

# **Quantitative Social Science**

---

An Introduction

**KOSUKE IMAI**

PRINCETON UNIVERSITY PRESS  
**Princeton and Oxford**

# Contents

---

List of Tables	xiii
List of Figures	xv
Preface	xvii
<b>1 Introduction</b>	<b>1</b>
1.1 Overview of the Book	3
1.2 How to Use this Book	7
1.3 Introduction to R	10
1.3.1 Arithmetic Operations	10
1.3.2 Objects	12
1.3.3 Vectors	14
1.3.4 Functions	16
1.3.5 Data Files	20
1.3.6 Saving Objects	23
1.3.7 Packages	24
1.3.8 Programming and Learning Tips	25
1.4 Summary	27
1.5 Exercises	28
1.5.1 Bias in Self-Reported Turnout	28
1.5.2 Understanding World Population Dynamics	29
<b>2 Causality</b>	<b>32</b>
2.1 Racial Discrimination in the Labor Market	32
2.2 Subsetting the Data in R	36
2.2.1 Logical Values and Operators	37
2.2.2 Relational Operators	39
2.2.3 Subsetting	40
2.2.4 Simple Conditional Statements	43
2.2.5 Factor Variables	44
2.3 Causal Effects and the Counterfactual	46

2.4	Randomized Controlled Trials	48
2.4.1	The Role of Randomization	49
2.4.2	Social Pressure and Voter Turnout	51
2.5	Observational Studies	54
2.5.1	Minimum Wage and Unemployment	54
2.5.2	Confounding Bias	57
2.5.3	Before-and-After and Difference-in-Differences Designs	60
2.6	Descriptive Statistics for a Single Variable	63
2.6.1	Quantiles	63
2.6.2	Standard Deviation	66
2.7	Summary	68
2.8	Exercises	69
2.8.1	Efficacy of Small Class Size in Early Education	69
2.8.2	Changing Minds on Gay Marriage	71
2.8.3	Success of Leader Assassination as a Natural Experiment	73
<b>3</b>	<b>Measurement</b>	<b>75</b>
3.1	Measuring Civilian Victimization during Wartime	75
3.2	Handling Missing Data in R	78
3.3	Visualizing the Univariate Distribution	80
3.3.1	Bar Plot	80
3.3.2	Histogram	81
3.3.3	Box Plot	85
3.3.4	Printing and Saving Graphs	87
3.4	Survey Sampling	88
3.4.1	The Role of Randomization	89
3.4.2	Nonresponse and Other Sources of Bias	93
3.5	Measuring Political Polarization	96
3.6	Summarizing Bivariate Relationships	97
3.6.1	Scatter Plot	98
3.6.2	Correlation	101
3.6.3	Quantile–Quantile Plot	105
3.7	Clustering	108
3.7.1	Matrix in R	108
3.7.2	List in R	110
3.7.3	The <i>k</i> -Means Algorithm	111
3.8	Summary	115
3.9	Exercises	116
3.9.1	Changing Minds on Gay Marriage: Revisited	116
3.9.2	Political Efficacy in China and Mexico	118
3.9.3	Voting in the United Nations General Assembly	120
<b>4</b>	<b>Prediction</b>	<b>123</b>
4.1	Predicting Election Outcomes	123
4.1.1	Loops in R	124

4.1.2	General Conditional Statements in R	127
4.1.3	Poll Predictions	130
4.2	Linear Regression	139
4.2.1	Facial Appearance and Election Outcomes	139
4.2.2	Correlation and Scatter Plots	141
4.2.3	Least Squares	143
4.2.4	Regression towards the Mean	148
4.2.5	Merging Data Sets in R	149
4.2.6	Model Fit	156
4.3	Regression and Causation	161
4.3.1	Randomized Experiments	162
4.3.2	Regression with Multiple Predictors	165
4.3.3	Heterogenous Treatment Effects	170
4.3.4	Regression Discontinuity Design	176
4.4	Summary	181
4.5	Exercises	182
4.5.1	Prediction Based on Betting Markets	182
4.5.2	Election and Conditional Cash Transfer Program in Mexico	184
4.5.3	Government Transfer and Poverty Reduction in Brazil	187
<b>5</b>	<b>Discovery</b>	<b>189</b>
5.1	Textual Data	189
5.1.1	The Disputed Authorship of <i>The Federalist Papers</i>	189
5.1.2	Document-Term Matrix	194
5.1.3	Topic Discovery	195
5.1.4	Authorship Prediction	200
5.1.5	Cross Validation	202
5.2	Network Data	205
5.2.1	Marriage Network in Renaissance Florence	205
5.2.2	Undirected Graph and Centrality Measures	207
5.2.3	Twitter-Following Network	211
5.2.4	Directed Graph and Centrality	213
5.3	Spatial Data	220
5.3.1	The 1854 Cholera Outbreak in London	220
5.3.2	Spatial Data in R	223
5.3.3	Colors in R	226
5.3.4	US Presidential Elections	228
5.3.5	Expansion of Walmart	231
5.3.6	Animation in R	233
5.4	Summary	235
5.5	Exercises	236
5.5.1	Analyzing the Preambles of Constitutions	236
5.5.2	International Trade Network	238
5.5.3	Mapping US Presidential Election Results over Time	239

<b>6</b>	<b>Probability</b>	<b>242</b>
6.1	Probability	242
6.1.1	Frequentist versus Bayesian	242
6.1.2	Definition and Axioms	244
6.1.3	Permutations	247
6.1.4	Sampling with and without Replacement	250
6.1.5	Combinations	252
6.2	Conditional Probability	254
6.2.1	Conditional, Marginal, and Joint Probabilities	254
6.2.2	Independence	261
6.2.3	Bayes' Rule	266
6.2.4	Predicting Race Using Surname and Residence Location	268
6.3	Random Variables and Probability Distributions	277
6.3.1	Random Variables	278
6.3.2	Bernoulli and Uniform Distributions	278
6.3.3	Binomial Distribution	282
6.3.4	Normal Distribution	286
6.3.5	Expectation and Variance	292
6.3.6	Predicting Election Outcomes with Uncertainty	296
6.4	Large Sample Theorems	300
6.4.1	The Law of Large Numbers	300
6.4.2	The Central Limit Theorem	302
6.5	Summary	306
6.6	Exercises	307
6.6.1	The Mathematics of Enigma	307
6.6.2	A Probability Model for Betting Market Election Prediction	309
6.6.3	Election Fraud in Russia	310
<b>7</b>	<b>Uncertainty</b>	<b>314</b>
7.1	Estimation	314
7.1.1	Unbiasedness and Consistency	315
7.1.2	Standard Error	322
7.1.3	Confidence Intervals	326
7.1.4	Margin of Error and Sample Size Calculation in Polls	332
7.1.5	Analysis of Randomized Controlled Trials	336
7.1.6	Analysis Based on Student's $t$ -Distribution	339
7.2	Hypothesis Testing	342
7.2.1	Tea-Tasting Experiment	342
7.2.2	The General Framework	346
7.2.3	One-Sample Tests	350
7.2.4	Two-Sample Tests	356
7.2.5	Pitfalls of Hypothesis Testing	361
7.2.6	Power Analysis	363
7.3	Linear Regression Model with Uncertainty	370
7.3.1	Linear Regression as a Generative Model	370
7.3.2	Unbiasedness of Estimated Coefficients	375

7.3.3 Standard Errors of Estimated Coefficients	378
7.3.4 Inference about Coefficients	380
7.3.5 Inference about Predictions	384
7.4 Summary	389
7.5 Exercises	390
7.5.1 Sex Ratio and the Price of Agricultural Crops in China	390
7.5.2 File Drawer and Publication Bias in Academic Research	392
7.5.3 The 1932 German Election in the Weimar Republic	394
<b>8 Next</b>	<b>397</b>
General Index	401
R Index	406