size

Consider the asymptotic performance of \( \hat{m}(x) \) in (16.3) when there are \( J \) possible values of \( X \). For simplicity, imagine stratified sampling with \( n/J \) observations in each stratum, i.e., \( n/J \) observations with \( X_i = x \) for each possible \( x \). Then \( \hat{m}(x) \) is the average of \( n/J \) iid random variables (the corresponding \( Y_i \)). Although \( n/J < n \), as \( n \to \infty \), so also \( n/J \to \infty \) since \( J \) is a fixed constant. So the convergence rate remains the same.

**Discussion Question 16.4** (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 \( N_x \equiv \sum_{i=1}^{n} \mathbb{I}\{X_i = x\} \). Let \( p_x \equiv P(X = x) > 0 \). Assume \((Y_i, X_i)\) are sampled iid. Hint: DQ 16.2 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 \( E(N_x) \)?

e) If \( p_x \) is fixed as \( n \to \infty \), then explain why \( \lim_{n \to \infty} E(N_x) = \infty \).

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 \( E(N_x) \to 1 \) as \( n \to \infty \).

The \( N_x \) in DQ 16.4 can be called a **local sample size** or **effective sample size**. That is, \( \hat{m}(x) \) is effectively only using \( N_x \) observations, even though there are \( n \) observations in the sample. If you imagine a scatterplot, \( N_x \) is the number of data points above the value \( x \) on the horizontal \( x \)-axis, i.e., the points “local” to (near) \( x \). (This “local” terminology may make more sense later with continuous \( X \).)

**Discussion Question 16.5** (local sample size rate). Consider the estimator \( \hat{m}(x) \) in (16.3). Let \( N_x \equiv \sum_{i=1}^{n} \mathbb{I}\{X_i = x\} \). Let \( p_x \equiv P(X = x) > 0 \). Assume \((Y_i, X_i)\) are sampled iid. 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 \( \lceil n/J_n \rceil \). (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 its limit?

Having \( J_n \to \infty \) in DQ 16.5 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 simply mathematical approximations to help us better understand finite-sample performance, which is ultimately what we care about. If \( J_n = J \), then in the asymptotic approximation with \( n \to \infty \), \( J_n/n \to 0 \), i.e., the number of possible values is very small compared to the number of observations. If your dataset has \( n = 100 \) and \( J = 88 \), \( J/n = 0.88 \) is not close to zero, so this is probably a bad approximation. Instead, an approximation with \( J_n \propto n \) may be better; then asymptotically \( J_n/n \to c \in (0,1] \). How do we know if \( J_n \propto n \) provides a better approximation for our finite-sample dataset than \( J_n \propto n^{4/5} \)? That’s more difficult to determine, as are many questions about how to apply asymptotic approximations in practice, like the classic question, “How big does \( n \) need to be?”

### 16.3.3 Bias–Variance Tradeoff

There is again a tradeoff like in Section 16.2.3 In the extreme, if we pool all the data, then the “local” sample size is back to \( n \), which reduces the variance, but at the expense of 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: \( \{1, 2\}, \{3, 4\}, \text{etc.} \). Modifying (16.3), for odd \( x \),

\[
\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i \in \{x, x+1\}\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i \in \{x, x+1\}\}} \to E(Y \mid X \in \{x, x+1\}). \tag{16.4}
\]

If \( m(x) \) and \( m(x+1) \) are very similar, then \( E(Y \mid X \in \{x, x+1\}) \approx E(Y \mid X = x) \), so the additional bias is small. If \( N_x + N_{x+1} \) is significantly bigger than \( N_x \), then the variance reduction is significant. For example, if \( N_x = N_{x+1} \), then the new (paired) local sample size is twice as big as before, so (if conditional variances are the same) the new variance is half as big.

**Discussion Question 16.6** (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 \( \text{Var}(Y \mid X = x) = \sigma^2 \), a constant (unrelated to \( x \)). Consider the estimator

\[
\hat{m}_h(1) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{1 \leq X_i \leq h\}}{\sum_{i=1}^{n} \mathbb{1}\{1 \leq X_i \leq h\}}.
\]
16.4. **CONTINUOUS REGRESSOR: INTRODUCTION**

Imagine $X^*$ is continuous, and interest is in $m^*(x^*) = \text{E}(Y \mid X^* = x^*)$.

### 16.4.1 Estimation with Discretization

The estimator in (16.3) cannot be used. Since $X^*$ is continuous, $P(X^* = x^*) = 0$ for any $x$, so the denominator $\text{plim}$ is zero. More directly, the denominator of $\hat{m}^*(x^*)$ is zero wp1
(with probability 1) since \( P(X^*_i = x^*) = 0 \). That is, with probability 1, zero observations have \( X^*_i = x^* \), so the local sample size is \( N_{x^*} = 0 \).

There are different ways to proceed, but I give only one example here for intuition, saving details for Section 16.5 and following sections.

One approach is to discretize \( X^* \) into a discrete random variable \( X \) and then apply Section 16.3. Let \( d(\cdot): \mathbb{R} \rightarrow \mathcal{X} \) be the discretization function that turns \( X^* \) values into \( X \) values, where \( \mathcal{X} \) is the discrete support of \( X \) (like \( \{1, \ldots, J\} \)). For example, if \( X^* \in \mathbb{R} \geq 0 \) is age in years (including decimal values), then one possible discretization is \( X = d(X^*) = \lfloor X^*/10 \rfloor \), which is usually how people (at least adults) answer the question, “How old are you?” If nobody lives to (say) 120 years old, then this discretization has \( J = 120 \) possible values, \( \mathcal{X} = \{0, 1, 2, \ldots, 119\} \). Alternatively, the discrete \( X \) could be measured in decades, like \( X = d(X^*) = \lfloor X^*/10 \rfloor \). Then there are \( J = 12 \) possible values, \( \mathcal{X} = \{0, 1, \ldots, 11\} \).

In this age example, the “years” and “decades” discretizations are nested. That is, there is a many-to-one (surjective) mapping from “years” values to “decades” values. Specifically, if \( x_y \) is years and \( x_d \) is decades, then \( x_d = \lfloor x_y/10 \rfloor \). Consequently, the local sample size for a particular \( x_d \) value equals the sum of local sample sizes for the corresponding \( x_y \) values.

To (approximately) estimate \( m^*(x^*) \), estimate \( m(d(x^*)) = E(Y \mid X = d(x^*)) \) using Section 16.3.

### 16.4.2 Bias–Variance Tradeoff

Discretization has essentially the same effect as using \( h > 1 \) in DQ 16.6. Discretizing with smaller \( J \) generally yields larger local sample sizes, meaning smaller variance of \( \hat{m}(d(x^*)) \). However, smaller \( J \) also generally means larger bias because more \( x^* \) with potentially different \( m^*(x^*) \) are pooled together. That is, there is a bias–variance tradeoff: pooling more observations together decreases variance but increases squared bias. Depending whether the decrease or increase is faster, it may decrease or increase MSE.

The main difference with DQ 16.6 is that now an unbiased estimator is generally impossible. It is impossible to only use the subsample with \( X^*_i = x^* \) because the local sample size is zero (wp1). Thus, we are forced to incur at least some bias.

### 16.4.3 Types of “Discretization”

There are three (or more?) different ways to “discretize” or pool observations. Each forms a branch within the local approach to nonparametric regression: the partitioning, kernel (local polynomial), and \( k \)-nearest neighbor approaches.

First, as above, a single discretization function \( d(\cdot) \) can be applied, and then for the discretized \( X \), \( E(Y \mid X = x) \) estimated for all \( x \in \mathcal{X} \). Such a function could be defined as placing the original \( x^* \) into different bins. In the age example, the decade bins are \( b_1 = [0, 10) \), \( b_2 = [10, 20) \), etc., up to \( b_{12} = [110, 120) \). If \( x^* \) is in bin \( b_j \), then \( d(x^*) = j \). Equivalently, this can be described as a partition of the support of \( X^* \), i.e.,
16.5. LOCAL CONSTANT REGRESSION

a set of mutually exclusive intervals whose union equals the original support. This is the foundation for partitioning estimators; e.g., see Cattaneo and Farrell (2013).

Second, for each \( x \), a bin centered at \( x \) could be defined. Instead of creating a single partition used for all \( x^* \), the bin definition depends on \( x^* \). For example, let \( h > 0 \) be the bin width, more generally called the bandwidth. For a given \( x^* \), the bin is defined as \([x^* - h/2, x^* + h/2]\). The estimator is then

\[
\hat{m}(x^*) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i \in [x^* - h/2, x^* + h/2]\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i \in [x^* - h/2, x^* + h/2]\}},
\]

(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 defining a bin around \( x^* \), the \( k \) closest observations can be used. That is, using \(|X_i - x^*|\) to define how “close” an observation is, the \( k \) observations with the smallest \(|X_i - x^*|\) are used to estimate \( \hat{m}(x^*) \). As before, the corresponding \( Y_i \) values can be averaged. The local sample size is always \( k \), but the bias may be larger if the \( X_i \) are more spread out. This is the core of the k-nearest neighbor (kNN) approach. The symmetric kNN estimator takes \( k/2 \) observations with \( X_i > x^* \) and \( k/2 \) with \( X_i < x^* \).

16.5 Local Constant Regression

For this section (and the rest of the chapter), let \( X \) denote a continuous random variable. As before, \( Y \) is the outcome (either discrete or continuous), \((Y_i, X_i)\) are sampled iid, and \( m(x) = \text{E}(Y \mid X = x) \).

Building from Section 16.4.3, consider the following local constant regression estimator. Given bandwidth \( h > 0 \) and value \( x \),

\[
\hat{m}_h(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}\{x - h/2 \leq X_i \leq x + h/2\}}{\sum_{i=1}^{n} \mathbb{1}\{x - h/2 \leq X_i \leq x + h/2\}}.
\]

(16.6)

That is, for the subsample of observations whose \( X_i \) is “close” to \( x \) (within \( h/2 \)), the corresponding \( Y_i \) values are averaged.

As you’ve seen, the bandwidth \( h \) affects both bias and variance. Larger \( h \) increases the local sample size, decreasing the variance. However, larger \( h \) also increases bias by including \( X_i \) farther from \( x \).

Discussion Question 16.9 (local constant local sample size). Let \( X_i \overset{iid}{\sim} \text{Unif}(0, 1) \). Let \( N_x = \sum_{i=1}^{n} \mathbb{1}\{x - h/2 \leq X_i \leq x + h/2\} \).

a) What’s \( \text{E}(N_x) \) if \( n = 100, x = 0.5, h = 0.4 \)?
b) What’s \( \text{E}(N_x) \) if \( n = 100, x = 0.5, h = 0.2 \)?
c) What’s \( \text{E}(N_x) \) as a function of \( n, x, \) and \( h \)?

The follow DQs consider models with no error terms to focus on understanding the bias of the local constant estimator.
Discussion Question 16.10 (local constant flexibility). Consider the local constant regression estimator in [16.6]. Let \( Y_i = \sin(X_i) \) with no error term, so \( m(x) = \sin(x) \), \( 0 \leq x \leq 2\pi \). 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 \( 0 < \hat{m}_h(x_1) < \sin(x_1), \ \hat{m}_h(x_2) = \sin(x_2), \) and \( 0 > \hat{m}_h(x_3) > \sin(x_3) \).

c) Let \( h = \pi \). For each \( j = 1, 2, 3 \), explain whether \( \hat{m}_h(x_j) \) 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 \( X_i \overset{iid}{\sim} \text{Unif}(0,2\pi) \)? Why/not?

f) Qualitatively, generally: does the estimator \( \hat{m}_h(x) \) become more or less flexible as \( h \downarrow 0 \)? Why?

Discussion Question 16.11 (local constant boundary). Consider the same setup of DQ [16.10] 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 [16.10] 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 16.12 (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 \( Y_i = 1 \{ x_0 - \epsilon \leq X_i \leq x_0 + \epsilon \} \) with no error term, so \( m(x) = 1 \{ x_0 - \epsilon \leq x \leq x_0 + \epsilon \} \) 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 \( \epsilon \), what’s the probability of sampling a dataset with \( \hat{m}_h(0.5) = 0 \)? Why: 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 \( \hat{m}_h(0.5) = 0 \) with \( h \geq 1 \), can using a smaller \( h \) help? Why/not?

d) Let \( h \geq 1 \). Given \( n \) and \( \epsilon \), 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?
16.5. LOCAL CONSTANT REGRESSION

Discussion Question 16.13 (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 16.14 (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.
CHAPTER 16. LOCAL (KERNEL) REGRESSION

Assumption A16.2 (finite variance). For all \( x \) in the support of \( X \), \( \text{Var}(Y \mid X = x) < \infty \). Slightly different: \( \mathbb{E}(|Y|^{2+\delta} \mid X) < \infty \) 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 \( m(x) = \mathbb{E}(Y \mid X = x) \). Let Assumptions A16.1–A16.5 hold. Then,

\[
\frac{\sqrt{nh}}{h} \left[ \hat{m}_h(x) - m(x) - h^2 B(x) \right] \overset{d}{\to} N(0, V(x)),
\]

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 16.12 and 16.13. If the CEF has a jump discontinuity, then the bias can be severe. If the CEF is 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 16.12 since \( \epsilon > 0 \), but the estimator is very bad with small \( n \). Additionally, DQ 16.12 could be modified so that \( m(\cdot) \) is technically twice continuously differentiable but looks basically the same as the discontinuous version (by letting the derivatives be very very large). Such a modification won’t affect the finite-sample performance, even though it technically satisfies A16.3.

The use of A16.4 was seen in DQ 16.11. 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 16.10 for \( x = \pi/2 \) and \( x = 3\pi/2 \). There, \( m''(\pi/2) > 0 \) caused positive bias, while \( m''(3\pi/2) < 0 \) caused negative 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. It’s fun to understand why that term arises, but not of first-order importance. Most important is to never use a local constant estimator because of this unnecessary bias.
16.5. LOCAL CONSTANT REGRESSION

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, if the asymptotic approximation is

$$\sqrt{nh}(\hat{m}_h(x) - m(x) - h^2B(x)) \sim N(0,V(x)),$$

then approximately

$$\hat{m}_h(x) - m(x) - h^2B(x) \sim N(0,V(x)/(nh)).$$

Rearranging further,

$$\hat{m}_h(x) \sim N(m(x) + h^2B(x),V(x)/(nh)).$$

That is, the sampling distribution of $\hat{m}_h(x)$ is approximately normal with mean $m(x) + h^2B(x)$ and variance $V(x)/(nh)$. (Technical note: the bias of the limiting distribution is not necessarily the limit of the bias, nor is the variance of the limiting distribution necessarily the limit of the variance; e.g., the IV regression estimator is asymptotically normal but in some cases has infinite variance for any $n$. One consequence is that the term “asymptotic bias” sometimes means the limit of the bias, but sometimes means the bias of the limit.)

From (16.8), the corresponding asymptotic MSE (AMSE) can be calculated. The variance is $V(x)/(nh)$, and the bias is $h^2B(x)$, so the AMSE (variance plus squared bias) is

$$AMSE_x(h) = V(x)/(nh) + h^4[B(x)]^2.$$

Using (16.9), given $V(x)$ and $B(x)$, there is some $h$ that minimizes AMSE. There is an interior solution: as $h \downarrow 0$, the variance grows to infinity, and if $h \to \infty$, the squared bias goes to infinity. Writing the AMSE-optimal bandwidth as $h_*$, the first-order condition (FOC) is

$$0 = \frac{\partial}{\partial h} AMSE_x(h) \bigg|_{h=h_*} = -h_*^{-2}V(x)/n + 4h_*^3[B(x)]^2,$$

$$h_*^{-2}V(x)/n = 4h_*^3[B(x)]^2,$$

$$h_*^{-5} = 4n[B(x)]^2/V(x),$$

$$h_* = n^{-1/5} \left( \frac{V(x)}{4[B(x)]^2} \right)^{1/5}. \quad (16.10)$$

Discussion Question 16.15 (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 16.8 was qualitatively similar.

a) $n$

b) $V(x)$

c) $B(x)$

Discussion Question 16.16 (local constant bandwidth 3). Equation (16.10) may help here. Explain both mathematically and intuitively.
CHAPTER 16. LOCAL (KERNEL) REGRESSION

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, since 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 (A)MSE-optimal bandwidth rate, like $h \propto n^{-1/5}$ in (16.10). “Undersmoothing” means smoothing less, which means smaller $h$. That is, $h \propto n^{-r}$ for $r > 1/5$ is undersmoothing here. “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 more smoothing (larger $h$) means less flexibility.

Discussion Question 16.17 (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 $\hat{m}_h(x) \pm 2n^{-1/5}$.

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?

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

Imagine you’re an undergraduate student again, and somebody asks you whether it’s better to run OLS with the regression model $Y = \gamma_0 + V$ or $Y = \beta_0 + \beta_1 X + U$. Your reasoning might be incorrect, but you’d probably recommend the second option.

Imagine you’re yourself again, and somebody asks you whether it’s better with a local sample to run OLS with only an intercept or to regress $Y$ on $(1, X)$. The second option is local linear regression. That is, it uses the same local sample as local constant regression, but then it runs OLS on the local sample and takes the OLS fitted value at $x$ to be the CEF estimate at $x$. Local linear regression turns out to be better than local constant regression.

Discussion Question 16.18 (local linear boundary bias). Consider $Y_i = X_i, X_i \sim Unif(0, 1)$. Hint: draw a picture.
16.6. LOCAL LINEAR REGRESSION

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 $\hat{m}_h(1)$?

You should always prefer local linear regression to local constant regression, although this may seem counterintuitive. Intuitively, we keep talking about the bias–variance tradeoff, how there’s a tension between the effects of model flexibility on bias and variance; so doesn’t the more flexible local linear estimator merely reduce bias at the expense of variance? However, theoretical results show the asymptotic variance is actually the same, even though the bias is indeed “smaller” (see below) for the local linear estimator.

Specifically, instead of the local constant bias of $h^2[m''(x) + 2m'(x)f_X'(x)/f_X(x)]/24$, the local linear bias is $h^2m''(x)/24$. This is actually not “smaller” in the sense of being always closer to zero; in certain lucky cases, the additional term in the local constant bias could actually cancel out the shared term, making the local constant bias smaller in absolute value. But, most people (including me) would prefer to have only one source of bias instead of two, so many people just say the local linear bias is smaller (without quotes). The local linear estimator also removes the design bias of the local constant estimator, i.e., the bias term that depends on $f_X(\cdot)$.

To write the local linear estimator, first write the local constant estimator as a least squares minimization:

$$\hat{\beta}_0(x) = \arg\min_{b_0} \sum_{i=1}^n \mathbb{1}\{x-h/2 \leq X_i \leq x+h/2\}(Y_i - b_0)^2, \quad \hat{m}_h(x) = \hat{\beta}_0(x).$$ \hspace{1cm} (16.11)

For local linear regression, the residual changes from $Y_i - b_0$ to $Y_i - b_0 - b_1X_i$. The regressor $X_i$ is centered at $x$ to become $X_i - x$, so that $\hat{\beta}_0$ remains the CEF estimator $\hat{m}_h(x)$:

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \arg\min_{(b_0, b_1)} \sum_{i=1}^n \mathbb{1}\{x-h/2 \leq X_i \leq x+h/2\}(Y_i - b_0 - b_1X_i)^2, \quad \hat{m}_h(x) = \hat{\beta}_0(x), \quad \hat{m}'_h(x) = \hat{\beta}_1(x).$$ \hspace{1cm} (16.12)

Sometimes $(X_i - x)/h$ is used instead of $X_i - x$, but it only affects the scaling of $\hat{\beta}_1$.

An estimator of the CEF slope $m'(x)$ is readily available. Specifically, using (16.12), $\hat{m}'_h(x) = \hat{\beta}_1$. This is another advantage over local constant estimation.
200  

CHAPTER 16. LOCAL (KERNEL) REGRESSION

16.7 Local Polynomial Regression

Why stop at linear? In principle, once we have a local sample, we could use any sort of regression, linear or nonlinear.

For local polynomial regression, odd-numbered polynomial degrees are always better for the same reason that local linear is better than local constant. That is, the degree $2p + 1$ local polynomial estimator has the same asymptotic variance as the degree $2p$ estimator, but "smaller" bias (one fewer term).

The polynomial degree helps determine the model flexibility, but the bandwidth could already control flexibility. Higher degree means more flexibility, so less bias and more variance; but smaller bandwidth has the same effect. Consequently, although sometimes local cubic regression is used, the local linear estimator is most commonly used.

16.8 Kernel Regression

The local regression approach can be extended further to local kernel regression, although this is not as important of an extension. Generally, the kernel function does not affect AMSE as much as the bandwidth choice or using local linear instead of local constant regression. If you want to skip this section, that’s fine; just use an “Epanechnikov kernel” when available.

16.8.1 Intuition

The principle of local regression is that observations with $X_i$ closer to $x$ have more information about $m(x)$ than observations with $X_i$ farther from $x$. Thus, the local sample includes observations with $|X_i - x| \leq h/2$ and excludes those farther than $h/2$.

Consider running weighted least squares (WLS) with the full sample of $n$ observations. Let $W_i$ be the weight given to observation $i$. Local regression sets $W_i = 1 \{ |X_i - x| \leq h/2 \}$. This is weird for two reasons. First, it treats all observations within $h/2$ equally. In the previous example, $X_i = 100$ and $X_i = 120$ are given equal weight, even though $X_i = 100$ is clearly more information about $m(100)$. Second, it drops off discontinuously at $h/2$. In the previous example, even though $X_i = 120$ and $X_i = 121$ are similarly far from $x = 100$, one observation has weight 1 while the other has weight 0.

The natural solution is to use weights that peak at $X_i = x$ and then decrease toward zero for $X_i$ farther from $x$. Such weighting functions are called kernel functions. (This is different that the “kernel” in Chapter 14.) Kernels maps $X_i - x$ to a weight. Their maximum is at $X_i - x = 0$, and they are zero (or at least very close to zero) for $X_i$ far from $x$. The kernel itself defines the shape of the weight mapping, and the bandwidth determines how fast the weight goes to zero as $X_i$ deviates from $x$. 
16.8. KERNEL REGRESSION

16.8.2 Local Linear Regression: Uniform Kernel

The local linear regression estimator from (16.12) can be written in terms of the uniform kernel. The uniform kernel is

\[ K(u) = \mathbb{1}\{-1/2 \leq u \leq 1/2\}. \]  

Let \( K(\cdot) \) be the uniform kernel function \( K(u) = \mathbb{1}\{-1/2 \leq u \leq 1/2\} \). Adding bandwidth \( h \), \( K(u/h) = \mathbb{1}\{-h/2 \leq u/h \leq h/2\} = \mathbb{1}\{-h/2 \leq u \leq h/2\} \). Then, (16.12) becomes

\[
(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \arg \min_{(b_0, b_1)} \sum_{i=1}^{n} \mathbb{1}\{x - h/2 \leq X_i \leq x + h/2\}(Y_i - b_0 - b_1 X_i)^2
\]

(16.14)

The CEF estimate is again \( \hat{m}_{K,h}(x) = \hat{\beta}_0(x) \). The uniform kernel \( K(\cdot) \) in (16.14) can be replaced by any other kernel, such as those below.

16.8.3 Other Second-Order Kernels

There are many, many possible kernels, but some are more commonly used. The following are all symmetric second-order kernels because with \( r = 2 \) they satisfy \( K(-u) = K(u) \) (symmetry) and

\[
1 = \int_{\mathbb{R}} K(u) \, du, \\
0 = \int_{\mathbb{R}} u^j K(u) \, du \text{ for all } j = 1, \ldots, r - 1, \\
0 < \mu_r \equiv \int_{\mathbb{R}} u^r K(u) \, du < \infty.
\]

(16.15)

Most second-order kernels also satisfy \( K(u) \geq 0 \) for all \( u \in \mathbb{R} \), so \( K(\cdot) \) is a PDF.

Here are a few others common kernels. The Epanechnikov kernel is the AMSE-optimal choice and makes you sound fancy:

\[
K(u) = \mathbb{1}\{|u| \leq 1\}(3/4)(1 - u^2).
\]

(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:

\[
K(u) = \mathbb{1}\{|u| \leq 1\}(1 - |u|).
\]

(16.17)

The Gaussian kernel is simply the standard normal (Gaussian) PDF:

\[
K(u) = (2\pi)^{-1/2} \exp\left(-u^2/2\right).
\]

(16.18)

The Gaussian kernel differs from the preceding kernels by having \( K(u) > 0 \) for all \( u \in \mathbb{R} \), but \( K(4) = 0.0001 \) and \( K(6) < 10^{-9} \), so the practical effect is negligible, other than ensuring the local sample is not empty.
16.8.4 Effect on AMSE

The local linear kernel regression estimator’s asymptotic bias and variance depend on the kernel as well as the bandwidth. Below, let \( \kappa_2 \equiv \int_{\mathbb{R}} [K(u)]^2 \, du \); \( \mu_2 \) is from (16.15). The AMSE is

\[
\text{AMSE}_x(h) = \frac{V(x)}{(nh)^2} + h^4 [B(x)]^2,
\]

\[
V(x) = \frac{\text{Var}(Y \mid X = x)}{f_X(x)^2} \kappa_2, \quad B(x) = (1/2)m''(x) \mu_2.
\] (16.19)

The AMSE is minimized by \( h_* \) solving

\[
0 = 4h^3 [B(x)]^2 - h_*^{-2} V(x)/n,
\]

\[
h_*^{-5} = 4n[B(x)]^2 / V(x),
\]

\[
h_* = n^{-1/5} \left( \frac{V(x)}{4[B(x)]^2} \right)^{1/5}.
\] (16.20)

Plugging the AMSE-optimal \( h_* \) in (16.20) back into the AMSE in (16.19), the best possible AMSE is

\[
\text{AMSE}_x(h_*) = \frac{V(x)}{(nh_*)^2} + h_*^4 [B(x)]^2
\]

\[
= \frac{[V(x)/n]n^{1/5}V(x)^{-1/5}4^{1/5}[B(x)]^{2/5}}{n^{-4/5}[V(x)]^{4/5}4^{-4/5}[B(x)]^{-8/5}[B(x)]^2}
\]

\[
= n^{-4/5}4^{1/5}[B(x)]^{2/5}[V(x)]^{4/5} + n^{-4/5}4^{-4/5}[B(x)]^{2/5}[V(x)]^{4/5}
\]

\[
= n^{-4/5}\{B(x)[V(x)]^2\}^{2/5}(4^{1/5} + 4^{-4/5})
\]

\[
= n^{-4/5}\{(1/2)m''(x)\mu_2\kappa_2^2[\text{Var}(Y \mid X = x)/f_X(x)^2]\}^{2/5}(4^{1/5} + 4^{-4/5})
\]

\[
= n^{-4/5}((\mu_2\kappa_2^2)^{2/5}C(x),
\]

where \( C(x) \) gathers the other terms. The main points are that the AMSE is proportional to \( n^{-4/5} \) and depends on the kernel \( K(\cdot) \) through \( \mu_2\kappa_2^2 \).

Consequently, the AMSE-optimal second-order kernel minimizes \( \mu_2\kappa_2^2 \) subject to \( K(u) \geq 0 \), \( \int_{\mathbb{R}} K(u) \, du = 1 \), \( K(-u) = K(u) \). This minimization problem is solved by the Epanechnikov kernel; e.g., see Pagan and Ullah (1999, p. 27).

Table 16.1 shows the \( \mu_2 \) and \( \kappa_2 \) of aforementioned kernels. The entries 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 all have \( \mu_2 = 1 \), then the AMSE (given \( h_* \)) ranking is the same as the \( \kappa_2 \) ranking. Table 16.1 shows that the Epanechnikov has the smallest \( \kappa_2 \) among the normalized \( \mu_2 = 1 \), as claimed earlier. Nonetheless, its \( \kappa_2 \) is not that much smaller (quantitatively) than the worst \( \kappa_2 \), that of the uniform kernel.
16.9. LINEAR SMOOTHER

Table 16.1: Various kernels’ $\mu_2$ and $\kappa_2$.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>$\mu_2$</th>
<th>$\kappa_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform</td>
<td>1/12</td>
<td>1</td>
</tr>
<tr>
<td>triangle</td>
<td>1/6</td>
<td>2/3</td>
</tr>
<tr>
<td>Epanechnikov</td>
<td>1/5</td>
<td>3/5</td>
</tr>
<tr>
<td>Gaussian</td>
<td>1</td>
<td>$(4\pi)^{-1/2} \approx 0.282$</td>
</tr>
<tr>
<td>uniform*</td>
<td>1</td>
<td>$1/(2\sqrt{3}) \approx 0.289$</td>
</tr>
<tr>
<td>triangle*</td>
<td>1</td>
<td>$\sqrt{2}/(3\sqrt{3}) \approx 0.272$</td>
</tr>
<tr>
<td>Epanechnikov*</td>
<td>1</td>
<td>$3/(5\sqrt{5}) \approx 0.268$</td>
</tr>
</tbody>
</table>

16.8.5 Higher-Order Kernels

Under stronger smoothness assumptions, the Epanechnikov kernel is outperformed by higher-order kernels that reduce bias. These higher-order kernels satisfy \((16.15)\) with $r > 2$. If $m(\cdot)$ is assumed to have (at least) $r$ derivatives, then an $r$th-order kernel can in theory use this to make the bias “smaller,” meaning proportional to a larger power of $h$. With $r = 2$, the bias is proportional to $h^2$. With $r = 4$, this drops to $h^4$, which is smaller than $h^2$ since $h \downarrow 0$. The trick is that higher-order $K(\cdot)$ cancel out higher-order terms in a Taylor expansion of $m(\cdot)$ around $x$. This comes from the property $0 = \int_{\mathbb{R}} u^j K(u) \, du$ in \((16.15)\). One interesting quirk of higher-order kernels is they require $K(u) < 0$ for some $u$, i.e., certain observations have negative weight in the locally weighted regression. As before, 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 can be an important issue since the promised small bias leads to a larger AMSE-optimal $h$. 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 $X$. They are so called because (for any $x$) the estimated $\hat{m}(x)$ is a linear combination of the $Y_i$. Perhaps surprisingly, many estimators are linear smoothers.

**Definition 16.1** (linear smoother). Estimator $\hat{m}(\tilde{x})$ is a linear smoother if it may be expressed as a linear combination of the $Y_i$ with linear combination weights $W_i(\tilde{x})$:

$$\hat{m}(\tilde{x}) = \sum_{i=1}^{n} W_i(\tilde{x})Y_i, \quad (16.21)$$

where $W_i(\tilde{x})$ may depend on $\tilde{x}$, $i$, and all the observed $(X_1, \ldots, X_n)$, but may not depend on any $Y_i$. 
CHAPTER 16. LOCAL (KERNEL) REGRESSION

Discussion Question 16.19 (\( \bar{Y} \) as a linear smoother). Show that the sample mean \( \bar{Y} \) is a linear smoother. That is, since there is no \( x \) here, determine the \( W_i \) such that \( \hat{Y} = \bar{Y} = \sum_{i=1}^{n} W_i Y_i \).

Discussion Question 16.20 (OLS as a linear smoother). Show that OLS is a linear smoother. That is, determine \( W_i(\tilde{x}) \) such that (16.21) is the OLS prediction. Hint: \( \hat{m}(\tilde{x}) = \tilde{x}' \hat{\beta} \), and write \( \hat{\beta} \) in summation notation; then move everything inside the final summation (that involves \( Y_i \)).

The local constant kernel regression estimator that extends (16.6) is readily seen to be a linear smoother. This estimator is also called the Nadaraya–Watson estimator [Nadaraya, 1964a; Watson, 1964]. Replacing the uniform kernel implicit in (16.6),

\[
\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K((X_i - x)/h)Y_i}{\sum_{i=1}^{n} K((X_i - x)/h)} = \sum_{i=1}^{n} W_i(x)Y_i, \quad W_i(x) = \frac{K((X_i - x)/h)}{\sum_{i=1}^{n} K((X_i - x)/h)}.
\]  

The local polynomial kernel regression estimator is just weighted OLS, so the linear smoother representation is similar to DQ 16.20. Let \( X = (1, X, X^2, \ldots, X^p)' \). Once \( \hat{\beta}(x) \) is estimated, the CEF estimate is \( \hat{\beta}_0(x) = (1, 0, \ldots, 0) \hat{\beta}(x) \). In turn, \( \hat{\beta}(x) \) is linear in \( Y_i \):

\[
\hat{\beta}(x) = \left( \sum_{i=1}^{n} K((X_i - x)/h)X_iX_i' \right)^{-1} \sum_{i=1}^{n} K((X_i - x)/h)X_iY_i.
\]  

(16.23)

Moving the \( (1, 0, \ldots, 0) \) and the \((\cdot)^{-1}\) inside the last summation yields the linear smoother representation. That is,

\[
\hat{m}_{K,h}(x) = \sum_{i=1}^{n} W_i(x)Y_i, \quad W_i(x) = (1, 0, \ldots, 0) \left( \sum_{i=1}^{n} K((X_i - x)/h)X_iX_i' \right)^{-1} K((X_i - x)/h)X_i.
\]  

(16.24)
Chapter 17

Series and Sieves

Optional resources for this chapter

- Textbook: Li and Racine (2007) Chapter 15
- Textbook: James et al. (2013) Chapter 7 (“Moving Beyond Linearity”), including §7.5 (“Smoothing Splines”)
- Original sieve paper: Grenander (1981)
- R: built-in package `splines` (R Core Team, 2019)
- R: built-in package `stats` (R Core Team, 2019) has functions like `smooth.spline`.

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 instrumental variables. Like before, model selection (Chapter 18) is critical for good performance in practice.

Although not mentioned here, the method of sieves can be applied to maximum likelihood, too.

Let \( m(\cdot) \) denote the CEF: \( m(x) = \text{E}(Y \mid X = x) \), where in this chapter \( X \) is scalar.

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, $E(Y \mid X = x) = E(Y)$, so estimating the CEF is equivalent to estimating the unconditional mean. The CEF model $m(x) = \beta_0$ can be estimate by OLS, yielding $\hat{\beta}_0 = \bar{Y}$, the sample average.

17.1.2 Binary Regressor

See Chapter 6 of Kaplan (2020b).

Let $X \in \{0, 1\}$. Estimating $m(x) = \beta_0 + \beta_1 x$ by OLS is consistent for the true CEF. Specifically, $\hat{\beta}_0 \xrightarrow{p} E(Y \mid X = 0)$ and $\hat{\beta}_1 \xrightarrow{p} E(Y \mid X = 1) - E(Y \mid X = 0)$, so $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \xrightarrow{p} m(x)$ for both $x = 0, 1$.

Another way to think of this is in terms of approximation. Generally, OLS estimates the population linear projection, which is the "best" linear approximation (BLA) of the CEF. If the BLA equals the CEF (no approximation error), then OLS estimates the true CEF. With $X \in \{0, 1\}$, there are only two points on $m(\cdot)$, so a straight line always fits. That is, there is no approximation error.

There could be approximation error if the model $m(x) = \beta_0$ is used instead. Then the OLS estimator is $\hat{m}(x) = \hat{\beta}_0 \xrightarrow{p} E(Y)$ for both $x = 0, 1$. If $m(0) \neq m(1)$, then this estimator is biased even asymptotically (i.e., not consistent).

However, if $m(0) = m(1) = E(Y)$, then this estimator is better. It has no bias, and it has lower variance since it uses all $n$ observations to estimate both $\hat{m}(0)$ and $\hat{m}(1)$. 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.

Even in the simple case of $X \in \{0, 1\}$, there is a tradeoff between bias (approximation error) and variance.

17.1.3 Trinary and More

See Chapter 7 of Kaplan (2020b).

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 “slopes” of the CEF, $m(2) - m(1)$ and $m(3) - m(2)$, also differ from the BLA slope.

You probably know how to fix this problem: add another regressor. There are multiple ways to achieve this. Basically anything works as long as the regressors are linearly independent (i.e., no perfect multicollinearity). There could be an intercept along with dummies $1\{X = 2\}$ and $1\{X = 3\}$. There could be no intercept and dummies for all three values. There could be an intercept along with $X$ and $X^2$. That is, since there are only three points on $m(\cdot)$, they can be perfectly approximated by a quadratic function of $X$. 
17.2 POLYNOMIAL SERIES

As before, though, there is a bias–variance tradeoff: including $X^2$ can both reduce bias (approximation error) and increase variance.

With $J$ possible values of $X$, a regression with $J$ coefficients could remove the approximation error completely, so the BLA is the CEF. For example, a degree $J-1$ polynomial would suffice.

If $J = n$ (or more), then this strategy breaks down. A degree $n-1$ polynomial can perfectly fit all $n$ observations in the data. This clearly suffers from overfitting. The approximation error may be minimized, but the variance is huge.

If $J = n$, then 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. Maybe we should not use the polynomial structure, but another flexible structure.

17.2 Polynomial Series

Nonparametric regression using polynomial terms is not recommended in practice, but it’s familiar and helps intuition.

Similar to Section 17.1.3, consider approximating the CEF by a degree $J - 1$ polynomial,

$$J - 1 \sum_{j=0}^{J-1} \beta_j x^j.$$  \hfill (17.1)

After $J$ is chosen, OLS can be used to 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 $J$. From [Weierstrass (1885)], letting $g(\cdot)$ denote a (possibly infinite-degree) polynomial function, given continuous $m: [a,b] \mapsto \mathbb{R}$,

$$\forall \epsilon > 0, \quad \exists g(\cdot) \text{ s.t. } \|m(\cdot) - g(\cdot)\|_\infty < \epsilon,$$  \hfill (17.2)

where $\|\cdot\|_\infty$ is the $L^\infty$ norm on $[a,b]$,

$$\|h(\cdot)\|_\infty = \sup_{x \in [a,b]} |h(x)|.$$  

That is, the approximation error can be made below any $\epsilon > 0$ for large enough $J$.

The key smoothing parameter that determines model flexibility is $J$, the number of terms. Larger $J$ increases flexibility, thus reducing bias but increasing variance. Thus, larger $J$ is analogous to smaller bandwidth for kernel regression. Clearly $J = n$ is too big ("perfect" overfitting) and usually $J = 1$ is too small (intercept-only model). 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

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^J} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{J-1} \beta_j x^j)^2,$$  \hfill (17.3)
minimizing over \( b \in \mathbb{R}^J \). Equivalently, consider minimization over functions within the set of degree \( J - 1 \) polynomials. Then,

\[
\hat{m}(\cdot) = \arg \min_{g(\cdot) \in M_J} \sum_{i=1}^{n} [Y_i - g(X_i)]^2, \quad M_J \equiv \{ g(\cdot) : g(x) = \sum_{j=0}^{J-1} \beta_j x^j \}.
\]  

(17.4)

Here, \( M_J \) is the set of all degree \( J - 1 \) polynomials, \( g(\cdot) \) is a generic function, and \( \hat{m}(\cdot) \) is the estimated CEF. This is the same estimate as before, i.e., \( \hat{m}(x) = \sum_{j=0}^{J-1} \hat{\beta}_j x^j \), 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 and “best” linear predictor (which are population objects, not related to sample data) and the pseudo-true parameter in quasi-maximum likelihood. (The BLA and BLP are described in Sections 7.4–7.5 of Kaplan (2020b).) Specifically, consider the function in \( M_J \) that’s “closest” to the true \( m(\cdot) \):

\[
m^*(\cdot) = \arg \min_{g(\cdot) \in M_J} E[(m(X) - g(X))^2].
\]  

(17.5)

That is, among all functions in \( M_J \), \( m^*(\cdot) \) 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 \( Q(g(\cdot)) = E[(m(X) - g(X))^2] \), with \( m^*(\cdot) \) as in (17.5). Note \( Q(m(\cdot)) = 0 \). As in (17.5), \( Q(m^*(\cdot)) \) is a measure of approximation error, where \( m^*(\cdot) = \arg \inf_{g(\cdot) \in M_J} Q(g(\cdot)) \). As in (17.4), let \( J_3 \) 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 (17.4),

\[
\hat{m}(\cdot) = \arg \min_{g(\cdot) \in M_n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2, \quad M_n \equiv \{ g(\cdot) : g(x) = \sum_{j=1}^{J_n} \beta_j \phi_j(x) \}.
\]  

(17.6)
where generally $J_n \to \infty$ 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 $M_n$ is called a sieve space, with the sequence of $M_n$ forming a sieve. This approach of minimization over a sequence of sieve spaces whose approximation error decreases to zero is known as the method of sieves (Grenander, 1981). This special case where the first $J_n$ terms in a basis form the sieve space is called series regression.

The choice of a 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 16.4.3), 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 $f, g \in M_n$, then $\lambda f + (1-\lambda)g \in M_n$ 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 since, 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. Instead of a single basis, there is a collection of bases called a library, or more generally some set of functions called a dictionary. These have redundancy in that one element may be a linear combination of other elements, which is not true of a basis. This highly nonlinear approach is more flexible than the nonlinear approach. Neural networks are an example of highly nonlinear approximation with a dictionary. For more on nonlinear and highly nonlinear approximation, see White (2006), “Approximate Nonlinear Forecasting Methods.”

### 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.
CHAPTER 17. SERIES AND SIEVES

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 (non-intercept) coefficient values according to their squared magnitude. Assume the regression model includes an intercept $\beta_0$ along with $k$ non-constant regressors, $X_1$ through $X_k$. Given penalty parameter $\lambda$ (like $\lambda = 0.1$), the ridge estimator solves

$$
(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) = \arg \min_{\beta_0, \beta_1, \ldots, \beta_k} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_1 - \cdots - \beta_k X_k)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 .
$$

(17.7)

If $\lambda = 0$, then the penalty is zero, and the ridge estimator is simply OLS, minimizing the SSR. If $\lambda = \infty$, then all $\hat{\beta}_1$ through $\hat{\beta}_k$ equal zero. The SSR may not be good, but making any $\hat{\beta}_j \neq 0$ would incur an infinite penalty, so it is never worth the reduction in SSR. Thus, the ridge estimator can be seen as a shrinkage estimator that “shrinks” all the $\hat{\beta}_j$ toward zero. When $\lambda = 0$, there is no shrinkage; as $\lambda$ increases, there is more shrinkage. Ridge regression can be written in multiple equivalent ways; these are insightful but beyond our scope.

As another example, lasso (Knight and Fu, 2000; Tibshirani, 1996) replaces the $\beta_j^2$ in (17.7) with absolute values, $|b_j|$. That is, given $\lambda$, lasso solves

$$
(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) = \arg \min_{\beta_0, \beta_1, \ldots, \beta_k} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_1 - \cdots - \beta_k X_k)^2 + \lambda \sum_{j=1}^{k} |b_j| .
$$

(17.8)

Again, the intercept $\beta_0$ 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 $\lambda = \infty$), whereas with lasso, usually some coefficients are shrunk all the way to $\hat{\beta}_j = 0$ exactly. In this way, lasso can be seen as performing model selection, “selecting” only the $X_j$ for which $\hat{\beta}_j \neq 0$. Indeed, lasso is an acronym (although it is usually written in lowercase), standing for the “least absolute shrinkage and selection operator.”

More generally, ridge and lasso can be seen as 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 $\gamma$. As special cases, ridge has $\gamma = 2$, and lasso has $\gamma = 1$.

Another generalization of ridge and lasso is called the elastic net. There, both $|b_j|$ and $b_j^2$ are penalized, possibly in different amounts, to try to get the best of both worlds.
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, since OLS is a linear smoother, so is series regression.
CHAPTER 17. SERIES AND SIEVES
Chapter 18

Model Selection

Optional resources for this chapter

- Textbook: Kaplan (2020b) 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
- 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: 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

As [Box (1979, p. 2)](https://en.wikipedia.org/wiki/All_models_are_wrong) famously wrote, “All models are wrong but some are useful.” 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. That is, the most useful model finds the happy medium between “too rigid” and “too flexible,” balancing the consequences of bias and variance appropriately. The goal of model selection is to choose the most useful model.

1See [https://en.wikipedia.org/wiki/All_models_are_wrong](https://en.wikipedia.org/wiki/All_models_are_wrong) for additional discussion.
There are two main approaches to model selection. One general approach (cross-validation) uses some observations to estimate each model and 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.

Historically, most model selection approaches focused on prediction accuracy under quadratic loss. Since the CEF is the best predictor under quadratic loss, this is also helpful for estimating the CEF. However, it is not necessarily optimal for estimating structural models or treatment effects.

18.1 Purpose

The famous quote from Box (1979) begs the question: useful for what? Just as a method cannot be simply “robust” but must be robust to something, a model cannot be simply “useful” but must be useful for something. Taking this question seriously has led to some innovative and practically useful model selection procedures, like those 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 give the best structural function estimate. 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, including many kernel and sieve estimators, have a linear smoother representation:

\[
\hat{m}(x) = \sum_{i=1}^{n} W_i(x)Y_i, \tag{18.1}
\]

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 \( \mathbf{W} \) have row \( i \), column \( j \) entry \( W_{ij} = W_j(X_i) \), so

\[
\hat{\mathbf{Y}} = \mathbf{W}\mathbf{Y}, \tag{18.3}
\]
18.3 BAD APPROACHES

where \( Y = (Y_1, \ldots, Y_n)' \) and \( \hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)' \).

To develop intuition, consider OLS. With OLS, \( W \) is the “hat matrix” or projection matrix \( X (X'X)^{-1}X' \). That is, \( \hat{Y} \) is the linear projection of \( Y \) onto the column space of \( X \), which is the \( n \times k \) matrix with row \( i \) equal to \( X_i' \). The natural measure of model flexibility for OLS is the number of terms (including the intercept) in the regression. The number of terms is \( k \), the number of columns in \( X \). However, \( W \) is \( n \times n \), so if we only know \( W \), it is not clear how to measure the flexibility. Thankfully, the trace is \( k \):

\[
\text{tr}(W) = \text{tr}(X (X'X)^{-1}X') = \text{tr}((X'X)^{-1}X'X) = \text{tr}(I_k) = k, \tag{18.4}
\]

the trace of the \( k \times k \) identity matrix.

**Discussion Question 18.1** (OLS flexibility 1). Consider OLS with \( k \) regressors (including the intercept), assumed to be all linearly independent (i.e., 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 \( \text{tr}(W) \)? Why?

c) What is \( W \)? 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 \( W \)? Hint: you can either use \( X = (1, 1, \ldots, 1)' \), or use \( \hat{Y} = WY \).

c) What is \( k \)? Why?

d) What is \( \text{tr}(W) \)? Why?

Equation (18.4) suggests \( \text{tr}(W) \) 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, the extreme of overfitting. That is, \( \hat{Y}_i = Y_i \) for all \( i = 1, \ldots, n \). Since \( \hat{Y}_i = \sum_{j=1}^n W_{ij}X_jY_j \), such an estimator must have \( W_{ii}(X_i) = 1 \) for \( i = 1, \ldots, n \), and \( W_{ij}(X_i) = 0 \) for \( j \neq i \). That is, \( W \) is the \( n \times n \) identity matrix, so \( \hat{Y} = WY = I_nY = Y \).

More generally, the main diagonal terms \( W_{ii} = W_i(X_i) \) help capture overfitting. They capture the influence of \( Y_i \) in \( \hat{Y}_i \) since

\[
\hat{Y}_i = W_{ii}Y_i + \sum_{j \neq i} W_{ij}Y_j. \tag{18.5}
\]

Many model selection procedures explicitly or implicitly use \( \text{tr}(W) \) 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 \( \alpha \)” Another
problem is sensitivity to the choice of null (vs. alternative) and choosing \( \alpha \). 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 \( n = 500 \), the true model suffers greatly from overfitting.

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 just tell us to use the most flexible model possible. This suffers overfitting; the bias is small, but the variance is huge. 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 \[18.13\].

18.4 Analytic Plug-in Approach

Equation \[16.20\] gives the AMSE-optimal bandwidth \( h_\ast \) for the local linear kernel regression estimator; isn’t that good enough? Indeed, the asymptotic approximation is probably good enough for most empirical examples. However, \( h_\ast \) is infeasible: it depends on unknown population objects like \( f_X(x) \), \( \text{Var}(Y \mid X = x) \), and most vexingly \( m^\prime\prime(x) \). So \( h_\ast \) cannot directly be used in practice.

Why don’t we just estimate the unknown terms? In principle, this is possible, yielding
18.5 CROSS-VALIDATION

a plug-in bandwidth. However, large estimation error may cause the plug-in bandwidth to differ significantly from the infeasible $h_\ast$. For example, with nonparametric regression, the goal is to learn the unknown $m(\cdot)$ without imposing a parametric form. This is a difficult estimation problem. It is even more difficult to estimate the second derivative $m''(\cdot)$, which is required for the plug-in bandwidth. Even if the plug-in bandwidth has the optimal power of $n$, in finite samples the “constant” term may have a big effect, too.

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.2 (Important, plus it’s a good pun.) Incidentally, it is also an extremely useful website.

18.5.1 Training and Validation Paradigm

Discussion Question 18.3 (validation SSR). Consider Figure 18.2. The (colored) lines are CEF estimates 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 (e.g., through kernel smoothing or an estimated coefficient). This is cheating! If we can use $Y_i$, then we should just predict $\hat{Y}_i = Y_i$. But we know this is (extreme) overfitting.

The cross-validation (CV) approach can be described as separating the observations used to generate $\hat{Y}_i$ from $Y_i$. That is, it doesn’t allow cheating. The subset of observations used to generate the predictions is often called the training sample. These predictions

2https://stats.stackexchange.com
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.

There are many ways to partition a sample of $n$ observations into training and validation samples. Multiple such partitions can be used to evaluate models. There are different names for the different types of CV using different partitioning strategies.

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), since 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, $\hat{m}_{-i}(-)$ 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, Let $s$ denote the “model” and write $\hat{m}_{-i}(X_i; s)$. For example, $s$ could be a kernel regression bandwidth, or the number of series regression terms. The LOOCV criterion is then

$$\text{LOOCV}(s) = \sum_{i=1}^{n} [Y_i - \hat{m}_{-i}(X_i; s)]^2. \quad (18.6)$$
18.5. CROSS-VALIDATION

Let $S$ be a set of possible candidate models. Then LOOCV chooses the model

$$s^* = \arg \min_{s \in S} \text{LOOCV}(s).$$  \hfill (18.7)

Scaling $\text{LOOCV}(s)$ by constants does not change $s^*$, so (18.6) could be multiplied by $1/n$.

LOOCV can take a long time to compute. Using brute force, LOOCV requires computing $\text{LOOCV}(s)$ for each $s \in S$, 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. The LOOCV criterion can be written in terms of the full-sample (not LOOCV) SSR. In the notation of Section 18.2,

$$\text{LOOCV}(s) = \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{m}(X_i; s) \right)^2,$$  \hfill (18.8)

a function of the usual full-sample residuals $Y_i - \hat{m}(X_i)$. See Section 18.A for the derivation.

LOOCV can also be interpreted in terms of SSR penalization. The usual SSR (on which $R^2$ is based) is

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$  \hfill (18.9)

As seen in (18.8), minimizing LOOCV is equivalent to minimizing a sum of squared penalized residuals. Recall from (18.5) that $W_i(X_i)$ measures the influence of $Y_i$ in $\hat{Y}_i$, which captures the flexibility of the model. Minimizing (18.8) is equivalent to minimizing

$$\sum_{i=1}^{n} p_i(Y_i - \hat{Y}_i)^2, \quad p_i \equiv 1/[1 - W_i(X_i)]^2,$$  \hfill (18.10)

where $p_i$ is the penalty for flexibility of $\hat{Y}_i$, and $(Y_i - \hat{Y}_i)^2$ is the usual squared residual. The penalty is in the range $p_i \in [1, \infty)$ if $0 \leq W_i(X_i) < 1$, with $p_i$ increasing in $W_i(X_i)$. For example, when $W_i(X_i) = 0$, then $p_i = 1$, the smallest possible value, and the squared residual is used unpenalized. When $W_i(X_i)$ is close to 1, then $p_i$ is very large, so there is a big penalty on the squared residual.

This penalty helps manifest the bias–variance tradeoff. Unlike with the unpenalized SSR, there is a tension in (18.8): more flexibility decreases the squared residual but increases the penalty. LOOCV only views additional flexibility as beneficial if the squared residuals decrease more than the flexibility penalty increases.

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:

\[
GCV = \frac{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{[1 - \text{tr}(W)/n]^2} = n^{-1}\text{SSR}/[1 - \text{tr}(W)/n]^2. \tag{18.11}
\]

The penalty comes from

\[
\frac{1}{n} \sum_{i=1}^{n} [1 - W_i(X_i)] = \frac{1}{n} (n - \sum_{i=1}^{n} W_{ii}) = 1 - \text{tr}(W)/n. \tag{18.12}
\]

Scaling by \( n \) does not change the minimizer, so GCV could also be written as

\[
\text{SSR}/[n - \text{tr}(W)]^2. \tag{18.13}
\]

This is actually similar to the SSR adjustment in the adjusted \( R^2 \), but here the penalty is squared.

**Discussion Question 18.4** (model selection and the CIA). Imagine using OLS with the structural model \( Y = \beta_0 + \beta_1 X_1 + X_2' \gamma + U \), and we want the conditional independence assumption \((U \perp \perp X_1 \mid X_2)\) to hold such that \( \beta_1 \) is the average structural effect of \( X_1 \) on \( Y \). Or, more realistically, we expect some omitted variable bias (OVB), but we hope it is small. More directly: we want to choose the variables in \( X_2 \) in order to minimize the MSE of \( \hat{\beta}_1 \), which (as a reminder) is the squared bias of \( \hat{\beta}_1 \) plus its squared standard error.

a) What about GCV might be good for picking \( X_2 \) 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 can be generalized 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 \( i \). 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.

### 18.5.5 \( k \)-fold CV

If \( d \) is a significant fraction of \( n \), then leave-\( d \)-out CV is usually called **\( k \)-fold cross-validation** where \( k = n/d \). For whatever reason, \( k = 5 \) is very popular. That is, split the sample into 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 sum of squared (cross-validated) residuals.

As noted in Section 18.7, this has somewhat different performance than LOOCV.
18.6. INFORMATION CRITERIA

18.5.6 Time Series

All of these are inappropriate for time series, but the same paradigm 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 since we do know $Y_{t+1}$, we can compare the forecast $\hat{Y}_{t+1}$ with the actual $Y_{t+1}$. Then we can pretend we live at time $t + 1$ to generate $\hat{Y}_{t+2}$, and compare with the true $Y_{t+2}$, and so on, up to comparing $\hat{Y}_T$ and $Y_T$.

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:

$$
AIC = -2 \ln(\mathcal{L}) + \frac{2k}{n} \text{ or } n \ln(SSR) + \frac{2k}{n},
$$

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). 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:

\[
AIC = -2\ln(L) + \ln(n)k \text{ or } n\ln(SSR) + \ln(n)k.
\] (18.15)

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 BIC penalty than 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 \(\ln(L) = 2\); Model B has \(\ln(L) = 4\); Model D has \(\ln(L) = 5\). Let \(n = 55\), so \(\ln(n) \approx 4.\) Refer to (18.14) and (18.15).

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 are many other information criteria. Just pick a letter of the alphabet, and add “IC” after it. There is GIC, FIC, (M)RIC, EAIC, NIC, and more; e.g., see the long list on page 223 of Shao (1997).

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 $2^{pn}$ possible models. 

Shao (1997) considers multiple properties. If the true CEF involves only a finite number of regressors, then it’s possible (with large enough $n$ and $p_n$) for a model selection procedure to choose the correct model. However, with nonparametric regression, generally the true CEF is assumed to involve an infinite-term basis expansion, so there is no correct model for any $n$ and $p_n$, no matter how large. Even if there is no “correct” model, one of the candidate models is still “best” according to some definition of accuracy. A model selection procedure is consistent if it selects the “best” model with probability approaching one (as $n \to \infty$). A weaker but still useful property is if the accuracy of the selected model approaches (as $n \to \infty$) the accuracy of the “best” model, which Shao (1997) refers to as asymptotic loss efficiency.

Shao (1997) defines a GIC$_{\lambda_n}$ model selection procedure that nests Mallows’ $C_p$, GIC, FPE$\lambda$, and others, along with AIC and BIC with a slight modification. On page 235, Shao (1997) delineates thee classes of model selection procedures and qualitatively compares their performance in different settings. Class 1 includes GIC$_2$, 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

Model averaging is beyond the (current) scope of this text, but you should be aware of the basic idea. 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 more, see for example Chapter 7 of Claeskens and Hjort (2008), who discuss both frequentist and Bayesian model averaging.

More generally, ensemble methods try to combine multiple (often simpler) models or predictions into a more sophisticated (and hopefully more accurate) 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

\[
\hat{m}_{-i}(X_i) = \frac{\sum_{j \neq i} W_j(X_i)Y_j}{\sum_{j \neq i} W_j(X_i)}.
\]

Then, since \( 1 = \sum_{j=1}^{n} W_j(X_i) = W_i(X_i) + \sum_{j \neq i} W_j(X_i) \),

\[
\hat{m}_{-i}(X_i) = \frac{\hat{m}(X_i) - W_i(X_i)Y_i}{1 - W_i(X_i)},
\]

\[
\text{LOOCV} = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{m}_{-i}(X_i)]^2 = n^{-1} \sum_{i=1}^{n} \left[ Y_i - \frac{\hat{m}(X_i) - W_i(X_i)Y_i}{1 - W_i(X_i)} \right]^2
\]

\[
= n^{-1} \sum_{i=1}^{n} \left( \frac{Y_i - \hat{m}(X_i)}{1 - W_i(X_i)} \right)^2.
\]
Chapter 19

Multiple Regressors

Unlike with OLS, it is not trivial to simply add another regressor to a nonparametric regression model. This chapter describes the difficulty and some approaches.

19.1 Curse of Dimensionality

Discussion Question 19.1 (curse of dimensionality 0). Let $n = 1000$. Let $X = (X_1, \ldots, X_6)$, 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 $E(Y \mid X = x)$ for all possible values of $x$, by taking the sample mean of the values $Y_i$ for which $X_i = x$; i.e., taking the sample mean within each “cell.” Denote subsample sizes as $n_x = \sum_{i=1}^{n_i} 1 \{X_i = x\}$.

a) Among all possible samples, what’s the largest possible value of $\min_x n_x$?

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 $m(x_0) = E(Y \mid X = x_0)$. Let $x_0 = 0.05$ with bandwidth $h = 0.1$. Let $n$ denote sample size.

a) Let $X_i \sim \text{Unif}(0, 1)$. 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 $P(X_i \in [x_0 - h/2, x_0 + h/2])$?

b) Let $X_i \sim \text{Unif}([0, 1]^2)$, the uniform distribution over the unit square $[0, 1] \times [0, 1]$; i.e., the PDF of $X_i$ is $f(x) = 1$ if $x \in [0, 1]^2$ and $f(x) = 0$ elsewhere. What’s the probability that a single $X_i$ 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 $X_i$ 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 $X_i$ is uniformly distributed over the $d$-dimensional hyper-cube $[0, 1]^d$?
e) For general \( h \), assuming \( x_0 \leq h \leq 1 - x_0 \), what’s \( P(X_i \in [0,1]^d) \) if again \( X_i \) is uniformly distributed over \([0,1]^d\)?

**Discussion Question 19.3** (curse of dimensionality 2). Continue from DQ\textsuperscript{19.2}.

a) Let \( N_0 = \sum_{i=1}^{n} \mathbb{1}\{X_i \in [x_0 - h/2, x_0 + h/2]^d\} \) be the local sample size. Explain why the mean local sample size \( \mathbb{E}(N_0) \) is proportional to \( nh^d \) with iid sampling.

Hint: \( B_i \equiv \mathbb{1}\{X_i \in \text{window}\} \) 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 \( p = P(X_i \in \text{window}) \); 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?

**Discussion Question 19.4** (curse of dimensionality 3). Continue from DQs\textsuperscript{19.2} and \textsuperscript{19.3}.

a) With \( d = 1 \), what does the linear smoother matrix \( W \) look like for the local constant estimator?

b) With \( d = 10 \), what does \( W \) look like? In particular, describe how it differs from the \( W \) when \( d = 1 \).

c) With \( d = 1 \), what does the matrix \( W \) for the linear-in-variables OLS estimator look like (qualitatively)? In particular, how does it differ from the local constant \( W \)?

d) With \( d = 10 \), what does the linear-in-variables OLS \( W \) look like? How does it differ from the \( d = 1 \) OLS \( W \) and from the \( d = 10 \) local constant \( W \)?

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 since \( h \downarrow 0 \). 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. Sieve estimators suffer similarly in larger dimensions.

Parametric CEF models do not suffer this curse because they impose enough structure that all observations are informative about \( m(x) \), no matter how far away \( X_i \) is from \( x \).

To alleviate the curse of dimensionality, one approach is to impose more structure than a fully nonparametric CEF but less than a parametric CEF. Some examples follow in Sections\textsuperscript{19.2,19.4} Alternatively, with a finite-dimensional object of interest, the fully nonparametric CEF could still result in a \( \sqrt{n} \) convergence rate.
19.2 Additive Model

One way to impose structure is to exclude interaction terms between certain pairs of regressors. If $X = (X_1, X_2)$, then instead of the fully nonparametric CEF $m(x_1, x_2)$, excluding interactions yields $m(x_1, x_2) = g_1(x_1) + g_2(x_2)$. More generally, with $k$ regressors in $x = (x_1, \ldots, x_k)'$ and no interactions, the additive model is

$$m(x) = \sum_{j=1}^{k} g_j(x_j). \quad (19.1)$$

Although this initially seems more difficult since there are $k$ unknown functions $g_j(\cdot)$ instead of a single unknown $m(\cdot)$, it is easier since each is a function of a scalar.

Some interactions could be maintained. For example,

$$m(x) = g_1(x_1) + g_2(x_2) + g_3(x_3, \ldots, x_k), \quad m(x) = \sum_{j=1}^{k/2} g_j(x_{2j-1}, x_{2j}), \quad \text{etc.} \quad (19.2)$$

Discussion Question 19.5 (additive wage model). Consider a CEF model of log wage in terms of education, experience, sex, and IQ score.

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 sex nonparametrically? Explain.

19.3 Partially Linear Model

Another alternative is to specify part of $m(\cdot)$ parametrically. Let $x = (x_1', x_2')'$. The partially linear model (PLM) has

$$m(x) = x_1' \beta + g(x_2). \quad (19.3)$$

The interpretation of (19.3) is simple if interest is only in the slopes for $x_1$, i.e., if $x_2$ only contains control variables. That is, the slope wrt any element of $x_1$ is the corresponding element of $\beta$.

The $\sqrt{n}$ convergence rate for $\hat{\beta}$ 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 $x_1$ variable, and it imposes linearity in certain dimensions. The vector $x_1$ 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 $x_2$ control variables, and it imposes a parametric structure of the part of the CEF involving $x_1$.

In principle, the PLM can be combined with the additive model approach, e.g.,

$$m(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \sum_{j=3}^{k} g_j(x_j). \quad (19.4)$$
CHAPTER 19. MULTIPLE REGRESSORS

If the entire function $m(\cdot)$ is of interest, then this helps. If only $\beta$ 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 $x$ to enter through a (single) linear index of the form $x'\beta$:

$$m(x) = h(x'\beta),$$

(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); 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.
Exercises

Exercise E19.1. Find a paper published in a respectable economics journal\footnote{E.g., in top 500 of https://ideas.repec.org/top/top.journals.all.html} 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. 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 published in a respectable economics journal\footnote{E.g., in top 500 of https://ideas.repec.org/top/top.journals.all.html} 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.

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 (RIMSE); you can just average the MSE over a grid of \( x \) 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 published in a respectable economics journal\(^3\) that reports a result that can be replicated with OLS; provide a link to the paper. 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?

\(^3\)E.g., in top 500 of [https://ideas.repec.org/top/top.journals.all.html](https://ideas.repec.org/top/top.journals.all.html)
Chapter 20

Nonparametric Regression in R

This chapter contains a few simple examples with different packages in R.

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 \texttt{ns()} in the \texttt{splines} package, which is part of core R \cite{R}. 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 \texttt{caret} package \cite{caret} 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 since 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))
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 (penalty) 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.

```r
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, df=n)  # GCV
rets[[2]] <- smooth.spline(x=df$X, y=df$Y, cv=TRUE, df=n)   # 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)
```

Figure 20.2: Natural cubic B-spline fit chosen by GCV.

```r
SSR <- sum(ret$residuals^2)
GCVs[iDF] <- (n/(n-ret$rank))^2 * SSR
}
(dfstar <- DFs[which.min(cvRMSEs)])
print(GCVs)
## [1] 138 136 121 117 120 123
print(cvRMSEs)
## [1] 1.17 1.16 1.12 1.06 1.09 1.09 1.07
```

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 (penalty) 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.

```r
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, df=n)  # GCV
rets[[2]] <- smooth.spline(x=df$X, y=df$Y, cv=TRUE, df=n)   # 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)
```
Figure 20.2: Smoothing spline estimates with same data but different penalties.

\[
\begin{align*}
xx & \leftarrow \text{seq(from=-3, to=3, by=0.005)} \\
mhat\text{LOOCV} & \leftarrow \text{predict(rets[[2]], x=xx)} \ #\text{has x and y}
\end{align*}
\]

20.2 Local Polynomial 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.

```r
library(KernSmooth)
set.seed(112358)
CEF <- function(x)sin(1.5*x)
n <- 100; X <- rnorm(n); Y <- CEF(X)+rnorm(n)
```

20.2. LOCAL POLYNOMIAL REGRESSION
CHAPTER 20. NONPARAMETRIC REGRESSION IN R

Figure 20.3: Local polynomial regression example, package KernSmooth.

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(2,3,3,1))
  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, “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.

```r
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=1^(-2:0))
f <- train(formula=Y~X, data=df, method='nnet', maxit=100,
    tuneGrid=tg, trace=F, metric='RMSE',
    trControl=trC)
b <- nf$bestTune
rets[[2]] <- nnet(formula=Y-X, data=df, linout=TRUE,
    size=b$tsize, decay=b$td, trace=FALSE)
mhat1 <- predict(object=rets[[1]],
    newdata=data.frame(X=xx))
```

Figure 20.4: Local linear regression example, package np.
20.4 Multiple Regressors

The following code shows examples of partially linear and additive nonparametric regression. You can call `plot()` on the returned objects.

```r
library(mgcv); library(splines)
set.seed(112358)
n <- 50
gladd <- function(x) { x^2 }
glplm <- function(x) { x }
CEFadd <- function(x1,x2,x3) { gladd(x1)+x2^2+x3^2 }
CEFplm <- function(x1,x2,x3) { glplm(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)
```

Figure 20.5: Random forest and neural network regression.

```r
mhat2 <- predict(object=rets[[2]],
newdata=data.frame(X=xx))
```
Chapter 21

Regression Discontinuity

Optional resources for this chapter

- Background (potential outcomes framework): Section 4.3 of Kaplan (2020b)
- Key papers: Thistlethwaite and Campbell (1960), Hahn, Todd, and Van der Klaauw (2001)
- Inference under weak identification: Feir, Lemieux, and Marmer (2016)
- R: packages rdrobust (Calonico, Cattaneo, Farrell, and Titunik 2020; Calonico, Cattaneo, and Titunik 2015) and rdd (Dimmery 2016)
- Stata: commands rdrobust, rdbwselect, and rdbinselect, as described in Stata Journal articles (Calonico, Cattaneo, Farrell, and Titunik 2017; Calonico, Cattaneo, and Titunik 2014a). Stata: command DCdensity for McCrary (2008) test

Regression discontinuity has been very popular in practice, which has led to a growth in methodological and theoretical regression discontinuity papers. Which means: I am missing lots of interesting, recent work. The links above include some extensions, but
below I only cover the core idea.
Proofs are in the chapter appendix.

21.1 Introduction

The approach of regression discontinuity (RD), or regression discontinuity design (RDD), exploits a known discontinuity to estimate a “local” treatment effect. The idea was originally proposed by Thistlethwaite and Campbell (1960) but only became popular in economics shortly before Hahn, Todd, and Van der Klaauw (2001) formally introduced RD to the econometrics literature. Its simplicity and relative transparency have made it popular. The usual downside of reduced-form methods applies: external validity (generalization) is limited. Recent work has extended the object of interest from a (local) average treatment effect to quantile and distributional effects.

Methodologically, RD is interesting: a clever identification argument (i.e., how to recover causality from the joint distribution of observables), nonparametric estimation, and need for inference beyond the usual CI of estimator plus/minus twice standard error.

The original application is well-known and provides intuition. The research question is the treatment effect of a college scholarship (on college graduation or other outcomes). The policy awarding scholarships has a discontinuity. Specifically, there is a precise test score cutoff, above which all students receive a scholarship, and below which none do. Let $z_0$ denote the score cutoff, say $z_0 = 100$.

Discussion Question 21.1 (RD: scholarship estimand). Consider the college scholarship example, where $Y = 1$ if the student went on to graduate from college (and $Y = 0$ if not), $Z$ is test score, and the scholarship (treatment) dummy is $X = 1\{Z \geq z_0\}$.

a) Imagine you want to learn if the current scholarship program has any (positive) effect on graduation rate. In terms of potential outcomes $(Y_1, Y_0)$, describe a useful object of interest, explaining why it helps answer your research question (effect of current policy).

b) Imagine the current program seems to have a big positive causal effect, and you are in charge of deciding whether or not to expand the program by lowering $z_0$ (so more students qualify). Describe an object of interest that helps you learn about the causal effect of expanding the scholarship program, explaining why it’s useful.


a) Consider estimating a linear probability model (LPM), using OLS to regress $Y$ on $(1, X)$. Can the coefficient on $X$ be interpreted as any kind of causal effect? If yes, explain which and why; if no, explain why not.

b) Repeat [a] but with $(1, X, Z)$ instead of $(1, X)$.

The core idea is that we think students scoring $Z = 99$ are extremely similar to students scoring $Z = 101$. That is, we believe the two-point difference $(101 - 99 = 2)$
is mostly random noise: the $Z = 99$ student got less sleep because their neighbor’s dog barked, or the $Z = 101$ student was (very slightly) inspired by seeing a pretty flower, or something. The difference is almost entirely due to such noise rather than underlying, persistent differences among the students.

It is “as if” students were simply randomized intro treatment ($Z = 101$) and control ($Z = 99$) groups, except instead of explicit randomization, the randomization mechanism is the neighbor’s dog, i.e., a quasi-experiment. Consequently, average treatment effects should be estimable by comparing treated and untreated outcomes. Similarly, quantile and distributional treatment effects can be estimated.

Of course, students with $Z \approx z_0$ may be affected very differently by college scholarships than students with $Z = 50$ or $Z = 150$. That is, the “average treatment effect” is only for the very specific group of students who score very near $z_0 = 100$; it is a “local” average treatment effect, but for a different “local” supopulation than the usual LATE (Imbens and Angrist, 1994).

**Discussion Question 21.3** (RD: scholarship heterogeneity). Consider the same setup as in DQ 21.1.

a) What would you guess is the relationship between the scholarship effect and $Z$? E.g., if we looked at groups of students with low $Z$, medium $Z$, and high $Z$, which group(s) would you expect to benefit the most, and why?

b) How does the ATE for the $Z = z_0$ subpopulation compare with the estimand you proposed in DQ 21.1[b]?

The “ATE for the $Z = z_0$ subpopulation” $E(Y_1 - Y_0 \mid Z = z_0)$ is similar to but different than the **marginal treatment effect** (MTE). For details on the MTE, see [Heckman and Vytlacil (2001, 2007)] and references therein. Similarly, both define an ATE for a marginal subpopulation that is affected by a marginal policy change. The difference is the MTE’s subpopulation is defined by the value of unobservables affecting the decision to select into treatment.

### 21.2 Notation and Terminology

I refer to units of observation $i = 1, \ldots, n$ as “individuals,” but as usual they may be countries, firms, hospitals, schools, etc. Similarly, the “treatment” could be essentially any policy (although SUTVA may be more plausible for some than others): a college scholarship, lower sales tax rate, different criminal sentencing law, etc.

As in the example, let $Z$ denote the **running variable** (also called **forcing variable**) such as a student’s test score. Also as in the example, let $z_0$ denote the cutoff (or “threshold”) at which the treatment probability jumps. Below, $Z$ is assumed to have a continuous distribution, at least near $z_0$. Of course, something like test score is probably measured as an integer number of points, i.e., is discrete rather than continuous. Even then, the intuition for the RD approach is the same, and in finite samples with many
possible score values, the difference is probably negligible. But, I think there are now some formal results on RD with a discrete running variable.

Let \( X = 1 \) if an individual is treated and \( X = 0 \) if untreated.

Identification results are for various treatment effects in the potential outcomes framework, a.k.a. the (Neyman) Rubin Causal Model (RCM). For a review, see Section 4.3 of Kaplan (2020b), or Section 21.2 of Wooldridge (2010) and my supplemental notes.

Each individual has a pair of treated and untreated potential outcomes \((Y_1, Y_0)\), but only one outcome is observed. This means we can never directly observe the treatment effect \( Y_1 - Y_0 \). More specifically, the observed outcome is

\[
Y = X Y_1 + (1 - X) Y_0
\]

where \( A \) is the random intercept, and random slope \( B \) is the treatment effect. The intercept and slope are upper case to emphasize they are random variables; i.e., each individual has her own \((Y, X, Z, A, B)\), although \( B \) is unobservable and \( A \) is only sometimes observable.

There are two types of RD, but really one is just a special case of the other. In the sharp design, all units above the cutoff are treated, and all below are untreated: \( X = 1 \{Z \geq z_0\} \). Put differently, the conditional treatment probability \( P(X = 1 \mid Z = z) \) jumps from 0 to 1 at \( z_0 \). In the fuzzy design, the probability of treatment for individuals above the cutoff is higher than the probability of treatment for individuals below the cutoff. Here, \( P(X = 1 \mid Z = z) \) still jumps at \( z_0 \), but not necessarily from 0 to 1; maybe from 0.2 to 0.7, for example. Thus, sharp RD is a special case of fuzzy RD with probabilities 0% and 100%. Sometimes a government’s or institution’s policy sounds like it has a clear cutoff (sharp RD), but in practice the enforcement is fuzzy (as often seen in data).

Let

\[
\begin{align*}
x^+ &\equiv \lim_{z \downarrow z_0} \mathbb{E}[X \mid Z = z], \quad x^- \equiv \lim_{z \uparrow z_0} \mathbb{E}[X \mid Z = z], \\
y^+ &\equiv \lim_{z \downarrow z_0} \mathbb{E}[Y \mid Z = z], \quad y^- \equiv \lim_{z \uparrow z_0} \mathbb{E}[Y \mid Z = z].
\end{align*}
\]

Thus, \( x^+ - x^- \) quantifies the jump in treatment probability at \( z_0 \), and \( y^+ - y^- \) quantifies the corresponding jump in mean outcome at \( z_0 \).

### 21.3 Identification: Constant Effect

The formal results below follow Hahn, Todd, and Van der Klaauw (2001).

\(^1\)http://faculty.missouri.edu/kaplandm/9473/Kaplan_9473_teaching_notes.pdf
21.3. IDENTIFICATION: CONSTANT EFFECT

To identify a causal (treatment) effect, we must be sure that the observed discontinuity in outcomes is driven only by the discontinuity in treatment probability. That is, if we see the conditional mean observed outcome \( E(Y \mid Z = z) \) jump at \( z_0 \), where \( P(X = 1 \mid Z = z) \) also jumps, we still need to rule out other possible causal mechanisms. If something else jumps at \( z_0 \), then perhaps the jump in \( Y \) is not caused entirely by the jump in treatment. For example, if for some reason the direct effect of \( Z \) on \( Y \) were also discontinuous at \( z_0 \), then we couldn’t disentangle the treatment effect of \( X \) from the direct effect of \( Z \) itself.

Assumptions A21.1 and A21.2 are sufficient for identification of a constant (common, homogeneous) treatment effect that’s the same for all individuals. That is, all individuals have the exact same \( \beta \equiv Y_1 - Y_0 \). In the notation of (21.1), \( B = \beta \) for all individuals. As usual, “identification” means that the treatment effect \( \beta \) can be expressed in terms of the joint distribution of observable variables \((X, Y, Z)\). That is, if we know the joint distribution of observables (which asymptotically we do), then we can uniquely derive the corresponding treatment effect \( \beta \).

The assumption of a constant treatment effect is never plausible, but it simplifies the proof to focus on the intuition.

Below, A21.1 is Assumption RD (including footnote 3) of Hahn, Todd, and Van der Klaauw (2001); A21.2 is Assumption A1 of Hahn, Todd, and Van der Klaauw (2001), which is similar to Assumption 2.1 of Imbens and Lemieux (2008); and A21.3 is stated in Theorem 1 of Hahn, Todd, and Van der Klaauw (2001). Finally, Theorem 21.1 is Theorem 1 of Hahn, Todd, and Van der Klaauw (2001).

Assumption A21.1 (RD conditional probability discontinuity). Using the notation of Section 21.2, assume \( Z \) has positive density in a neighborhood around \( z_0 \). The limits defined in (21.2) exist and are not equal to each other, i.e., \( x^+ \neq x^- \).

Assumption A21.2 (RD running variable). In (21.1), \( E(Y_0 \mid Z = z) \) is continuous in \( z \) at \( z_0 \).

Assumption A21.3 (RD constant treatment effect). In (21.1), \( B = \beta \) for constant \( \beta \in \mathbb{R} \).


\[
\beta = \frac{y^+ - y^-}{x^+ - x^-},
\]

(D21.4)

Discussion Question 21.4 (RD: sharp vs. fuzzy). Recall “sharp RD” means \( X = \mathbb{I}\{Z \geq z_0\} \). Consider the RHS of (21.4), \( (y^+ - y^-)/(x^+ - x^-) \).

a) For sharp RD, what’s \( x^+ \) ? \( x^- \) ? \( x^+ - x^- \) ? Explain why.

b) Imagine you know \( z_0 \), and you observe \((Y_i, Z_i)\) but not \( X_i \). Since you can’t estimate \( x^+ \) and \( x^- \), you assume the policy is perfectly enforced (sharp RD). However, in reality, it’s not perfectly enforced (fuzzy RD). In which direction is your (sharp) RD estimator biased? Why?
c) Imagine $\mathbb{E}[X \mid Z = z]$ is increasing in $z$ near $z_0$, but the policy is so loosely enforced that there is no jump discontinuity at $z_0$. Explain both mathematically and intuitively why the RD identification approach fails.

### 21.4 Identification: “Local” ATE

Without the constant treatment effect (A21.3), a type of “local” average treatment effect (ATE) can be identified. To review ATE, see Section 4.3.4 of Kaplan (2020b), or Section 21.2 of Wooldridge (2010); to review LATE (Imbens and Angrist, 1994), see Section 21.4.3 of Wooldridge (2010) and my supplemental notes. The idea is similar to the traditional LATE estimand for IV estimation, but instead of the ATE for the subpopulation of compliers, it’s the ATE for the subpopulation with $Z = z_0$. That is, the object of interest is

$$\mathbb{E}[B \mid Z = z_0]. \quad (21.5)$$

Assumption [A21.3] is replaced by a pair of assumptions, [A21.4] and [A21.5]. The latter is similar to a selection on observables or conditional independence or unconfoundedness assumption often made in the treatment effects literature (e.g., Section 6.6.2 of Kaplan (2020b)), but here it is much weaker because it need only hold near $z_0$. That is, we hope the discontinuity gives us a sort of locally-randomized experiment.

Below, A21.4 is Assumption A2 of Hahn, Todd, and Van der Klaauw (2001); the stronger form of A21.5 is stated as part of Theorem 2 in Hahn, Todd, and Van der Klaauw (2001); and Theorem 21.2 is Theorem 2 of Hahn, Todd, and Van der Klaauw (2001).

**Assumption A21.4** (RD variable treatment effect). In (21.1), $\mathbb{E}[B \mid Z = z]$ is a continuous function of $z$ at $z_0$.

**Assumption A21.5** (RD local conditional independence). In a neighborhood of $z_0$, $X$ is independent of $B$ conditional on $Z$. More weakly, conditional mean independence is sufficient: for arbitrarily small $\delta > 0$, for $z \in [z_0 - \delta, z_0 + \delta]$, $\mathbb{E}[B \mid Z = z, X] = \mathbb{E}[B \mid Z = z]$.


$$\mathbb{E}[B \mid Z = z_0] = \mathbb{E}[y^+ - y^-]_{x^+ - x^-}. \quad (21.6)$$

### 21.5 Identification: Local LATE


---

http://faculty.missouri.edu/kaplandm/9473/Kaplan_9473_teaching_notes.pdf
This new assumption and result have closer parallels to the usual LATE of Imbens and Angrist (1994). From Section 21.4.3 of Wooldridge (2010) and my supplemental notes, for the usual LATE, the subpopulation of compliers can be written in terms of individuals’ “potential treatments” $X_z$ for $z = 0, 1$, where $z$ is the value of the binary instrument and $X = ZX_1 + (1 - Z)X_0$ is the observed treatment. “Compliers” have $X_z = z$, or equivalently $X_1 - X_0 = 1$. “Defiers” have $X_z = 1 - z$, or equivalently $X_1 - X_0 = -1$, so the “no defiers” assumption can be written as a “monotonicity” assumption that $X_1 \geq X_0$.

In RD, the running variable $Z$ acts as an instrument that produces exogenous variation in treatment $X$. See the 2SLS interpretation in (4.9) of Imbens and Lemieux (2008), for example. The “compliers” are the subpopulation whose treatment status changes discontinuously at $z_0$. Assumption A21.6(ii) is a (local) “no defiers” assumption, written in terms of monotonicity.

Assumption A21.6 is a variant of Assumption A3 of Hahn, Todd, and Van der Klaauw (2001), the latter half of which is similar to Assumption 2.3 of Imbens and Lemieux (2008).

**Assumption A21.6** (RD local independence and monotonicity). Consider treatment as a function $X(\cdot)$ of $z$, so observed treatment status is $X(Z)$. There exists arbitrarily small $\delta > 0$ such that the following hold, defining the interval $N_\delta = [z_0 - \delta, z_0 + \delta]$. i) $(B, X(z))$ is jointly independent of $Z$ near $z_0$; more precisely, the joint distribution of $B$ and $\{X(r)\}_{r \in N_\delta}$ is the same conditional on any $Z \in N_\delta$. ii) Local monotonicity: for $z_0 - \delta < z' < z'' < z_0 + \delta$, $X(z'') \geq X(z')$.

Theorem 21.3 is Theorem 3 of Hahn, Todd, and Van der Klaauw (2001), although they seem to have forgotten the local-to-$z_0$ conditioning event in their (5). I have written the local-to-$z_0$ condition as $Z \in N_\delta$. It could be replaced by $Z = z$ for any $z \in N_\delta$ since A21.6 assumes the distribution of $B$ is the same conditional on any $Z = z \in N_\delta$. Specifically, to make the result look more similar to Theorem 21.2, one could write $Z = z_0$.


$$
\lim_{e \downarrow 0} E[Y_1 - Y_0 \mid X(z_0 + e) - X(z_0 - e) = 1, Z \in N_\delta] = \frac{y^+ - y^-}{x^+ - x^-}.
$$

(D21.7)

**Discussion Question 21.5** (RD: local L/ATE). Recall your proposal for part (b) of DQ 21.1, i.e., what you’d like to learn to inform your decision about expanding the scholarship program.

a) In the scholarship example, what’s the difference between the LHS (estimand) in Theorem 21.2 and the LHS (estimand) in Theorem 21.3?

b) Consider a fuzzy RD example where families are eligible for some government assistance program if household income $Z$ is below $z_0$, but not all eligible families actually apply to get the benefit ($X = 1$). You wonder what would be the effect of increasing the program by increasing $z_0$ slightly (to make more families eligible). Discuss whether the LHS/estimand in Theorem 21.2 or Theorem 21.3 seems more helpful.

[http://faculty.missouri.edu/kaplandm/9473/Kaplan_9473_teaching_notes.pdf](http://faculty.missouri.edu/kaplandm/9473/Kaplan_9473_teaching_notes.pdf)
21.6 Manipulation of the Running Variable

Continuing the college scholarship example, imagine students were allowed to retake the test multiple times, but only the most recent score is observed in the data. Consider a student who scores just below $z_0$. If the student is lazy (or rich?), then they won’t bother to retake the test. But if the student is highly motivated and diligent, then they might keep retaking the test until the score just above $z_0$ to earn a scholarship. The core of RD identification is that students with scores just below $z_0$ are essentially the same as students with scores just above $z_0$. But here, students just below $z_0$ are generally lazier than those just above $z_0$. Intuitively, this seems problematic: if the outcome of interest is something like college completion or college GPA, then laziness has a direct (negative) effect, so it is impossible to disentangle the scholarship effect from the laziness effect. This has been a problem with some RD analysis of the GED test; see Jepsen, Mueser, and Troske (2016).

More generally, the problem is called manipulation of the running variable. As with the word “treatment,” “manipulate” should be interpreted broadly. In the scholarship example, students could not “manipulate their score” in the sense of cheating, but they could take actions to change their observed $Z$. In other contexts: somebody could “manipulate” their observed earnings (by adjusting work hours, lying, etc.) to become eligible for a means-tested government program, or a family could move to a better school district (or to benefit from other geography-based policies), and so on. Formally, manipulation likely violates A21.2 and/or A21.4.

An implication of manipulation (though not manipulation itself) is tested by the McCrary (2008) test. Consider again the scholarship example, where $X = 1 \{Z \geq z_0\}$. If there is manipulation, then there should be many scores of exactly $Z = z_0$, and very few with $Z = z_0 - 1$. Thus, the distribution of $Z$ can alert us to manipulation, if it is discontinuous at $z_0$. See McCrary (2008) for details.

For additional specification tests and such, see Section 7 of Imbens and Lemieux (2008).

Discussion Question 21.6 (RD: manipulation). Imagine $Z$ is the longitude (east-west location) of an individual’s house. (To abstract some: assume their office, kids’ school, etc., also share the exact same $Z$.) Let $z_0$ be the location of a time zone boundary. Let $Y$ be hours of sleep per night. So, $y^+$ is the average sleep hours for people just east of the boundary, and $y^-$ is the average sleep hours for people just west of the boundary. Both people see the same sunset, but the clocks on the west side read one hour earlier than the clocks on the east side. According to the clocks, people on the east side have one more hour of daylight in the evening but the sun rises one hour later.

a) Without worrying about manipulation: what could we learn from $y^+ - y^-$ (the sharp RD estimator)?

b) What type of person would want to live just east of the boundary?

---

4This has been done; e.g., [https://doi.org/10.1016/j.jhealeco.2019.03.007](https://doi.org/10.1016/j.jhealeco.2019.03.007)
c) Given (b), why/how might/not the RD estimator be biased?  

d) Would the distribution of $Z$ show evidence of “manipulation”? Why/not?  

21.7 Estimation  

Consistent estimation of the RD estimand requires consistent estimators of $y^+, y^-, x^+$, and $x^-$. (Only $y^+$ and $y^-$ are needed for sharp RD since $x^+ - x^- = 1$.) From the definitions in (21.2) and (21.3), these are all certain points on certain CEFs. The relevant CEFs are $E(Y \mid Z = z)$ and $E(X \mid Z = z)$. The point of evaluation initially appears to be the interior point $z_0$. However, due to the discontinuity at $z_0$, it behaves like a boundary point. For example, when estimating $y^+$ or $x^+$, observations with $Z_i \geq z_0$ are useful but those with $Z_i < z_0$ are useless. Similarly, for $y^-$ or $x^-$, only observations with $Z_i < z_0$ are used, so effectively $z_0$ is a boundary point.  

For nonparametric CEF estimation at a boundary point, local linear estimation works well. The local constant estimator has notably more bias. Local cubic would be fine, too. Chapter 16 describes local linear and local polynomial estimation.  

Method 21.1 (RD estimation). Estimate $x^+$, $x^-$, $y^+$, and $y^-$ as defined in (21.2) and (21.3), and plug into the expression from Theorem 21.1, Theorem 21.2, or Theorem 21.3 depending on your assumptions.  

Corollary 21.4 (RD consistency). Method 21.1 is consistent for the respective objects of interest in Theorems 21.1–21.3 under the respective assumptions, assuming the respective nonparametric estimators of $x^+$, $x^-$, $y^+$, and $y^-$ are consistent.  

Discussion Question 21.7 (RD: bandwidth). Imagine $Z$ is household income, and families are eligible for some government program if $Z < z_0 = $50,000/yr. You run an RD estimator using local linear regression; the function reports that the optimal bandwidth is $h = $21,498/yr, i.e., $y^-$ and $x^-$ are estimated by OLS on observations with $z_0 - h \leq Z_i < z_0$, and $y^+$ and $x^+$ are estimated by OLS on observations with $z_0 \leq Z_i \leq z_0 + h$.  

a) Is this estimator consistent? Why/not?  

b) Explain any relationship between this $h$ and the $\delta$ from A21.6 including why it is/not ok to have $h > \delta$.  

21.8 Inference  

See links at beginning of chapter.
Exercises

Exercise E21.1.  a. Find a published paper with RD results.

b. Replicate one of their main effect estimates and confidence intervals. (Using their code is fine and encouraged.)

c. Test for manipulation of the running variable.

d. Construct a confidence interval using a method not used in the original code. Compare with the CI that you replicated.
Appendix to Chapter 21

21.A Proofs

Proof of Theorem 21.1. From A21.3 and (21.1), $Y = A + XB = A + X\beta$. For $e > 0$,

$$
E[Y | Z = z_0 + e] - E[Y | Z = z_0 - e] = E[A + X\beta | Z = z_0 + e] - E[A + X\beta | Z = z_0 - e] = \{E[A | Z = z_0 + e] - E[A | Z = z_0 - e]\} + \beta \{E[X | Z = z_0 + e] - E[X | Z = z_0 - e]\}.
$$

Taking the limit at $e \downarrow 0$ and solving for $\beta$ yields the result. For the LHS, by (21.3),

$$
\lim_{e \downarrow 0} \{E[Y | Z = z_0 + e] - E[Y | Z = z_0 - e]\} = y^+ - y^-.
$$

For the first RHS term, by A21.2,

$$
\lim_{e \downarrow 0} \{E[A | Z = z_0 + e] - E[A | Z = z_0 - e]\} = E[A | Z = z_0] - E[A | Z = z_0] = 0.
$$

For the second RHS term, by (21.2),

$$
\lim_{e \downarrow 0} \{E[XB | Z = z_0 + e] - E[XB | Z = z_0 - e]\} = x^+ - x^-.
$$

Thus, altogether,

$$
y^+ - y^- = 0 + \beta \{x^+ - x^-\} \implies \beta = (y^+ - y^-)/(x^+ - x^-),
$$

(21.8)

where the denominator is assumed non-zero by A21.1.

Proof of Theorem 21.2. Starting like the proof of Theorem 21.1,

$$
E[Y | Z = z_0 + e] - E[Y | Z = z_0 - e] = E[A + X\beta | Z = z_0 + e] - E[A + X\beta | Z = z_0 - e] = \{E[A | Z = z_0 + e] - E[A | Z = z_0 - e]\} + \{E[XB | Z = z_0 + e] - E[XB | Z = z_0 - e]\}.
$$
Continuing to follow the proof of Theorem 21.1, using (21.3) and A21.2 and taking \( e \downarrow 0 \),
\[
y^+ - y^- = 0 + \lim_{e \downarrow 0} \{E[XB \mid Z = z_0 + e] - E[XB \mid Z = z_0 - e]\}.
\] (21.9)

As \( e \downarrow 0 \), eventually \( z_0 + e \) and \( z_0 - e \) enter (and remain inside) the neighborhood defined in A21.5. Thus, by A21.5, \( E[B \mid Z = z_0 + e, X] = E[B \mid Z = z_0 + e] \), and same for \( z_0 - e \). Using the law of iterated expectations,
\[
E[XB \mid Z = z_0 + e] = E\{E[XB \mid Z = z_0 + e, X] \mid Z = z_0 + e\} = E\{X E[B \mid Z = z_0 + e, X] \mid Z = z_0 + e\} = E\{X E[B \mid Z = z_0 + e] \mid Z = z_0 + e\} = E[B \mid Z = z_0 + e]E[X \mid Z = z_0 + e],
\] (21.10)
and same for \( z_0 - e \). By A21.4
\[
\lim_{e \downarrow 0} E[B \mid Z = z_0 \pm e] = E[B \mid Z = z_0].
\]

As in the proof of Theorem 21.1 by A21.1
\[
\lim_{e \downarrow 0} E[X \mid Z = z_0 + e] = x^+, \quad \lim_{e \downarrow 0} E[X \mid Z = z_0 - e] = x^-.
\] (21.11)

Applying (21.9) and (21.10) to (21.9),
\[
y^+ - y^- = E[B \mid Z = z_0](x^+ - x^-) \implies E[B \mid Z = z_0] = (y^+ - y^-)/(x^+ - x^-),
\] (21.12)
where the denominator is assumed non-zero by A21.1.

Proof of Theorem 21.3. Write \( Y = Y_0 + X(Z)B \), so
\[
E[Y \mid Z = z] = E[Y_0 + X(Z)B \mid Z = z] = E[Y_0 \mid Z = z] + E[X(z)B \mid Z = z].
\]

For \( z' < z'' \),
\[
E[Y \mid Z = z''] - E[Y \mid Z = z'] = \{E[Y_0 \mid Z = z''] - E[Y_0 \mid Z = z']\} + \{E[X(z'')B \mid Z = z''] - E[X(z')B \mid Z = z']\}.
\] (21.13)

Below, the second RHS term is simplified using the assumptions, and then limits are taken as \( e \downarrow 0 \) with \( z' = z_0 - e \) and \( z'' = z_0 + e \).

As in A21.6 let \( z_0 - \delta < z' < z'' < z_0 + \delta \). From A21.6, the distribution of \( X(r)B \mid Z = z \) is the same for any \( z \in \mathcal{N}_\delta \). Then,
\[
E[X(z'')B \mid Z = z''] - E[X(z')B \mid Z = z'] = E[X(z'')B \mid Z \in \mathcal{N}_\delta] - E[X(z')B \mid Z \in \mathcal{N}_\delta] = E\{[X(z'') - X(z')]B \mid Z \in \mathcal{N}_\delta\}.
\] (21.14)
21.A. PROOFS

Since \( X \in \{0, 1\} \), there are four possible pairs of values of \((X(z'), X(z''))\): \((0, 0)\), \((1, 1)\), \((0, 1)\), and \((1, 0)\). However, the monotonicity in \([21.6]\) precludes \((1, 0)\). Since \(1-1 = 0-0 = 0\), this leaves only two possible values of \(X(z'') - X(z')\): zero and one. If it’s zero, then \([X(z'') - X(z')]B = 0\), regardless of \(B\). If it’s one, then \([X(z'') - X(z')]B = B\).

Thus,

\[ E[(X(z'') - X(z'))B \mid Z \in \mathcal{N}_\delta] = P(X(z'') - X(z') = 1 \mid Z \in \mathcal{N}_\delta) \times E[B \mid X(z'') - X(z') = 1, Z \in \mathcal{N}_\delta]. \tag{21.15} \]

Plugging (21.15) into (21.14), and plugging that into (21.13),

\[
E[Y \mid Z = z''] - E[Y \mid Z = z'] = \{E[Y_0 \mid Z = z''] - E[Y_0 \mid Z = z']\}
+ P(X(z'') - X(z') = 1 \mid Z \in \mathcal{N}_\delta) \times E[B \mid X(z'') - X(z') = 1, Z \in \mathcal{N}_\delta]. \tag{21.16}
\]

Further,

\[
E[X(Z) \mid Z = z''] - E[X(Z) \mid Z = z'] = E[X(z'') \mid Z = z''] - E[X(z') \mid Z = z']
= E[X(z'') \mid Z \in \mathcal{N}_\delta] - E[X(z') \mid Z \in \mathcal{N}_\delta]
= E[X(z'') - X(z') \mid Z \in \mathcal{N}_\delta].
\]

Applying the same logic as for (21.15), this further simplifies; altogether,

\[ E[X(Z) \mid Z = z''] - E[X(Z) \mid Z = z'] = P(X(z'') - X(z') = 1 \mid Z \in \mathcal{N}_\delta). \tag{21.17} \]

Plugging (21.17) into (21.16) and dividing through,

\[
\frac{E[Y \mid Z = z''] - E[Y \mid Z = z']}{E[X \mid Z = z''] - E[X \mid Z = z']} = \frac{E[Y_0 \mid Z = z''] - E[Y_0 \mid Z = z']}{E[X \mid Z = z''] - E[X \mid Z = z']}
+ E[B \mid X(z'') - X(z') = 1, Z \in \mathcal{N}_\delta]. \tag{21.18}
\]

Plugging in \(z' = z_0 - e\) and \(z'' = z_0 + e\) and taking \(e \downarrow 0\), the numerator of the first RHS term goes to zero by [21.2] as in the proofs of Theorems [21.1] and [21.2]. Also as in those proofs, the limit of the LHS is \((y^+ - y^-)/(x^+ - x^-)\), yielding the result. \(\square\)
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? First: why do we assume we know the joint distribution of all observable variables? With iid sampling, this distribution can be learned asymptotically (essentially a vector version of the Glivenko–Cantelli theorem in Section 8.2). That is, with large enough $n$, we essentially “know” the joint distribution of observables. However, that may not be enough to learn $\theta$. For example, “correlation does not imply causation”: without additional assumptions, we cannot learn about causal parameters $\theta$ from regression slopes or even the full joint distribution of $(Y, X)$. With other 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 can be learned asymptotically, so they are a more reasonable starting place for identification.

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 $\theta$. For example, the population median is identified: the distribution of $Y$ uniquely determines the median $Q_{0.5}(Y)$.

If the population distribution of observables does not uniquely determine $\theta$, but we can still learn something about $\theta$, then it’s called **partial identification** or **set identification**. (I’ve been convinced that “set identification” is the better 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 $\theta$ 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. Below, I try to 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 22

Missing Data

Optional resources for this chapter

• MAR: 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/]

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.

For notation, imagine observations $i = 1, \ldots, n$, with $Y_i$ and $X_i$ as usual. Let $S_i = 1$ if both $Y_i$ and $X_i$ are observed for individual $i$, and $S_i = 0$ if either $Y_i$ or $X_i$ 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 $(Y_i, X_i)$.

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., 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,” a range of estimates (or “bounds”) can be computed. 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 data to be missing, and the “bounds” approach is explored.
CHAPTER 22. MISSING DATA

22.1 Best Case: MCAR

When does complete case analysis provide consistent estimators? It depends on the type of missing data and on the object of interest.

For example, data missing completely at random (MCAR) means whether or not a value is missing is unrelated to either \( Y \) or \( X \). In our notation,

\[
S_i \perp \perp Y_i, X_i \quad \text{or equivalently} \quad P(S_i = 1 \mid Y_i, X_i) = P(S_i = 1).
\] (22.1)

If we simply remove observations with \( S_i = 0 \) and use only observations with \( S_i = 1 \), then the missing data does not cause any bias. (However, it decreases the sample size, which increases variance; there is potential for imputation to improve precision.) Selecting observations with \( S_i = 1 \) is essentially taking a random sample from within our original random sample. With MCAR, it does not matter what we want to estimate; any estimator that would work for the full sample also works for the complete case subsample.

Figure 22.1 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 \( X_i \), but not everyone reports \( Y_i \); but the decision whether or not to report \( Y_i \) is unrelated to \( Y_i \) or \( X_i \), i.e., MCAR holds as in (22.1). 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 full sample including both observed and unobserved \( Y_i \) values; this is the data we would have had if everybody reported both \( X_i \) and \( Y_i \). The black OLS estimate and sample mean are thus what we would estimate if there were no missing data problem; this is not feasible in practice, but it is a helpful benchmark for comparison. The blue (complete case) and black estimates are extremely similar, illustrating the lack of bias caused by MCAR.
22.2 Fixable: MAR

Without MCAR, things are more complicated, so we’ll consider a simple example. Let $Y$ denote an individual’s annual earnings; $X = 1$ indicates a college degree, $X = 0$ no degree. We have survey data, $(Y_i, X_i)$ for $i = 1, \ldots, n$, but some people skipped the earnings question, i.e., $Y_i$ may be missing. We define $S_i = 1$ if both $Y_i$ and $X_i$ are observed, and $S_i = 0$ if $Y_i$ is missing.

Consider a weaker, conditional version of MCAR, where independence only holds within each subpopulation. Formally,

$$ S \perp Y \mid X \quad (22.2) $$

(“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 for the $X = 1$ subpopulation. This is related to the concept of missing at random (MAR), where missingness ($S = 0$) is random (unrelated to the missing values) conditional on certain variables that are always observed.

22.2.1 Complete Case Estimation

Given (22.2), can we consistently estimate $E(Y)$ by averaging only the observed $Y_i$ values? Figure 22.2 (right panel) illustrates the following example in which the complete case sample mean is biased. Imagine everyone with $X_i = 0$ reports $Y_i$, but many individuals with $X_i = 1$ do not report $Y_i$. Since individuals with a college degree tend to have higher earnings, the missing $Y_i$ values tend to be high. If we only average observed $Y_i$, then our estimate of $E(Y)$ has downward bias. That is, if we were able to observe everybody’s $Y_i$, then our sample mean would be higher. Even though missingness is random within the $X_i = 1$ group, there is still bias.

Continuing the same example, what if instead we want to regress $Y$ on $X$ using OLS? Figure 22.2 shows that this is fine. In the example, no data are missing when $X = 0$, so
we focus on the \( X = 1 \) (college) subpopulation, but the same arguments would apply to \( X = 0 \). From (22.2), within the \( X = 1 \) subpopulation, missingness of \( Y \) is unrelated to the value of \( Y \), so the complete case subpopulation mean earnings \( \mathbb{E}(Y \mid X = 1, S = 1) \) is the same as the true subpopulation mean earnings \( \mathbb{E}(Y \mid X = 1) \). Mathematically, \( S \perp Y \mid X \) means that once we condition on \( X \), the distribution of \( Y \) does not depend on the value of \( S \). That is, the conditional distribution of \( Y \) given \( X = 1 \) among individuals in the population who would report their earnings \((S = 1)\) is identical to the conditional distribution of \( Y \) given \( X = 1 \) among individuals who would not report their earnings \((S = 0)\). Thus, \[
\text{CEF among "reporters"} = \text{CEF for everyone}
\]
\[
\mathbb{E}(Y \mid X = 1, S = 1) = \mathbb{E}(Y \mid X = 1).
\] (22.3)

Similarly, \( \mathbb{E}(Y \mid X = 0, S = 1) = \mathbb{E}(Y \mid X = 0) \).

The result in (22.3) is an identification result. We cannot directly estimate the CEF \( \mathbb{E}(Y) \) because the random variable \( Y \) is not observed for everybody. This is similar to the problem of not being able to directly estimate the ATE \( \mathbb{E}(Y^1 - Y^0) \) because either \( Y^1 \) or \( Y^0 \) (the potential outcomes) is always unobserved. That is, ATE identification can be seen as a missing data problem: either \( Y^0 \) or \( Y^1 \) is missing for every individual. Just like in that case, we need an identification result to link the population object \( \mathbb{E}(Y \mid X = x) \), which involves a variable \( Y \) that is not always observed, to another population object that involves only observed variables. This is what (22.3) shows.

In sum, given (22.2), complete case OLS can consistently estimate the population CEF. First, complete case analysis can consistently estimate the CEF among individuals who report their earnings \((S = 1)\), i.e., \( \mathbb{E}(Y \mid X = x, S = 1) \). Second, since (22.2) equates this CEF with the overall CEF \( \mathbb{E}(Y \mid X = x) \), complete case analysis also consistently estimates the overall CEF.

### 22.2.2 Inverse Probability Weighting

It seems suspicious that we can consistently estimate all the conditional means of \( Y \) but not the unconditional mean. In fact, although the complete case sample average is not a consistent estimator of \( \mathbb{E}(Y) \), the conditional sample averages can be combined to consistently estimate \( \mathbb{E}(Y) \). The key is another identification argument:

\[
\mathbb{E}(Y) = \mathbb{E}(Y \mid X = 0) P(X = 0) + \mathbb{E}(Y \mid X = 1) P(X = 1)
= \mathbb{E}(Y \mid X = 0, S = 1) P(X = 0) + \mathbb{E}(Y \mid X = 1, S = 1) P(X = 1).
\] (22.4)

All four terms on the right-hand side are observable since \( X \) is always observable (and \( Y \) is always observable conditional on \( S = 1 \)). The probability \( P(X = 1) \) 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 \( P(X = 0) \). As noted before, complete case OLS can estimate \( \mathbb{E}(Y \mid X = 0, S = 1) \) and \( \mathbb{E}(Y \mid X = 1, S = 1) \), with \( \hat{\beta}_0 \) and \( \hat{\beta}_0 + \hat{\beta}_1 \), respectively. Then we can just plug in the estimates of the four right-hand side terms in (22.4) to estimate \( \mathbb{E}(Y) \).
22.2. FIXABLE: MAR

This approach is a special case of inverse probability weighting (IPW). Consider Figure 22.2. 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

\[
\frac{1}{8} (Y_1 + Y_2 + Y_3 + Y_4 + Y_5 + Y_6 + Y_7 + Y_8) = \frac{1}{8} (Y_1 + Y_2 + Y_3 + Y_4 + 4Y_5).
\]

(22.5)

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.

The weight of 4 on \( Y_5 \) in (22.5) 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,

\[
\hat{P}(S = 1 \mid X = 1) = \frac{1}{4}, \quad \frac{1}{\hat{P}(S = 1 \mid X = 1)} = 4,
\]

(22.6)

where the inverse probability 4 is the weight that appears in (22.5). The expression \( \hat{P}(S = 1 \mid X = 1) \) is the estimated probability of having an observable \( Y \) value \((S = 1)\) given \( X = 1 \).

More generally, the IPW estimator of \( E(Y) \) is

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{S_i Y_i}{\hat{P}(S = 1 \mid X = X_i)}.
\]

(22.7)

22.2.3 Linear Projection Estimation

If instead of a single binary \( X \) there is general vector \( X \), then complete case OLS may not be consistent for the population linear projection. Generally, OLS only estimates the linear projection, not the CEF (e.g., Hansen 2020a, §2.18). In that case, even if (22.2) holds, different amounts of missing \( Y_i \) can lead to (very) different OLS slope estimates.

For example, imagine the possible values of \( X \) are \( X = 0, 1, 2 \), and the CEF values are \( m(0) = m(1) = 40, m(2) = 60 \). Since (22.2) does not restrict the relationship between \( X \) and \( S \), it could be that everybody with \( X_i = 0 \) or \( X_i = 1 \) reports her \( Y_i \), but nobody with \( X_i = 2 \) does. In that case, OLS estimates a slope near zero since \( m(0) = m(1) \). However, it could also be that nobody with \( X_i = 0 \) reports \( Y_i \), but everybody with \( X_i = 1 \) or \( X_i = 2 \)
does. In that case, OLS estimates a slope near 20 since \( m(2) - m(1) = 60 - 40 = 20 \). Both examples satisfy (22.2), but clearly the OLS \( \hat{\beta}_1 \) is very different in each case. Analogously, in the population, the linear projection with only observable \((Y, X)\) usually differs from the linear projection with all \((Y, X)\). That is, even in the population, conditioning on \( S = 1 \) affects the linear projection if \( X \) has a different distribution in the \( S = 1 \) and \( S = 0 \) subpopulations. Put differently, if we have a misspecified CEF model (that’s really just a linear projection model), then complete case OLS is sensitive to missing data.

However, nonparametric CEF estimators are not subject to this problem. They are actually estimating the CEF, not just the linear projection. Unlike the linear projection, the CEF is not sensitive to the marginal distribution of \( X \).

22.3 Worst Case: Non-Ignorable

The term “non-ignorable” suggests we can’t ignore the missing data problem. Section 22.3.1 explains why, and Section 22.3.2 suggests one way to cope.

22.3.1 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.

For example, imagine \( X = 1 \) if somebody has a college degree and \( X = 0 \) otherwise, and \( Y \) is annual earnings, which is generally higher when \( X = 1 \) than \( X = 0 \). Imagine 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. Since most people with very high earnings (\( Y \)) have a college degree (\( X = 1 \)), this disproportionately affects our estimate of \( E(Y \mid X = 1) \). Specifically, our estimate is lower than it would be if we were able to see all the very high values; i.e., there is downward (negative) bias. We cannot predict the missing \( Y \) values using \( X \), because our predictions would also be biased downward. Perhaps also people with very low earnings are less likely to report \( Y \). These are disproportionately \( X = 0 \) individuals, so this biases our estimate of \( E(Y \mid X = 0) \) upward (positive bias). If the estimate of \( E(Y \mid X = 0) \) is biased upward and/or the estimate of \( E(Y \mid X = 1) \) is biased downward, then the estimate of the slope \( \beta_1 = E(Y \mid X = 1) - E(Y \mid X = 0) \) is biased downward.

Figure 22.3 illustrates the case where the highest-earning individuals do not report \( Y_i \) but everyone else does. This mostly affects \( X_i = 1 \) individuals, but also the very highest \( Y_i \) in the no-college group. With complete case analysis, both the OLS slope and sample mean are biased downward. The OLS intercept is very slightly downward biased, too, since the top \( Y_i \) when \( X_i = 0 \) are missing. Since we cannot consistently estimate the true CEF, the IPW approach cannot consistently estimate the population mean, either.
22.3. WORST CASE: NON-IGNORABLE

Methods to perfectly fix the problem of non-ignorable missing data are beyond our scope. However, one way to address the problem is discussed next.

22.3.2 Worst-Case Bounds

Discussion Question 22.1 (point ID w/ missing data: mean). Let scalar rv $Y^*$ have CDF $F^*_{Y}(.).$ Let $S = 1$ if $Y^*$ is observed, and $S = 0$ if $Y^*$ is missing. Individuals are randomly sample 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 $S_i = 0$.

a) If $P(S = 1) = 1$, then is $E(Y^*)$ point identified? Why/not? (Be precise even if you think it’s obvious!) Hint: $E(A) = E(A | B = 1) P(B = 1) + E(A | B = 0) P(B = 0)$.

b) Same as [a] but if $P(S = 1) < 1$.

Discussion Question 22.2 (point ID w/ missing data: median). Same as DQ 22.1 but for the median of $Y^*$ instead of the mean.

Identification refers to population distributions and values, but some intuition can be developed thinking about samples. (Recall samples can be thought of as discrete population distributions.) Hence, DQs 22.3 and 22.4 is not about “identification” per se but hopefully illuminates a relevant concept.

Discussion Question 22.3 (bounds for binary sample mean). Let binary unobserved $Y^* = 1\{\text{employed}\}$. You observe $n = 5$ values of $Y_i$: $(1, 1, 0, 1, \text{NA})$. (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 $\bar{Y}^* = (1/5)(Y^*_1 + Y^*_2 + Y^*_3 + Y^*_4 + Y^*_5)$?

d) What’s the largest possible value of $\bar{Y}^*$?

Figure 22.3: Non-ignorable missing data: bias of both OLS and sample mean.
Discussion Question 22.4 (bounds for sample mean education). Let unobserved $Y^*$ be years of education, where $0 \leq Y^* \leq 21$. You observe $n = 5$ values of $Y$: $(12, 12, 11, 18, \text{NA})$. (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 $\bar{Y}^* \equiv \frac{1}{5}(Y_1^* + Y_2^* + Y_3^* + Y_4^* + Y_5^*)$?

d) What’s the largest possible value of $\bar{Y}^*$?

Discussion Question 22.5 (bounds for sample median). Consider the “true” sample median $\hat{Q}_{0.5}(Y^*)$. (With $n = 5$, this is the “middle” observation when sorted from low to high.)

a) In the setup of DQ 22.3, what are the smallest and largest possible values of the sample median of $Y^*$?

b) In the setup of DQ 22.4, 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 $Y_5 = \text{NA}$ (missing). Is it still possible to get a lower and/or upper bound for the sample median? How?

The type of bounds in DQs 22.3 and 22.4 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 22.6 tries to use our insights from DQ 22.3 to find bounds for the population employment probability, i.e., find a partial identification result.

Discussion Question 22.6 (bounds for binary population mean). Same as DQ 22.3, but in the population (which is trickier). As usual, we assume we know the joint distribution of all (fully) observable variables, i.e., the joint distribution of $(Y, S)$ but not $Y^*$. Recall that this implies we also know the marginal and conditional distributions of $Y$ and $S$. What are the worst-case bounds on $P(Y^* = 1)$, i.e., bounds that do not assume anything about why/how data are missing? Hint: $E(Y^*) = E(Y^* | S = 1)P(S = 1) + E(Y^* | S = 0)P(S = 0)$. 
Discussion Question 22.7 (bounds for population mean). Let $Y^* \in \mathbb{R}$; the joint population distribution of $(Y, S)$ is known.

a) What are bounds on $E(Y^*)$ if $Y^*$ is a binary employment indicator? Hint: recall DQ 22.3

b) What are bounds on $E(Y^*)$ if $Y^*$ is years of education? Hint: recall DQ 22.4

c) Same but $Y^*$ is consumption.

d) Same but for any variable $Y^*$.

Discussion Question 22.8 (bounds for population median). Let $Y^* \in \mathbb{R}$; the joint population distribution of $(Y, S)$ is known. Hint: recall DQ 22.5

a) What are bounds on $Q_{0.5}(Y^*)$ 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 $Y^*$.

Discussion Question 22.9 (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 $P(W^* = 1)$?

b) Bounds on $P(Y^* \leq y)$?

c) Bounds on $F^*_Y(\cdot)$? (A lower bound for a function is itself a function, as is the upper bound function.)

Discussion Question 22.10 (IQR bounds). Assume $F_1(\cdot) \leq F^*(\cdot) \leq F_2(\cdot)$.

a) What are bounds for $Q_{0.75}(Y^*)$?

b) Bounds for $Q_{0.25}(Y^*)$?

c) Bounds for the interquartile range, $Q_{0.75}(Y^*) - Q_{0.25}(Y^*)$?

Stoye [2010] gives more results like DQ 22.10

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.
22.4 Summary

We looked at a single, simple example, and saw the effect of different reasons why data might be missing. With MCAR, there was no bias in complete case analysis. With the MAR-type condition, the sample mean was biased, and OLS estimates of the population linear projection could be biased, but CEF estimation was fine (and could be used to estimate the sample mean with the IPW approach). With non-ignorable missingness, everything was biased, but bounds could be computed. So, if a significant proportion of observations have missing values, it is important to think critically about why those values are missing before just doing complete case analysis.

22.5 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.

```r
n <- 5
Y <- rnorm(n); X <- rnorm(n)
r <- lm(Y~X)       # no hint of missing/dropped obs
coef(r)            # still no hint

## (Intercept)     X
## -/zero.alt3./zero.alt3412 /zero.alt3.9548

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
## 1.691 0.358 -0.808
```
22.5. R CODE

```r
predict(lm(Y~X, na.action=na.exclude))  # fill in NA
##  1  2  3  4  5
## 1.691 NA NA 0.358 -0.808

lm(Y~X, na.action=na.fail)  # give an error if NAs in data
## Error in na.fail.default(structure(list(Y = c(1.28856597618863, NA, 0.72210318661022, :
##  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(structure(list(Y = c(1.28856597618863, NA, 0.72210318661022, :
##  missing values in object
```
Exercises

**Exercise E22.1.** 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 $\hat{Q}_{0.5}(Y^*)$? 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, $\hat{\beta}_{\text{OLS}} = \hat{E}(Y \mid X = 1) - \hat{E}(Y \mid X = 0)$.

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, $\hat{\beta}_\tau = \hat{Q}_\tau(Y \mid X = 1) - \hat{Q}_\tau(Y \mid X = 0)$.

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 23

Interval Data

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 23.1** (interval-valued mean: identification). The following is somewhat similar to DQs 22.1 and 22.6. Continue 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 $E(Y^*)$ point identified from the distribution of $(Y_1, Y_2)$? Why not?
   b) What are bounds for $E(Y^*)$ given the distribution of $(Y_1, Y_2)$?

**Discussion Question 23.2** (interval-valued mean: estimation). Continue the setup and notation from DQ 23.1.
   a) Given DQ 23.1, 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 23.3** (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 $m(x) = E(Y^* | X = x)$ for a single $x$?
   b) What are bounds for the CEF (as a function), $m(\cdot)$?

**Discussion Question 23.4** (interval regression: estimation). Continue from DQ 23.3.
   a) How could you estimate your proposed bounds from DQ 23.3?
   b) How do you interpret your estimated bounds? Is it like a uniform confidence band, or something else?
**Discussion Question 23.5** (interval regression slope: identification). Use your results from DQ 23.3 below. Let $x_2 > x_1$ be two points of evaluation. Assume $m(\cdot)$ is continuously differentiable.

a) What are bounds for $E(Y^* \mid X = x_2) - E(Y^* \mid X = x_1)$?

b) What are the limits of your bounds as $x_2 \downarrow x_1$?

c) What are bounds for $m'(x)$? Hint: $m'(x) = \lim_{x_2 \downarrow x_1} \frac{m(x_2) - m(x_1)}{(x_2 - x_1)}$. 


EXERCISES

Exercises

Exercise E23.1. Assume scalar $X \geq 0$. Define vector $X = (1, X)'$. The observables are $(Y_1, Y_2, X)$ with $Y_1 \leq Y^* \leq Y_2$ for latent $Y^*$. Assume iid sampling. For vectors $a$ and $b$, let $a \leq b$ mean that the inequality holds element-wise, i.e., $a_j \leq b_j$ for each $j = 1, \ldots, \dim(a)$.

a. Consider the linear projection model in error form, $Y^* = X'\beta + U$ with $E(XU) = 0$, where $\beta = (\beta_0, \beta_1)'$. Show that $E(XY^*) = E(XX')\beta$.

b. Propose bounds for $E(XY^*)$ using any feature(s) of the population joint distribution of observables. Hint: separately consider the two elements of the vector $E(XY^*) = [E(Y^*), E(XY^*)]'$, and remember $X \geq 0$.

c. Denote your bounds in part (b) as $\mu_L$ and $\mu_U$, where $\mu_L \leq E(XY^*) \leq \mu_U$. Given part (a), your bounds imply

$$\mu_L \leq E(XX')\beta \leq \mu_U.$$  (23.1)

Re-write (23.1) 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 $\beta_0$, the vertical axis is $\beta_1$; 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 $\beta_1$, can we get bounds on $\beta_1$ using this approach? You don’t have to derive such bounds, just explain why or why not.

f. With $\beta \in \mathbb{R}^2$, the bounds generate four lines in $\mathbb{R}^2$ that determine a quadrilateral subset of $\mathbb{R}^2$ containing all values of $\beta$ consistent with the joint distribution of observables, i.e., the identified set for $\beta$. Without necessarily solving for the bounds exactly, what is the corresponding geometry of the bounds when instead $\beta \in \mathbb{R}^k$ for general $k > 2$? Is it still possible to get bounds for a single coefficient like $\beta_1$? Think about the structure of the generalization of (23.1) in that case; how many equations, what shape, etc. Hint: to develop intuition, you could start with $k = 3$.

Exercise E23.2. 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 $\tau$-CQF as $q_\tau(\cdot)$, where $q_\tau(x) = Q_\tau(Y^* \mid X = x)$.

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, \[ \frac{E(Y^* \mid Z = 1) - E(Y^* \mid Z = 0)}{E(X \mid Z = 1) - E(X \mid Z = 0)} \]. For simplicity, assume the denominator is strictly positive. Propose bounds for the LATE.

g. Propose estimators of your LATE bounds.
Chapter 24

Ordinal Data

Much of this chapter is based on Kaplan and Zhuo (2019). R code: http://faculty.missouri.edu/kaplandm

24.1 Latent Variable Framework

Imagine you observe an ordinal random variable $H$: its value is one of multiple 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 (SRHS), where $H = 1$ means “poor” health, $H = 2$ means “fair,” $H = 3$ means “good,” $H = 4$ means “very good,” and $H = 5$ means “excellent.” Again, the numbers have no cardinal meaning; it’s just shorter than writing out “very good” each time. Other ordinal variables include bond ratings, self-reported happiness, political indexes, school ratings, etc.

Imagine the health category $H$ is reported based on an underlying, unobserved (latent), continuously distributed variable $H^*$. There are thresholds $\gamma_j$ with $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_J = \infty$ such that $H = j$ iff $\gamma_{j-1} < H^* \leq \gamma_j$. In the running example, there are $J = 5$ categories.

The following CDF notation is used. The CDF of $H$ is $F(\cdot)$. That is, $F(j) = P(H \leq j)$ for $j = 1, \ldots, J$, with $F(J) = 1$ by definition. The CDF of $H^*$ is $F^*(\cdot)$. That is, $F^*(r) = P(H^* \leq r)$ for $r \in \mathbb{R}$.

The CDF of $H$ is $F(\cdot)$, and the CDF of $H^*$ is $F^*(\cdot)$. That is,

$$F(j) \equiv P(H \leq j) \text{ for } j \in \{1, \ldots, J\}, \quad F^*(r) \equiv P(H^* \leq r) \text{ for } r \in \mathbb{R}. \quad (24.1)$$

Discussion Question 24.1 (ordinal: known thresholds 1). Consider the health example above, with the same notation. Assume (unrealistically) that the $\gamma_j$ are known.

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 $F(1) = F^*(\gamma_1)$. Hint: write out $F(1)$ as a probability involving $H$, and then write out $F^*(\gamma_1)$ as a probability involving $H^*$.  

273
c) Is $F^*(r)$ point identified (i.e., uniquely determined by $F(\cdot)$) for any other $r$? Explain.

**Discussion Question 24.2** (ordinal: known thresholds 2). Continue from DQ \[24.1\]

a) Consider $r < \gamma_1$. 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 bounds for $F^*(r)$ than $[0, 1]$? 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 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 $r \in \mathbb{R}$.

Now assume we have two populations that we wish to compare. We observe iid samples from each independently, so for identification, assume we know the marginal distributions of the ordinal health variables, say $H$ and $G$.

Whether for health, income, or other variables, there are two types of “inequality” people talk about. One type is “between-group” inequality: is one group (subpopulation) “better” than another? For example, if somebody says there’s great health inequality across education levels, it means something like college-educated individuals are overall “healthier” than below-college-educated individuals, etc. Or if somebody says there’s great income inequality across occupational type, it means certain job types pay better (overall) than other job types. Another type is “within-group” inequality: within a group (subpopulation), how different are the healthiest and least healthy individuals? For example, if somebody says there’s great health inequality in the U.S., it means (roughly) there are some people who are very healthy but some who are very unhealthy. Or if somebody says income inequality is high in the U.S., it means there’s a big spread between the upper end of the income distribution and the lower end. The two types of inequality are related: if there is much between-group inequality, then the population containing such groups has high within-group inequality. Sections \[24.2\] and \[24.3\] consider these two types of inequality in the latent variables.

### 24.2 Between-Group Inequality: Stochastic Dominance

Continuing the health example, imagine we want to learn if $H^*$ is “healthier” than $G^*$. That is, we wish to learn about between-group inequality. For now, assume the $\gamma_j$ are the same for both populations.

One definition of “healthier” is in terms of first-order stochastic dominance (SD1). (Recall Chapter \[10\]) SD1 of $G^*$ over $H^*$ is written $G^* \text{ SD}_1 H^*$; we say $G^*$ (first-order) stochastically dominates $H^*$. Usually (below) this means “weak” SD1, but it is impossible to distinguish weak and strong SD1 from data anyway, so the difference is not important here. There are three ways to characterize $G^* \text{ SD}_1 H^*$. First, $G^*$ has higher expected utility: $E[u(G^*)] \geq E[u(H^*)]$ 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
24.3. WITHIN-GROUP INEQUALITY: DISPERSION

275

r ∈ ℝ. Third, the quantiles of G* are all higher than the corresponding quantiles of H*:
\( Q_\tau(G^*) ≥ Q_\tau(H^*) \) for all \( \tau ∈ [0, 1] \), or \( Q_{G^*}(\cdot) ≥ Q_{H^*}(\cdot) \).

A weaker version of SD1 is restricted SD1. Being weaker makes it less helpful economically but more tractable statistically. In particular, it does not require knowledge of distributions’ tails, which are especially difficult to learn about statistically. The restricted SD1 concept is due to [Atkinson 1987, Condition I, p. 751]. If \( F_{G^*}(r) ≤ F_{H^*}(r) \) for all \( r ∈ [r^-, r^+] \), then there is restricted SD1 of G* over H* on the interval \([r^-, r^+]\). As a special case, if the interval is ℝ, then there is (unrestricted) SD1.

**Discussion Question 24.3** (testable implication). Continue to assume the same \( γ_j \) are used for both 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 \( F_G(j) ≤ F_H(j) \) for all \( j \), i.e., G SD1 H. Does this imply that G* SD1 H*?, i.e., that \( F_{G^*}(r) ≤ F_{H^*}(r) \) for all \( r ∈ ℝ \)? Why/not? Hint: focus on the range \([γ_1, γ_2]\), and which latent CDFs are consistent with the bounds implied by the ordinal CDF values \( F_G(1) = F_G(2) = F_H(1) = F_H(2) \).

b) Imagine \( F_G(r) ≤ F_{H^*}(r) \) for all \( r ∈ ℝ \), i.e., G* SD1 H*. Does this imply that G SD1 H, i.e., that \( F_G(j) ≤ F_H(j) \) for all \( j = 1, \ldots, J \)?

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 24.4** (super SD1). Imagine \( F_G(j + 1) ≤ F_H(j) \) for all \( j \). What (if anything) does this imply about restricted SD1 between G* and H* over the interval \([γ_1, γ_J]\)? Hint: draw a picture of bounds for the latent CDFs.

24.3 Within-Group Inequality: Dispersion

Now we wish to learn about the dispersion of latent health. In particular, we wish to learn if G* or H* is more dispersed (has more within-group inequality). Dispersion could be measured by variance, but variance is sensitive to the extreme tails, and the ordinal distributions of G and H contain no information about the tails of G* and H*. 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.

Continue to assume G and H share the same (unobservable) thresholds \( γ_j \). For simplicity of discussing quantiles, assume both \( F_G(\cdot) \) and \( F_H(\cdot) \) are strictly increasing, so the quantile function is the inverse CDF.

**Discussion Question 24.5** (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 $\tau_2$-quantile of continuous random variable $Y^*$, and you know its $\tau_1$-quantile, but nothing else. Given $\tau_1 \leq \tau \leq \tau_2$, explain why $Q_{\tau_1}(Y^*) \leq Q_{\tau}(Y^*) \leq Q_{\tau_2}(Y^*)$.

d) Let $F_G(1) < \tau < F_G(2)$. Using (a)–(c) provide bounds for the $\tau$-quantile of $G^*$.

The following are less directly related to DQ 24.6, but helpful for intuition.

e) What is the $F_H(j)$-quantile of $H^*$? Why?

f) If $F_G^c(r) \leq F_H^c(r)$ for all $r \in \mathbb{R}$, which distribution has the bigger $\tau$-quantile? Why?

Discussion Question 24.6 (single crossing). Your thoughts from DQ 24.5 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? Why?

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 $Q_b(Y^*) - Q_a(Y^*)$ is the $b$–$a$ interquantile range of $Y^*$. ) That is, show that $Q_{F_H(2)}(H^*) - Q_{F_H(1)}(H^*)$ is greater than $Q_{F_H(2)}(G^*) - Q_{F_H(1)}(G^*)$.

Discussion Question 24.7 (single crossing with shift). Now let $G$ and $H$ have different thresholds. The thresholds for $G$ are $\gamma_{jG}$, and the $H$ thresholds are $\gamma_{jH} = \gamma_{jG} + \Delta$, where possibly $\Delta \neq 0$ but $\Delta$ does not depend on $j$. Do the results from DQ 24.6 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?

Other variations of these results are provided by Kaplan and Zhuo (2019). Some impose additional restrictions, such as latent symmetry or a location–scale family. We also discuss frequentist and Bayesian inference and provide some code. “Regression” is also discussed but not detailed.

24.4 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 $Y = 1\{Y^* > 0\}$, 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 $\gamma_{j}$ 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. And since latent variables are by definition unobserved, it's impossible to learn the true shape of their distributions.

### 24.5 Inequality Indexes

There is a literature mostly (it seems) in health economics about computing an inequality index from an observed ordinal distribution. The advantage of this approach is you get a single number that (supposedly) measures how much inequality there is in the distribution. Thus, you can definitively compare any two ordinal distributions, as opposed to Kaplan and Zhuo (2019), who develop only a partial ordering (certain pairs of distributions cannot be ordered as having more and less inequality). However, there are many such indexes, 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 health variable has only $J$ possible different values (i.e., not continuous), which does not seem realistic.

There is an excellent Stata .ado program `ineqord` available from SSC that calculates a slew of ordinal inequality indexes.
Exercises

Exercise E24.1.  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 $X$ 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 $X$ 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 “$X$” is $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$.

c. Compare the two ordinal distributions with some of the methods from [Kaplan and Zhuo (2019)], optionally using the R code at [http://faculty.missouri.edu/kaplandm]. Most importantly, verbally interpret your results clearly in terms of the underlying latent distributions.

d. Discuss similarities and differences between the original results and your new results.
Bibliography


BIBLIOGRAPHY


Corrected 12th printing, January 13, 2017. 185 205 210 213
BIBLIOGRAPHY


Imbens, Guido W. and Thomas Lemieux. 2008. “Regression discontinuity designs: A guide to practice.” *Journal of Econometrics* 142 (2):615–635. URL [https://doi.org/10.1016/j.jeconom.2007.05.001](https://doi.org/10.1016/j.jeconom.2007.05.001) [239, 243, 245, 246]


Kaplan, David M. 2015. “Improved quantile inference via fixed-smoothing asymptotics


———. XXX. “Frequentist properties of Bayesian inequality tests.” Journal of Econometrics XXX (XXX):XXX. URL XXX


Knight, Keith and Wenjiang Fu. 2000. “Asymptotics for Lasso-Type Estimators.” Annals
BIBLIOGRAPHY

Analysis 91 (1):74–89. URL https://doi.org/10.1016/j.jmva.2004.05.006  
University Press. URL https://doi.org/10.1017/CBO9780511754098.  
Kolmogorov, Andrey Nikolaevich. 1933. “Sulla determinazione empirica di una legge di  
Konishi, Sadanori and Genshiro Kitagawa. 2008. Information Criteria and Statistical  
ed. URL http://books.google.com/books?id=Y7vSVW3ebSwC.  
Li, Qi and Jeffrey Scott Racine. 2007. Nonparametric Econometrics: Theory and Practice.  
Princeton University Press.  
Liu, Xin. 2020. “Panel Quantile Regression with Time-Invariant Rank.” Working paper,  
University of Missouri.  
BIBLIOGRAPHY


Index

k-fold cross-validation, 220

ABC, see approximate bootstrap confidence
additively separable, 83
AIC, see Akaike information criterion
Akaike information criterion, 221
analogy principle, 61, 130
Anderson–Darling, 98
approximate bootstrap confidence, 145
asymptotic loss efficiency, 223
ATE, see average treatment effect
average treatment effect, 78

bandwidth, 193
infeasible, 216
pilot, 217
plug-in, 217
basis, 209
basis function, 209
Bayesian information criterion, 222
BCa, see bootstrap
belief, 155
Bernoulli distribution, 158
Bernstein–von Mises theorems, 167
bias–variance tradeoff, 187
BIC, see Bayesian information criterion
bin, 192
Bonferroni adjustment, 118
Bonferroni correction, 118
bootstrap
Bayesian, 144
bias-corrected and accelerated, 145
circular block, 150
double, 141
empirical, 134
exchangeable weights, 143
m-out-of-n, 144
moving blocks, 149
multinomial, 134, 143
nonparametric, 134
pairs, 134
parametric, 144
percentile, 140
stationary, 150
wild, 145
bootstrap world, 133
boundary point, 194
breakdown point, 65
bridge estimator, 210

CEF, see conditional expectation function
censored, 62
check function, 59
complete case analysis, 257
conditional expectation function, 244
conditional independence, 244
conditional quantile function, 69
conjugacy, 159, 162
conjugate prior, 162
consistent, 223
contrapositive, 50, 51
converse, 50
coverage probability, 136
CP, see coverage probability
INDEX

CQF, see conditional quantile function
Cramér–von Mises, 98
cross-validation, 217
curse of dimensionality, 228
CV, see cross-validation
CvM or CM, see Cramér–von Mises
design bias, 199
dictionary, 209
Dirichlet distribution, 162
Dirichlet process, 165
Dirichlet–multinomial model, 163
discretize, 192
ECDF, see empirical CDF
effective degrees of freedom, 215
effective dimension, 215
effective number of parameters, 215
effective sample size, 188, 189
EL, see empirical likelihood
elastic net, 210
empirical CDF, 98
empirical likelihood, 156
ensemble, 223
equal-tailed, 136
equivalent number of parameters, 215
equivariance, 82
exchangeable, 144
expected loss, 59
false discovery proportion, 119
false discovery rate, 119
familywise error rate, 118
FCLT, see functional central limit theorem
FDR, see false discovery rate
FIC, see focused information criterion
first-order stochastic dominance, 111
focused information criterion, 222
forcing variable, 241
functional central limit theorem, 105
FWER, see familywise error rate
strong control of, 118
weak control of, 118
Gaussian process, 74
GCV, see generalized cross-validation
generalized cross-validation, 219
GOF, see goodness-of-fit
goodness-of-fit, 98
identification, 63, 255
partial, 255
point, 255
set, 255
identified, 63
if, 49
if and only if, 50
implied by, 49
implies, 49
imputation, 257
impute, 62
incidental parameters problem, 151
index, 230
inequality index, 277
infeasible bandwidth, see bandwidth
infinite-dimensional parameter, 182
information criterion, 221
instrumental variables quantile regression, 85
inverse, 50
inverse probability weighting, 261
IPW, see inverse probability weighting
IVQR, see instrumental variables quantile regression
k-nearest neighbor, 193
kernel
Bartlett, 201
Epanechnikov, 201
Gaussian, 201
higher-order, 203
of PDF, 162
second-order, 201
tent, 201
triangle, 201
uniform, 201
kernel functions, 200
kernel regression, 193
kNN, see k-nearest neighbor
knot, 209
Kolmogorov–Smirnov, 98
KS, see Kolmogorov–Smirnov
lasso, 210
latent, 273
leave-d-out cross-validation, 220
leave-one-out cross-validation, 218
library, 209
likelihood, 155
linear smoothers, 203
local constant, 193
local linear regression, 198
local polynomial regression, 193
local quantile treatment effect, 89
local sample size, 188, 189
LOOCV, see leave-one-out cross-validation
loss function, 59
LQTE, see local quantile treatment effect
manipulation of the running variable, 246
MAR, see missing at random
marginal treatment effect, 241
MCAR, see missing completely at random
mean squared prediction error, 59
method of sieves, 209
missing at random, 259
missing completely at random, 258
model averaging, 223
model selection, 179
monotonicity, 82
MSPE, see mean squared prediction error
MTE, see marginal treatment effect
MTP, see multiple testing procedure
multiple testing procedure, 118
necessary, 49
non-ignorable, 262
nonparametric, 182
nonparametric regression
local approach, 192
nonseparable, 83
only if, 49
optimal stopping, 152
order statistics, 62
ordinal, 273
overfitting, 179
oversmoothing, 198
parametric, 182
partial ordering, 112
partially linear model, 229
partition, 192
partitioning estimators, 193
percentile, see quantile permutation test, 108
pivotal, 137
PLM, see partially linear model
plug-in principle, 61, 130
pointwise confidence band, 101
posterior, 155
posterior expected loss, 160
posterior mean, 160
potential outcome, 77
potential outcomes, 242
pre-test procedure, 119
prior, 155
improper, 164
matching, 164
proper, 164
prior elicitation, 156
product kernel, 230
QTE, see quantile treatment effect
quantile, 57
quantile function, 58
quantile index, 57
quantile level, 57
quantile treatment effect, 79

R
  arguments, 35
  coerce, 31
  command prompt, 29
  comments, 29
  console, 29
  counter, 41
  data frame, 31
  data type, 30
  double, 30
  editor pane, 29
  else if, 40
  errors, 42
  escape character, 35
  for loops, 41
  hard coded, 43
  if-else statement, 40
  index, 32
  list, 31
  logical, 31
  lossless, 38
  lossy, 38
  packages, 28
  panes, 28
  parameters, 35
  plots, 29
  return, 35
  try-catch statements, 42
  variables, 30
  warnings, 42
  while loops, 41
  working directory, 36
randomization test, 108
RCM, see Rubin Causal Model
RD, see regression discontinuity
RDD, see regression discontinuity
treatment effect, 78

regression discontinuity design, see regression discontinuity
ridge regression, 210
risk, 59
root, 138
root method, 138
Rubin Causal Model, 242
running variable, 241

sample analog, 61
sample path, 105
sample quantile, 61
save.results(), 30
SD1, see first-order stochastic dominance
SD2, see second-order stochastic dominance
second-order stochastic dominance, 114
selection on observables, 244
semi-nonparametric, 182
semiparametric, 182
series regression, 208, 209
shrinkage estimator, 210
sieve, 209
sieve space, 209
single index model, 230
smoothing parameter, 207
statistics, 2
stepdown procedure, 119
stochastic dominance
  restricted, 275
  stronger, 49
Studentized, 137
subsampling, 146
sufficient, 49
symmetric, 137
tensor product basis, 230
test inversion, 100
tick function, 59
top-coding, 62
training sample, 217

regression discontinuity,
  fuzzy design, 242
  sharp design, 242
INDEX

unconditional quantile regression, 83
unconfoundedness, 244
undersmoothing, 198
uniform confidence band, 74, 101
UQR, see unconditional quantile regression
validation sample, 218
weaker, 49
with replacement, 134
worst-case bounds, 264