

Peng Ding

A First Course in Causal Inference



*Lecture notes for my “Causal Inference” course
at the University of California Berkeley*



Contents

Preface	xv
Acronyms	xvii
I Introduction	1
1 Correlation, Association, and the Yule–Simpson Paradox	3
1.1 Traditional view of statistics	3
1.2 Some commonly-used measures of association	4
1.2.1 Correlation and regression	4
1.2.2 Contingency tables	5
1.3 An example of the Yule–Simpson Paradox	7
1.3.1 Data	7
1.3.2 Explanation	8
1.3.3 Geometry of the Yule–Simpson Paradox	9
1.4 The Berkeley graduate school admission data	10
1.5 Homework Problems	12
2 Potential Outcomes	15
2.1 Experimentalists' view of causal inference	15
2.2 Formal notation of potential outcomes	16
2.2.1 Causal effects, subgroups, and the non-existence of Yule–Simpson Paradox	18
2.2.2 Subtlety of experimental unit	18
2.3 Treatment assignment mechanism	19
2.4 Homework Problems	20
II Randomized experiments	23
3 The Completely Randomized Experiment and the Fisher Randomization Test	25
3.1 CRE	25
3.2 FRT	26
3.3 Canonical choices of the test statistic	28
3.4 A case study of the LaLonde experimental data	33
3.5 Some history of randomized experiments and FRT	35
3.5.1 James Lind's experiment	35

3.5.2	Lady tasting tea	37
3.5.3	Two Fisherian principles for experiments	37
3.6	Discussion	38
3.6.1	Other sharp null hypotheses and confidence intervals .	38
3.6.2	Other test statistics	39
3.6.3	Final remarks	39
3.7	Homework Problems	40
4	Neymanian Repeated Sampling Inference in Completely Randomized Experiments	43
4.1	Finite population quantities	43
4.2	Neyman (1923)'s theorem	44
4.3	Proofs	47
4.4	Regression analysis of the CRE	49
4.5	Examples	49
4.5.1	Simulation	49
4.5.2	Heavy-tailed outcome and failure of Normal approximations	52
4.5.3	Application	52
4.6	Homework Problems	55
5	Stratification and Post-Stratification in Randomized Experiments	59
5.1	Stratification	59
5.2	FRT	61
5.2.1	Theory	61
5.2.2	An application	63
5.3	Neymanian inference	64
5.3.1	Point and interval estimation	64
5.3.2	Numerical examples	66
5.3.3	Comparing the SRE and the CRE	68
5.4	Post-stratification in a CRE	71
5.4.1	Meinert et al. (1970)'s Example	72
5.4.2	Chong et al. (2016)'s Example	73
5.5	Practical questions	74
5.6	Homework Problems	75
6	Rerandomization and Regression Adjustment	79
6.1	Rerandomization	79
6.1.1	Experimental design	79
6.1.2	Statistical inference	81
6.2	Regression adjustment	83
6.2.1	Covariate-adjusted FRT	83
6.2.2	Analysis of covariance and extensions	84
6.2.2.1	Some heuristics for Lin (2013)'s results	85

6.2.2.2	Understanding Lin (2013)'s estimator via predicting the potential outcomes	87
6.2.2.3	Understanding Lin (2013)'s estimator via adjusting for covariate imbalance	88
6.2.3	Some additional remarks on regression adjustment	88
6.2.3.1	Duality between ReM and regression adjustment	88
6.2.3.2	Equivalence of regression adjustment and post-stratification	89
6.2.3.3	Difference-in-difference as a special case of covariate adjustment $\hat{\tau}(\beta_1, \beta_0)$	89
6.2.4	Extension to the SRE	90
6.3	Unification, combination, and comparison	90
6.4	Simulation	91
6.5	Final remarks	91
6.6	Homework Problems	92
7	Matched-Pairs Experiment	95
7.1	Design of the experiment and potential outcomes	95
7.2	FRT	96
7.3	Neymanian inference	99
7.4	Covariate adjustment	101
7.4.1	FRT	101
7.4.2	Regression adjustment	102
7.5	Examples	104
7.5.1	Darwin's data comparing cross-fertilizing and self-fertilizing on the height of corns	104
7.5.2	Children's television workshop experiment data	105
7.6	Comparing the MPE and CRE	106
7.7	Extension to the general matched experiment	107
7.7.1	FRT	108
7.7.2	Estimating the average of the within-strata effects	108
7.7.3	A more general causal estimand	109
7.8	Homework Problems	110
8	Unification of the Fisherian and Neymanian Inferences in Randomized Experiments	113
8.1	Testing strong and weak null hypotheses in the CRE	113
8.2	Covariate-adjusted FRTs in the CRE	115
8.3	General recommendations	115
8.4	A case study	116
8.5	Homework Problems	117

9 Bridging Finite and Super Population Causal Inference	121
9.1 CRE	122
9.2 SRE	123
9.3 Homework Problems	124
III Observational studies	125
10 Observational Studies, Selection Bias, and Nonparametric Identification of Causal Effects	127
10.1 Motivating Examples	127
10.2 Causal effects and selection bias under the potential outcomes framework	128
10.3 Sufficient conditions for nonparametric identification	130
10.3.1 Identification	130
10.3.2 Plausibility of the assumption	133
10.4 Two simple estimation strategies and their limitations	134
10.4.1 Stratification or standardization based on discrete covariates	134
10.4.2 Outcome regression	134
10.5 Homework Problems	136
11 The Central Role of the Propensity Score in Observational Studies for Causal Effects	139
11.1 The propensity score as a dimension reduction tool	140
11.1.1 Theory	140
11.1.2 Propensity score stratification	141
11.1.3 Application	142
11.2 Propensity score weighting	144
11.2.1 Theory	144
11.2.2 Inverse propensity score weighting estimators	145
11.2.3 A problem of weighting and a fundamental problem of causal inference	146
11.2.4 Application	146
11.3 The propensity score as a balancing score	147
11.3.1 Theory	147
11.3.2 Covariate balance check	148
11.4 Homework Problems	149
12 The Doubly Robust or the Augmented Inverse Propensity Score Weighting Estimator for the Average Causal Effect	153
12.1 The doubly robust estimator	154
12.1.1 Population version	154
12.1.2 Sample version	155
12.2 More intuition and theory for the doubly robust estimator	156
12.2.1 Reducing the variance of the IPW estimator	156
12.2.2 Reducing the bias of the outcome regression estimator	157

Contents ix

12.3 Examples	157
12.3.1 Summary of some canonical estimators for τ	157
12.3.2 Simulation	159
12.3.3 Applications	160
12.4 Some further discussion	161
12.5 Homework problems	161
13 The Average Causal Effect on the Treated Units and Other Estimands	163
13.1 Nonparametric identification of τ_T	163
13.2 Inverse propensity score weighting and doubly robust estimation of τ_T	166
13.3 An example	168
13.4 Other estimands	169
13.5 Homework Problems	171
14 Using the Propensity Score in Regressions for Causal Effects	175
14.1 Regressions with the propensity score as a covariate	175
14.2 Regressions weighted by the inverse of the propensity score .	178
14.2.1 Average causal effect	178
14.2.2 Average causal effect on the treated units	180
14.3 Homework problems	182
15 Matching in Observational Studies	185
15.1 A simple starting point: many more control units	185
15.2 A more complicated but realistic scenario	186
15.3 Matching estimator for the average causal effect	188
15.3.1 Point estimation and bias correction	188
15.3.2 Connection with the doubly robust estimators	189
15.4 Matching estimator for the average causal effect on the treated	190
15.5 A case study	192
15.5.1 Experimental data	192
15.5.2 Observational data	193
15.5.3 Covariate balance checks	194
15.6 Discussion	196
15.7 Homework Problems	196
IV Difficulties and challenges of observational studies	199
16 Difficulties of Unconfoundedness in Observational Studies for Causal Effects	201
16.1 Some basics of the causal diagram	201
16.2 Assessing ignorability	202
16.2.1 Using negative outcomes	202
16.2.2 Using negative exposures	204

16.2.3 Summary	204
16.3 Problems of over adjustment	205
16.3.1 M-bias	205
16.3.2 Z-bias	207
16.3.3 What covariates should we adjust for in observational studies?	209
16.4 Homework Problems	210
17 E-Value: Evidence for Causation in Observational Studies with Unmeasured Confounding	211
17.1 Cornfield-type sensitivity analysis	211
17.2 E-value	214
17.3 A classic example	216
17.4 Extensions	217
17.4.1 E-value and Bradford Hill's criteria for causation	217
17.4.2 E-value after logistic regression	218
17.4.3 Non-zero true causal effect	220
17.5 Critiques and responses	220
17.5.1 E-value is just a monotone transformation of the risk ratio	221
17.5.2 Calibration of the E-value	221
17.5.3 It works the best for a binary outcome and the risk ratio	222
17.6 Homework Problems	222
18 Sensitivity Analysis for the Average Causal Effect with Unmeasured Confounding	225
18.1 Introduction	225
18.2 Manski-type worse-case bounds on the average causal effect without assumptions	226
18.3 Sensitivity analysis for the average causal effect	228
18.3.1 Identification formulas	228
18.4 Example	230
18.5 Homework Problems	230
19 Rosenbaum-Style <i>p</i>-Values for Matched Observational Studies with Unmeasured Confounding	233
19.1 The model for sensitivity analysis with matched data	233
19.2 Worst-case <i>p</i> -values under Rosenbaum's sensitivity model	234
19.3 Revisiting the LaLonde data	235
19.4 Homework Problems	239
20 Overlap in Observational Studies: Difficulties and Opportunities	241
20.1 Implications of overlap	241
20.1.1 Trimming in the presence of limited overlap	242
20.1.2 Outcome modeling in the presence of limited overlap	242

<i>Contents</i>	xi
20.2 Causal inference with no overlap: regression discontinuity	243
20.2.1 Examples and graphical diagnostics	243
20.2.2 A mathematical formulation of regression discontinuity	245
20.2.3 Regressions near the boundary	246
20.2.4 An example	248
20.2.5 Problems of regression discontinuity	249
20.3 Homework Problems	250
V Instrumental variables	251
21 An Experimental Perspective	253
21.1 Encouragement Design and Noncompliance	253
21.2 Latent Compliance Status and Effects	254
21.2.1 Nonparametric identification	254
21.2.2 Estimation	257
21.3 Covariates	258
21.3.1 Covariate adjustment in complete randomization	258
21.3.2 Covariates in conditional randomization or unconfounded observational studies	259
21.4 Weak IV	259
21.5 Application	261
21.6 Interpreting the Complier Average Causal Effect	262
21.7 Homework problems	263
22 Disentangle Mixture Distributions and Instrumental Variable Inequalities	267
22.1 Disentangle Mixture Distributions and Instrumental Variable Inequalities	267
22.2 Testable implications	270
22.3 Examples	270
22.4 Homework problems	273
23 An Econometric Perspective	279
23.1 Examples of studies with IVs	280
23.2 Brief Review of the Ordinary Least Squares	281
23.3 Linear Instrumental Variable Model	282
23.4 The Just-Identified Case	284
23.5 The Over-Identified Case	284
23.6 A Special Case: A Single IV for a Single Endogenous Treatment	286
23.6.1 Two-stage least squares	287
23.6.2 Indirect least squares	287
23.6.3 Weak IV	288
23.7 Application	288
23.8 Homework	290

24 Application of the Instrumental Variable Method: Fuzzy Regression Discontinuity	293
24.1 Motivating examples	293
24.2 Mathematical formulation	294
24.3 Application	296
24.3.1 Re-analyzing Asher and Novosad (2020)'s data	296
24.3.2 Re-analyzing Li et al. (2015)'s data	298
24.4 Discussion	299
24.5 Homework Problems	299
25 Application of the Instrumental Variable Method: Mendelian Randomization	301
25.1 Background and motivation	301
25.2 MR based on summary statistics	303
25.2.1 Fixed-effect estimator	303
25.2.2 Egger regression	304
25.3 An example	305
25.4 Critiques of the analysis based on Mendelian randomization	307
25.5 Homework Problems	308
VI Causal Mechanisms with Post-Treatment Variables	309
26 Principal Stratification	311
26.1 Motivating Examples	311
26.2 The Problem of Conditioning on the Post-Treatment Variable	312
26.3 Conditioning on the Potential Values of the Post-Treatment Variable	314
26.4 Statistical Inference and Its Difficulty	315
26.4.1 Special case: truncation by death with binary outcome	316
26.4.2 An application	317
26.4.3 Extensions	318
26.5 Principal score method	318
26.5.1 Principal score method under strong monotonicity . .	319
26.5.2 Extensions	320
26.6 Other methods	320
26.7 Homework problems	321
27 Mediation Analysis: Natural Direct and Indirect Effects	323
27.1 Motivating Examples	323
27.2 Nested Potential Outcomes	324
27.2.1 Natural Direct and Indirect Effects	324
27.2.2 Metaphysics or Science	326
27.3 The Mediation Formula	328
27.4 The Mediation Formula Under Linear Models	331
27.4.1 The Baron–Kenny Method	332

<i>Contents</i>	xiii
27.4.2 An Example	333
27.5 Sensitivity analysis	334
27.6 Homework problems	335
28 Controlled Direct Effect	339
28.1 Identification and estimation of the controlled direct effect	339
28.2 Discussion	341
28.3 Homework problems	342
29 Time-Varying Treatment and Confounding	345
29.1 Basic setup and sequential ignorability	345
29.2 g-formula and outcome modeling	347
29.2.1 Plug-in estimation based on outcome modeling	348
29.2.2 Recursive estimation based on outcome modeling	349
29.3 Inverse propensity score weighting	350
29.4 Multiple time points	352
29.4.1 Marginal structural model	352
29.4.2 Structural nested model	353
29.5 Homework problems	356
VII Appendices	361
A1 Probability and Statistics	363
A1.1 Probability	363
A1.1.1 Tower property and variance decomposition	363
A1.1.2 Limiting theorems	363
A1.1.3 Delta method	364
A1.2 Statistical inference	365
A1.2.1 Point estimation	365
A1.2.2 Confidence interval	366
A1.2.3 Hypothesis testing	366
A1.2.4 Wald-type confidence interval and test	367
A1.2.5 Duality between constructing confidence sets and testing null hypotheses	367
A1.3 Inference with 2×2 tables	368
A1.3.1 Fisher's exact test	368
A1.3.2 Estimation with 2×2 tables	368
A1.4 Two famous problems in statistics	369
A1.4.1 Behrens–Fisher problem	369
A1.4.2 Fieller–Creasy problem	370
A1.5 Bootstrap	371
A1.6 Homework problems	372

A2Linear and Logistic Regressions	373
A2.1 Population ordinary least squares	373
A2.2 Sample OLS	374
A2.3 Frisch–Waugh–Lovell Theorem	375
A2.4 Linear model	376
A2.5 Weighted least squares	377
A2.6 Logistic regression	377
A2.6.1 Model	377
A2.6.2 Maximum likelihood estimate	378
A2.6.3 Extension to the case-control study	379
A2.6.4 Logistic regression with weights	379
A2.7 Homework problems	379
A3Some Useful Lemmas for Simple Random Sampling	381
A3.1 Lemmas	381
A3.2 Proofs	383
A3.3 Comments on the literature	385
A3.4 Homework Problems	385

Preface

I developed the lecture notes based on my “Causal Inference” course at the University of California Berkeley over the past seven years. Since half of the students were undergraduate, my lecture notes only require basic knowledge of probability theory, statistical inference, and linear and logistic regressions.

I am grateful for the constructive comments from many students. If you identify any errors, please feel free to email me.



Acronyms

acronym	full name	first chapter
RD	risk difference	1
RR	risk ratio or relative risk	1
OR	odds ratio	1
RCT	randomized controlled trial	1
BMI	body mass index	2
SUTVA	stable unit treatment value assumption	2
ACE	average causal effect	2
CRE	completely randomized experiment	3
BRE	Bernoulli randomized experiment	3
IID	independent and identically distributed	3 and A1
FRT	Fisher randomization test	3
OLS	ordinary least squares	4 and A2
EHW	Eicker–Huber–White (robust standard error)	4 and A2
SRE	stratified randomized experiment	5
ReM	rerandomization using the Mahalanobis distance	6
ANCOVA	analysis of covariance	6
LASSO	least absolute shrinkage and selection operator	6
MPE	matched-pairs experiment	7
NHANES	National Health and Nutrition Examination Survey	10
IPW	inverse propensity score weighting	11
HT	Horvitz–Thompson	11
WLS	weighted least squares	14 and A2
IV	instrumental variable	21
ITT	intention-to-treat (analysis)	21
CACE	complier average causal effect	21
LATE	local average treatment effect	21
TSLS	two-stage least squares	23
ILS	indirect least squares	23
MR	Mendelian randomization	25
SNP	single nucleotide polymorphism	25
NDE	natural direct effect	27
NIE	natural indirect effect	27
CDE	controlled direct effect	29
MSM	marginal structural model	29
FWL	Frisch–Waugh–Lovell (theorem)	A2
MLE	maximum likelihood estimate	A2



Part I

Introduction



1

Correlation, Association, and the Yule–Simpson Paradox

Causality is central to human knowledge. Two famous quotes from ancient Greeks are below.

I would rather discover one causal law than be King of Persia.

— Democritus

We do not have knowledge of a thing until we grasped its cause.

— Aristotle

However, the major part of classic statistics is about association rather than causation. This chapter will review some basic association measures and point out their fundamental limitations.

1.1 Traditional view of statistics

A traditional view of statistics is to infer correlation or association among variables. Based on this view, there is no role for causal inference in statistics. Two famous aphorisms based on this view are below:

- “Correlation does not imply causation.”
- “You cannot prove causality with statistics.”

This book has a very different view: statistics is crucial for understanding causality. The main focus of this book is to introduce the formal language for causal inference and develop statistical methods to estimate causal effects in randomized experiments and observational studies.

1.2 Some commonly-used measures of association

1.2.1 Correlation and regression

The Pearson correlation coefficient between two random variables Z and Y is

$$\rho_{ZY} = \frac{\text{cov}(Z, Y)}{\sqrt{\text{var}(Z)\text{var}(Y)}},$$

which measures the linear dependence of Z and Y .

The linear regression of Y on Z is the model

$$Y = \alpha + \beta Z + \varepsilon, \quad (1.1)$$

where $E(\varepsilon) = 0$ and $E(\varepsilon Z) = 0$. We can show that the regression coefficient β equals

$$\beta = \frac{\text{cov}(Z, Y)}{\text{var}(Z)} = \rho_{ZY} \sqrt{\frac{\text{var}(Y)}{\text{var}(Z)}}.$$

So β and ρ_{ZY} always have the same sign.

We can also define multiple regression of Y on Z and X :

$$Y = \alpha + \beta Z + \gamma X + \varepsilon, \quad (1.2)$$

where $E(\varepsilon) = 0$, $E(\varepsilon Z) = 0$ and $E(\varepsilon X) = 0$. We usually interpret β as the “effect” of Z on Y , *holding X constant* or *conditioning on X* or *controlling for X*. Chapter A2 reviews the basics of linear regression.

More interestingly, the β ’s in the above two regressions (1.1) and (1.2) can be different; they can even have different signs. The following R code reanalyzed the LaLonde observational data used by Hainmueller (2012). The main question of interest is the “causal effect” of a job training program on earning. The regression controlling for all covariates gives coefficient 1067.5461 for `treat`, while the regression not controlling for any covariates gives coefficient -8506.4954 for `treat`.

```
> dat <- read.table("cps1re74.csv", header = TRUE)
> dat$u74 <- as.numeric(dat$re74==0)
> dat$u75 <- as.numeric(dat$re75==0)
>
> ## linear regression on the outcome
> lmoutcome = lm(re78 ~ ., data = dat)
> summary(lmoutcome)$coef[2, 1:2]
  Estimate Std. Error
  1067.5461    554.0595
>
> lmoutcome = lm(re78 ~ treat, data = dat)
> summary(lmoutcome)$coef[2, 1:2]
  Estimate Std. Error
  -8506.4954   712.7664
```

1.2.2 Contingency tables

We can represent the joint distribution of two binary variables Z and Y by a two-by-two contingency table. With $p_{zy} = \text{pr}(Z = z, Y = y)$, we can summarize the joint distribution in the following table:

	$Y = 1$	$Y = 0$
$Z = 1$	p_{11}	p_{10}
$Z = 0$	p_{01}	p_{00}

Viewing Z as the treatment or exposure and Y as the outcome, we can define the risk difference as

$$\begin{aligned} \text{RD} &= \text{pr}(Y = 1 \mid Z = 1) - \text{pr}(Y = 1 \mid Z = 0) \\ &= \frac{p_{11}}{p_{11} + p_{10}} - \frac{p_{01}}{p_{01} + p_{00}}, \end{aligned}$$

the risk ratio as

$$\begin{aligned} \text{RR} &= \frac{\text{pr}(Y = 1 \mid Z = 1)}{\text{pr}(Y = 1 \mid Z = 0)} \\ &= \frac{p_{11}}{p_{11} + p_{10}} / \frac{p_{01}}{p_{01} + p_{00}}, \end{aligned}$$

and the odds ratio¹ as

$$\begin{aligned} \text{OR} &= \frac{\text{pr}(Y = 1 \mid Z = 1) / \text{pr}(Y = 0 \mid Z = 1)}{\text{pr}(Y = 1 \mid Z = 0) / \text{pr}(Y = 0 \mid Z = 0)} \\ &= \frac{\frac{p_{11}}{p_{11} + p_{10}} / \frac{p_{10}}{p_{11} + p_{10}}}{\frac{p_{01}}{p_{01} + p_{00}} / \frac{p_{00}}{p_{01} + p_{00}}} \\ &= \frac{p_{11}p_{00}}{p_{10}p_{01}}. \end{aligned}$$

The terminologies risk difference, risk ratio, and odds ratio come from epidemiology. Because the outcomes in epidemiology are often diseases, it is natural to use the name “risk” for the probability of having diseases.

We have the following simple facts for these measures.

Proposition 1.1 (1) The following statements are all equivalent²: $Z \perp\!\!\!\perp Y$, $\text{RD} = 0$, $\text{RR} = 1$, and $\text{OR} = 1$. (2) If p_{zy} 's are all positive, then $\text{RD} > 0$ is equivalent to $\text{RR} > 1$ and is also equivalent to $\text{OR} > 1$ (3) $\text{OR} \approx \text{RR}$ if $\text{pr}(Y = 1 \mid Z = 1)$ and $\text{pr}(Y = 1 \mid Z = 0)$ are small.

¹In probability theory, the odds of an event is defined as the ratio of the probability that the event happens over the probability that the event does not happen.

²This book uses the notation $\perp\!\!\!\perp$ to denote independence or conditional independence of random variables. The notation is due to Dawid (1979).

I leave the proofs of statements (1) and (2) as a homework problem. Statement (3) is informal. The approximation holds because the odds $p/(1 - p)$ is close to the probability p for rare diseases with $p \approx 0$: by Taylor expansion $p/(1 - p) = p + p^2 + \dots \approx p$. In epidemiology, if the outcome represents the occurrence of a rare disease, then it is reasonable to assume that $\text{pr}(Y = 1 | X = 1)$ and $\text{pr}(Y = 1 | X = 0)$ are small.

We can also define conditional versions of the RD, RR, and OR if the probabilities are replaced by the conditional probabilities given another variable X , i.e., $\text{pr}(Y = 1 | Z = 1, X = x)$ and $\text{pr}(Y = 1 | Z = 0, X = x)$.

With frequencies $n_{zy} = \#\{i : Z_i = z, Y_i = y\}$, we can summarize the observed data in the following two-by-two table:

	$Y = 1$	$Y = 0$
$Z = 1$	n_{11}	n_{10}
$Z = 0$	n_{01}	n_{00}

We can estimate RD, RR, and OR by replacing the true probabilities by the sample proportions. In R, functions `fisher.test` performs exact test and `chisq.test` performs asymptotic test for $Z \perp\!\!\!\perp Y$ based on a two-by-two table of observed data.

Example 1.1 Bertrand and Mullainathan (2004) conducted a randomized experiment on resumes to study the effect of perceived race on callbacks for interviews. They randomly assigned African-American- or White-sounding names on fictitious resumes to help-wanted ads in Boston and Chicago newspapers. The following two-by-two table summarizes perceived race and callback:

```
> resume = read.csv("resume.csv")
> Alltable = table(resume$race, resume$call)
> Alltable

      0      1
black 2278 157
white 2200 235
```

The two rows have the same total count, so it is apparent that White names received more callbacks. Fisher's exact test below shows that this difference is statistically significant.

```
> fisher.test(Alltable)

Fisher's Exact Test for Count Data
```

```
data: Alltable
p-value = 4.759e-05
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
1.249828 1.925573
sample estimates:
```

```
odds ratio
1.549732
```

1.3 An example of the Yule–Simpson Paradox

1.3.1 Data

The classic Kidney stone example is from Charig et al. (1986), where Z is the treatment with 1 for an open surgical procedure and 0 for a small puncture, and Y is the outcome with 1 for success and 0 for failure. The treatment and outcome data can be summarized in the following two-by-two table:

	$Y = 1$	$Y = 0$
$Z = 1$	273	77
$Z = 0$	289	61

The estimated RD is

$$\widehat{RD} = \frac{273}{273 + 77} - \frac{289}{289 + 61} = 78\% - 83\% = -5\% < 0.$$

Treatment 0 seems better, that is, the small puncture leads to higher successful rate compared to the open surgical procedure.

However, the data were not from a randomized controlled trial (RCT)³. Patients receiving treatment 1 can be very different from patients receiving treatment 0. A “lurking variable” in this study is the severity of the case: some patients have smaller stones but some patients have larger stones. We can split the data according to the size of the stones.

For patients with smaller stones, the treatment and outcome data can be summarized in the following two-by-two table:

	$Y = 1$	$Y = 0$
$Z = 1$	81	6
$Z = 0$	234	36

For patients with larger stones, the treatment and outcome data can be summarized in the following two-by-two table:

	$Y = 1$	$Y = 0$
$Z = 1$	192	71
$Z = 0$	55	25

³In an RCT, patients are randomly assigned to the treatment arms. Part II of this book will focus on RCTs.

The latter two tables must add up to the first table:

$$81 + 192 = 273, \quad 6 + 71 = 77, \quad 234 + 55 = 289, \quad 36 + 25 = 61.$$

From the table for patients with smaller stones, the estimated RD is

$$\widehat{RD}_{\text{smaller}} = \frac{81}{81+6} - \frac{234}{234+36} = 93\% - 87\% = 6\% > 0,$$

suggesting that treatment 1 is better. From the table for patients with larger stones, the estimated RD is

$$\widehat{RD}_{\text{larger}} = \frac{192}{192+71} - \frac{55}{55+25} = 73\% - 69\% = 4\% > 0,$$

also suggesting that treatment 1 is better.

The above data analysis leads to

$$\widehat{RD} < 0, \quad \widehat{RD}_{\text{smaller}} > 0, \quad \widehat{RD}_{\text{larger}} > 0.$$

Informally, treatment 1 is better for both patients with smaller and larger stones, but treatment 1 is worse for the whole population. This interpretation is quite confusing if the goal is to infer the treatment effect. In statistics, this is called the Yule–Simpson or Simpson’s Paradox in which the marginal association has the opposite sign to the conditional associations at all levels.

1.3.2 Explanation

Let X be the binary indicator with $X = 1$ for smaller stones and $X = 0$ for larger stones. Let us first take a look at the X – Z relationship by comparing the probabilities of receiving treatment 1 among patients with smaller and larger stones:

$$\begin{aligned} & \widehat{\text{pr}}(Z = 1 | X = 1) - \widehat{\text{pr}}(Z = 1 | X = 0) \\ &= \frac{81 + 6}{81 + 6 + 234 + 36} - \frac{192 + 71}{192 + 71 + 55 + 25} \\ &= 24\% - 77\% \\ &= -53\% < 0. \end{aligned}$$

So patients with larger stones tend to take treatment 1. Statistically, X and Z have negative association.

Let us then take a look at the X – Y relationship by comparing the probabilities of success among patients with smaller and larger stones: under treatment 1,

$$\begin{aligned} & \widehat{\text{pr}}(Y = 1 | Z = 1, X = 1) - \widehat{\text{pr}}(Y = 1 | Z = 1, X = 0) \\ &= \frac{81}{81 + 6} - \frac{192}{192 + 71} \\ &= 93\% - 73\% \\ &= 20\% > 0; \end{aligned}$$

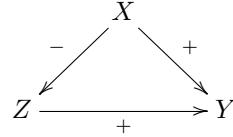


FIGURE 1.1: A diagram for the kidney stone example. The signs indicate the associations of two variables, conditioning on other variables pointing to the downstream variable.

under treatment 0,

$$\begin{aligned}
 & \widehat{\text{pr}}(Y = 1 \mid Z = 0, X = 1) - \widehat{\text{pr}}(Y = 1 \mid Z = 0, X = 0) \\
 &= \frac{234}{234 + 36} - \frac{55}{55 + 25} \\
 &= 87\% - 69\% \\
 &= 18\% > 0.
 \end{aligned}$$

So under both treatment levels, patients with smaller stones have higher success probabilities. Statistically, X and Y have positive association conditional on both treatment levels.

We can summarize the qualitative associations in the diagram in Figure 1.1. In technical terms, the treatment has a positive direct path and a more negative indirect path to the outcome, so the overall association is negative between the treatment and outcome. In plain English, when less effective treatment 0 is applied more frequently to the less severe cases, it can appear to be a more effective treatment.

1.3.3 Geometry of the Yule–Simpson Paradox

Assume that the 2×2 table based on the aggregated data has counts

whole population	$Y = 1$	$Y = 0$
$Z = 1$	n_{11}	n_{10}
$Z = 0$	n_{01}	n_{00}

The two 2×2 tables based on subgroups with $X = 1$ and $X = 0$ have counts

subpopulation $X = 1$	$Y = 1$	$Y = 0$
$Z = 1$	$n_{11 1}$	$n_{10 1}$
$Z = 0$	$n_{01 1}$	$n_{00 1}$

subpopulation $X = 0$	$Y = 1$	$Y = 0$
$Z = 1$	$n_{11 0}$	$n_{10 0}$
$Z = 0$	$n_{01 0}$	$n_{00 0}$

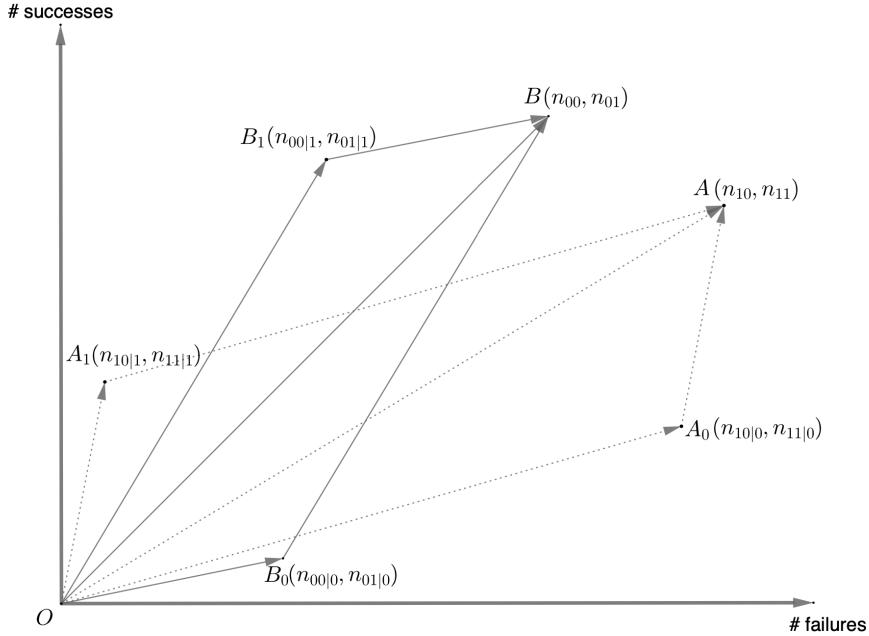


FIGURE 1.2: Geometry of the Yule–Simpson Paradox

Figure 1.2 shows the geometry of the Yule–Simpson Paradox. The y-axis shows the count of successes with $Y = 1$ and the x-axis shows the count of failures with $Y = 0$. The two parallelograms corresponds to aggregating the counts of successes and failures under two treatment levels. The slope of OA_1 is larger than that of OB_1 , and the slope of OA_0 is larger than that of OB_0 . So the treatment seems beneficial to the outcome within both levels of X . However, the slope of OA is smaller than that of OB . So the treatment seems harmful to the outcome for the whole population. The Yule–Simpson Paradox arises.

1.4 The Berkeley graduate school admission data

Bickel et al. (1975) investigated the admission rates of male and female students into the graduate school of Berkeley. The R package **datasets** contains the original data **UCBAdmissions**. The raw data by the six largest departments are shown below:

```
> library(datasets)
```

```
> UCBAdmissions = aperm(UCBAdmissions, c(2, 1, 3))
> UCBAdmissions
, , Dept = A

      Admit
Gender   Admitted Rejected
Male       512      313
Female      89       19

, , Dept = B

      Admit
Gender   Admitted Rejected
Male       353      207
Female      17        8

, , Dept = C

      Admit
Gender   Admitted Rejected
Male       120      205
Female      202      391

, , Dept = D

      Admit
Gender   Admitted Rejected
Male       138      279
Female      131      244

, , Dept = E

      Admit
Gender   Admitted Rejected
Male       53       138
Female      94      299

, , Dept = F

      Admit
Gender   Admitted Rejected
Male       22       351
Female      24      317
```

Aggregating the data over departments, we have a simple two-by-two table:

```
> UCBAdmissions.sum = apply(UCBAdmissions, c(1, 2), sum)
> UCBAdmissions.sum
      Admit
Gender   Admitted Rejected
```

Male	1198	1493
Female	557	1278

The following function, building upon `chisq.test`, have a two-by-two table as the input and the estimated RD and *p*-value as output:

```
> risk.difference = function(tb2)
+ {
+   p1      = tb2[1, 1]/(tb2[1, 1] + tb2[1, 2])
+   p2      = tb2[2, 1]/(tb2[2, 1] + tb2[2, 2])
+   testp   = chisq.test(tb2)
+
+   return(list(p.diff = p1 - p2,
+              pv = testp$p.value))
+ }
```

With this function, we find large and significant difference between the admission rates of male and female students:

```
> risk.difference(UCBAdmissions.sum)
$p.diff
[1] 0.1416454

$pv
[1] 1.055797e-21
```

Stratifying on the departments, we find smaller and insignificant differences between the admission rates of male and female students. In department A, the difference is significant but negative.

```
> P.diff = rep(0, 6)
> PV     = rep(0, 6)
> for(dd in 1:6)
+ {
+   department = risk.difference(UCBAdmissions[, , dd])
+   P.diff[dd] = department$p.diff
+   PV[dd]    = department$pv
+ }
>
> round(P.diff, 2)
[1] -0.20 -0.05  0.03 -0.02  0.04 -0.01
> round(PV, 2)
[1]  0.00  0.77  0.43  0.64  0.37  0.64
```

1.5 Homework Problems

1.1 Independence in two-by-two tables

Prove (1) and (2) in Proposition 1.1.

1.2 Correlation and partial correlation

Consider a three-dimensional Normal random vector:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{pmatrix} \right).$$

The correlation coefficient between X and Y is ρ_{XY} . There are many equivalent definitions of the partial correlation coefficient. For a multivariate Normal vector, let $\rho_{XY|Z}$ denote the partial correlation coefficient between X and Y given Z , which is defined as their correlation coefficient in the conditional distribution $(X, Y) | Z$. Show that

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}}$$

Give an example with $\rho_{XY} > 0$ and $\rho_{XY|Z} < 0$.

Remark: This is the Yule–Simpson Paradox for a Normal random vector.

1.3 Specification searches

Section 1.2.1 re-analyzes the data used by Hainmueller (2012) with R code in `LalondeRegression.R`. In total, the data contain 10 covariates and therefore $2^{10} = 1024$ possible subsets of covariates in the linear regression. Run 1024 linear regressions with all possible subsets of covariates, and report the regression coefficients of the treatment. How many are positively significant, how many are negatively significant, and how many are not significant? You can also report other interesting findings from these regressions.

1.4 More on racial discrimination

Section 1.2.2 re-analyzes the data collected by Bertrand and Mullainathan (2004) with R code in `resume.R`. Conduct analyses separately for males and females. What do you find from these subgroup analyses?

1.5 Recommended reading

Bickel et al. (1975) is the original paper for the paradox reported in Section 1.4.



2

Potential Outcomes

2.1 Experimentalists' view of causal inference

Rubin (1975) and Holland (1986) made up the aphorism:

no causation without manipulation.

Not everybody agrees with this point of view. However, it is quite helpful to clarify ambiguity in thinking about causal relationships. This book follows this view and defines causal effects using the potential outcomes framework (Neyman, 1923; Rubin, 1974). In this framework, an experiment, or at least a thought experiment, has an intervention, a manipulation, or a treatment, and we are interested in its effect on an outcome or multiple outcomes.

Example 2.1 *If we are interested in the effect of taking aspirin or not on the relief of head ache, the intervention is taking aspirin.*

Example 2.2 *If we are interested in the effect of participating in a job training program or not on employment and wage, the intervention is participating in a job training program.*

Example 2.3 *If we are interested in the effect of studying in a small classroom or a large classroom on standardized test scores, the intervention is studying in a small classroom.*

Example 2.4 *Gerber et al. (2008) were interested in the effect of different get-out-to-vote messages on the voting behavior. The intervention is different get-out-to-vote messages.*

Example 2.5 *Pearl (2018) claimed that we could infer the effect of obesity on life span. A popular measure of obesity is the body mass index (BMI), defined as the body mass divided by the square of the body height in units of kg/m². So the intervention can be BMI.*

However, there are different levels of ambiguity of the interventions above. The meanings of interventions in Examples 2.1–2.4 are relatively clear, but the meaning of intervention on BMI in Example 2.5 is less clear. In particular, we can imagine different versions of BMI reduction: healthier diet, more physical

exercise, bariatric surgery, etc. These different versions of intervention can have quite different effects on the outcome. In this book, we will view the intervention in Example 2.5 as ill-defined without further clarifications.

Another ill-defined intervention is race. Racial discrimination is an important issue in labor market, but it is not easy to imagine an experiment to change the race of any experimental unit. Bertrand and Mullainathan (2004) give an interesting experiment that partially answers the question.

Example 2.6 *Bertrand and Mullainathan (2004) randomly change the names on the resumes, and compare the callback rates of resumes with African-American- or White-sounding names. For each resume, the intervention is the binary indicator of African-American- or White-sounding name, and the outcome is the binary indicator of callback. We have analyzed the following two-by-two table in Section 1.2.2:*

	callback	no callback
African-American	157	2278
White	235	2200

From the above, we can compare the the probabilities of being called back among African-American- and White-sounding names:

$$\frac{157}{2278 + 157} - \frac{235}{2200 + 235} = 6.45\% - 9.65\% = -3.20\% < 0$$

with p-value from the Fisher exact test much smaller than 0.001.

In Bertrand and Mullainathan (2004)'s experiment, the treatment is the perceived race which can be manipulated by experimenters. They design an experiment to answer a well-defined causal question.

2.2 Formal notation of potential outcomes

Consider a study with n experimental units indexed by $i = 1, \dots, n$. As a starting point, we focus on a treatment with two levels: 1 for the treatment and 0 for the control. For each unit i , the outcome of interest Y has two versions:

$$Y_i(1) \text{ and } Y_i(0),$$

which are potential outcomes under the hypothetical interventions 1 and 0. Neyman (1923) first used this notation. It seems intuitive but has some hidden assumptions. Rubin (1980) made the following clarifications on the hidden assumptions.

Assumption 2.1 (no interference) *Unit i 's potential outcomes do not depend on other units' treatments. This is sometimes called the no-interference assumption.*

Assumption 2.2 (consistency) *There are no other versions of the treatment. Equivalently, we require that the treatment level be well defined, or have no ambiguity at least for the outcome of interest. This is sometimes called the consistency assumption.*

Assumption 2.1 can be violated in infectious diseases or network experiments. For instance, if some of my friends receive flu shots, my chance of getting the flu decrease even if I do not receive the flu shot; if my friends see an ad on Facebook, my chance of buying that product increase even if I do not see the ad. It is an active research area to study situations with interfering units in modern causal inference literature.

Assumption 2.2 can be violated for treatment with complex components. For instance, when studying the effect of cigarette smoking on lung cancer, the type of cigarettes may matter; when studying the effect of college education on income, the type and major of college education may matter.

Rubin (1980) called the Assumptions 2.1 and 2.2 above together the Stable Unit Treatment Value Assumption (SUTVA).

Assumption 2.3 (SUTVA) *Both Assumptions 2.1 and 2.2 hold.*

Under SUTVA, Rubin (2005) called the $n \times 2$ matrix of potential outcomes the Science Table:

i	$Y_i(1)$	$Y_i(0)$
1	$Y_1(1)$	$Y_1(0)$
2	$Y_2(1)$	$Y_2(0)$
\vdots	\vdots	\vdots
n	$Y_n(1)$	$Y_n(0)$

Due to Neyman and Rubin's fundamental contribution to statistical causal inference, the potential outcomes framework is sometimes called the Neyman model, the Neyman–Rubin model, or the Rubin Causal Model.

Causal effects are functions of the Science Table. Inferring individual causal effects

$$\tau_i = Y_i(1) - Y_i(0)$$

is fundamentally challenging because we can only observe either $Y_i(1)$ or $Y_i(0)$ for each unit i , that is, we can observe only half of the Science Table. As a starting point, most parts of the book focus on the average causal effect (ACE):

$$\tau = n^{-1} \sum_{i=1}^n \{Y_i(1) - Y_i(0)\} = n^{-1} \sum_{i=1}^n Y_i(1) - n^{-1} \sum_{i=1}^n Y_i(0).$$

But we can easily extend our discussion to many other parameters (also called estimands).

2.2.1 Causal effects, subgroups, and the non-existence of Yule–Simpson Paradox

If we have two subgroups defined by a binary variable x_i , we can define the subgroup causal effects as

$$\tau_x = \frac{\sum_{i=1}^n I(x_i = x) \{Y_i(1) - Y_i(0)\}}{\sum_{i=1}^n I(x_i = x)}, \quad (x = 0, 1)$$

where $I(\cdot)$ is the indicator function. A simple identity is that

$$\tau = \pi_1 \tau_1 + \pi_0 \tau_0$$

where $\pi_x = \sum_{i=1}^n I(x_i = x)/n$ is the proportion of units with $x_i = x$ ($x = 0, 1$). Therefore, if $\tau_1 > 0$ and $\tau_0 > 0$, we must have $\tau > 0$. The Yule–Simpson Paradox thus cannot happen to causal effects.

2.2.2 Subtlety of experimental unit

I end this section with a subtlety related to the definition of the experimental unit. Simply speaking, the experimental unit can be different from the physical unit. For example, if I did not take aspirin before and my headache did not go away, but I take aspirin now and my headache goes away, you might think that we can observe my potential outcomes under both control and treatment. Let i index myself, and let $Y = 1$ denote the indicator of no headache. Then, the above heuristic suggests that $Y_i(0) = 0$ and $Y_i(1) = 1$, so it seems that aspirin kills my headache. But this logic is very wrong because of the misunderstanding of the definition of the experimental unit. At different time points, I, the same physical person, become two distinct experiment units, indexed by “ i , before” and “ i , after”. Therefore, we have four potential outcomes

$$Y_{i,\text{before}}(0) = 0, \quad Y_{i,\text{before}}(1) = ?, \quad Y_{i,\text{after}}(0) = ?, \quad Y_{i,\text{after}}(1) = 1,$$

with two of them observed and two of them missing. The individual causal effects

$$Y_{i,\text{before}}(1) - Y_{i,\text{before}}(0) = ? - 0 \text{ and } Y_{i,\text{after}}(1) - Y_{i,\text{after}}(0) = 1 - ?$$

are unknown. It is possible that my headache goes away even if I do not take aspirin:

$$Y_{i,\text{after}}(0) = 1, \quad Y_{i,\text{after}}(1) = 1$$

which implies zero effect; it is also possible that my headache does not go away if I do not take aspirin:

$$Y_{i,\text{after}}(0) = 0, \quad Y_{i,\text{after}}(1) = 1$$

which implies a positive effect of aspirin.

The wrong heuristic argument might get the right answer if the control potential outcomes are stable at the before and after periods: $Y_{i,\text{before}}(0) = Y_{i,\text{after}}(0) = 0$. But this assumption is rather strong and fundamentally untestable.

2.3 Treatment assignment mechanism

Let Z_i be the binary treatment indicator for unit i , vectorized as $\mathbf{Z} = (Z_1, \dots, Z_n)$. The observed outcome of unit i is a function of the potential outcomes and the treatment indicator:

$$Y_i = \begin{cases} Y_i(1), & \text{if } Z_i = 1 \\ Y_i(0), & \text{if } Z_i = 0 \end{cases} \quad (2.1)$$

$$= Z_i Y_i(1) + (1 - Z_i) Y_i(0) \quad (2.2)$$

$$= Y_i(0) + Z_i \{Y_i(1) - Y_i(0)\} \quad (2.3)$$

$$= Y_i(0) + Z_i \tau_i. \quad (2.4)$$

Equation (2.1) is the definition of the observed outcome. Equation (2.2) is equivalent to (2.1). It is a trivial fact, but Judea Pearl viewed it as the fundamental bridge between the potential outcomes and the observed outcome. Equations (2.3) and (2.4) highlight the fact that the individual causal effect $\tau_i = Y_i(1) - Y_i(0)$ can be heterogeneous across units.

The experiment only reveals one of unit i 's potential outcomes with the other one missing:

$$\begin{aligned} Y_i^{\text{mis}} &= \begin{cases} Y_i(0), & \text{if } Z_i = 1 \\ Y_i(1), & \text{if } Z_i = 0 \end{cases} \\ &= Z_i Y_i(0) + (1 - Z_i) Y_i(1). \end{aligned}$$

The missing potential outcome correspond to the opposite treatment level of unit i . For this reason, the potential outcomes framework is also called the counterfactual framework. This name can be confusing because before the experiment, both potential outcomes are observable, and after the experiment, one potential outcomes is actually observed.

The treatment assignment mechanism, i.e., the probability distribution of \mathbf{Z} , plays an important role in inferring causal effects. The following simple numerical examples illustrate this point. We first generate potential outcomes from Normal distributions with the average causal effect close to -0.5 .

```
> n = 500
> Y0 = rnorm(n)
> tau = - 0.5 + Y0
> Y1 = Y0 + tau
```

A perfect doctor assigns the treatment to the patient if s/he knows that the individual causal effect is non-negative. This results in a positive difference in means of the observed outcomes:

```
> Z = (tau >= 0)
> Y = Z*Y1 + (1 - Z)*Y0
```

```
> mean(Y[Z==1]) - mean(Y[Z==0])
[1] 2.166509
```

A clueless doctor does not know any information about the individual causal effects and assigns the treatment to patients by flipping a fair coin. This results in a difference in means of the observed outcomes close to the true average causal effect:

```
> Z = rbinom(n, 1, 0.5)
> Y = Z*Y1 + (1 - Z)*Y0
> mean(Y[Z==1]) - mean(Y[Z==0])
[1] -0.552064
```

The above examples are hypothetical since no doctors perfectly know the individual causal effects. However, the examples do demonstrate the crucial role of the treatment assignment mechanism. This book will organize the topics based on the treatment assignment mechanism.

2.4 Homework Problems

2.1 A perfect doctor

Following the first perfect doctor example in Section 2.3, assume the potential outcomes are random variables generated from

$$Y(0) \sim N(0, 1), \quad \tau = -0.5 + Y(0), \quad Y(1) = Y(0) + \tau.$$

The binary treatment is determined by the treatment effect as $Z = 1(\tau \geq 0)$, and the observed outcome is determined by the potential outcomes and the treatment by $Y = ZY(1) + (1 - Z)Y(0)$. Calculate the difference in means

$$E(Y | Z = 1) - E(Y | Z = 0).$$

Hint: The mean of a truncated Normal random variable equals

$$E(X | a < X < b) = \mu - \sigma \frac{\phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)},$$

where $X \sim N(\mu, \sigma^2)$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density and cumulative distribution functions of a standard Normal random variable.

2.2 Nonlinear causal estimands

With potential outcomes $\{(Y_i(1), Y_i(0)\}_{i=1}^n$ for n units under the treatment and control, the difference in means equals the mean of the individual treatment effects:

$$\bar{Y}(1) - \bar{Y}(0) = n^{-1} \sum_{i=1}^n \{Y_i(1) - Y_i(0)\}.$$

Therefore, the average treatment effect is a linear causal estimand.

Other estimands may not be linear. For instance, we can define the median treatment effect as

$$\delta_1 = \text{median}\{(Y_i(1))_{i=1}^n - \text{median}\{(Y_i(0))_{i=1}^n\},$$

which is, in general, different from the median of the individual treatment effect

$$\delta_2 = \text{median}\{(Y_i(1) - Y_i(0))_{i=1}^n\}.$$

1. Give numerical examples which have $\delta_1 = \delta_2$, $\delta_1 > \delta_2$, and $\delta_1 < \delta_2$.
2. Which estimand makes more sense, δ_1 or δ_2 ? Why? Use examples to justify your conclusion. If you feel that both δ_1 and δ_2 can make sense in different applications, you can also give examples to justify both estimands.

2.3 Average and individual effects

Give a numerical example in which $\tau = n^{-1} \sum_{i=1}^n \{Y_i(1) - Y_i(0)\} > 0$ but the proportion of units with $Y_i(1) > Y_i(0)$ is smaller than 0.5. That is, the average causal effect is positive, but the treatment benefits less than half of the units.

2.4 Recommended reading

Holland (1986) is a classic review article on statistical causal inference. It popularized the name “Rubin Causal Model” for the potential outcomes framework. At the University of California Berkeley, we call it the “Neyman Model” for obvious reasons.



Part II

Randomized experiments



3

The Completely Randomized Experiment and the Fisher Randomization Test

The potential outcomes framework has intrinsic connections with randomized experiments. Understanding causal inference with various randomized experiments is fundamental and quite helpful for understanding causal inference in more complicated non-experimental studies.

Part II of this book focuses on randomized experiments. This chapter focuses on the simplest experiment, the completely randomized experiment (CRE).

3.1 CRE

Consider an experiment with n units, with n_1 receiving the treatment and n_0 receiving the control. We can define the CRE based on its treatment assignment mechanism¹.

Definition 3.1 (CRE) *A CRE has the treatment assignment mechanism:*

$$\text{pr}(\mathbf{Z} = \mathbf{z}) = 1 / \binom{n}{n_1},$$

where $\mathbf{z} = (z_1, \dots, z_n)$ satisfies $\sum_{i=1}^n z_i = n_1$ and $\sum_{i=1}^n (1 - z_i) = n_0$.

In Definition 3.1, we view the potential outcome vector under treatment $\mathbf{Y}(1) = (Y_1(1), \dots, Y_n(1))$ and the potential outcome vector under control $\mathbf{Y}(0) = (Y_1(0), \dots, Y_n(0))$ are both fixed. Even if we view them as random, we can condition on them and the treatment assignment mechanism becomes

$$\text{pr}\{\mathbf{Z} = \mathbf{z} \mid \mathbf{Y}(1), \mathbf{Y}(0)\} = 1 / \binom{n}{n_1}$$

¹Readers may think that a CRE has Z_i 's as independent and identically distributed (IID) Bernoulli random variables with probability π , in which n_1 is a $\text{Binomial}(n, \pi)$ random variable. This is called the Bernoulli randomized experiment (BRE), which reduces to the CRE if we condition on (n_1, n_0) . I will give more details for the BRE in Problem 4.7 in Chapter 4.

because $\mathbf{Z} \perp\!\!\!\perp \{\mathbf{Y}(1), \mathbf{Y}(0)\}$ in a CRE. In a CRE, the treatment vector \mathbf{Z} is from a random permutation of n_1 1's and n_0 0's.

In his seminal book *Design of Experiments*, Fisher (1935) pointed out the following advantages of randomization:

1. It creates comparable treatment and control groups on average.
2. It serves as a “reasoned basis” for statistical inference.

Point 1 is intuitive because the random treatment assignment does not bias toward the treatment or the control. Most people understand point 1 well. Point 2 is more subtle. What Fisher meant is that randomization justifies a statistical test, which is now called the Fisher Randomization Test (FRT). This chapter illustrates the basic idea of the FRT under a CRE.

3.2 FRT

Fisher (1935) was interested in testing the following null hypothesis:

$$H_{0F} : Y_i(1) = Y_i(0) \text{ for all units } i = 1, \dots, n.$$

Rubin (1980) called it the *sharp null hypothesis* in the sense that it can determine all the potential outcomes based on the observed data: $\mathbf{Y}(1) = \mathbf{Y}(0) = \mathbf{Y} = (Y_1, \dots, Y_n)$, the vector of the observed outcomes. It is also called the *strong null hypothesis* (e.g., Wu and Ding, 2021).

Conceptually, under H_{0F} , the FRT works for any test statistic

$$T = T(\mathbf{Z}, \mathbf{Y}), \quad (3.1)$$

which is a function of the observed data. The observed outcome vector \mathbf{Y} is fixed under H_{0F} , so the only random component in the test statistic T is the treatment vector \mathbf{Z} . The experimenter determines the distribution of \mathbf{Z} , which in turn determines the distribution of T under H_{0F} . This is the basis for calculating the *p*-value. I will give more details below.

In a CRE, \mathbf{Z} is uniform over the set

$$\{\mathbf{z}^1, \dots, \mathbf{z}^M\}$$

where $M = \binom{n}{n_1}$, and the \mathbf{z}^m 's are all possible vectors with n_1 1's and n_0 0's. For instance, with $n = 5$ and $n_1 = 3$, we can enumerate $M = \binom{5}{3} = 10$ vectors as follows:

```
> permutation10 = function(n, n1){
+   M = choose(n, n1)
+   treat.index = combn(n, n1)
+   Z = matrix(0, n, M)
```

```

+   for(m in 1:M){
+     treat = treat.index[, m]
+     Z[treat, m] = 1
+   }
+ Z
+ }
>
> permutation10(5, 3)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    1    1    1    1    1    0    0    0    0
[2,]    1    1    1    0    0    0    1    1    1    0
[3,]    1    0    0    1    1    0    1    1    0    1
[4,]    0    1    0    1    0    1    1    0    1    1
[5,]    0    0    1    0    1    1    0    1    1    1

```

As a consequence, T is uniform over the set (with possible duplications)

$$\{T(\mathbf{z}^1, \mathbf{Y}), \dots, T(\mathbf{z}^M, \mathbf{Y})\}.$$

That is, the distribution of T is known due to the design of the CRE. We will call this distribution of T the *randomization distribution*.

If larger values are more extreme for T , we can use the following tail probability to measure the extremeness of the test statistic with respect to its randomization distribution:

$$p_{\text{FRT}} = M^{-1} \sum_{m=1}^M I\{T(\mathbf{z}^m, \mathbf{Y}) \geq T(\mathbf{Z}, \mathbf{Y})\}, \quad (3.2)$$

which is called the *p*-value by Fisher. Figure 3.1 illustrates the computational process of p_{FRT} .

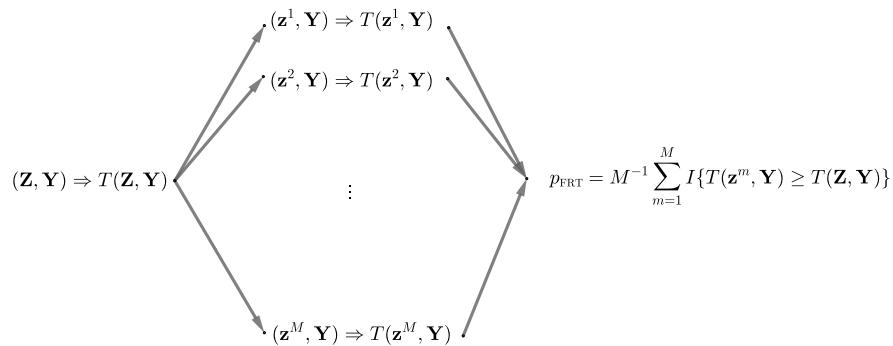


FIGURE 3.1: Illustration of the FRT

The *p*-value, p_{FRT} , in (3.2) works for any choice of test statistic and any outcome-generating process. It also extends naturally to any experiments,

which will be a topic repeatedly discussed in the following chapters. Importantly, it is finite-sample exact in the sense² that under H_{0F} ,

$$\text{pr}(p_{\text{FRT}} \leq u) \leq u \quad \text{for all } 0 \leq u \leq 1. \quad (3.3)$$

In practice, M is often too large (e.g., with $n = 100, n_1 = 50$, we have $M > 10^{29}$), and it is computationally infeasible to enumerate all possible values of the treatment vector. We often approximate p_{FRT} by Monte Carlo. To be more specific, we take simple random draws from the possible values of the treatment vector, or, equivalently, we randomly permute \mathbf{Z} , and approximate p_{FRT} by

$$\hat{p}_{\text{FRT}} = R^{-1} \sum_{r=1}^R I\{T(\mathbf{z}^r, \mathbf{Y}) \geq T(\mathbf{Z}, \mathbf{Y})\}, \quad (3.4)$$

where the \mathbf{z}^r 's are the R random permutations of \mathbf{Z} . The p -value in (3.4) has Monte Carlo error decreasing fast with an increasing R ; see Problem 3.2. Because the calculation of the p -value in (3.4) involves permutations of \mathbf{Z} , the FRT is sometimes called the *permutation test* in the context of the CRE. However, the idea of FRT is more general than the permutation test in more complex experiments.

3.3 Canonical choices of the test statistic

From the above discussion, the FRT generates finite-sample exact p -value for any choice of test statistic. This is a feature of the FRT. However, this feature should not encourage arbitrary choice of the test statistic. Intuitively, we must choose test statistics that give information for the possible violations of H_{0F} . Below I will review some canonical choices.

Example 3.1 (difference-in-means) *The difference-in-means statistic is*

$$\hat{\tau} = \hat{Y}(1) - \hat{Y}(0)$$

where

$$\hat{Y}(1) = n_1^{-1} \sum_{Z_i=1} Y_i = n_1^{-1} \sum_{i=1}^n Z_i Y_i$$

²This is the standard definition of the p -value in mathematical statistics. The inequality is often due to discreteness of the test statistic, and when the equality holds, the p -value is Uniform(0, 1) under the null hypothesis. Let $F(\cdot)$ be the distribution function of $T(\mathbf{Z}, \mathbf{Y})$. Even though it is a step function, we assume that it is continuous and strictly increasing as if it is the distribution function of a continuous random variable taking values on the whole real line. So $p_{\text{FRT}} = 1 - F(T)$, and

$$\text{pr}(p_{\text{FRT}} \leq u) = \text{pr}\{1 - F(T) \leq u\} = \text{pr}\{T \geq F^{-1}(1 - u)\} = 1 - F(F^{-1}(1 - u)) = u.$$

The discreteness of T does cause some technical issues in the proof, yielding an inequality instead of an equality. I leave the technical details in Problem 3.1.

is the sample mean of the outcomes under the treatment and

$$\hat{Y}(0) = n_0^{-1} \sum_{Z_i=0} Y_i = n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i$$

is the sample mean of the outcomes under the control, respectively. Under H_{0F} , it has mean

$$E(\hat{\tau}) = n_1^{-1} \sum_{i=1}^n E(Z_i) Y_i - n_0^{-1} \sum_{i=1}^n E(1 - Z_i) Y_i = 0$$

and variance

$$\begin{aligned} \text{var}(\hat{\tau}) &= \text{var} \left\{ n_1^{-1} \sum_{i=1}^n Z_i Y_i - n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i \right\} \\ &= \text{var} \left(\frac{n}{n_0 n_1} \sum_{i=1}^n Z_i Y_i \right) \\ &=_* \frac{n^2}{n_0^2} \left(1 - \frac{n_1}{n} \right) \frac{s^2}{n_1} \\ &= \frac{n}{n_1 n_0} s^2, \end{aligned}$$

where $=_*$ follows from Lemma A3.2 for simple random sampling with

$$\bar{Y} = n^{-1} \sum_{i=1}^n Y_i, \quad s^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Furthermore, the randomization distribution of $\hat{\tau}$ is approximately Normal due to the finite population central limit theorem in Lemma A3.4:

$$\frac{\hat{\tau}}{\sqrt{\frac{n}{n_1 n_0} s^2}} \rightarrow N(0, 1) \tag{3.5}$$

in distribution. Since s^2 is fixed under H_{0F} , it is equivalent to use

$$\frac{\hat{\tau}}{\sqrt{\frac{n}{n_1 n_0} s^2}}$$

as the test statistic in the FRT, which is asymptotically Normal as shown above. Then we can calculate an approximate p-value.

The observed data are $\{Y_i : Z_i = 1\}$ and $\{Y_i : Z_i = 0\}$, so the problem is essentially a two-sample problem. Under the assumption of IID Normal

outcomes (see Section A1.4.1), the classic two-sample t -test assuming equal variance is based on

$$\sqrt{\frac{\hat{\tau}}{\frac{n}{n_1 n_0(n-2)} \left[\sum_{Z_i=1} \{Y_i - \hat{Y}(1)\}^2 + \sum_{Z_i=0} \{Y_i - \hat{Y}(0)\}^2 \right]}} \sim t_{n-2}. \quad (3.6)$$

Based on some algebra (see Problem 3.8), we have the expansion

$$(n-1)s^2 = \sum_{Z_i=1} \{Y_i - \hat{Y}(1)\}^2 + \sum_{Z_i=0} \{Y_i - \hat{Y}(0)\}^2 + \frac{n_1 n_0}{n} \hat{\tau}^2. \quad (3.7)$$

With a large sample size n , we can ignore the difference between $N(0, 1)$ and t_{n-2} and the difference between $n - 1$ and $n - 2$. Moreover, under H_{0F} , $\hat{\tau}$ converges to zero in probability, so $n_1 n_0 / n \hat{\tau}^2$ can be ignored asymptotically. Therefore, under H_{0F} , the approximate p -value in Example 3.1 is close to the p -value from the classic two-sample t -test assuming equal variance, which can be calculated by `t.test` with `var.equal = TRUE`. Under alternative hypotheses with nonzero τ , the additional term $\frac{n_1 n_0}{n} \hat{\tau}^2$ in the above expansion can make the FRT less powerful than the usual t -test.

Based on the above discussion, the FRT with $\hat{\tau}$ effectively uses a pooled variance ignoring the heteroskedasticity between these two groups. In classical statistics, the two-sample problem with heteroskedastic Normal outcomes is called the Behrens–Fisher problem (see Section A1.4.1). In the Behrens–Fisher problem, a standard choice of the test statistic is the studentized statistic below.

Example 3.2 (studentized statistic) *The studentized statistic is*

$$t_{\text{unequal}} = \frac{\hat{Y}(1) - \hat{Y}(0)}{\sqrt{\frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}}},$$

where

$$\hat{S}^2(1) = (n_1 - 1)^{-1} \sum_{Z_i=1} \{Y_i - \hat{Y}(1)\}^2, \quad \hat{S}^2(0) = (n_0 - 1)^{-1} \sum_{Z_i=0} \{Y_i - \hat{Y}(0)\}^2$$

are the sample variances of the observed outcomes under the treatment and control, respectively. Under H_{0F} , the finite population central limit theorem again implies that t is asymptotically Normal:

$$t \rightarrow N(0, 1)$$

in distribution. Then we can calculate an approximate p -value which is close to the p -value from `t.test` with `var.equal = FALSE`.

An extremely important point is that the FRT justifies the traditional t -tests using `t.test` with either `var.equal = TRUE` or `var.equal = FALSE`, even if the underlying distributions are not Normal. Standard statistics textbooks motivate the t -tests based on the Normality assumption, but the assumption is too strong. Fortunately, the t -test procedures can still be used as long as the finite population central limit theorems hold. Even if we do not believe the central limit theorems, we can still use $\hat{\tau}$ and t as test statistics in the FRT to obtain finite-sample exact p -values.

We will motivate this studentized statistic from another perspective in Chapter 8. The theory shows that using t in FRT is more robust to heteroskedasticity across the two groups.

The following test statistic is robust to outliers resulting from heavy-tailed outcome data.

Example 3.3 (Wilcoxon rank sum) *The difference-in-means statistic uses the original outcomes, and its sampling distribution depends on the second moments of the outcomes. This makes it sensitive to outliers. Another popular test statistic is based on the ranks of the pooled observed outcomes. Let R_i denote the rank of Y_i in the pooled samples \mathbf{Y} :*

$$R_i = \#\{j : Y_j \leq Y_i\}.$$

The Wilcoxon rank sum statistic is the sum of the ranks under treatment:

$$W = \sum_{i=1}^n Z_i R_i.$$

For algebraic simplicity, we assume that there are no ties in the outcomes, although the FRT can be applied regardless of the existence of ties. For the case with ties, see Lehmann (1975, Chapter 1 Section 4). Because the sum of the ranks of the pooled samples are fixed at $1 + 2 + \dots + n = n(n+1)/2$, the Wilcoxon statistic is equivalent to the difference in the means of the ranks under treatment and control. Under H_{0F} , the R_i 's are fixed, so W has mean

$$E(W) = \sum_{i=1}^n E(Z_i) R_i = \frac{n_1}{n} \sum_{i=1}^n i = \frac{n_1}{n} \times \frac{n(n+1)}{2} = \frac{n_1(n+1)}{2}$$

and variance

$$\begin{aligned}
\text{var}(W) &= \text{var} \left(n_1 \frac{1}{n_1} \sum_{i=1}^n Z_i R_i \right) \\
&=_* n_1^2 \left(1 - \frac{n_1}{n} \right) \frac{1}{n_1} \frac{1}{n-1} \sum_{i=1}^n \left(R_i - \frac{n+1}{2} \right)^2 \\
&= \frac{n_1 n_0}{n(n-1)} \sum_{i=1}^n \left(i - \frac{n+1}{2} \right)^2 \\
&= \frac{n_1 n_0}{n(n-1)} \left\{ \sum_{i=1}^n i^2 - n \left(\frac{n+1}{2} \right)^2 \right\} \\
&= \frac{n_1 n_0}{n(n-1)} \left\{ \frac{n(n+1)(2n+1)}{6} - n \left(\frac{n+1}{2} \right)^2 \right\} \\
&= \frac{n_1 n_0 (n+1)}{12},
\end{aligned}$$

where $=_*$ follows from Lemma A3.2. Furthermore, under H_{0F} , the finite population central limit theorem ensures that the randomization distribution of $\hat{\tau}$ is approximately Normal:

$$\frac{\sum_{i=1}^n Z_i R_i - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_0 (n+1)}{12}}} \rightarrow N(0, 1) \quad (3.8)$$

in distribution. Based on (3.8), we can conduct an asymptotic test. In R, the function `wilcox.test` can compute both exact and asymptotic p-values based on the statistic $W - n_1(n_1 + 1)/2$. Based on some asymptotic analyses, Lehmann (1975) showed that the FRT using W has reasonable powers over a wide range of data generating processes.

Example 3.4 (Kolmogorov–Smirnov statistic) The treatment may affect the outcome in different ways. It seems natural to summarize the treatment outcomes and control outcomes based on the empirical distributions:

$$\hat{F}_1(y) = n_1^{-1} \sum_{i=1}^n Z_i I(Y_i \leq y), \quad \hat{F}_0(y) = n_0^{-1} \sum_{i=1}^n (1 - Z_i) I(Y_i \leq y).$$

Comparing these two empirical distributions yields the famous Kolmogorov–Smirnov statistic

$$D = \max_y |\hat{F}_1(y) - \hat{F}_0(y)|.$$

It is a challenging mathematics problem to derive the distribution of D . With large sample sizes, its distribution function converges to

$$\text{pr} \left(\frac{n_1 n_0}{n} D \leq x \right) \rightarrow \frac{\sqrt{2\pi}}{x} \sum_{j=1}^{\infty} e^{-(2j-1)^2 \pi^2 / (8x^2)},$$

based on which we calculate an asymptotic p -value (Van der Vaart, 2000). In R, `ks.test` can compute both exact and asymptotic p -values.

3.4 A case study of the LaLonde experimental data

I use LaLonde (1986)'s experimental data to illustrate the FRT. The data are available in the `Matching` package (Sekhon, 2011):

```
> library(Matching)
> data(lalonde)
> z = lalonde$treat
> y = lalonde$re78
```

Figure 3.2 shows the histograms of the outcomes under the treatment and control.

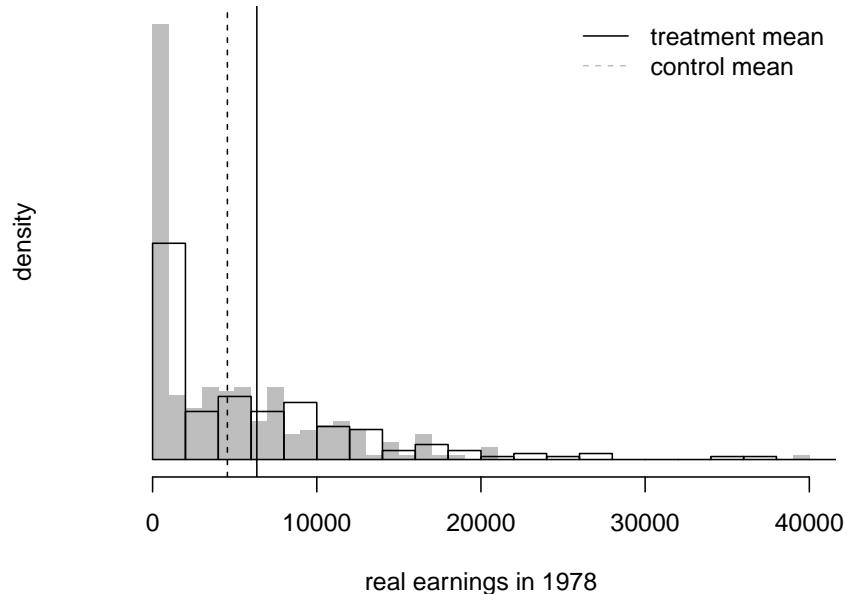


FIGURE 3.2: Histograms of the outcomes in the LaLonde experimental data: treatment in white and control in grey

The following code computes the observed values of the test statistics using existing functions:

```
> tauhat = t.test(y[z == 1], y[z == 0],
+                   var.equal = TRUE)$statistic
> tauhat
      t
2.835321
> student = t.test(y[z == 1], y[z == 0],
+                     var.equal = FALSE)$statistic
> student
      t
2.674146
> W = wilcox.test(y[z == 1], y[z == 0])$statistic
> W
      W
27402.5
> D = ks.test(y[z == 1], y[z == 0])$statistic
> D
      D
0.1321206
```

By randomly permuting the treatment vector, we can obtain the Monte Carlo approximation of the randomization distributions of the test statistics, stored in four vectors `Tauhat`, `Student`, `Wilcox`, and `Ks`.

```
> MC = 10^4
> Tauhat = rep(0, MC)
> Student = rep(0, MC)
> Wilcox = rep(0, MC)
> Ks = rep(0, MC)
> for(mc in 1:MC)
+ {
+     zperm = sample(z)
+     Tauhat [mc] = t.test(y[zperm == 1], y[zperm == 0],
+                           var.equal = TRUE)$statistic
+     Student [mc] = t.test(y[zperm == 1], y[zperm == 0],
+                           var.equal = FALSE)$statistic
+     Wilcox [mc] = wilcox.test(y[zperm == 1], y[zperm == 0])$statistic
+     Ks [mc] = ks.test(y[zperm == 1], y[zperm == 0])$statistic
+ }
```

The one-sided p -values based on the FRT are all smaller than 0.05:

```
> exact.pv = c(mean(Tauhat >= tauhat),
+               mean(Student >= student),
+               mean(Wilcox >= W),
+               mean(Ks >= D))
> round(exact.pv, 3)
[1] 0.002 0.002 0.006 0.040
```

Without using Monte Carlo, we can also compute the asymptotic p -values which are all smaller than 0.05:

```
> asym.pv = c(t.test(y[z == 1], y[z == 0],
+                     var.equal = TRUE)$p.value,
+                     t.test(y[z == 1], y[z == 0],
+                     var.equal = FALSE)$p.value,
+                     wilcox.test(y[z == 1], y[z == 0])$p.value,
+                     ks.test(y[z == 1], y[z == 0])$p.value)
> round(asym.pv, 3)
[1] 0.005 0.008 0.011 0.046
```

The differences between the p -values are due to the asymptotic approximations as well as the fact that the default choices for `t.test` and `wilcox.test` are two-sided tests.

Figure 3.3 shows the histograms of the randomization distributions of four test statistics, as well as their corresponding observed values. For the first three test statistics, the Normal approximations works quite well even though the underlying outcome data distribution is far from Normal. In general, a figure like Figure 3.3 can give very clear information for testing the sharp null hypothesis. Recently, Bind and Rubin (2020) proposes, in the title of their paper, that “when possible, report a Fisher-exact p -value and display its underlying null randomization distribution.”

3.5 Some history of randomized experiments and FRT

3.5.1 James Lind’s experiment

James Lind (1716–1794) was a Scottish doctor and a pioneer of naval hygiene in the Royal Navy. At his time, scurvy was a major cause of death among sailors. He conducted one of the earliest randomized experiments with a clear documentation of the details, and concluded that citrus fruits cured scurvy before the discovery of Vitamin C.

In Lind (1753), he described the following randomized experiment with 12 patients of scurvy assigned to 6 groups. With some simplifications, the 6 groups are:

1. two received a quart of cider every day;
2. two received twenty-five drops of sulfuric acid three times every day;
3. two received two spoonfuls of vinegar three times every day;
4. two received half a pint of seawater every day;
5. two received two oranges and one lemon every day;
6. two received a spicy paste plus a drink of barley water every day.

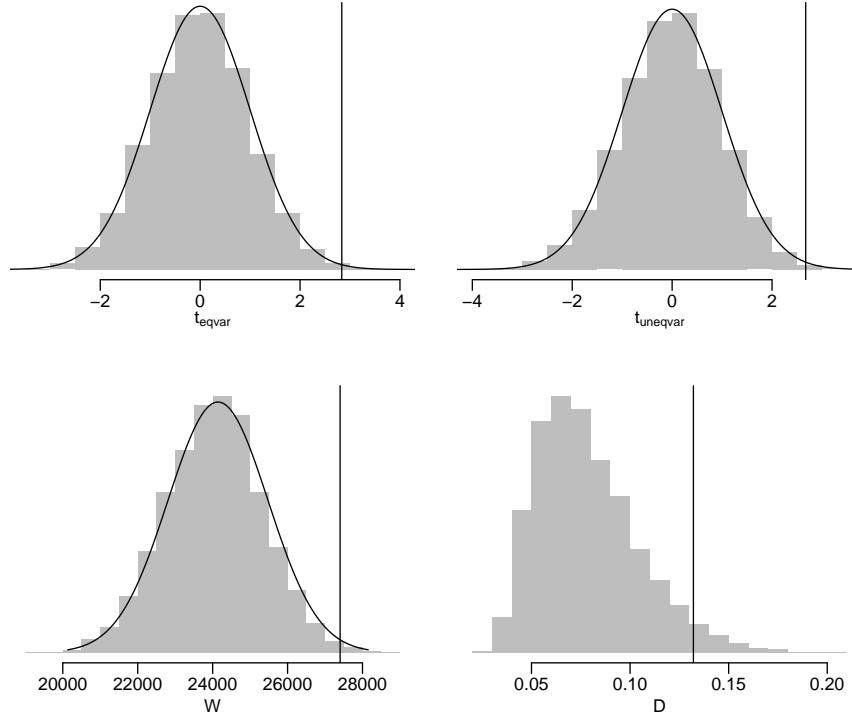


FIGURE 3.3: The randomization distributions of four test statistics based on the LaLonde experimental data

After six days, patients in the fifth group recovered, but patients in other groups did not. If we simplify the treatment as

$$Z_i = 1(\text{unit } i \text{ received citrus fruits})$$

and the outcome as

$$Y_i = 1(\text{unit } i \text{ recovered after six days}),$$

then we have a 2×2 table

	$Y_i = 1$	$Y_i = 0$
$Z_i = 1$	2	0
$Z_i = 0$	0	10

This is the extremest possible 2×2 table we can observe under this experiment, and the data contain strong evidence for the positive effect of citrus fruits for curing scurvy. Statistically, how do we measure the strength of the evidence?

Following the logic of the FRT, if the treatment has no effect at all (under H_{0F}), the extreme 2×2 table will occur with probability

$$\frac{1}{\binom{12}{2}} = \frac{1}{66} = 0.015$$

which is the p_{FRT} . This seems a surprise under H_{0F} : we can easily reject H_{0F} at the level 0.05.

3.5.2 Lady tasting tea

Fisher (1935) described the following famous experiment of *Lady Tasting Tea*³. A lady claimed that she could tell the difference between the two ways of making milk tea: one with milk added first, and the other with tea added first. This might sound odd to most people. As a statistician, Fisher designed an experiment to test whether the lady could tell the difference between the two ways of making milk tea.

He made 8 cups of tea, 4 with milk added first and the other 4 four with tea added first. Then he presented these 8 cups of tea in a random order to the lady, and asked the lady to pick up the 4 with milk added first. The final experiment result can be summarized in the following 2×2 table

	milk first (lady)	tea first (lady)	column sum
milk first (Fisher)	X	$4 - X$	4
tea first (Fisher)	$4 - X$	X	4
row sum	4	4	8

The X can be 0, 1, 2, 3, 4. In the real experiment, $X = 4$, which is the most extreme data, strongly suggesting that the lady could tell the difference of the two ways of making milk tea. Again, how do we measure the strength of the evidence?

Under the null hypothesis that the lady could not tell the difference, only one of the $\binom{8}{4} = 70$ possible orders yields the 2×2 table with $X = 4$. So the p -value is

$$p_{FRT} = \frac{1}{70} = 0.014.$$

Given the significance level 0.05, we reject the null hypothesis.

3.5.3 Two Fisherian principles for experiments

In the above two examples in Sections 3.5.1 and 3.5.2, the p_{FRT} 's are justified by the randomization of the experiments. This highlights the first Fisherian principle of experiments: *randomization*.

Moreover, the above two experiments are in some sense the smallest possible experiments that can yield statistically meaningful results. For instance,

³It becomes the title of a book on the modern history of statistics by Salsburg (2001)

if Lind only assign one patient to each of the six groups, then the smallest p -value is

$$\frac{1}{\binom{6}{1}} = \frac{1}{6} = 0.167;$$

if Fisher only made 6 cups of tea, 3 with milk added first and the other 3 four with tea added first, then the smallest p -value is

$$\frac{1}{\binom{6}{3}} = \frac{1}{20} = 0.05.$$

We can never reject the null hypotheses at the level of 0.05. This highlights the second Fisherian principle of experiments: *replications*.

Chapter 5 will discuss the third Fisherian principle of experiments: *blocking*.

3.6 Discussion

3.6.1 Other sharp null hypotheses and confidence intervals

I focus on the sharp null hypothesis H_{0F} above. In fact, the logic of the FRT also works for other sharp null hypotheses. For instance, we can test

$$H_0(\boldsymbol{\tau}) : Y_i(1) - Y_i(0) = \tau_i \text{ for all } i = 1, \dots, n$$

for a known vector $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n)$. Because the individual causal effects are all known under $H_0(\boldsymbol{\tau})$, we can impute all missing potential outcomes based on the observed data. With known potential outcomes, the distribution of any test statistic is completely determined by the treatment assignment mechanism, and therefore, we can compute the corresponding p_{FRT} as a function of $\boldsymbol{\tau}$, denoted by $p_{FRT}(\boldsymbol{\tau})$. If we can specify all possible $\boldsymbol{\tau}$'s, then we can compute a series of $p_{FRT}(\boldsymbol{\tau})$'s. By duality of hypothesis testing and confidence set (see Section A1.2.5), we can obtain a $(1 - \alpha)$ -level confidence set for the average causal effect:

$$\left\{ \boldsymbol{\tau} = n^{-1} \sum_{i=1}^n \boldsymbol{\tau}_i : p_{FRT}(\boldsymbol{\tau}) \geq \alpha \right\}.$$

Although this strategy is conceptually straightforward, it has practical complexities due to the large number of all possible $\boldsymbol{\tau}$'s. In the special case of a binary outcome, Rigdon and Hudgens (2015) and Li and Ding (2016) proposed some computationally feasible approaches to constructing confidence intervals for $\boldsymbol{\tau}$ based on the FRT. For general unbounded outcomes, this strategy is often computationally infeasible.

A canonical simplification is to consider a subclass of the sharp null hypotheses with constant individual causal effects:

$$H_0(c) : Y_i(1) - Y_i(0) = c \text{ for all } i = 1, \dots, n$$

for a known constant c . Given c , we can compute $p_{\text{FRT}}(c)$. By duality, we can obtain a $(1 - \alpha)$ -level confidence set for the average causal effect:

$$\{c : p_{\text{FRT}}(c) \geq \alpha\}.$$

Because this procedure only involves one-dimensional search, it is computationally feasible. However, it is often criticized that the constant individual causal effect assumption is too strong which does not hold for a binary outcome in particular.

3.6.2 Other test statistics

The FRT is a general strategy that is applicable in any randomized experiments with any test statistic. I give several examples of test statistics in Section 3.3. In fact, the definition of a test statistic can be much more general. For instance, with pre-treatment covariate matrix \mathbf{X} with the i th row being X_i for unit i ($i = 1, \dots, n$)⁴, we can allow the test statistic $T(\mathbf{Z}, \mathbf{Y}, \mathbf{X})$ to be a function of the treatment vector, outcome vector, and the covariate matrix. Problem 3.6 gives an example.

3.6.3 Final remarks

For a general experiment, the probability distribution of \mathbf{Z} is not uniform over all possible permutations of n_1 1's and n_0 0's. But its distribution is completely known by the experimenter. Therefore, we can always simulate its distribution which in turn implies the distribution of any test statistic under the sharp null hypothesis. A finite-sample exact p -value follows from (3.2). I will discuss other experiments in the subsequent chapters and I want to emphasize that the FRT works beyond the specific experiments discussed in this book.

The FRT works with any test statistic. However, this does answer the practical question of how to choose a test statistic in the data analysis. If the goal is to find surprise with respect to the sharp null hypothesis, it is desirable to choose a test statistic that yields high power under alternative hypotheses. In general, no test statistic can dominate others in terms of power

⁴In causal inference, we call X_i a covariate if it is not affected by the treatment. That is, if the covariate has two potential outcomes $X_i(1)$ and $X_i(0)$, then they must satisfy $X_i(1) = X_i(0)$. Standard statistics books often do not distinguish the treatment and covariates because they often appear on the right-hand side of a regression model for the outcome. They are both called covariates in those statistical models. This book distinguishes the treatment and covariates because they play different roles in causal inference.

because power depends on the alternative hypothesis. The four test statistics in Section 3.3 are motivated by different alternative hypotheses. For instance, $\hat{\tau}$ and t are motivated by an alternative hypothesis with nonzero average treatment effect; W is motivated by an alternative hypothesis with a constant causal effect with outliers. Specifying a working alternative hypothesis is often helpful for constructing a test statistic although it does not have to be precise to guarantee the validity of the FRT. Problems 3.6 and 3.7 illustrate the idea of using a working alternative hypothesis or statistical model to construct test statistics.

3.7 Homework Problems

3.1 Exactness of p_{FRT}

Prove (3.2).

3.2 Monte Carlo error of \hat{p}_{FRT}

Given data, p_{FRT} is a fixed number while its Monte Carlo estimator \hat{p}_{FRT} as in (3.4) is random. Show that

$$E_{\text{mc}}(\hat{p}_{\text{FRT}}) = p_{\text{FRT}}$$

and

$$\text{var}_{\text{mc}}(\hat{p}_{\text{FRT}}) \leq \frac{1}{4R},$$

where the subscript ‘‘mc’’ signifies the randomness due to Monte Carlo, that is, \hat{p}_{FRT} is random because \mathbf{z}^r ’s are R independent random draws from all possible values of \mathbf{Z} .

Remark: p_{FRT} is random because \mathbf{Z} is random. But in this problem, we condition on data, so p_{FRT} becomes a fixed number. \hat{p}_{FRT} is random because the \mathbf{z}^r ’s are random permutations of \mathbf{Z} .

Problem 3.2 shows that \hat{p}_{FRT} is unbiased for p_{FRT} over the Monte Carlo randomness and gives an upper bound on the variance of \hat{p}_{FRT} . Luo et al. (2021, Theorem 2) gives a more delicate bound on the Monte Carlo error.

3.3 A finite-sample valid Monte Carlo approximation of p_{FRT}

Although \hat{p}_{FRT} is unbiased for p_{FRT} , it may not be a valid p -value in the sense that $\text{pr}(\hat{p}_{\text{FRT}} \leq u) \leq u$ for all $u \in (0, 1)$ due to Monte Carlo error with a finite R . The following modified Monte Carlo approximation is. Phipson and Smyth (2010) pointed out this trick in the permutation test.

Define

$$\tilde{p}_{\text{FRT}} = \frac{1 + \sum_{r=1}^R I\{T(\mathbf{z}^r, \mathbf{Y}) \geq T(\mathbf{Z}, \mathbf{Y})\}}{1 + R}$$

where the \mathbf{z}^r 's the R random permutations of \mathbf{Z} . Show that with an arbitrary R , the Monte Carlo approximation \tilde{p}_{FRT} is always a finite-sample valid p -value in the sense that $\text{pr}(\tilde{p}_{\text{FRT}} \leq u) \leq u$ for all $u \in (0, 1)$.

Hint: You can use the following two basic probability results to prove the claim in Problem 3.3. First, for two Binomial random variables $X_1 \sim \text{Binomial}(R, p_1)$ and $X_2 \sim \text{Binomial}(R, p_2)$ with $p_1 \geq p_2$, we have $\text{pr}(X_1 \leq x) \leq \text{pr}(X_2 \leq x)$ for all x . Second, if $p \sim \text{Uniform}(0, 1)$ and $X \mid p \sim \text{Binomial}(R, p)$, then, marginally, X is a uniform random variable over $\{0, 1, \dots, R\}$.

3.4 Fisher's exact test

Consider a CRE with a binary outcome, with data summarized in the following 2×2 table:

	$Y = 1$	$Y = 0$	total
$Z = 1$	n_{11}	n_{10}	n_1
$Z = 0$	n_{01}	n_{00}	n_0

Under H_{0F} , show that any test statistic is a function of n_{11} and other non-random fixed constants, and the exact distribution of n_{11} is Hypergeometric. Specify the parameters for the Hypergeometric distribution.

Remark: Barnard (1947) and Ding and Dasgupta (2016) pointed out the equivalence of Fisher's exact test (reviewed in Section A1.3.1) and the FRT under a CRE with a binary outcome.

3.5 More details for lady tasting tea

Recall Section 3.5.2. Calculate $\text{pr}(X = k)$ for $k = 0, 1, 2, 3, 4$.

3.6 Covariate-adjusted FRT

This problem gives more details for Section 3.6.2.

Section 3.4 re-analyzed the LaLonde experimental data using the FRT. The R code `FRTLalonde.R` implemented the FRT with four test statistics. With additional covariates, the FRT can be more general with at least the following two additional strategies. Under the potential outcomes framework, all potential outcomes and covariates are fixed numbers.

First, we can use test statistics based on residuals from the linear regression. Run a linear regression of the outcomes on the covariates, and obtain the residuals (i.e., treat the residuals as the pseudo "outcomes"). Then define the four test statistics based on the residuals. Conduct the FRT using these four new test statistics. Report the corresponding p -values.

Second, we can define the test statistic as the coefficient in the linear regression of the outcomes on the treatment and covariates. Conduct the FRT using this test statistic. Report the corresponding p -value.

Why are the five p -values from the above two strategies finite-sample exact? Justify them.

3.7 FRT with a generalized linear model

Use the same dataset as Problem 3.6 but change the outcome to a binary indicator whether `re78` is positive or not. Run logistic regression of the outcome on the treatment and covariates. Is the coefficient of the treatment significant and what is the p -value? Calculate the p -value from the FRT with the coefficient of the treatment as the test statistic.

3.8 An algebraic detail

Verify (3.7)

3.9 Recommended reading

Bind and Rubin (2020) is a recently paper advocating the use of p -values as well as the display of the corresponding randomization distributions in analyzing complex experiments.

4

Neymanian Repeated Sampling Inference in Completely Randomized Experiments

In his seminal paper, Neyman (1923) not only proposed to use the notation of potential outcomes but also derived rigorous mathematical results for making inference of the average causal effect under a CRE. In contrast to Fisher's idea of calculating the p -value under the sharp null hypothesis, Neyman (1923) proposed an unbiased point estimator and a conservative confidence interval based on the sampling distribution of the point estimator. This chapter will introduce Neyman (1923)'s fundamental results, which are very important for understanding later chapters in Part II of this book.

4.1 Finite population quantities

Consider a CRE with n units, where n_1 of them receive the treatment and n_0 of them receive the control. For unit $i = 1, \dots, n$, we have potential outcomes $Y_i(1)$ and $Y_i(0)$, and individual effect $\tau_i = Y_i(1) - Y_i(0)$. The potential outcomes have finite population means

$$\bar{Y}(1) = n^{-1} \sum_{i=1}^n Y_i(1), \quad \bar{Y}(0) = n^{-1} \sum_{i=1}^n Y_i(0),$$

variances¹

$$S^2(1) = (n-1)^{-1} \sum_{i=1}^n \{Y_i(1) - \bar{Y}(1)\}^2, \quad S^2(0) = (n-1)^{-1} \sum_{i=1}^n \{Y_i(0) - \bar{Y}(0)\}^2,$$

and covariance

$$S(1, 0) = (n-1)^{-1} \sum_{i=1}^n \{Y_i(1) - \bar{Y}(1)\} \{Y_i(0) - \bar{Y}(0)\}.$$

¹Here the divisor $n - 1$ makes the formulas more elegant. Changing the divisor to n complicates the formulas but does not change the results fundamentally. With large n , the difference is minor.

The individual effects have mean

$$\tau = n^{-1} \sum_{i=1}^n \tau_i = \bar{Y}(1) - \bar{Y}(0).$$

and variance

$$S^2(\tau) = (n-1)^{-1} \sum_{i=1}^n (\tau_i - \tau)^2.$$

We have the following relationship between the variances and covariance.

Lemma 4.1 $2S(1, 0) = S^2(1) + S^2(0) - S^2(\tau)$.

The proof of Lemma 4.1 follows from elementary algebra. I leave it as Problem 4.1.

These fixed quantities are functions of the Science Table $\{Y_i(1), Y_i(0)\}_{i=1}^n$. We are interested in estimating the average causal effect τ based on the data $(Z_i, Y_i)_{i=1}^n$ from a CRE.

4.2 Neyman (1923)'s theorem

Based on the observed outcomes, we can calculate the sample means

$$\hat{\bar{Y}}(1) = n_1^{-1} \sum_{i=1}^n Z_i Y_i, \quad \hat{\bar{Y}}(0) = n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i,$$

the sample variances

$$\hat{S}^2(1) = (n_1 - 1)^{-1} \sum_{i=1}^n Z_i \{Y_i - \hat{\bar{Y}}(1)\}^2, \quad \hat{S}^2(0) = (n_0 - 1)^{-1} \sum_{i=1}^n (1 - Z_i) \{Y_i - \hat{\bar{Y}}(0)\}^2.$$

But there are no sample versions of $S(1, 0)$ and $S^2(\tau)$ because the potential outcomes $Y_i(1)$ and $Y_i(0)$ are never jointly observed for each unit i . Neyman (1923) proved the following theorem.

Theorem 4.1 *Under a CRE,*

1. *the difference-in-means estimator $\hat{\tau} = \hat{\bar{Y}}(1) - \hat{\bar{Y}}(0)$ is unbiased for τ :*

$$E(\hat{\tau}) = \tau;$$

2. *$\hat{\tau}$ has variance*

$$\text{var}(\hat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n} \tag{4.1}$$

$$= \frac{n_0}{n_1 n} S^2(1) + \frac{n_1}{n_0 n} S^2(0) + \frac{2}{n} S(1, 0); \tag{4.2}$$

3. the variance estimator

$$\hat{V} = \frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}$$

is conservative for estimating $\text{var}(\hat{\tau})$:

$$E(\hat{V}) - \text{var}(\hat{\tau}) = \frac{S^2(\tau)}{n} \geq 0$$

with the equality holding if and only if $\tau_i = \tau$ for all units.

I will present the proof of Theorem 4.1 in Section 4.3. It is important to clarify the meanings of $E(\cdot)$ and $\text{var}(\cdot)$ in Theorem 4.1. The potential outcomes are all fixed numbers, and only the treatment indicators Z_i 's are random. Therefore, the expectations and variances are all over the randomness of the Z_i 's, which are random permutations of n_1 1's and n_0 0's. Figure 4.1 illustrates the randomness of $\hat{\tau}$, which is a discrete uniform distribution over $\{\hat{\tau}^1, \dots, \hat{\tau}^M\}$ induced by $M = \binom{n}{n_1}$ possible treatment allocations. Compare Figure 4.1 with Figure 3.1 to see the key differences between the FRT and Neyman (1923)'s theorem:

1. the FRT works for any test statistic but Neyman (1923)'s theorem is only about the difference in means. Although we could derive the properties of other estimators similar to Neyman (1923)'s theorem, this mathematical exercise is often quite challenging for general estimators;
2. in Figure 3.1, the observed outcome vector \mathbf{Y} is fixed but in Figure 4.1, the observed outcome vector $\mathbf{Y}(\mathbf{z}^m)$ changes as \mathbf{z}^m changes;
3. the $T(\mathbf{z}^m, \mathbf{Y})$'s are all computable based on the observed data, but the $\hat{\tau}^m$'s are hypothetical values because not all potential outcomes are known.

The point estimator is standard but it has a non-trivial variance under the potential outcomes framework with a CRE. The variance formula (4.1) differs from the classic variance formula for difference in means² because it not only depends on the finite population variances of the potential outcomes but also depends on the finite population variance of the individual effects, or, equivalently, the finite population covariance of the potential outcomes.

²In the classic two-sample problem, the outcomes under treatment are IID draws from a distribution with mean μ_1 and variance σ_1^2 , and the outcomes under control are IID draws from a distribution with mean μ_0 and variance σ_0^2 . Under this assumption, we have

$$\text{var}(\hat{\tau}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}.$$

Here, $\text{var}(\cdot)$ is over the randomness of the outcomes. This variance formula does not involve a third term that depends on the variance of the individual causal effects.

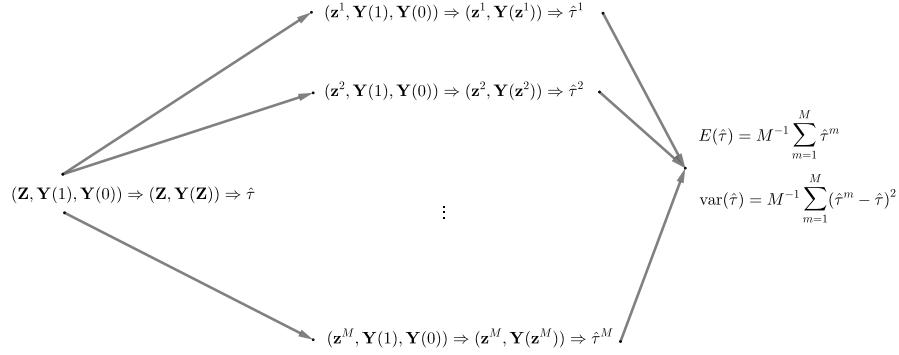


FIGURE 4.1: Illustration of Neyman (1923)'s theorem

Unfortunately, $S^2(\tau)$ and $S(1, 0)$ are not identifiable from the data because $Y_i(1)$ and $Y_i(0)$ are never jointly observed.

Due to the fundamental problem of missing one potential outcome, we can at most obtain a conservative variance estimator. In statistics, the definition of the confidence interval allows for over coverage and thus conservativeness in variance estimation. This may be not a good idea in some applications, for example, studies on side effects of drugs.

The formula (4.1) is a little puzzling in that the more heterogeneous the individual effects are the smaller the variability of $\hat{\tau}$ is. Section 4.5.1 will use numerical examples to verify (4.1). What is the intuition here? I give an explanation based on the equivalent form (4.2). Compare the case with positively correlated potential outcomes and the case with negatively correlated potential outcomes. Although the treatment group is a simple random sample from the finite population of n units, it is possible to observe relatively large treatment potential outcomes in a realized experiment. If this happens, then those control units have relatively small treatment potential outcomes. Consequently, if $S(1, 0) > 0$, then the control potential outcomes are relatively small; if $S(1, 0) < 0$, then the control potential outcomes are relatively large. Therefore, $\hat{\tau}$ tends to larger when the potential outcomes are positively correlated, resulting in more extreme values of $\hat{\tau}$. So the variance of $\hat{\tau}$ is larger when the potential outcomes are positively correlated.

Li and Ding (2017, Theorem 5 and Proposition 3) further proved the following asymptotic Normality of $\hat{\tau}$ based on the finite population central limit theorem.

Theorem 4.2 *Let $n \rightarrow \infty$ and $n_1 \rightarrow \infty$. If n_1/n has a limiting value in $(0, 1)$, $\{S^2(1), S^2(0), S(1, 0)\}$ have limiting values, and*

$$\max_{1 \leq i \leq n} \{Y_i(1) - \bar{Y}(1)\}^2/n \rightarrow 0, \quad \max_{1 \leq i \leq n} \{Y_i(0) - \bar{Y}(0)\}^2/n \rightarrow 0,$$

then

$$\frac{\hat{\tau} - \tau}{\sqrt{\text{var}(\hat{\tau})}} \rightarrow N(0, 1)$$

in distribution, and

$$\hat{S}^2(1) \rightarrow S^2(1), \quad \hat{S}^2(0) \rightarrow S^2(0)$$

in probability.

The proof of Theorem 4.2 is technical and beyond the scope of this book. It ensures that the sampling distribution of $\hat{\tau}$ can be approximated by Normal distribution with large sample size and some regularity conditions. Moreover, it ensures that the sample variances of the outcomes are consistent for the population variances, which further ensures that the probability limit of Neyman (1923)'s variance estimator is larger than the true variance of $\hat{\tau}$. This justifies a conservative large-sample confidence interval for τ :

$$\hat{\tau} \pm z_{1-\alpha/2} \sqrt{\hat{V}},$$

which is the same as the confidence interval for the standard two-sample problem asymptotically. This confidence interval covers τ with probability at least at large as $1 - \alpha$ when the sample size is large enough. By duality, the confidence interval implies a test for $H_{0N} : \tau = 0$.

The conservativeness of Neyman (1923)'s confidence interval for τ is not a big problem if under reporting the treatment effect is not a big problem. It can be problematic if the outcomes measure the side effects of a treatment. In medical experiments, under reporting the side effects of a new drug can have severe consequences.

4.3 Proofs

In this section, I will prove Theorem 4.1.

First, the unbiasedness of $\hat{\tau}$ follows from the representation

$$\begin{aligned}\hat{\tau} &= n_1^{-1} \sum_{i=1}^n Z_i Y_i - n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i \\ &= n_1^{-1} \sum_{i=1}^n Z_i Y_i(1) - n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i(0)\end{aligned}$$

and the linearity of the expectation:

$$\begin{aligned}
E(\hat{\tau}) &= E \left\{ n_1^{-1} \sum_{i=1}^n Z_i Y_i(1) - n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i(0) \right\} \\
&= n_1^{-1} \sum_{i=1}^n E(Z_i) Y_i(1) - n_0^{-1} \sum_{i=1}^n E(1 - Z_i) Y_i(0) \\
&= n_1^{-1} \sum_{i=1}^n \frac{n_1}{n} Y_i(1) - n_0^{-1} \sum_{i=1}^n \frac{n_0}{n} Y_i(0) \\
&= n^{-1} \sum_{i=1}^n Y_i(1) - n^{-1} \sum_{i=1}^n Y_i(0) \\
&= \tau.
\end{aligned}$$

Second, we can further write $\hat{\tau}$ as

$$\hat{\tau} = \sum_{i=1}^n Z_i \left\{ \frac{Y_i(1)}{n_1} + \frac{Y_i(0)}{n_0} \right\} - n_0^{-1} \sum_{i=1}^n Y_i(0).$$

The variance of $\hat{\tau}$ follows from Lemma A3.2 of simple random sampling:

$$\begin{aligned}
\text{var}(\hat{\tau}) &= \frac{n_1 n_0}{n(n-1)} \sum_{i=1}^n \left\{ \frac{Y_i(1)}{n_1} + \frac{Y_i(0)}{n_0} - \bar{Y}(1) - \bar{Y}(0) \right\}^2 \\
&= \frac{n_1 n_0}{n(n-1)} \left[\frac{1}{n_1^2} \sum_{i=1}^n \{Y_i(1) - \bar{Y}(1)\}^2 + \frac{1}{n_0^2} \sum_{i=1}^n \{Y_i(0) - \bar{Y}(0)\}^2 \right. \\
&\quad \left. + \frac{2}{n_1 n_0} \sum_{i=1}^n \{Y_i(1) - \bar{Y}(1)\} \{Y_i(0) - \bar{Y}(0)\} \right] \\
&= \frac{n_0}{n_1 n} S^2(1) + \frac{n_1}{n_0 n} S^2(0) + \frac{2}{n} S(1, 0).
\end{aligned}$$

From Lemma 4.1, we can also write the variance as

$$\begin{aligned}
\text{var}(\hat{\tau}) &= \frac{n_0}{n_1 n} S^2(1) + \frac{n_1}{n_0 n} S^2(0) + \frac{1}{n} \{S^2(1) + S^2(0) - S^2(\tau)\} \\
&= \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n}.
\end{aligned}$$

Third, because the treatment group is a simple random sample of size n_1 from the n units, Lemma A3.3 ensures that the sample variance of $Y_i(1)$'s is unbiased for its population variance:

$$E\{\hat{S}^2(1)\} = S^2(1).$$

Similarly, $E\{\hat{S}^2(0)\} = S^2(0)$. Therefore, \hat{V} is unbiased for the first two terms in (4.1).

4.4 Regression analysis of the CRE

Practitioners often use regression-based inference for the average causal effect τ . A standard approach is to run the ordinary least squares (OLS) of the outcomes on the treatment indicators with an intercept

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{(a,b)} \sum_{i=1}^n (Y_i - a - bZ_i)^2,$$

and use the coefficient of the treatment $\hat{\beta}$ as the estimator for the average causal effect. We can show the coefficient $\hat{\beta}$ equals the difference in means:

$$\hat{\beta} = \hat{\tau}. \quad (4.3)$$

However, the usual variance estimator from the OLS, e.g., the output from the `lm` function of R, equals

$$\begin{aligned} \hat{V}_{\text{OLS}} &= \frac{N(N_1 - 1)}{(N - 2)N_1 N_0} \hat{S}^2(1) + \frac{N(N_0 - 1)}{(N - 2)N_1 N_0} \hat{S}^2(0) \\ &\approx \frac{\hat{S}^2(1)}{N_0} + \frac{\hat{S}^2(0)}{N_1}, \end{aligned} \quad (4.4)$$

where the approximation holds with large N_1 and N_0 . It differs from \hat{V} even with large N_1 and N_0 .

Fortunately, the Eicker–Huber–White (EHW) robust variance estimator is close to \hat{V} :

$$\begin{aligned} \hat{V}_{\text{EHW}} &= \frac{\hat{S}^2(1)}{N_1} \frac{N_1 - 1}{N_1} + \frac{\hat{S}^2(0)}{N_0} \frac{N_0 - 1}{N_0} \\ &\approx \frac{\hat{S}^2(1)}{N_1} + \frac{\hat{S}^2(0)}{N_0} \end{aligned} \quad (4.5)$$

where the approximation holds with large N_1 and N_0 . It is almost identical to \hat{V} . Moreover, the so-called HC2 variant of the EHW robust variance estimator is identical to \hat{V} . The `hccm` function in the `car` package returns the EHW robust variance estimator as well as its HC2 variant.

Problem 4.3 provides more technical details for (4.3)–(4.5).

4.5 Examples

4.5.1 Simulation

I first choose the sample size as $n = 100$ with 60 treated and 40 control units, and generate the potential outcomes with constant individual causal effects.

```

n = 100
n1 = 60
n0 = 40
y0 = rexp(n)
y0 = sort(y0, decreasing = TRUE)
y1 = y0 + 1

```

With the Science Table fixed, I repeated generate completely randomized experiments and apply Theorem 4.1 to obtain the point estimator, the conservative variance estimator, and the confidence interval based on the Normal approximation. The first panel of Figure 4.2 shows the histogram of $\hat{\tau} - \tau$ over 10^4 simulations.

I then change the potential outcome by sorting the control potential outcome in reverse order

```
y0 = sort(y0, decreasing = FALSE)
```

and repeat the above simulation. The second panel of Figure 4.2 shows the histogram of $\hat{\tau} - \tau$ over 10^4 simulations.

I finally permute the control potential outcomes

```
y0 = sample(y0)
```

and repeat the above simulation. The third panel of Figure 4.2 shows the histogram of $\hat{\tau} - \tau$ over 10^4 simulations.

Importantly, in the above three sets of simulations, the correlations between potential outcomes are different but the marginal distributions are the same. The following table compares the true variances, the conservative estimated variances, and the coverage rates of the 95% confidence intervals.

	constant	negative	independent
var	0.036	0.007	0.020
estimated var	0.036	0.036	0.036
coverege rate	0.947	1.000	0.989

The true variance depends on the correlation between the potential outcomes, with positively correlated potential outcomes corresponding to a larger sampling variance. This verifies (4.2). The estimated variances are almost identical because the formula of \hat{V} depends only on the marginal distributions of the potential outcomes. Due to the discrepancy between the true and estimated variances, the coverage rates differ across the three sets of simulations. Only with constant causal effects, the estimated variance is identical to the true variance, verifying point 3 of Theorem 4.1.

Figure 4.2 also shows the Normal density curves based on the central limit theorem for $\hat{\tau}$. They are very close to the histogram over simulations, verifying Theorem 4.2.

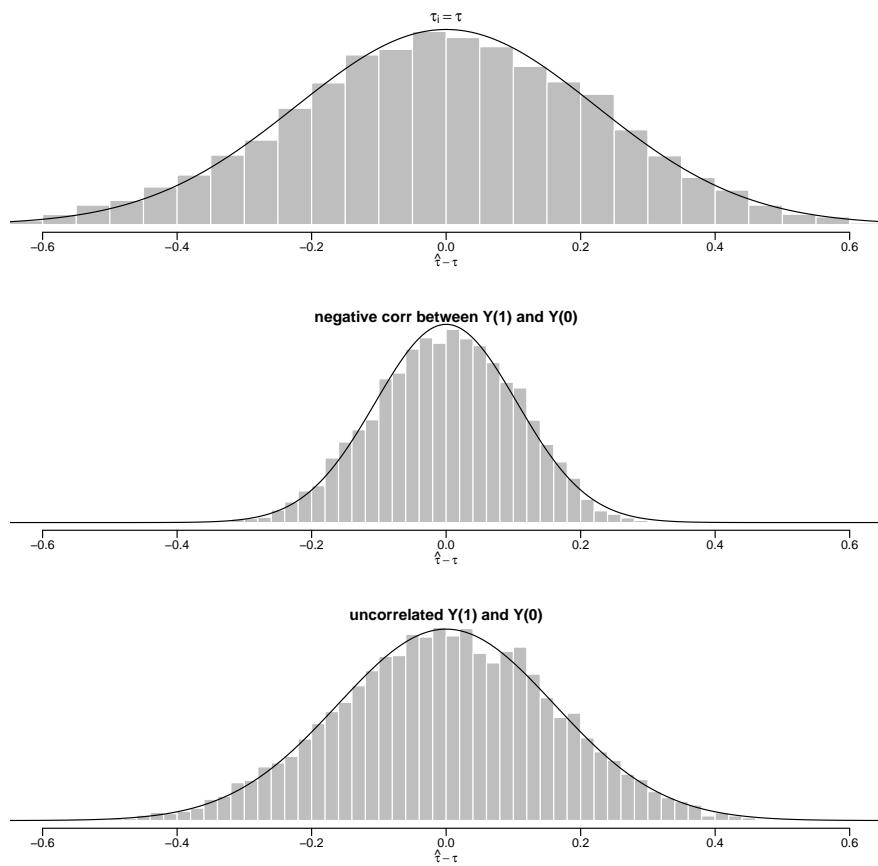


FIGURE 4.2: Sampling distribution of $\hat{\tau} - \tau$ with the same marginal but different joint distributions of the potential outcomes.

4.5.2 Heavy-tailed outcome and failure of Normal approximations

The central limit theorem of $\hat{\tau}$ in Theorem 4.2 holds under some regularity conditions. Those conditions will be violated with heavy-tailed potential outcomes. We can modify the above simulation studies to illustrate this point. Assume the individual causal effects are constant but the control potential outcomes are contaminated by a Cauchy component with probability 0.1, 0.3 or 0.5. The following code generates the potential outcomes with the probability of contamination being 0.1.

```
combination = rbinom(n, 1, 0.1)
y0 = (1 - combination)*rexp(n) + combination*rcauchy(n)
y1 = y0 + 1
```

Figures 4.3 and 4.4 show two realizations of the histograms of $\hat{\tau} - \tau$ with the corresponding Normal approximations. With heavy-tailed potential outcomes, the Normal approximations are quite poor. Moreover, unlike Figure 4.2, the histograms are quite sensitive to the random seed of the simulation.

4.5.3 Application

I again use the `lalonde` data to illustrate the theory.

```
> library(Matching)
> data(lalonde)
> z = lalonde$treat
> y = lalonde$re78
```

We can easily calculate the point estimator and standard error based on the formulas in Theorem 4.1:

```
> n1= sum(z)
> n0= length(z) - n1
> tauhat = mean(y[z==1]) - mean(y[z==0])
> vhat    = var(y[z==1])/n1 + var(y[z==0])/n0
> sehat   = sqrt(vhat)
> tauhat
[1] 1794.343
> sehat
[1] 670.9967
```

Practitioners often use ordinary least squares (OLS) to estimate the average causal effect which also gives a standard error.

```
> olsfit = lm(y ~ z)
> summary(olsfit)$coef[2, 1: 2]
  Estimate Std. Error
  1794.3431   632.8536
```

However, the above standard error seems too small compared to the one based

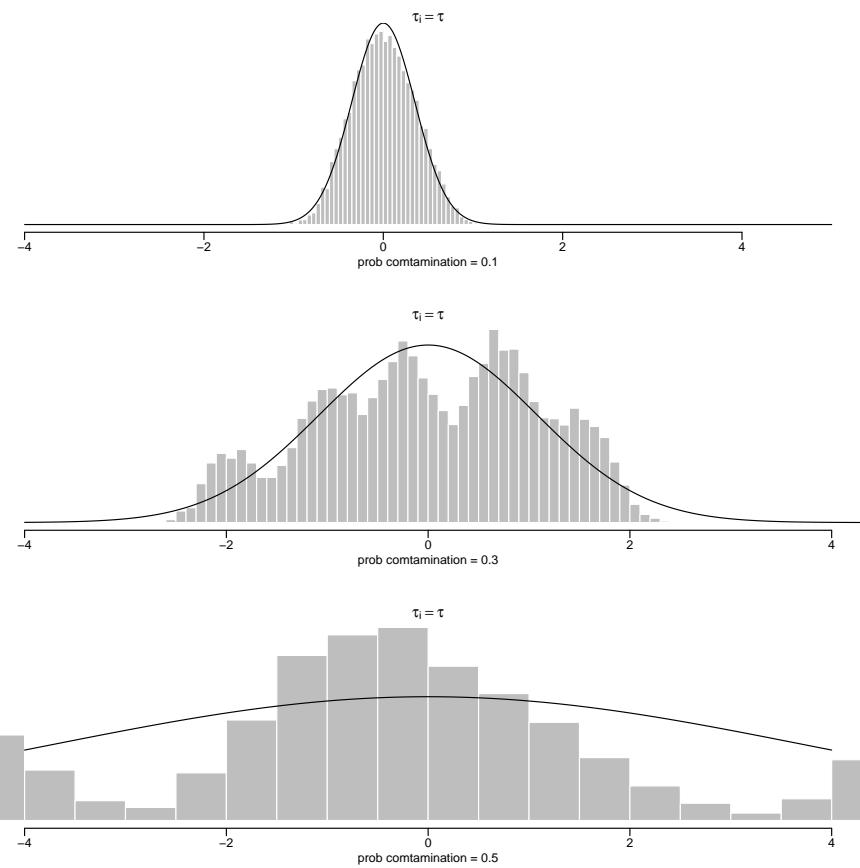


FIGURE 4.3: Sampling distribution of $\hat{\tau} - \tau$ with contaminated potential outcomes: realization one

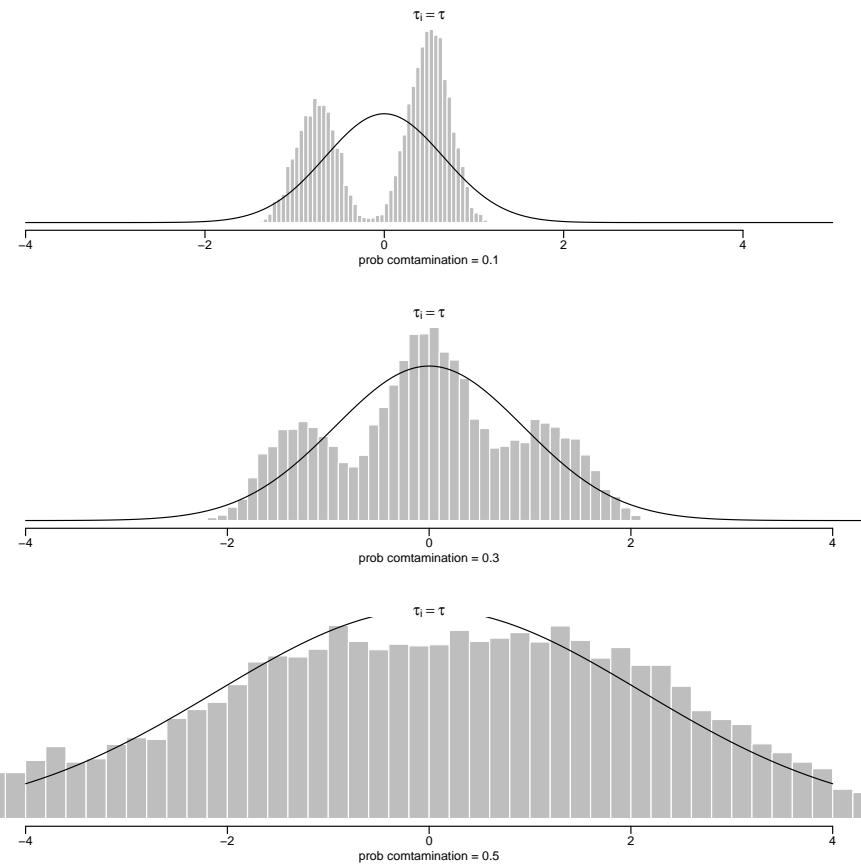


FIGURE 4.4: Sampling distribution of $\hat{\tau} - \tau$ with contaminated potential outcomes: realization two

on Theorem 4.1. However, this can be easily solved by using the Eicker–Huber–White robust standard error.

```
> library(car)
> sqrt(hccm(olsfit)[2, 2])
[1] 672.6823
> sqrt(hccm(olsfit, type = "hc0")[2, 2])
[1] 669.3155
> sqrt(hccm(olsfit, type = "hc2")[2, 2])
[1] 670.9967
```

Different versions of the robust standard error exist. They yield similar results if the sample size is large, with `hc2` yielding a standard error identical to Theorem 4.1. Problem 4.3 gives a theoretical explanation for the possible failure of the standard error based on OLS and the asymptotic validity of the Eicker–Huber–White robust standard error.

4.6 Homework Problems

4.1 Proof of Lemma 4.1

Prove Lemma 4.1.

4.2 Alternative proof of Theorem 4.1

Under a CRE, calculate

$$\text{var}\{\hat{Y}(1)\}, \quad \text{var}\{\hat{Y}(0)\}, \quad \text{cov}\{\hat{Y}(1), \hat{Y}(0)\}$$

and use these formulas to calculate $\text{var}(\hat{\tau})$.

Hint: Use the results in Chapter A3.

4.3 Neymanian inference and OLS

Prove (4.3)–(4.5). Moreover, prove that the HC2 variant of the EHW robust variance estimator recovers \hat{V} exactly.

Hint: Appendix A2 reviews some important technical results about OLS.

4.4 Treatment effect heterogeneity

Show that $S^2(\tau) = 0$ implies that $S^2(1) = S^2(0)$. Given a counterexample with $S^2(1) = S^2(0)$ but $S^2(\tau) \neq 0$.

Show that $S^2(1) < S^2(0)$ implies that

$$S(Y(0), \tau) = (n - 1) \sum_{i=1}^n \{Y_i(0) - \bar{Y}(0)\}(\tau_i - \tau) < 0.$$

Give a counterexample with $S^2(1) > S^2(0)$ but $S(Y(0), \tau) < 0$.

Remark: The first result states that no treatment effect heterogeneity implies equal variances in the treated and control potential outcomes. But the converse is not true. The second result states that if the treated potential outcome has larger variance than the control potential outcome, then the individual treatment effect is negatively correlated with the control potential outcome. But the converse is not true. Gerber and Green (2012, page 293) and (Ding et al., 2019, Appendix B.3) gave related discussions.

4.5 A better bound of the variance formula

Neyman (1923)'s conservative variance estimator essentially uses the following upper bound on the true variance:

$$\text{var}(\hat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n} \leq \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0},$$

which uses the trivial fact that $S^2(\tau) \geq 0$. Show the following upper bound

$$\text{var}(\hat{\tau}) \leq \frac{1}{n} \left\{ \sqrt{\frac{n_0}{n_1}} S(1) + \sqrt{\frac{n_1}{n_0}} S(0) \right\}^2. \quad (4.6)$$

When does the equality in (4.6) hold?

The upper bound (4.6) motivates another conservative variance estimator

$$\hat{V}' = \frac{1}{n} \left\{ \sqrt{\frac{n_0}{n_1}} \hat{S}(1) + \sqrt{\frac{n_1}{n_0}} \hat{S}(0) \right\}^2.$$

Section 4.5.1 used \hat{V} in the simulation with R code `NeymanCR.R`. Repeat the simulation with additional comparison with the variance estimator \hat{V}' and the associated confidence interval.

Remark: The upper bound (4.6) can be further improved. Aronow et al. (2014) derived the sharp upper bound for $\text{var}(\hat{\tau})$ using the Frechet–Hoeffding inequality. Those improvements are rarely used in practice mainly for two reasons. First, they are more complicated than \hat{V} which can be conveniently implemented by OLS. Second, the confidence interval based on \hat{V} also works under other formulations, for example, under a true linear model of the outcome on the treatment, but those improvements do not. Although they are theoretically interesting, those improvements have little practical impact.

4.6 Vector version of Neyman (1923)

The classic result of Neyman (1923) is about a scalar outcome. It is common to have multiple outcomes in practice. Therefore, we can extend the potential outcomes to vectors. We consider the average causal effect on a vector outcome $\mathbf{V} \in \mathbb{R}^K$,

$$\tau_{\mathbf{V}} = \frac{1}{n} \sum_{i=1}^n \{ \mathbf{V}_i(1) - \mathbf{V}_i(0) \},$$

where $\mathbf{V}_i(1)$ and $\mathbf{V}_i(0)$ are the potential outcomes of \mathbf{V} for unit i . The Neyman-type estimator for $\tau_{\mathbf{V}}$ is the difference between the sample mean vectors of the observed outcomes under treatment and control:

$$\hat{\tau}_{\mathbf{V}} = \bar{\mathbf{V}}_1 - \bar{\mathbf{V}}_0 = \frac{1}{n_1} \sum_{i=1}^n Z_i \mathbf{V}_i - \frac{1}{n_0} \sum_{i=1}^n (1 - Z_i) \mathbf{V}_i.$$

Consider a CRE. Show that $\hat{\tau}_{\mathbf{V}}$ is unbiased for $\tau_{\mathbf{V}}$. Find the covariance matrix of $\hat{\tau}_{\mathbf{V}}$. Find a (possibly conservative) estimator for the variance.

4.7 Inference in the BRE

Consider the BRE where the Z_i 's are IID Bernoulli(π) with $n_1 = \sum_{i=1}^n Z_i$ receiving the treatment and $n_0 = \sum_{i=1}^n (1 - Z_i)$ receiving the control.

First, we can use the FRT to analyze the BRE. How do we test H_{0F} in the CRE? Can we use the same FRT procedure as in the CRE if the actual experiment is the BRE? If yes, give a justification; if no, explain why.

Second, we can obtain point estimator for τ and find the associated variance estimator, as Neyman (1923) did for the CRE.

1. Is $\hat{\tau}$ unbiased for τ ? Is it consistent?
2. Find an unbiased estimator for τ .
3. Compare the variance of the above unbiased estimator and the asymptotic variance of $\hat{\tau}$.

Remark: The estimator $\hat{\tau}$ does not have finite variance but the variance of its asymptotic distribution is finite.

4.8 Recommended reading

Ding (2016) compared the Fisherian and Neymanian approaches to analyzing the CRE.



5

Stratification and Post-Stratification in Randomized Experiments

Block what you can and randomize what you cannot.

— George Box

This is the second most famous quote from George Box¹. This chapter will explain its meaning.

5.1 Stratification

A CRE may generate an undesired treatment allocation. Let us start with a completely randomized experiment with a discrete covariate $X_i \in \{1, \dots, K\}$, and define $n_{[k]} = \#\{i : X_i = k\}$ and $\pi_{[k]} = n_{[k]}/n$ as the number and proportion of units in stratum k ($k = 1, \dots, K$). A CRE assigns n_1 units to the treatment group and n_0 units to the control group, which results in

$$n_{[k]1} = \#\{i : X_i = k, Z_i = 1\}, \quad n_{[k]0} = \#\{i : X_i = k, Z_i = 0\}$$

units in the treatment and control groups within stratum k . With positive probability, $n_{[k]1}$ or $n_{[k]0}$ is zero for some k , that is, it is possible that some strata only have treated or control units. Even none of the $n_{[k]1}$'s or $n_{[k]0}$'s are zero, with high probability

$$\frac{n_{[k]1}}{n_1} - \frac{n_{[k]0}}{n_0} \neq 0, \tag{5.1}$$

and the magnitude can be quite large. So the proportions of units in stratum k are different across the treatment and control groups although on average their difference is zero:

$$\begin{aligned} & E \left(\frac{n_{[k]1}}{n_1} - \frac{n_{[k]0}}{n_0} \right) \\ &= E \left\{ n_1^{-1} \sum_{i=1}^n Z_i 1(X_i = k) - n_0^{-1} \sum_{i=1}^n (1 - Z_i) 1(X_i = k) \right\} \\ &= 0. \end{aligned}$$

¹His most famous quote is “all models are wrong but some are useful.”

When $n_{[k]1}/n_1 - n_{[k]0}/n_0$ is large for some strata with $X = k$, the treatment and control groups have undesirable covariate imbalance. Such covariate imbalance deteriorates the quality of the experiment, making it difficult to interpret the results of the experiment since the difference in the outcomes may be attributed to the treatment or the covariate imbalance.

How can we actively avoid covariate imbalance in the experiment? We can fix the $n_{[k]1}$'s or $n_{[k]0}$'s in advance and conduct stratified randomized experiments (SRE).

Definition 5.1 (SRE) *We conduct K independent CREs within the K strata of a discrete covariate X .*

In agricultural experiments, the SRE is also called the *randomized block design*, with the strata called the blocks. Analogously, *stratified randomization* is also called *block randomization*. The total number of randomizations in an SRE equals

$$\prod_{k=1}^K \binom{n_{[k]}}{n_{[k]1}},$$

and each feasible randomization has equal probability. Within stratum k , the proportion of units receiving the treatment is

$$e_{[k]} = \frac{n_{[k]1}}{n_{[k]}},$$

which is also called the *propensity score*, a conceptual that will play a central role in Part III of this book. An SRE is different from a CRE: first, all feasible randomizations in an SRE form a subset of all feasible randomizations in a CRE, so

$$\prod_{k=1}^K \binom{n_{[k]}}{n_{[k]1}} < \binom{n}{n_1};$$

second, $e_{[k]}$ is fixed in an SRE but random in a CRE.

For every unit i , we have potential outcomes $Y_i(1)$ and $Y_i(0)$, and individual causal effect $\tau_i = Y_i(1) - Y_i(0)$. For stratum k , we have stratum-specific average causal effect

$$\tau_{[k]} = n_{[k]}^{-1} \sum_{X_i=k} \tau_i.$$

The average causal effect is

$$\tau = n^{-1} \sum_{i=1}^n \tau_i = n^{-1} \sum_{k=1}^K \sum_{X_i=k} \tau_i = \sum_{k=1}^K \pi_{[k]} \tau_{[k]},$$

which is also the weighted average of the stratum-specific average causal effects.

If we are interested in $\tau_{[k]}$, then we can use the methods in Chapters 3 and 4 for the CRE within stratum k . Below I will discuss statistical inference for τ .

5.2 FRT

5.2.1 Theory

In parallel with the discussion of a CRE, I will start with the FRT in an SRE. The sharp null hypothesis is still

$$H_{0F} : Y_i(1) = Y_i(0) \text{ for all units } i = 1, \dots, n.$$

The fundamental idea of the FRT applies to any randomized experiment: we can use any test statistic which has a known distribution under H_{0F} and the SRE. However, we must be careful with two subtle issues. First, when we simulate the treatment vector, we must permute the treatment indicators within strata of X . The resulting FRT is sometimes called the *conditional randomization test* or *conditional permutation test*. Second, we should choose test statistics that can reflect the nature of the SRE. Below I give some canonical choices of the test statistic.

Example 5.1 (Stratified estimator) Motivated by estimating τ , we can use the following stratified estimator in the FRT:

$$\hat{\tau}_S = \sum_{k=1}^K \pi_{[k]} \hat{\tau}_{[k]},$$

where

$$\hat{\tau}_{[k]} = n_{[k]1}^{-1} \sum_{i=1}^n I(X_i = k, Z_i = 1) Y_i - n_{[k]0}^{-1} \sum_{i=1}^n I(X_i = k, Z_i = 0) Y_i$$

is the stratum-specific difference-in-means within stratum k .

Example 5.2 (Studentized stratified estimator) Motivated by the studentized statistic in the simple two-sample problem, we can use the following studentized statistic for the stratified estimator in the FRT:

$$t_S = \frac{\hat{\tau}_S}{\sqrt{\hat{V}_S}},$$

with

$$\hat{V}_S = \sum_{k=1}^K \pi_{[k]}^2 \left(\frac{\hat{S}_{[k]}^2(1)}{n_{[k]1}} + \frac{\hat{S}_{[k]}^2(0)}{n_{[k]0}} \right)$$

where $\hat{S}_{[k]}^2(1)$ and $\hat{S}_{[k]}^2(0)$ are the stratum-specific sample variances of the outcomes under treatment and control, respectively. The exact form of this statistic is motivated by the Neymanian perspective discussed in Section 5.3.

Example 5.3 (Combining Wilcoxon rank-sum statistics) We first compute the Wilcoxon rank sum statistic $W_{[k]}$ within stratum k and then combine them as

$$W_S = \sum_{k=1}^K c_{[k]} W_{[k]}.$$

Based on different asymptotic schemes and optimality criteria, Van Elteren (1960) proposed two weighting methods, one with

$$c_{[k]} = \frac{1}{n_{[k]1} n_{[k]0}},$$

and the other with

$$c_{[k]} = \frac{1}{n_{[k]} + 1}$$

The motivations for these weights appear to be quite technical, and other choices of weights may also be reasonable.

Example 5.4 (Hodges and Lehmann (1962)'s aligned rank statistic) Van Elteren (1960)'s statistic works well with a few large strata. However, it does not work well with many small strata since it does not make enough comparisons, potentially losing information in the data. Hodges and Lehmann (1962) proposed a test statistic that makes more comparisons across strata after standardizing the outcomes. They suggested first centering the outcomes as

$$\tilde{Y}_i = Y_i - \bar{Y}_{[k]}$$

with the stratum-specific mean $\bar{Y}_{[k]} = n_{[k]}^{-1} \sum_{X_i=k} Y_i$ if $X_i = k$, then obtaining the ranks $(\tilde{R}_1, \dots, \tilde{R}_n)$ of the pooled outcomes $(\tilde{Y}_1, \dots, \tilde{Y}_n)$, and finally constructing the test statistic

$$\tilde{W} = \sum_{i=1}^n Z_i \tilde{R}_i.$$

We can simulate the exact distributions of the above test statistics under the SRE. We can also calculate their means and variances and obtain the p -values based on Normal approximations.

After searching for a while, I failed to find detailed discussion of the Kolmogorov–Smirnov statistic for the SRE. Below is my proposal.

Example 5.5 (Kolmogorov–Smirnov statistic) We compute $D_{[k]}$, the maximum difference between the empirical distributions of the outcomes under treatment and control within stratum k . The final test statistic can be

$$D_S = \sum_{k=1}^K c_{[k]} D_{[k]}$$

or

$$D_{\max} = \max_{1 \leq k \leq K} c_{[k]} D_{[k]},$$

where $c_{[k]} = \sqrt{n_{[k]1} n_{[k]0} / n_{[k]}}$ is motivated by the limiting distribution of $D_{[k]}$ with $n_{[k]1}$ and $n_{[k]0}$ approach infinity (Van der Vaart, 2000). The statistics D_S and D_{\max} are more appropriate when all strata have large sample size. Another reasonable choice is

$$D = \max_y \left| \sum_{k=1}^K \pi_{[k]} \{ \hat{F}_{[k]1}(y) - \hat{F}_{[k]0}(y) \} \right|,$$

where $\hat{F}_{[k]1}(y)$ and $\hat{F}_{[k]0}(y)$ are the stratum-specific empirical distribution functions of the outcomes under treatment and control, respectively. The statistic D is appropriate in both the cases with large strata and the cases with many small strata.

5.2.2 An application

The Penn Bonus experiment as an example to illustrate the FRT in the SRE. The dataset used by Koenker and Xiao (2002) is from a job training program stratified on quarter, with the outcome being the duration before employed.

```
penndata = read.table("Penn46_ascii.txt")
z = penndata$treatment
y = log(penndata$duration)
block = penndata$quarter
```

I will focus on $\hat{\tau}_S$ and W_S , and leave the FRT with other statistics as exercise. The following function computes $\hat{\tau}_S$ and V :

```
stat_SRE = function(z, y, x)
{
  xlevels = unique(x)
  K = length(xlevels)
  PiK = rep(0, K)
  TauK = rep(0, K)
  WilcoxK = rep(0, K)
  for(k in 1:K)
  {
    xk = xlevels[k]
    zk = z[x == xk]
    yk = y[x == xk]
    PiK[k] = length(zk)/length(z)
    TauK[k] = mean(yk[zk==1]) - mean(yk[zk==0])
    WilcoxK[k] = wilcox.test(yk[zk==1], yk[zk==0])$statistic
  }
  return(c(sum(PiK*TauK), sum(WilcoxK/PiK)))
}
```

The following function generates a random treatment assignment in the SRE of the observed data:

```
zRandomSRE = function(z, x)
{
    xlevels = unique(x)
    K       = length(xlevels)
    zrandom = z
    for(k in 1:K)
    {
        xk = xlevels[k]
        zrandom[x == xk] = sample(z[x == xk])
    }

    return(zrandom)
}
```

Based on the above data and functions, we can easily simulate the randomization distributions of the test statistics (shown in Figure 5.1 with 10^4 Monte Carlo draws) and compute the p -values.

```
> MC = 10^4
> statSREMC = matrix(0, MC, 2)
> for(mc in 1:MC)
+ {
+     zrandom      = zRandomSRE(z, block)
+     statSREMC[mc, ] = stat_SRE(zrandom, y, block)
+ }
> mean(statSREMC[, 1] <= stat.obs[1])
[1] 0.0019
> mean(statSREMC[, 2] <= stat.obs[2])
[1] 5e-04
```

5.3 Neymanian inference

5.3.1 Point and interval estimation

Statistical inference for an SRE builds on the fact that it essentially consists of K independent CREs. Based on this, we can easily extend Neyman (1923)'s results to the SRE. Within stratum k , the difference-in-means $\hat{\tau}_{[k]}$ is unbiased for $\tau_{[k]}$ with variance

$$\text{var}(\hat{\tau}_{[k]}) = \frac{S_{[k]}^2(1)}{n_{[k]1}} + \frac{S_{[k]}^2(0)}{n_{[k]0}} - \frac{S_{[k]}^2(\tau)}{n_{[k]}},$$

where $S_{[k]}^2(1)$, $S_{[k]}^2(0)$ and $S_{[k]}^2(\tau)$ are the stratum-specific variances of potential outcomes and the individual treatment effects, respectively. Therefore, the

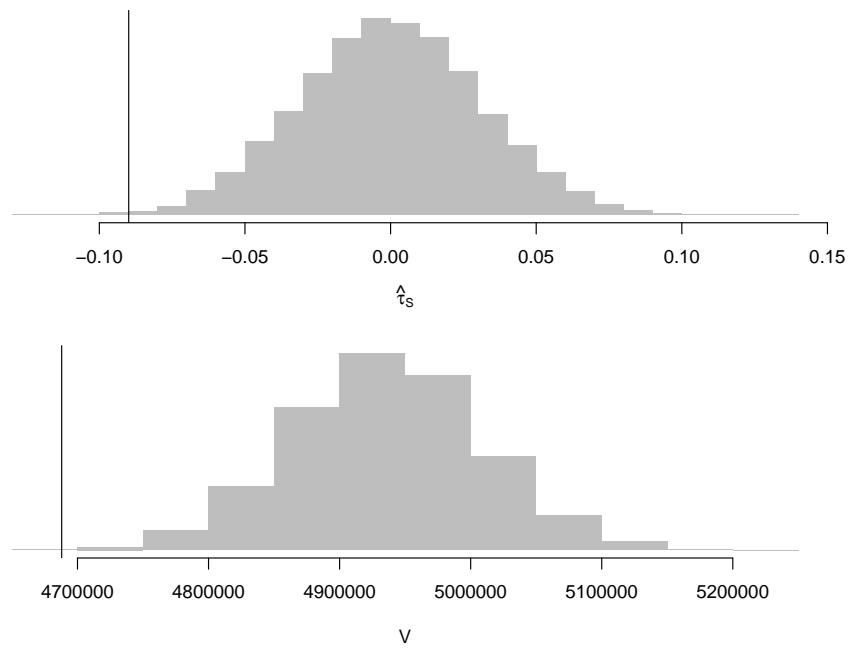


FIGURE 5.1: The randomization distributions of $\hat{\tau}_S$ and V based on the Penn Bonus experiment.

stratified estimator $\hat{\tau}_S = \sum_{k=1}^K \pi_{[k]} \hat{\tau}_{[k]}$ is unbiased for $\tau = \sum_{k=1}^K \pi_{[k]} \tau_{[k]}$ with variance

$$\text{var}(\hat{\tau}_S) = \sum_{k=1}^K \pi_{[k]}^2 \text{var}(\hat{\tau}_{[k]}).$$

If $n_{[k]1} \geq 2$ and $n_{[k]0} \geq 2$, then we can obtain the sample variances $\hat{S}_{[k]}^2(1)$ and $\hat{S}_{[k]}^2(0)$ of the outcomes within stratum k and construct a conservative variance estimator

$$\hat{V}_S = \sum_{k=1}^K \pi_{[k]}^2 \left(\frac{\hat{S}_{[k]}^2(1)}{n_{[k]1}} + \frac{\hat{S}_{[k]}^2(0)}{n_{[k]0}} \right),$$

where $\hat{S}_{[k]}^2(1)$ and $\hat{S}_{[k]}^2(0)$ are the stratum-specific sample variances of the outcomes under treatment and control, respectively. Based on a Normal approximation of $\hat{\tau}_S$, we can construct a Wald-type $1 - \alpha$ confidence interval for τ :

$$\hat{\tau}_S \pm z_{1-\alpha/2} \sqrt{\hat{V}_S}.$$

From a hypothesis testing perspective, under $H_{0N} : \tau = 0$, we can compare $t_S = \hat{\tau}_S / \sqrt{\hat{V}_S}$ with the standard Normal quantiles to obtain asymptotic p -values. The statistic t_S has appeared in Example 5.2 for the FRT. Similar to the discussion for the CRE, using t_S in the FRT yields finite-sample exact p -value under H_{0F} and asymptotically valid p -value under H_{0N} . Wu and Ding (2021) provided a justification for this claim.

Here I omit the technical details for the central limit theorem of $\hat{\tau}_S$. See Liu and Yang (2020) for a proof, which includes the two regimes with a few large strata and many small strata. I will illustrate this theoretical issues using a numerical example in Section 5.3.2.

5.3.2 Numerical examples

The following function computes the Neymanian point and variance estimators:

```
Neyman_SRE = function(z, y, x)
{
  xlevels = unique(x)
  K       = length(xlevels)
  PiK     = rep(0, K)
  TauK    = rep(0, K)
  varK    = rep(0, K)
  for(k in 1:K)
  {
    xk      = xlevels[k]
    zk      = z[x == xk]
    yk      = y[x == xk]
```

```

PiK[k]      = length(zk)/length(z)
TauK[k]     = mean(yk[zk==1]) - mean(yk[zk==0])
varK[k]     = var(yk[zk==1])/sum(zk) +
              var(yk[zk==0])/sum(1 - zk)
}

return(c(sum(PiK*TauK), sum(PiK^2*varK)))
}

```

The first simulation setting has $K = 5$ and each stratum has 80 units. **TauHat** and **VarHat** are the point and variance estimators over 10^4 simulations.

```

> K   = 5
> n   = 80
> n1  = 50
> n0  = 30
> x   = rep(1:K, each = n)
> y0  = rexp(n*K, rate = x)
> y1  = y0 + 1
> zb  = c(rep(1, n1), rep(0, n0))
> MC  = 10^4
> TauHat = rep(0, MC)
> VarHat = rep(0, MC)
> for(mc in 1:MC)
+ {
+   z   = replicate(K, sample(zb))
+   z   = as.vector(z)
+   y   = z*y1 + (1-z)*y0
+   est = Neyman_SRE(z, y, x)
+   TauHat[mc] = est[1]
+   VarHat[mc] = est[2]
+ }
> var(TauHat)
[1] 0.002248925
> mean(VarHat)
[1] 0.002266396

```

The upper panel of Figure 5.2 shows the histogram of the point estimator, which is symmetric and bell-shaped around the true parameter. From the above, the average value of the variance estimator is almost identical to the variance of the estimators because the individual causal effects are constant.

The first simulation setting has $K = 50$ and each stratum has 8 units.

```

> K   = 50
> n   = 8
> n1  = 5
> n0  = 3
> x   = rep(1:K, each = n)
> y0  = rexp(n*K, rate = log(x + 1))
> y1  = y0 + 1
> zb  = c(rep(1, n1), rep(0, n0))

```

```

> MC = 10^4
> TauHat = rep(0, MC)
> VarHat = rep(0, MC)
> for(mc in 1:MC)
+ {
+   z = replicate(K, sample(zb))
+   z = as.vector(z)
+   y = z*y1 + (1-z)*y0
+   est = Neyman_SRE(z, y, x)
+   TauHat[mc] = est[1]
+   VarHat[mc] = est[2]
+ }
>
> hist(TauHat, xlab = expression(hat(tau)[S]),
+       ylab = "", main = "many small strata",
+       border = FALSE, col = "grey",
+       breaks = 30, yaxt = 'n',
+       xlim = c(0.8, 1.2))
> abline(v = 1)
>
> var(TauHat)
[1] 0.0014443111
> mean(VarHat)
[1] 0.001473616

```

The lower panel of Figure 5.2 shows the histogram of the point estimator, which is symmetric and bell-shaped around the true parameter.

We finally use the Penn Bonus Experiment to illustrate the Neymanian inference in an SRE. Applying the function `Neyman_SRE` to the dataset, we obtain:

```

> est = Neyman_SRE(z, y, block)
> est[1]
[1] -0.08990646
> sqrt(est[2])
[1] 0.03079775

```

So the job training program significantly shortens the duration time before employment.

5.3.3 Comparing the SRE and the CRE

What are the benefits of the SRE compared to the CRE? I have motivated the SRE from the covariate balance perspective. In addition, I will show that better covariate balance in turn results in better estimation precision of the average causal effect. To make a fair comparison, I assume that $e_{[k]} = e$ for all k which ensures that $\hat{\tau} = \hat{\tau}_S$. I leave the proof of this result as Problem 5.1.

We now compare the sampling variances. The classic analysis of variance technique motivates the decomposition of the total variance into the summa-

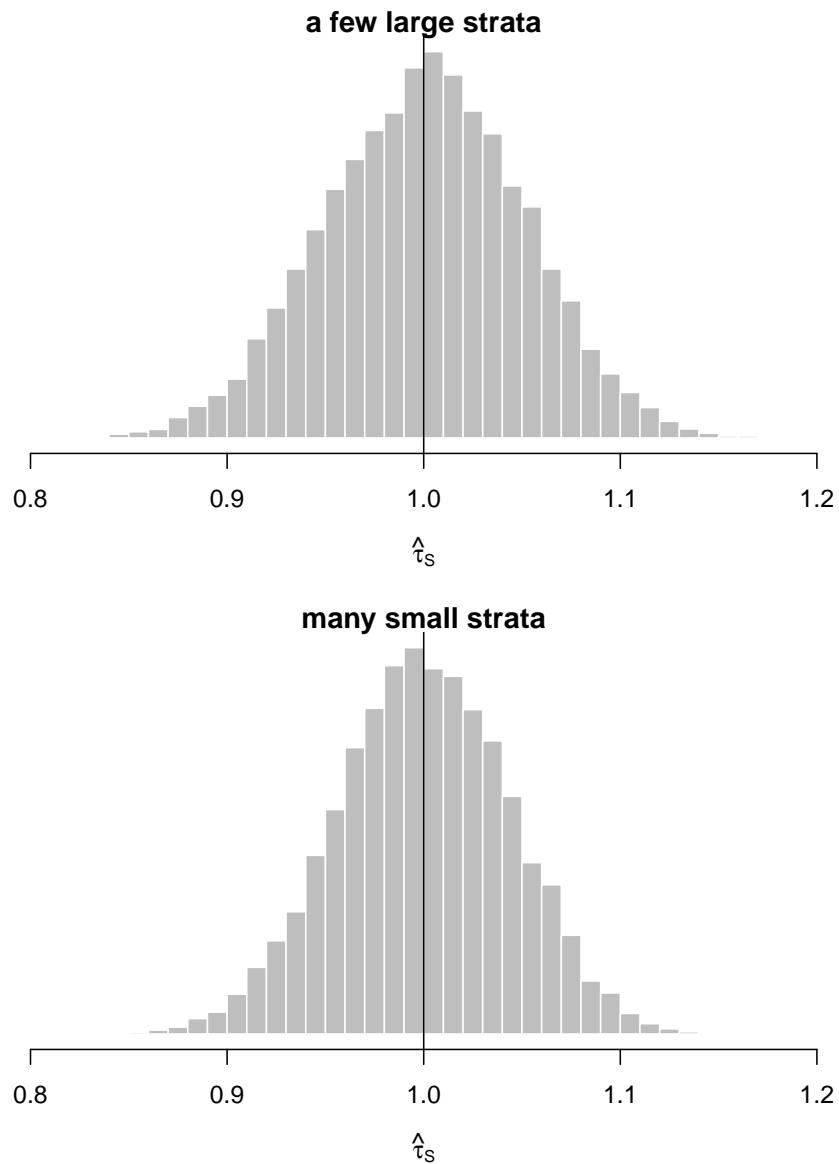


FIGURE 5.2: Normal approximations under two regimes

tion of the within-strata and between-strata variances, yielding

$$\begin{aligned}
S^2(1) &= (n-1)^{-1} \sum_{i=1}^n \{Y_i(1) - \bar{Y}(1)\}^2 \\
&= (n-1)^{-1} \sum_{k=1}^K \sum_{X_i=k} \{Y_i(1) - \bar{Y}_{[k]}(1) + \bar{Y}_{[k]}(1) - \bar{Y}(1)\}^2 \\
&= (n-1)^{-1} \sum_{k=1}^K \sum_{X_i=k} [\{Y_i(1) - \bar{Y}_{[k]}(1)\}^2 + \{\bar{Y}_{[k]}(1) - \bar{Y}(1)\}^2] \\
&= \sum_{k=1}^K \left[\frac{n_{[k]} - 1}{n-1} S_{[k]}^2(1) + \frac{n_{[k]}}{n-1} \{\bar{Y}_{[k]}(1) - \bar{Y}(1)\}^2 \right],
\end{aligned}$$

and similarly,

$$\begin{aligned}
S^2(0) &= \sum_{k=1}^K \left[\frac{n_{[k]} - 1}{n-1} S_{[k]}^2(0) + \frac{n_{[k]}}{n-1} \{\bar{Y}_{[k]}(0) - \bar{Y}(0)\}^2 \right], \\
S^2(\tau) &= \sum_{k=1}^K \left[\frac{n_{[k]} - 1}{n-1} S_{[k]}^2(\tau) + \frac{n_{[k]}}{n-1} \{\tau_{[k]} - \tau\}^2 \right].
\end{aligned}$$

With large strata, the variance of the difference-in-means estimator under complete randomization is approximately

$$\begin{aligned}
&\text{var}_{\text{CRE}}(\hat{\tau}) \\
&= \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n} \\
&\approx \sum_{k=1}^K \left[\frac{\pi_{[k]}}{n_1} S_{[k]}^2(1) + \frac{\pi_{[k]}}{n_0} S_{[k]}^2(0) - \frac{\pi_{[k]}}{n} S_{[k]}^2(\tau) \right] \\
&\quad + \sum_{k=1}^K \left[\frac{\pi_{[k]}}{n_1} \{\bar{Y}_{[k]}(1) - \bar{Y}(1)\}^2 + \frac{\pi_{[k]}}{n_0} \{\bar{Y}_{[k]}(0) - \bar{Y}(0)\}^2 - \frac{\pi_{[k]}}{n} \{\tau_{[k]} - \tau\}^2 \right].
\end{aligned}$$

The constant propensity scores assumption ensures

$$\pi_{[k]}/n_{[k]1} = 1/(ne), \quad \pi_{[k]}/n_{[k]0} = 1/\{n(1-e)\}, \quad \pi_{[k]}/n_{[k]} = 1/n,$$

which allow us to rewrite the variance of $\hat{\tau}_S$ under the SRE as

$$\begin{aligned}
\text{var}_{\text{SRE}}(\hat{\tau}_S) &= \sum_{k=1}^K \pi_{[k]}^2 \left[\frac{S_{[k]}^2(1)}{n_{[k]1}} + \frac{S_{[k]}^2(0)}{n_{[k]0}} - \frac{S_{[k]}^2(\tau)}{n_{[k]}} \right] \\
&= \sum_{k=1}^K \left[\frac{\pi_{[k]}}{n_1} S_{[k]}^2(1) + \frac{\pi_{[k]}}{n_0} S_{[k]}^2(0) - \frac{\pi_{[k]}}{n} S_{[k]}^2(\tau) \right].
\end{aligned}$$

Approximately, the difference between $\text{var}_{\text{CRE}}(\hat{\tau})$ and $\text{var}_{\text{SRE}}(\hat{\tau}_S)$ is

$$\begin{aligned} & \sum_{k=1}^K \left[\frac{\pi_{[k]}}{n_1} \{ \bar{Y}_{[k]}(1) - \bar{Y}(1) \}^2 + \frac{\pi_{[k]}}{n_0} \{ \bar{Y}_{[k]}(0) - \bar{Y}(0) \}^2 - \frac{\pi_{[k]}}{n} (\tau_{[k]} - \tau)^2 \right] \\ &= \sum_{k=1}^K \frac{\pi_{[k]}}{n} \left\{ \sqrt{\frac{n_0}{n_1}} \{ \bar{Y}_{[k]}(1) - \bar{Y}(1) \} + \sqrt{\frac{n_1}{n_0}} \{ \bar{Y}_{[k]}(0) - \bar{Y}(0) \} \right\}^2 \geq 0, \end{aligned}$$

which is non-negative. The difference is zero only in the extreme case that

$$\sqrt{\frac{n_0}{n_1}} \{ \bar{Y}_{[k]}(1) - \bar{Y}(1) \} + \sqrt{\frac{n_1}{n_0}} \{ \bar{Y}_{[k]}(0) - \bar{Y}(0) \} = 0$$

for $k = 1, \dots, K$. When the covariate is predictive to the potential outcomes, the above quantities are usually not all zeros, which ensure the efficiency gain of the SRE compared to the CRE. Only in the extreme cases that the covariate is not predictive at all, the large-sample efficiency gain is zero. In those cases, the SRE can even result in worse estimators in finite sample. The above discussion corroborates the quote from George Box at the beginning of this chapter.

I will end this section with several remarks. First, the above comparison is based on the sampling variance, and we can also compare the estimated variances under the SRE and the CRE. The results are similar. Second, increasing K improves efficiency, but this argument depends on the large strata assumption. So we face a tradeoff in practice. We cannot arbitrarily increase K , and the most extreme case is $n_{[k]1} = n_{[k]0} = 1$, which is called the matched pair experiment and will be discussed later.

5.4 Post-stratification in a CRE

In a CRE with a discrete covariate X , the numbers of units receiving the treatment and control are random within stratum k . In a SRE, these numbers are fixed. But if we conduct conditional inference given $\mathbf{n} = \{n_{[k]1}, n_{[k]0}\}_{k=1}^K$, then a CRE becomes a SRE. Mathematically, if none of the components of \mathbf{n} are zero, then

$$\text{pr}_{\text{CRE}}(\mathbf{Z} = \mathbf{z} \mid \mathbf{n}) = \frac{\text{pr}_{\text{CRE}}(\mathbf{Z} = \mathbf{z}, \mathbf{n})}{\text{pr}_{\text{CRE}}(\mathbf{n})} = \frac{1}{\prod_{k=1}^K \binom{n_{[k]}}{n_{[k]1}}}, \quad (5.2)$$

that is, the conditional distribution of \mathbf{Z} from a CRE given \mathbf{n} is identical to the distribution of \mathbf{Z} from an SRE. So conditional on \mathbf{n} , we can analyze a CRE with a discrete covariate X in the same way as in a SRE. In particular,

the FRT becomes a conditional FRT, and the Neymanian analysis becomes post-stratification:

$$\hat{\tau}_{PS} = \sum_{k=1}^K \pi_{[k]} \hat{\tau}_{[k]},$$

which has an identical form as $\hat{\tau}_S$. The variance of $\hat{\tau}_{PS}$ conditioning on \mathbf{n} is identical to the variance of $\hat{\tau}_S$ under the SRE.

Hennessy et al. (2016) used simulation to show that the conditional FRT is often more powerful than the unconditional one. Miratrix et al. (2013) used theory to show that in many cases, post-stratification improves efficiency compared to $\hat{\tau}$. However, the simulation is based on limited number of data generating processes, and the theory assumes all strata are large enough. We cannot go too extreme in the conditional FRT or post-stratification because with a larger K it is more likely that some $n_{[k]1}$ or $n_{[k]0}$ become zero. Small or zero values of $n_{[k]1}$ or $n_{[k]0}$ greatly reduces the number of randomizations in the FRT, possibly reducing the power dramatically. The problem for the Neymanian counterpart is more salient because we cannot even define $\hat{\tau}_{PS}$ and the corresponding variance estimator.

Stratification uses X in the design stage and post-stratification uses X in the analysis stage. They are duals. Asymptotically, their difference is small with large strata (Miratrix et al., 2013).

5.4.1 Meinert et al. (1970)'s Example

We use the data from a randomized trial from Meinert et al. (1970), which were also used by Rothman et al. (2008). The treatment is tolbutamide and the control is a placebo.

		Age < 55		Age ≥ 55			
		Surviving	Dead	Surviving	Dead		
$Z = 1$	98	8	$Z = 1$	76	22		
	115	5	$Z = 0$	69	16		
Total							
		Surviving		Dead			
$Z = 1$		174	30				
$Z = 0$		184	21				

The following table shows the estimates for two strata separately, the post-stratified estimator, and the crude estimator ignoring the binary covariate, as well as the corresponding standard errors.

	stratum 1	stratum 2	post-stratification	crude
est	-0.034	-0.036	-0.035	-0.045
se	0.031	0.060	0.032	0.033

Although the crude estimator and the post-stratification estimator do not

lead to fundamentally different results, the crude estimator is outside the range of the stratum-specific estimators while the post-stratification estimator is within the range.

5.4.2 Chong et al. (2016)'s Example

Chong et al. (2016) ran a randomized experiment in Peru to study the effect of supplemental iron pills on school performance. The experiment is stratified on `class_level`. I will only use a subset of the original data.

```
library("foreign")
dat_chong = read.dta("chong.dta")
use.vars = c("treatment",
            "gradesq34",
            "class_level",
            "anemic_base_re")
dat_physician = subset(dat_chong,
                        treatment != "Soccer Player",
                        select = use.vars)
dat_physician$z = (dat_physician$treatment == "Physician")
dat_physician$y = dat_physician$gradesq34
```

The treatment and control group sizes vary across five strata:

```
> table(dat_physician$z,
+        dat_physician$class_level)

      1   2   3   4   5
FALSE 15 19 16 12 10
TRUE  17 20 15 11 10
```

We can use the `Neyman_SRE` function defined before to compute the stratified estimator and its estimated variance.

```
tauS = with(dat_physician,
            Neyman_SRE(z, gradesq34, class_level))
```

An important additional covariate is the baseline anemic indicator which is quite important for predicting the outcome. Further conditioning the baseline anemic indicator, we have an experiment with $5 \times 2 = 10$ strata, with the treatment and control group sizes shown below.

```
> table(dat_physician$z,
+        dat_physician$class_level,
+        dat_physician$anemic_base_re)
, ,  = No

      1   2   3   4   5
FALSE 6 14 12  7  4
TRUE  8 12  9  5  6
```

```
, , = Yes
```

	1	2	3	4	5
FALSE	9	5	4	5	6
TRUE	9	8	6	6	4

Again we can use the `Neyman_SRE` function defined before to compute the post-stratified estimator and its estimated variance.

```
tausPS = with(dat_physician,
{
  sps = interaction(class_level, anemic_base_re)
  Neyman_SRE(z, gradesq34, sps)
})
```

The following table compares these two estimators. The post-stratified estimator yields a much smaller p -value.

	est	se	t.stat	p.value
stratify	0.406	0.202	2.005	0.045
stratify and post-stratify	0.463	0.190	2.434	0.015

This example illustrates that post-stratification can be used not only in the CRE but also in the SRE with additional discrete covariates.

5.5 Practical questions

How do we choose X to construct a SRE? Theoretically, X should be predictive to the potential outcomes. In some cases, the experimenter has enough background knowledge about the predictive covariates based on, for example, some pilot studies. Then the choice of X should be straightforward. In some other cases, this background knowledge may not be clear enough. Experimenters instead choose X based on logistic convenience, for example, X can be indicator for the study areas or the cohort of students.

The choose of K is a related problem. Theoretically, more stratification increases the estimation efficiency if all strata are large enough. However, extremely large K may even decrease the estimation efficiency. In simulation studies, we observe diminishing marginal returns of increasing K . Anecdotally, $K = 5$ often suffices for efficiency gain. Some experimenter prefers the most extreme version of the SRE with $K = n/2$. This results in the matched pair design, which will be discussed in Chapter 7 later.

Some experiments have multidimensional continuous covariates. Can the SRE still be used? If we have a pilot study, we can build a model for the potential outcome $Y(0)$ given those covariates, and then we can choose X as

a discretized version of the predictor $\hat{Y}(0)$. In general, if we do not have such a pilot study or we do not want to make ad hoc discretizations, we can use a more general strategy called rerandomization, which is the topic for Chapter 6.

5.6 Homework Problems

5.1 Consequence of the constant propensity score

Show that if $e_{[k]} = e$ for all $k = 1, \dots, K$, then $\hat{\tau} = \hat{\tau}_S$.

5.2 Consequence of constant individual causal effects

Assume that the individual causal effects are constant $\tau_i = \tau$ for all $i = 1, \dots, n$. Consider the following class of weighted estimator for τ :

$$\hat{\tau}_w = \sum_{k=1}^K w_{[k]} \hat{\tau}_{[k]},$$

where $w_{[k]} \geq 0$ for all k .

Find the condition on the $w_{[k]}$'s such that $\hat{\tau}_w$ is unbiased for τ . Among all unbiased estimators, find the one with the minimum variance.

5.3 FRT for the Project STAR data in Imbens and Rubin (2015)

Reanalyze the Project STAR data using the Fisher randomization test. Note that I use Z for the treatment indicator but Imbens and Rubin (2015) use W . Use $\hat{\tau}_S$, V and the aligned rank statistic in the Fisher randomization test. Compare the p -values.

```
treatment = list(c(1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,0,0),
                  c(1,1,0,0))
```

```

outcome = list(c(0.165,0.321,-0.197,0.236),
               c(0.918,-0.202,1.19,0.117),
               c(0.341,0.561,-0.059,-0.496,0.225),
               c(-0.024,-0.450,-1.104,-0.956),
               c(-0.258,-0.083,-0.126,0.106),
               c(1.151,0.707,0.597,-0.495),
               c(0.077,0.371,0.685,0.270),
               c(-0.870,-0.496,-0.444,0.392,-0.934,-0.633),
               c(-0.568,-1.189,-0.891,-0.856),
               c(-0.727,-0.580,-0.473,-0.807),
               c(-0.533,0.458,-0.383,0.313),
               c(1.001,0.102,0.484,0.474,0.140),
               c(0.855,0.509,0.205,0.296),
               c(0.618,0.978,0.742,0.175),
               c(-0.545,0.234,-0.434,-0.293),
               c(-0.240,-0.150,0.355,-0.130))

```

5.4 A multi-center trial

Gould (1998, Table 1) reported the following data from a multi-center trial:

```

> multicenter = read.csv("multicenter.csv")
> multicenter
   center n0 mean0  sd0  n1 mean1  sd1  n5 mean5  sd5
1       1   7  0.43  4.58   7 -5.43  5.53   8 -2.63  3.38
2       2  11  0.10  4.21  11 -2.59  3.95  12 -2.21  4.14
3       3   6  2.58  4.80   6 -3.94  4.25   7  1.29  7.39
4       4  10 -2.30  3.86  10 -1.23  5.17  10 -1.40  2.27
5       5  10  2.08  6.46  10 -6.70  7.45  10 -5.13  3.91
6       6   6  1.13  3.24   5  3.40  8.17   5 -1.59  3.19
7       7   5  1.20  7.85   6 -3.67  4.89   5 -1.40  2.61
8       8  12 -1.21  2.66  13  0.18  3.81  12 -4.08  6.32
9       9   8  1.13  5.28   8 -2.19  5.17   9 -1.96  5.84
10      10   9 -0.11  3.62  10 -2.00  5.35  10  0.60  3.53
11      11  15 -4.37  6.12  14 -2.68  5.34  15 -2.14  4.27
12      12   8 -1.06  5.27   9  0.44  4.39   9 -2.03  5.76
13      13  12 -0.08  3.32  12 -4.60  6.16  11 -6.22  5.33
14      14   9  0.00  5.20   9 -0.25  8.23   7 -3.29  5.12
15      15   6  1.83  5.85   7 -1.23  4.33   6 -1.00  2.61
16      16  14 -4.21  7.53  14 -2.10  5.78  12 -5.75  5.63
17      17  13  0.76  3.82  13  0.55  2.53  13 -0.63  5.41
18      18  15 -1.05  4.54  13  2.54  4.16  14 -2.80  2.89
19      19  15  2.07  4.88  15 -1.67  4.95  15 -3.43  4.71
20      20  11 -1.46  5.48  10 -1.99  5.63  10 -6.77  5.19
21      21   5  0.80  4.21   5 -3.35  4.73   5 -0.23  4.14
22      22  11 -2.92  5.42  10 -1.22  5.95  11 -4.45  6.65
23      23   9 -3.37  4.73   9 -1.38  4.17   7  0.57  2.70
24      24  12 -1.92  2.91  12 -0.66  3.55  12 -2.39  2.27
25      25   9 -3.89  4.76   9 -3.22  5.54   8 -1.23  4.91

```

```

26      26 15 -3.48 5.98 15 -2.13 3.25 14 -3.71 5.30
27      27 11 -1.91 6.49 12 -1.33 4.40 11 -1.52 4.68
28      28 10 -2.66 3.80 10 -1.29 3.18 10 -4.70 3.43
29      29 13 -0.77 4.73 13 -2.31 3.88 13 -0.47 4.95

```

This is a SRE with centers being the strata. The trial was conducted to study the efficacy and tolerability of finasteride, a drug for treating benign prostatic hyperplasia. Within each of the 29 centers, patients were randomized into three arms: control, finasteride 1mg, and finasteride 5mg. The above dataset provides summary statistics for the outcome, which is the change from baseline in total symptom score. The total symptom score is the sum of the responses to nine questions (score 0 to 4) about symptoms pertaining to various aspects of impaired urinary ability. The meanings of the columns are:

1. `center`: number of the center;
2. `n0`, `n1`, `n5`: sample sizes of the three arms;
3. `mean0`, `mean1`, `mean5`: mean of the outcome;
4. `sd0`, `sd1`, `sd5`: standard deviation of the outcome.

The individual-level outcomes are not reported so we cannot implement the FRT. However, the Neymanian inference only requires the summary statistics. Report the point estimators and variance estimators for comparing “finasteride 1mg” and “finasteride 5mg” to “control”, separately.

5.5 Data re-analyses

Re-analyze the LaLonde data used in `Neymanlalonde.R`. Conduct both Fisherian and Neymanian inferences.

The original experiment is a completely randomized experiment. Now we pretend that the original experiment is a stratified randomized experiment. First, re-analyze the data pretending that the experiment is stratified on the race (black, Hispanic or other). Second, re-analyze the data pretending that the experiment is stratified on marital status. Third, re-analyze the data pretending that the experiment is stratified on the indicator of high school diploma.

Compare with the results obtained under a completely randomized experiments.

5.6 Recommended reading

Miratrix et al. (2013) provided solid theory for post-stratification and compared it with stratification. A main theoretical result is that their difference is small asymptotically although they can differ in finite samples.



6

Rerandomization and Regression Adjustment

Stratification and post-stratification in Chapter 5 are duals for discrete covariates in the design and analysis of randomized experiments. How should we deal with multidimensional possibly continuous covariates? We can discretize continuous covariates, but this is not an ideal strategy with many covariates. Rerandomization and regression adjustment are duals for general covariates, which are the topics for this chapter.

The following table summarizes the topics of Chapters 5 and 6:

	design	analysis
discrete covariate	stratification	post-stratification
general covariate	rerandomization	regression adjustment

6.1 Rerandomization

6.1.1 Experimental design

Again we consider a finite population of n units, where n_1 of them receive the treatment and n_0 of them receive the control. Let $\mathbf{Z} = (Z_1, \dots, Z_n)$ be the treatment vector for these units. Unit i has covariate $X_i \in \mathbb{R}^K$ which can have continuous or binary components. Concatenate them as $\mathbf{X} = (X_1, \dots, X_n)$ and center them at mean zero $\bar{X} = n^{-1} \sum_{i=1}^n X_i = 0$ without loss of generality.

The CRE balances the covariates in the treatment and control groups on average, for instance, the difference in means of the covariates

$$\hat{\tau}_X = n_1^{-1} \sum_{i=1}^n Z_i X_i - n_0^{-1} \sum_{i=1}^n (1 - Z_i) X_i$$

has mean zero under the CRE. However, it can result in undesirable covariate balance across the treatment and control groups in the realized treatment allocation, that is, the realized value of $\hat{\tau}_X$ is often not zero. Using the vector form of Neyman (1923) in Problem 4.6, we can show that

$$\text{cov}(\hat{\tau}_X) = \frac{1}{n_1} S_X^2 + \frac{1}{n_0} S_X^2 = \frac{n}{n_1 n_0} S_X^2,$$

where $S_X^2 = (n - 1)^{-1} \sum_{i=1}^n X_i X_i^\top$. The following Mahalanobis distance measures the difference between the treatment and control groups:

$$M = \hat{\tau}_X^\top \text{cov}(\hat{\tau}_X)^{-1} \hat{\tau}_X = \hat{\tau}_X^\top \left(\frac{n}{n_1 n_0} S_X^2 \right)^{-1} \hat{\tau}_X.$$

Technically the above formula of M is meaningful only if S_X^2 is invertible, which means that the columns of the covariate matrix are linearly independent. If a column can be represented by a linear combinations of other columns, it is redundant and should be dropped before the experiment. A nice feature of M is that it is invariant under non-degenerate linear transformations of X . Lemma 6.1 below summarizes the result with the proof relegated to Problem 6.2.

Lemma 6.1 *M remains the same if we transform X_i to $\alpha + BX_i$ for all units $i = 1, \dots, n$ where $\alpha \in \mathbb{R}^K$ and $B \in \mathbb{R}^{K \times K}$ is invertible.*

The finite population central limit theorem (Li and Ding, 2017) ensures that with large n , the Mahalanobis distance M is approximately χ_K^2 under the CRE. Therefore, it is likely that M has a large realized value under the CRE with asymptotic mean K and variance $2K$. Rerandomization avoids covariate imbalance by discarding the treatment allocations with large values of M . Below I give a formal definition of the rerandomization using the Mahalanobis distance (ReM), which was proposed by Cox (1982) and Morgan and Rubin (2012).

Definition 6.1 (ReM) *Draw \mathbf{Z} from CRE and accept it if and only if*

$$M \leq a,$$

for some predetermined constant $a > 0$.

Choosing a is like choosing the number of strata in the SRE, which is a non-trivial problem in practice. At one extreme, $a = \infty$, we just conduct the CRE. At the other extreme, $a = 0$, there are very few feasible treatment allocations, and consequently, the experiment has little randomness, rendering randomization-based inference useless. As a compromise, we choose a small but not extremely small a , for example, $a = 0.001$ or some upper quantile of a χ_K^2 distribution.

ReM uses the Mahalanobis distance as the balance criterion. We can consider general rerandomization with the balance criterion defined as a function of \mathbf{Z} and \mathbf{X} . For example, we can use the following criterion based on marginal tests for all coordinates of $X_i = (x_{i1}, \dots, x_{iK})^\top$. We accept \mathbf{Z} if and only if

$$\left| \frac{\hat{\tau}_{xk}}{\sqrt{\frac{n}{n_1 n_0} S_{xk}^2}} \right| \leq a \quad (k = 1, \dots, K) \tag{6.1}$$

for some predetermined constant $a > 0$. For example, a some upper quantile of a standard Normal distribution. ReM has many desirable properties. As mentioned above, it is invariant to linear transformation of the covariates. Moreover, it has nice geometric properties and elegant mathematical theory. This chapter will focus on ReM. See Zhao and Ding (2021b) for the theory for the rerandomization based on criterion (6.1) as well as other criteria.

6.1.2 Statistical inference

An important question is how to analyze the data under ReM. Bruhn and McKenzie (2009) and Morgan and Rubin (2012) argued that we can always use the FRT as long as we simulate \mathbf{Z} under the constraint $M \leq a$. This always yields finite-sample exact p -values under the sharp null hypothesis.

It is a challenging problem to derive the finite sample properties of ReM without assuming the sharp null hypothesis. Instead, Li et al. (2018b) derived the asymptotic distribution of the difference in means of the outcome $\hat{\tau}$ under ReM and the regularity conditions below.

Condition 6.1 *As $n \rightarrow \infty$,*

1. n_1/n and n_0/n have positive limits;
2. the finite population covariance of $\{X_i, Y_i(1), Y_i(0), \tau_i\}$ has limit;
3. $\max_{1 \leq i \leq n} |Y_i(1) - \bar{Y}(1)|^2/n \rightarrow 0$, $\max_{1 \leq i \leq n} |Y_i(0) - \bar{Y}(0)|^2/n \rightarrow 0$, and $\max_{1 \leq i \leq n} \|X_i\|^2/n \rightarrow 0$,

Below is the main theorem for ReM. Let

$$L_{K,a} \sim D_1 \mid \mathbf{D}^\top \mathbf{D} \leq a$$

where $\mathbf{D} = (D_1, \dots, D_K)$ follows a K -dimensional standard Normal distribution; let ε follows a univariate standard Normal distribution; $L_{K,a} \perp\!\!\!\perp \varepsilon$.

Theorem 6.1 *Under ReM with $M \leq a$ and Condition 6.1, we have¹*

$$\hat{\tau} - \tau \stackrel{d}{\sim} \sqrt{\text{var}(\tau)} \left\{ \sqrt{R^2} L_{K,a} + \sqrt{1 - R^2} \varepsilon \right\},$$

where

$$\text{var}(\hat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n}$$

is Neyman (1923)'s variance formula proved in Chapter 4, and

$$R^2 = \text{corr}^2(\hat{\tau}, \hat{\tau}_X)$$

¹The notation “ $A \stackrel{d}{\sim} B$ ” means that A and B have the same asymptotic distributions.

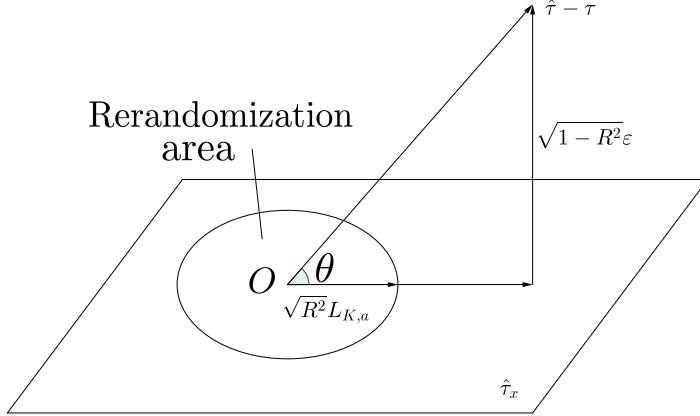


FIGURE 6.1: Geometry of ReM

is the squared multiple correlation coefficient² between $\hat{\tau}$ and $\hat{\tau}_X$ under the CRE.

Although the proof of Li et al. (2018b) is technical, the asymptotic distribution in Theorem 6.1 has clear geometric interpretation, as shown in Figure 6.1. It shows that $\hat{\tau}$ decomposes into a component that is a linear combination of $\hat{\tau}_X$ and a component that is orthogonal to $\hat{\tau}_X$. Geometrically, $\cos^2 \theta = R^2$, where θ is the angle between $\hat{\tau}$ and $\hat{\tau}_X$. ReM affects the first component but does not change the second component. The truncated Normal distribution $L_{K,a}$ is due to the restriction of ReM on the first component.

When $a = \infty$, the asymptotic distribution simplifies to the one under the CRE:

$$\hat{\tau} - \tau \stackrel{d}{\sim} \sqrt{\text{var}(\tau)} \varepsilon.$$

When the threshold a is close to zero, the the asymptotic distribution simplifies to

$$\hat{\tau} - \tau \stackrel{d}{\sim} \sqrt{\text{var}(\tau)(1 - R^2)} \varepsilon.$$

So with a small threshold a , the efficiency gain due to ReM depends on R^2 , which has the following equivalent form.

Proposition 6.1 *Under the CRE,*

$$R^2 = \text{corr}^2(\hat{\tau}, \hat{\tau}_X) = \frac{n_1^{-1} S^2(1 | x) + n_0^{-1} S^2(0 | x) - n^{-1} S^2(\tau | x)}{n_1^{-1} S^2(1) + n_0^{-1} S^2(0) - n^{-1} S^2(\tau)},$$

²The squared multiple correlation coefficient between a random variable y and a random vector X is defined as

$$R_{yX}^2 = \text{corr}^2(y, X) = \frac{\text{cov}(y, X)\text{cov}(X)^{-1}\text{cov}(X, y)}{\text{var}(y)}.$$

It extends the definition of the Pearson correlation coefficient and measures the linear dependence of y on X .

where $\{S^2(1), S^2(0), S^2(\tau)\}$ are the finite population variances of $\{Y_i(1), Y_i(0), \tau_i\}_{i=1}^n$, and $\{S^2(1 | x), S^2(0 | x), S^2(\tau | x)\}$ are the corresponding finite population variances of their linear projections on $(1, X_i)$.³ Under the constant causal effect assumption with $\tau_i = \tau$, R^2 reduces to the finite population squared multiple correlation between $Y_i(0)$ and X_i .

I leave the proof of Proposition 6.1 to Problem 6.4.

When $0 < a < \infty$, the asymptotic distribution has a more complicated form and is more concentrated at τ and thus the difference in means is more precise under ReM than under the CRE.

If we ignore the design of ReM and still use the confidence interval based on Neyman (1923)'s variance formula and the Normal approximation, it is overly conservative and overcovers τ even if the individual causal effects are constant. Li et al. (2018b) described how to construct confidence intervals based on Theorem 6.1. We omit the discussion here but will come back to the inference issue in Section 6.3.

6.2 Regression adjustment

What if we do not conduct rerandomization in the design stage but want to adjust for covariate imbalance in the analysis stage of the CRE? We will discuss several regression adjustment strategies.

6.2.1 Covariate-adjusted FRT

The covariates \mathbf{X} are all fixed, and furthermore, under H_{0F} , the observed outcomes are all fixed. Therefore, we can simulate the distribution of any test statistic $T(\mathbf{Z}, \mathbf{Y}, \mathbf{X})$ and calculate the p -value. The basic idea of the FRT remains the same in the presence additional covariates.

There are two general strategies to construct the test statistic, as summarized by Zhao and Ding (2021a). Problem 3.6 hints at both of them. I summarize them below:

- The first strategy is to construct the test statistic based on residuals from fitted statistical models. We can regress Y_i on X_i to obtain residual ε_i , and then treat ε_i as the pseudo outcome to construct test statistics.
- The second strategy is to use a regression coefficient as a test statistic. We

³For example, the linear projection of $Y_i(1)$ on $(1, X_i)$ is $\alpha_1 + \beta_1 X_i$ where

$$(\alpha_1, \beta_1) = \arg \min_{a, b} \sum_{i=1}^n \{Y_i(1) - a - b^T X_i\}^2.$$

can regress Y_i on (Z_i, X_i) to obtain the coefficient of Z_i as the test statistic. The rest of this section will review some test statistics based on OLS.

In strategy one, we only need to run regression once, but in strategy two, we need to run regression many times. In the above, “regression” is a generic term, which can be linear regression, logistic regression, or even machine learning algorithms. The FRT with any test statistics from these two strategies will be finite-sample exact under H_{0F} although they differ under alternative hypotheses.

6.2.2 Analysis of covariance and extensions

Now we turn to direct estimation of the average causal effect τ that adjusts for the observed covariates.

Historically, Fisher (1925) proposed to use the analysis of covariance (ANCOVA) to improve estimation efficiency. This remains a standard strategy in many fields. He suggested running the OLS of Y_i on (Z_i, X_i) and obtaining the coefficient of Z_i as an estimator for τ . Let $\hat{\tau}_F$ denote Fisher’s ANCOVA estimator.

A former Berkeley Statistics Professor, David Freedman, reanalyzed Fisher’s ANCOVA under Neyman (1923)’s potential outcomes framework. Freedman (2008a,b) found the following negative results:

1. $\hat{\tau}_F$ is biased, but the simple difference in means $\hat{\tau}$ is unbiased.
2. The asymptotic variance of $\hat{\tau}_F$ may be even larger than that of $\hat{\tau}$.
3. The standard error from the OLS is inconsistent for the true standard error of $\hat{\tau}_F$ under the CRE.

A Berkeley Ph.D. student, Winston Lin, wrote a thesis in response to Freedman’s critiques. Lin (2013) found the following positive results:

1. The bias of $\hat{\tau}_F$ is small in large samples, and it goes to zero as the sample size approaches infinity.
2. We can improve the asymptotic efficiency of both $\hat{\tau}$ and $\hat{\tau}_F$ by using the coefficient of Z_i in the OLS of Y_i on $(Z_i, X_i, Z_i \times X_i)$. Let $\hat{\tau}_L$ denote Lin (2013)’s estimator. Moreover, the EHW standard error is a conservative estimator for the true standard error of $\hat{\tau}_L$ under the CRE.
3. The EHW standard error⁴ for $\hat{\tau}_F$ in the OLS fit of Y_i on (Z_i, X_i) is

⁴Without covariates, the HC2 correction yields identical variance estimator as Neyman (1923)’s classic one. For coherence, we can also use the HC2 correction for Lin (2013)’s estimator with covariate adjustment. When the number of covariates is small compared to the sample size and the covariates do not contain outliers, the variants of the EHW standard error perform similarly to the original one. When the number of covariates is large compared to the sample size or the covariates contain outliers, the variants can outperform

a conservative estimator for the true standard error of $\hat{\tau}_F$ under the CRE.

6.2.2.1 Some heuristics for Lin (2013)'s results

Neyman (1923)'s result demonstrates that the variance of the difference-in-means estimator depends on the variances of the potential outcomes. Intuitively, we can reduce the variance of the estimator by reducing the variances of the outcomes. A simple family of linearly adjusted estimator is

$$\hat{\tau}(\beta_1, \beta_0) = n_1^{-1} \sum_{i=1}^n Z_i(Y_i - \beta_1^\top X_i) - n_0^{-1} \sum_{i=1}^n (1 - Z_i)(Y_i - \beta_0^\top X_i) \quad (6.2)$$

$$= \left\{ \hat{\bar{Y}}(1) - \beta_1^\top \hat{\bar{X}}(1) \right\} - \left\{ \hat{\bar{Y}}(0) - \beta_0^\top \hat{\bar{X}}(0) \right\}, \quad (6.3)$$

where $\{\hat{\bar{Y}}(1), \hat{\bar{Y}}(0)\}$ are the sample means of the outcomes, and $\{\hat{\bar{X}}(1), \hat{\bar{X}}(0)\}$ are the sample means of the covariates. This covariate-adjusted estimator $\hat{\tau}(\beta_1, \beta_0)$ tries to reduce the variance of $\hat{\tau}$ by residualizing the potential outcomes. It reduces to $\hat{\tau}$ with $\beta_1 = \beta_0 = 0$. It has mean τ for any fixed values of β_1 and β_0 because $\bar{X} = 0$. We are interested in finding the (β_1, β_0) that minimized the variance of $\hat{\tau}(\beta_1, \beta_0)$. This estimator is essentially the difference in means of the adjusted potential outcomes $\{Y_i(1) - \beta_1^\top X_i, Y_i(0) - \beta_0^\top X_i\}_{i=1}^n$. Applying Neyman (1923)'s result, this estimator has variance

$$\text{var}\{\hat{\tau}(\beta_1, \beta_0)\} = \frac{S^2(1; \beta_1)}{n_1} + \frac{S^2(0; \beta_1)}{n_0} - \frac{S^2(\tau; \beta_1, \beta_0)}{n},$$

where $S^2(z; \beta_1)$ ($z = 1, 0$) and $S^2(\tau; \beta_1, \beta_0)$ are the finite population variances of the adjusted potential outcomes and individual effects, respectively; moreover, a conservative variance estimate is

$$\hat{V}(\beta_1, \beta_0) = \frac{\hat{S}^2(1; \beta_1)}{n_1} + \frac{\hat{S}^2(0; \beta_1)}{n_0},$$

where

$$\hat{S}^2(1; \beta_1) = (n_1 - 1)^{-1} \sum_{i=1}^n Z_i \{Y_i - \gamma_1 - \beta_1^\top X_i\}^2,$$

$$\hat{S}^2(0; \beta_0) = (n_0 - 1)^{-1} \sum_{i=1}^n (1 - Z_i) \{Y_i - \gamma_0 - \beta_0^\top X_i\}^2$$

are the sample variances of the adjusted potential outcomes with γ_1 and γ_0

the original one. In those cases, Lei and Ding (2021) recommend using the HC3 variant of the EHW standard error. See Chapter A2 for more details of the EHW standard errors.

being the sample means of $Y_i - \beta_1^T X_i$ under treatment and $Y_i - \beta_0^T X_i$ under control. To minimize $\hat{V}(\beta_1, \beta_0)$, we need to solve two OLS problems:

$$\min_{\gamma_1, \beta_1} \sum_{i=1}^n Z_i \{Y_i - \gamma_1 - \beta_1^T X_i\}^2, \quad \min_{\gamma_0, \beta_0} \sum_{i=1}^n (1 - Z_i) \{Y_i - \gamma_0 - \beta_0^T X_i\}^2.$$

We run OLS of Y_i on X_i for the treatment and control groups separately and obtain $(\hat{\gamma}_1, \hat{\beta}_1)$ and $(\hat{\gamma}_0, \hat{\beta}_0)$. The final estimator is

$$\begin{aligned}\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0) &= n_1^{-1} \sum_{i=1}^n Z_i (Y_i - \hat{\beta}_1^T X_i) - n_0^{-1} \sum_{i=1}^n (1 - Z_i) (Y_i - \hat{\beta}_0^T X_i) \\ &= \left\{ \hat{Y}(1) - \hat{\beta}_1^T \hat{X}(1) \right\} - \left\{ \hat{Y}(0) - \hat{\beta}_0^T \hat{X}(0) \right\}.\end{aligned}$$

From the properties of the OLS fits (see (A2.3)), we know

$$\hat{Y}(1) = \hat{\gamma}_1 + \hat{\beta}_1^T \hat{X}(1), \quad \hat{Y}(0) = \hat{\gamma}_0 + \hat{\beta}_0^T \hat{X}(0).$$

Therefore, we can rewrite the estimator as

$$\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0) = \hat{\gamma}_1 - \hat{\gamma}_0 \tag{6.4}$$

The equivalent form in (6.4) suggests that we can obtain $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ from a single OLS fit below.

Proposition 6.2 *The estimator $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ in (6.4) equals the coefficient of Z_i in the OLS fit of Y_i on $(Z_i, X_i, Z_i \times X_i)$, which is $\hat{\tau}_L$ introduced before.*

I leave the proof of Proposition 6.2 to Problem 6.5, which is a pure algebra fact.

Based on the discussion above, a conservative variance estimator for $\hat{\tau}_L$ is

$$\begin{aligned}\hat{V}(\hat{\beta}_1, \hat{\beta}_0) &= \frac{1}{n_1(n_1 - 1)} \sum_{i=1}^n Z_i (Y_i - \hat{\gamma}_1 - \hat{\beta}_1^T X_i)^2 \\ &\quad + \frac{1}{n_0(n_0 - 1)} \sum_{i=1}^n (1 - Z_i) (Y_i - \hat{\gamma}_0 - \hat{\beta}_0^T X_i)^2.\end{aligned}$$

Based on quite technical calculations, Lin (2013) further showed that the EHW standard error from the OLS in Proposition 6.2 is almost identical to $\hat{V}(\hat{\beta}_1, \hat{\beta}_0)$ which is a conservative estimator of the true standard error of $\hat{\tau}_L$ under the CRE. Intuitively, this is because we do not assume that the linear model is correctly specified, and the EHW standard error is robust to model misspecification.

There is a subtle issue with the discussion above. The variance formula $\text{var}\{\hat{\tau}(\beta_1, \beta_0)\}$ works for fixed (β_1, β_0) , but the estimator $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ uses two estimated coefficients $(\hat{\beta}_1, \hat{\beta}_0)$. The additional uncertainty in the estimated coefficients may cause finite-sample bias in the final estimator. Lin (2013) showed that the issue goes away asymptotically. However, his theory requires a large sample size and some regularity conditions on the potential outcomes and covariates.

TABLE 6.1: Predicting the potential outcomes

X	Z	$Y(1)$	$Y(0)$	$\hat{Y}(1)$	$\hat{Y}(0)$
X_1	1	$Y_1(1)$?	$\hat{\mu}_1(X_1)$	$\hat{\mu}_0(X_1)$
⋮					
X_{n_1}	1	$Y_{n_1}(1)$?	$\hat{\mu}_1(X_{n_1})$	$\hat{\mu}_0(X_{n_1})$
X_{n_1+1}	0	?	$Y_{n_1+1}(0)$	$\hat{\mu}_1(X_{n_1+1})$	$\hat{\mu}_0(X_{n_1+1})$
⋮					
X_n	0	?	$Y_n(0)$	$\hat{\mu}_1(X_n)$	$\hat{\mu}_0(X_n)$

6.2.2.2 Understanding Lin (2013)'s estimator via predicting the potential outcomes

We can view Lin (2013)'s estimator as a *predictive estimator* based on OLS fits of the potential outcomes. We build a prediction model for $Y(1)$ based on X using the data from the treatment group:

$$\hat{\mu}_1(x) = \hat{\gamma}_1 + \hat{\beta}_1^T x. \quad (6.5)$$

Similarly, we build a prediction model for $Y(0)$ based on X using the data from the control group:

$$\hat{\mu}_0(x) = \hat{\gamma}_0 + \hat{\beta}_0^T x. \quad (6.6)$$

If we predict the missing potential outcomes, then we have the following predictive estimator:

$$\hat{\tau}_{\text{pred}} = n^{-1} \left\{ \sum_{Z_i=1} Y_i + \sum_{Z_i=0} \hat{\mu}_1(X_i) - \sum_{Z_i=1} \hat{\mu}_0(X_i) - \sum_{Z_i=0} Y_i \right\}. \quad (6.7)$$

We can verify that with (6.5) and (6.6), the predictive estimator equals Lin (2013)'s estimator:

$$\hat{\tau}_{\text{pred}} = \hat{\tau}_L. \quad (6.8)$$

If we predict all potential outcomes even if they are observed, we have the following *projective estimator*:

$$\hat{\tau}_{\text{proj}} = n^{-1} \sum_{i=1}^n \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}. \quad (6.9)$$

We can verify that with (6.5) and (6.6), the projective estimator equals Lin (2013)'s estimator:

$$\hat{\tau}_{\text{proj}} = \hat{\tau}_L. \quad (6.10)$$

I leave the proofs of (6.8) and (6.10) to Problem 6.6.

The more general formulas (6.7) and (6.9) are well defined with other predictors of the potential outcomes. To make connections with Lin (2013)'s estimator, I focus on the linear predictors here. They can be quite general, including much more complicated machine learning algorithms. However, constructing point estimator is just the first step in analyzing the CRE. A more important second step is to quantify the uncertainty associated with the estimator, which depends on the properties of the predictors of the potential outcomes. Nevertheless, without doing additional theoretical analysis, we can always use (6.7) and (6.9) as the test statistics in the FRT.

6.2.2.3 Understanding Lin (2013)'s estimator via adjusting for covariate imbalance

The linearly-adjusted estimator has an equivalent form

$$\hat{\tau}(\beta_1, \beta_0) = \hat{\tau} - \gamma^T \hat{\tau}_X \quad (6.11)$$

where $\gamma = \frac{n_0}{n} \beta_1 + \frac{n_1}{n} \beta_0$, so we can also write it as $\hat{\tau}(\gamma) = \hat{\tau}(\beta_1, \beta_0)$. Similarly, Lin (2013)'s estimator has an equivalent form

$$\hat{\tau}_L = \hat{\tau} - \hat{\gamma}^T \hat{\tau}_X, \quad (6.12)$$

where $\hat{\gamma} = \frac{n_0}{n} \hat{\beta}_1 + \frac{n_1}{n} \hat{\beta}_0$. I leave the proofs of (6.11) and (6.12) to Problem 6.7. The forms (6.11) and (6.12) are the mathematical statements of “adjusting for the covariate imbalance.” They essentially subtract some linear combinations of the difference in means of the covariates. Since $\hat{\tau}$ and $\hat{\tau}_X$ are correlated, the covariate adjustment with an appropriate γ reduces the variance of $\hat{\tau}$. Another interesting feature of (6.11) and (6.12) is that the final estimators depend only on γ or $\hat{\gamma}$, so the choice of the β -coefficients are not unique. Therefore, Lin (2013)'s estimator is just one of the optimal estimators, but it can be easily implemented via the standard OLS with the EHW standard error.

6.2.3 Some additional remarks on regression adjustment

6.2.3.1 Duality between ReM and regression adjustment

Li et al. (2018b) pointed out that ReM and Lin (2013)'s regression adjustment are duals in using covariates in the design and analysis stages of the experiment. To be more specific, when a is small, the asymptotic distribution of $\hat{\tau}$ under ReM is almost identical to the asymptotic distribution of $\hat{\tau}_L$ under the CRE. So ReM uses covariates in the design stage and Lin (2013)'s regression adjustment uses covariates in the analysis stage, achieving nearly the same asymptotic efficiency gain when a is small.

6.2.3.2 Equivalence of regression adjustment and post-stratification

If we have discrete covariate C_i with K categories, we can create $K-1$ centered dummy variables

$$X_i = (I(C_i = 1) - \pi_{[1]}, \dots, I(C_i = K-1) - \pi_{[K-1]}).$$

In this case, Lin (2013)'s regression adjustment is equivalent to post-stratification, as summarized by the following proposition.

Proposition 6.3 $\hat{\tau}_L$ based in X_i is numerically identical to the post-stratification estimator based on C_i .

I leave the proof of Proposition 6.3 as Problem 6.9.

6.2.3.3 Difference-in-difference as a special case of covariate adjustment $\hat{\tau}(\beta_1, \beta_0)$

An important covariate X in many studies is the lagged outcome before the treatment. For instance, the covariate X is the pre-test score if the outcome Y is the post-test score in educational research; the covariate X is the log wage before the job training program if the outcome Y is the log wage after the job training program. With the lagged outcome X as a covariate, a popular estimator is the *gain score* or *difference-in-difference* estimator with $\beta_1 = \beta_0 = 1$ in (6.2) and (6.3):

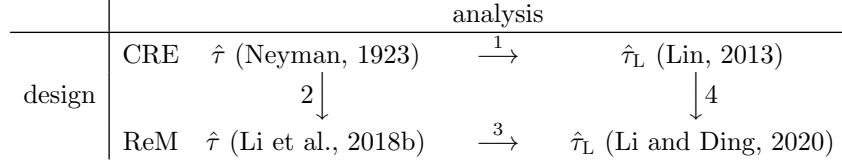
$$\begin{aligned}\hat{\tau}(1, 1) &= n_1^{-1} \sum_{i=1}^n Z_i(Y_i - X_i) - n_0^{-1} \sum_{i=1}^n (1 - Z_i)(Y_i - X_i) \\ &= \left\{ \hat{\bar{Y}}(1) - \hat{\bar{Y}}(0) \right\} - \left\{ \hat{\bar{X}}(1) - \hat{\bar{X}}(0) \right\}.\end{aligned}$$

The first form of $\hat{\tau}(1, 1)$ justifies the name *gain score* because it is essentially the difference in means of the gain score $g_i = Y_i - X_i$. The second form of $\hat{\tau}(1, 1)$ justifies the name *difference-in-difference* because it is the difference between two differences in means. This estimator is different from Lin (2013)'s estimator: it fixes $\beta_1 = \beta_0 = 1$ in advance while Lin (2013)'s estimator involves two estimated β 's. It is unbiased with a conservative variance estimator

$$\begin{aligned}\hat{V}(1, 1) &= \frac{1}{n_1(n_1 - 1)} \sum_{i=1}^n Z_i \{g_i - \hat{\bar{g}}(1)\}^2 \\ &\quad + \frac{1}{n_0(n_0 - 1)} \sum_{i=1}^n (1 - Z_i) \{g_i - \hat{\bar{g}}(0)\}^2,\end{aligned}$$

where $\hat{\bar{g}}(1)$ and $\hat{\bar{g}}(0)$ are the sample means of the gain score $g_i = Y_i - X_i$ under treatment and control, respectively. When the lagged outcome is a strong predictor of the outcome, the gain score $g_i = Y_i - X_i$ often has much smaller variance than the outcome itself. In this case, $\hat{\tau}(1, 1)$ often greatly reduces the variance of the simple difference in means of the outcome.

TABLE 6.2: Design and analysis of experiments



6.2.4 Extension to the SRE

It is possible that we have an experiment stratified on a discrete variable C and observe additional covariates X . If all strata are large, then we can obtain Lin (2013)'s estimators within strata $\hat{\tau}_{L,[k]}$ and obtain the final estimator as

$$\hat{\tau}_{L,S} = \sum_{k=1}^K \pi_{[k]} \hat{\tau}_{L,[k]}.$$

A conservative variance estimator is

$$\hat{V}_{L,S} = \sum_{k=1}^K \pi_{[k]}^2 \hat{V}_{EHW,[k]},$$

where $\hat{V}_{EHW,[k]}$ is the EHW variance estimator from the OLS fit of the outcome on the treatment indicator, the covariates, and their interactions within stratum k . Importantly, we need to center covariates by their stratum-specific means.

6.3 Unification, combination, and comparison

Li and Ding (2020) unified the literature and showed that we can combine rerandomization and regression adjustment. That is, if we rerandomize in the design stage, we can use Lin (2013)'s estimator with the EHW standard error in the analysis stage. The combination of rerandomization and regression adjustment improves covariate balance in the design stage and estimation efficiency in the analysis stage.

Table 6.2 summarizes the literature from Neyman (1923) to Li and Ding (2020). Arrow 1 illustrates the efficiency gain of covariate adjustment in the CRE: asymptotically, $\hat{\tau}_L$ has smaller variance than $\hat{\tau}$. Arrow 2 illustrates the efficiency gain of the ReM: asymptotically, $\hat{\tau}$ has narrower quantile range under the ReM than under the CRE. Arrows 3 and 4 illustrate the benefits of the combination.

6.4 Simulation

Angrist et al. (2009) conducted an experiment to evaluate different strategies to improve academic performance among college freshmen. Here I use a subset of the original data, focusing on the control group and the treatment group offered academic support services and financial incentives for good grades. The outcome is the GPA at the end of the first year, and two covariates are the gender and baseline GPA. The following table summarizes the results based on the unadjusted and adjusted estimators. The adjusted estimator has smaller standard error although it gives the same insignificant result as the unadjusted estimator.

	estimate	s.e.	t-stat	p-value
Neyman	0.054	0.076	0.719	0.472
Lin	0.075	0.072	1.036	0.300

I also use this dataset to conduct simulation studies to evaluate the four design and analysis strategies summarized in Table 6.2. I fit quadratic functions of the outcome on the covariates and use them to impute all the missing potential outcomes, separately for the treated and control groups. To show the improvement of ReM and regression adjustment, I also rescale the error terms by 0.1 and 0.25 to increase the signal to noise ratio. With the imputed Science Table, I generate 2000 treatments, obtain the observed data, and calculate the estimators. In the simulation, the “true” outcome model is nonlinear, but we still use linear adjustment for estimation. By doing this, we can evaluate the properties of the estimators when the linear model is misspecified.

Figure 6.2 shows the violin plots of the four combinations, subtracting the true τ from the estimates. As predicted by the theory, all estimators are nearly unbiased, and both ReM and regression adjustment improve efficiency. They are more effective when the noise level is smaller.

6.5 Final remarks

With a continuous outcome, Fisher’s ANCOVA has been the standard approach for many years. Lin (2013)’s improvement has better theoretical properties even if the linear model is misspecified. With a binary outcome, it is common to use the coefficient of the treatment in the logistic regression of the observed outcome on the treatment indicator and covariates to estimate the causal effects. However, Freedman (2008c) showed that this logistic regression does not have nice properties under the potential outcomes framework. Even if the logistic model is correct, the coefficient estimates the conditional odds

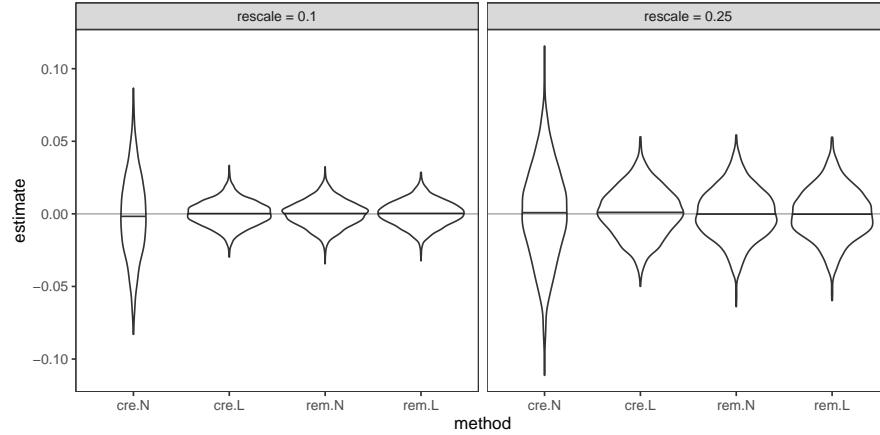


FIGURE 6.2: Simulation with 2000 Monte Carlo replicates and $a = 0.05$ for the ReM

ratio which may not be the parameter of interest; when the logistic model is incorrect, it is even harder to interpret the coefficient. From the discussion above, if the parameter of interest is the average causal effect, we can still use Lin (2013)'s estimator to analyze the binary outcome data in the CRE. Guo and Basse (2023) extend Lin (2013)'s theory to allow for using generalized linear models to construct estimators for the average causal effect under the potential outcomes framework.

Other extensions of Lin (2013)'s theory focus on high dimensional covariates. Bloniarz et al. (2016) focus on the regime with many covariates than the sample size, and under the sparsity assumption, they suggest replacing the OLS fits by the least absolute shrinkage and selection operator (LASSO) fits (Tibshirani, 1996) of the outcome on the treatment, covariates and their interactions. Lei and Ding (2021) focus on the regime with a diverging number of covariates without assuming sparsity, and under certain regularity conditions, they show that Lin (2013)'s estimator is still consistent and asymptotically Normal. Wager et al. (2016) propose to use machine learning methods to analyze high dimensional experimental data.

6.6 Homework Problems

6.1 FRT under ReM

Describe the FRT under ReM.

6.2 Invariance of the Mahalanobis Distance

Prove Lemma 6.1.

6.3 Bias of the difference-in-means estimator under rerandomization

Assume that we draw $\mathbf{Z} = (Z_1, \dots, Z_n)$ from a CRE and accept it if and only if $\phi(\mathbf{Z}, \mathbf{X}) = 1$, where ϕ is a predetermined balance criterion. Show that if $n_1 = n_0$ and

$$\phi(\mathbf{Z}, \mathbf{X}) = \phi(\mathbf{1}_n - \mathbf{Z}, \mathbf{X}), \quad (6.13)$$

then $\hat{\tau}$ is unbiased for τ . Verify that rerandomization using the Mahalanobis distance satisfies (6.13) if $n_1 = n_0$. Give a counterexample that $\hat{\tau}$ is biased for τ when these two conditions do not hold.

6.4 Equivalent form of R^2 in the CRE

Prove Proposition 6.1.

6.5 Lin's estimator for covariate adjustment

Prove Proposition 6.2.

6.6 Predictive and projective estimators

Prove (6.8) and (6.10).

6.7 Equivalent form of the covariate-adjusted estimator

Prove (6.11) and (6.12).

6.8 ANCOVA also adjusts for covariate imbalance

This problem gives a result for ANCOVA that is similar to (6.12).

Show that

$$\hat{\gamma}_{\mathbf{F}} = \hat{\tau} - \hat{\gamma}_{\mathbf{F}}^T \hat{\tau}_{\mathbf{X}},$$

where $\hat{\gamma}_{\mathbf{F}}$ is the coefficient of X_i in the OLS fit of Y_i on $(1, Z_i, X_i)$.

6.9 Regression adjustment / post-stratification of CRE

Prove Proposition 6.3.

Hint: Sometimes $\hat{\tau}_{PS}$ or $\hat{\tau}_L$ may not be well-defined. In those cases, we treat $\hat{\tau}_{PS}$ and $\hat{\tau}_L$ as equal. You can ignore this complexity in the proof.

6.10 More on the difference-in-difference estimator in the CRE

This problem gives more details for the difference-in-difference estimator in the CRE in Section 6.2.3.3.

Show that $\hat{\tau}(1, 1)$ is unbiased for τ , calculate its variance, and show that

$\hat{V}(1, 1)$ is a conservative estimator for the true variance of $\hat{\tau}(1, 1)$. When does $E\{\hat{V}(1, 1)\} = \text{var}\{\hat{\tau}(1, 1)\}$ hold?

Compare the variances of $\hat{\tau}(0, 0)$ and $\hat{\tau}(1, 1)$ to show that

$$\text{var}\{\hat{\tau}(0, 0)\} \geq \text{var}\{\hat{\tau}(1, 1)\}$$

if and only if

$$2\frac{n_0}{n}\beta_1 + 2\frac{n_1}{n}\beta_0 \geq 1,$$

where

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})\{Y_i(1) - \bar{Y}(1)\}}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \beta_0 = \frac{\sum_{i=1}^n (X_i - \bar{X})\{Y_i(0) - \bar{Y}(0)\}}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

are the coefficients of X_i in the OLS fits of $Y_i(1)$ and $Y_i(0)$ on $(1, X_i)$, respectively.

Remark: Gerber and Green (2012, page 28) discussed a special case of this problem with $n_1 = n_0$.

6.11 Data re-analyses

Re-analyze the data used in `SRE_Neyman_penn.R`. The analysis in Chapter 5 uses the treatment indicator, the outcome and the block indicator. Now we want to use all other covariates.

Conduct regression adjustments within strata of the experiment, and then combine these adjusted estimators to estimate the average causal effect. Report the point estimator, estimated standard error and 95% confidence interval. Compare them with those without regression adjustments.

6.12 Recommended reading

The title of this chapter is the same as that of Li and Ding (2020), which studied the roles of rerandomization and regression adjustment in the design and analysis stages of randomized experiments, respectively.

7

Matched-Pairs Experiment

The matched-pairs experiment (MPE) is the most extreme version of the SRE with only one treated unit and one control unit within each stratum. In this case, the strata are also called pairs. Although this type of experiment is a special case of the SRE discussed in Chapter 5, it has its own estimation and inference strategy. Moreover, it has many new features and it is closely related to the “matching” strategy in observational studies which will be covered in Chapter 15 later. So we discuss the MPE in this separate chapter.

7.1 Design of the experiment and potential outcomes

Consider an experiment with $2n$ units. If we have predictive covariates to the outcomes, we can pair units based on the similarity of covariates. With a scalar covariate, we can order units based on this covariate and then form pairs based on the adjacent units. With many covariates, we can define pairwise distances between units and then form pairs based on these distances. In this case, pair matching can be done using a greedy algorithm or an optimal nonbipartite matching algorithm. The greedy algorithm pairs the two units with the smallest distance, drop them from the pool of units, pair the two remaining units with the smallest distance, etc. The optimal nonbipartite matching algorithm divides the $2n$ units into n pairs of two units to minimize the sum of the within-pair distances. See Greevy et al. (2004) for more details of the computational aspect of the MPE. In this chapter, we assume that the pairs are formed based on the covariates, and discuss the subsequent design and analysis issues.

Let (i, j) index the unit j in pair i , where $i = 1, \dots, n$ and $j = 1, 2$. Unit (i, j) has potential outcomes $Y_{ij}(1)$ and $Y_{ij}(0)$ under the treatment and control, respectively. Within each pair, we randomly assign one unit to receive the treatment and the other to receive the control. Let

$$Z_i = \begin{cases} 1, & \text{if the first unit receives the treatment,} \\ 0, & \text{if the second unit receives the treatment.} \end{cases}$$

We can formally define MPE based on the treatment assignment mechanism.

Definition 7.1 (MPE) We have

$$(Z_i)_{i=1}^n \stackrel{IID}{\sim} \text{Bernoulli}(1/2). \quad (7.1)$$

The observed outcomes within pair i are

$$Y_{i1} = Z_i Y_{i1}(1) + (1 - Z_i) Y_{i1}(0) = \begin{cases} Y_{i1}(1), & \text{if } Z_i = 1; \\ Y_{i1}(0), & \text{if } Z_i = 0; \end{cases}$$

and

$$Y_{i2} = Z_i Y_{i2}(0) + (1 - Z_i) Y_{i2}(1) = \begin{cases} Y_{i2}(0), & \text{if } Z_i = 1; \\ Y_{i2}(1), & \text{if } Z_i = 0. \end{cases}$$

So the observed data are $(Z_i, Y_{i1}, Y_{i2})_{i=1}^n$.

7.2 FRT

Similar to the discussion before, we can always use the FRT to test the sharp null hypothesis:

$$H_{0F} : Y_{ij}(1) = Y_{ij}(0) \text{ for all } i = 1, \dots, n \text{ and } j = 1, 2.$$

When conducting the FRT, we need to simulate the distribution of (Z_i, \dots, Z_n) from (7.1). I will discuss some canonical choices of test statistics based on the within-pair differences between the treated and control outcomes:

$$\begin{aligned} \hat{\tau}_i &= \text{outcome under treatment} - \text{outcome under control (within pair } i) \\ &= (2Z_i - 1)(Y_{i1} - Y_{i2}) \\ &= S_i(Y_{i1} - Y_{i2}), \end{aligned}$$

where the $S_i = 2Z_i - 1$ are IID random signs with mean 0 and variance 1, for $i = 1, \dots, n$. Since the pairs with zero $\hat{\tau}_i$'s do not contribute to the randomization distribution, we drop those pairs in the discussion of the FRT.

Example 7.1 (paired t statistic) The average of the within-pair differences is

$$\hat{\tau} = n^{-1} \sum_{i=1}^n \hat{\tau}_i.$$

Under H_{0F} ,

$$E(\hat{\tau}) = 0$$

and

$$\text{var}(\hat{\tau}) = n^{-2} \sum_{i=1}^n \text{var}(\hat{\tau}_i) = n^{-2} \sum_{i=1}^n \text{var}(S_i)(Y_{i1} - Y_{i2})^2 = n^{-2} \sum_{i=1}^n \hat{\tau}_i^2.$$

Based on the CLT for the sum of independent random variables, we have the Normal approximation:

$$\frac{\hat{\tau}}{\sqrt{n^{-2} \sum_{i=1}^n \hat{\tau}_i^2}} \xrightarrow{d} N(0, 1).$$

We can use this Normal approximation to construct an asymptotic test. Many standard test books suggest using the following paired t statistic in the MPE:

$$t_{\text{pair}} = \frac{\hat{\tau}}{\sqrt{\{n(n-1)\}^{-1} \sum_{i=1}^n (\hat{\tau}_i - \hat{\tau})^2}},$$

which is almost identical to $\hat{\tau}$ with large n and small $\hat{\tau}$ under H_{0F} .

In classic statistics, the motivation for using t_{pair} is under a different framework. When $\hat{\tau}_i \stackrel{\text{IID}}{\sim} N(0, \sigma^2)$, we can show that $t_{\text{pair}} \sim t(n-1)$, i.e., the exact distribution of t_{pair} is t with degrees of freedom $n-1$, which is close to $N(0, 1)$ with a large n . The R function `t.test` with `paired=TRUE` can implement this test. With a large n , these procedures give similar results. The discussion in Example 7.1 gives another justification of the classic paired t test without assuming the Normality of the data.

Example 7.2 (Wilcoxon sign-rank statistic) Based on the ranks (R_1, \dots, R_n) of $(|\hat{\tau}_1|, \dots, |\hat{\tau}_n|)$, we can define a test statistic

$$W = \sum_{i=1}^n I(\hat{\tau}_i > 0) R_i.$$

Under H_{0F} ,

$$E(W) = \frac{1}{2} \sum_{i=1}^n R_i = \frac{1}{2} \sum_{i=1}^n i = \frac{n(n+1)}{4}$$

and

$$\text{var}(W) = \frac{1}{4} \sum_{i=1}^n R_i^2 = \frac{1}{4} \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{24}.$$

The CLT for the sum of independent random variables ensures the following Normal approximation:

$$\frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \xrightarrow{d} N(0, 1).$$

We can use this Normal approximation to construct an asymptotic test. The R function `wilcox.test` with `paired=TRUE` can implement these tests.

Example 7.3 (Kolmogorov–Smirnov-type statistic) Under H_{0F} , the absolute values ($|\hat{\tau}_1|, \dots, |\hat{\tau}_n|$) are fixed but their signs are random. So $(\hat{\tau}_1, \dots, \hat{\tau}_n)$ and $-(\hat{\tau}_1, \dots, \hat{\tau}_n)$ should have the same distribution. Let

$$\hat{F}(t) = n^{-1} \sum_{i=1}^n I(\hat{\tau}_i \leq t)$$

be the empirical distribution of $(\hat{\tau}_1, \dots, \hat{\tau}_n)$, and

$$1 - \hat{F}(-t-) = n^{-1} \sum_{i=1}^n I(-\hat{\tau}_i \leq t)$$

be the empirical distribution of $-(\hat{\tau}_1, \dots, \hat{\tau}_n)$, where $\hat{F}(-t-)$ is the left limit of the function $\hat{F}(\cdot)$ at $-t$. A Kolmogorov–Smirnov-type statistic is then

$$D = \max_t |\hat{F}(t) + \hat{F}(-t-) - 1|.$$

Butler (1969) proposed this test statistic and derived its exact and asymptotic distributions. Unfortunately, this is not implemented in standard software packages. Nevertheless, we can simulate its exact distribution and compute the *p*-value based on the FRT.¹

Example 7.4 (sign statistic) The sign statistic uses only the signs of the within-pair differences

$$\Delta = \sum_{i=1}^n I(\hat{\tau}_i > 0).$$

Under H_{0F} ,

$$I(\hat{\tau}_i > 0) \stackrel{IID}{\sim} Bernoulli(1/2)$$

and therefore

$$\Delta \sim Binomial(n, 1/2).$$

Based on this we have an exact Binomial test, which is implemented in the R function `binom.test` with `p=1/2`. Using the CLT, we can also conduct a test based on the following Normal approximation of the Binomial distribution:

$$\frac{\Delta - n/2}{\sqrt{n/4}} \xrightarrow{d} N(0, 1).$$

¹Butler (1969)'s proposed this test statistic under a slightly different framework. Given IID draws of $(\hat{\tau}_1, \dots, \hat{\tau}_n)$ from a distribution $F(y)$, if they are symmetrically distributed around 0, then

$$F(t) = \text{pr}(\hat{\tau}_i \leq t) = \text{pr}(-\hat{\tau}_i \leq t) = 1 - \text{pr}(\hat{\tau}_i < -t) = 1 - F(-t-).$$

Therefore, $\hat{F}(t) + \hat{F}(-t-) - 1$ measures the deviation from the null hypothesis of symmetry, which motivates the definition of D . A naive definition of the Kolmogorov–Smirnov-type statistic is to compare the empirical distributions of the outcomes under treatment and control as in Example 3.4. Using that definition, we effectively break the pairs. Although it can still be used in the FRT for the MPE, it does not capture the matched-pairs structure of the experiment.

TABLE 7.1: Counts of four types of pairs

	control outcome 1	control outcome 0
treated outcome 1	m_{11}	m_{10}
treated outcome 0	m_{01}	m_{00}

Example 7.5 (McNemar's statistic for a binary outcome) If the outcome is binary, we can summarize the data from the MPE in a more compact way. Given a pair, the treated outcome can be either 1 or 0 and the control outcome can be either 1 or 0, yielding a 2×2 table as in Table 7.1.

Under H_{0F} , the numbers of concordant pairs m_{11} and m_{00} are fixed, and $m_{10} + m_{01}$ is also fixed. So the only random component is m_{10} which has distribution

$$m_{10} \sim \text{Binomial}(m_{10} + m_{01}, 1/2).$$

This implies an exact test based on the Binomial distribution. The R function `mcnemar.test` gives an asymptotic test based on the Normal approximation of the Binomial distribution:

$$\frac{m_{10} - (m_{10} + m_{01})/2}{\sqrt{(m_{10} + m_{01})/4}} = \frac{m_{10} - m_{01}}{\sqrt{m_{10} + m_{01}}} \xrightarrow{d} N(0, 1).$$

Both the exact FRT and the asymptotic test do not depend on m_{11} or m_{00} . Only the numbers of discordant pairs matter in these tests.

7.3 Neymanian inference

The average causal effect within pair i is

$$\tau_i = \frac{1}{2} \{Y_{i1}(1) + Y_{i2}(1) - Y_{i1}(0) - Y_{i2}(0)\},$$

and the average causal effect for all units is

$$\tau = n^{-1} \sum_{i=1}^n \tau_i = (2n)^{-1} \sum_{i=1}^n \sum_{j=1}^2 \{Y_{ij}(1) - Y_{ij}(0)\}.$$

It is intuitive that $\hat{\tau}_i$ is unbiased for τ_i , so $\hat{\tau}$ is unbiased for τ . We can also calculate the variance of $\hat{\tau}$. I relegate the exact formula to a homework problem because the MPE is just a special case of the SRE.

However, we cannot follow the strategy of a SRE to estimate the variance of $\hat{\tau}$. The within-pair sample variances of the outcomes are not well defined because within each pair we have only one treated and one control unit. The data do not allow us to estimate the variance of $\hat{\tau}_i$ within pair i .

Is it possible to estimate the variance of $\hat{\tau}$ in the MPE? Let us forget about the MPE and change the perspective to the classic IID sampling. If the $\hat{\tau}_i$'s are IID with mean μ and σ^2 , then the variance of $\hat{\tau} = n^{-1} \sum_{i=1}^n \hat{\tau}_i$ is σ^2/n . An unbiased estimator for σ^2 is the sample variance $(n-1)^{-1} \sum_{i=1}^n (\hat{\tau}_i - \hat{\tau})^2$, so an unbiased estimator for $\text{var}(\hat{\tau})$ is

$$\hat{V} = \{n(n-1)\}^{-1} \sum_{i=1}^n (\hat{\tau}_i - \hat{\tau})^2.$$

The discussion also extends to the independent but not IID setting; see Problem A1.1 in Chapter A1. The above discussion seems a digression from the MPE which has completely different statistical assumptions. But at least it motivates a variance estimator \hat{V} , which uses the between-pair variance of $\hat{\tau}_i$ to estimate variance of $\hat{\tau}$. Of course, it is derived under different assumptions. Does it work for the MPE? Theorem 7.1 below is a positive result.

Theorem 7.1 *Under the MPE, \hat{V} is a conservative estimator for the true variance of $\hat{\tau}$:*

$$E(\hat{V}) - \text{var}(\hat{\tau}) = \{n(n-1)\}^{-1} \sum_{i=1}^n (\tau_i - \tau)^2 \geq 0.$$

If the τ_i 's are constant across pairs, then $E(\hat{V}) = \text{var}(\hat{\tau})$.

Theorem 7.1 states that under the MPE, \hat{V} is a conservative variance estimator in general and becomes unbiased if the average causal effects are constant across pairs. It is somewhat surprising because \hat{V} depends on the between-pair variance of the $\hat{\tau}_i$'s whereas $\text{var}(\hat{\tau})$ depends on the within-pair variance of each of $\hat{\tau}_i$. The proof below might provide some insights for this surprisingly result.

Proof of Theorem 7.1: Using the basic algebraic fact that $\sum_{i=1}^n (a_i - \bar{a})^2 = \sum_{i=1}^n a_i^2 - n\bar{a}^2$ in the following steps 2 and 5, we have

$$\begin{aligned} n(n-1)E(\hat{V}) &= E \left\{ \sum_{i=1}^n (\hat{\tau}_i - \hat{\tau})^2 \right\} \\ &= E \left(\sum_{i=1}^n \hat{\tau}_i^2 - n\hat{\tau}^2 \right) \\ &= \sum_{i=1}^n \{\text{var}(\hat{\tau}_i) + \tau_i^2\} - n\{\text{var}(\hat{\tau}) + \tau^2\} \\ &= \sum_{i=1}^n \text{var}(\hat{\tau}_i) - n\text{var}(\hat{\tau}) + \sum_{i=1}^n \tau_i^2 - n\tau^2 \\ &= n^2\text{var}(\hat{\tau}) - n\text{var}(\hat{\tau}) + \sum_{i=1}^n (\tau_i - \tau)^2. \end{aligned}$$

Therefore,

$$E(\hat{V}) = \text{var}(\hat{\tau}) + \{n(n-1)\}^{-1} \sum_{i=1}^n (\tau_i - \tau)^2 \geq \text{var}(\hat{\tau}).$$

□

Similar to the discussions for other experiments, the Neymanian approach relies on the large-sample approximation:

$$\frac{\hat{\tau} - \tau}{\sqrt{\text{var}(\hat{\tau})}} \rightarrow N(0, 1)$$

in distribution if $n \rightarrow \infty$ and some regularity conditions hold. Due to the over estimation of the variance, the Wald-type confidence interval

$$\hat{\tau} \pm z_{1-\alpha/2} \sqrt{\hat{V}}$$

covers τ with probability at least $1 - \alpha$.

Both the point estimator $\hat{\tau}$ and the variance estimator \hat{V} can be conveniently obtained by OLS, as shown in the proposition below.

Proposition 7.1 *$\hat{\tau}$ and \hat{V} are identical to the coefficient and variance estimator of the intercept from the OLS fit of the vector $(\hat{\tau}_1, \dots, \hat{\tau}_n)^\top$ on the intercept only.*

I leave the proof of Proposition 7.1 as Problem 7.3.

7.4 Covariate adjustment

7.4.1 FRT

Similar to the discussion in the CRE, there are two general strategies of covariate adjustment in the MPE. First, we can construct test statistics based on the residuals from a model fitting of the outcome on the covariates, since those residuals are fixed numbers under the sharp null hypothesis. A canonical choice is to fit OLS of all observed Y_{ij} 's on X_{ij} 's to obtain the residuals $\hat{\varepsilon}_{ij}$'s. We can then construct test statistics pretending that the $\hat{\varepsilon}_{ij}$'s are the observed outcomes. Rosenbaum (2002a) advocated this strategy in particular to the MPE.

Second, we can directly use some coefficients from model fitting as the test statistics. The discussion in the next subsection will suggest a choice of the test statistic for the second strategy.

7.4.2 Regression adjustment

Although we have matched on covariates in the design stage, it is possible that the matching is not perfect and sometimes we have additional covariates beyond those used in the pair-matching stage. In those cases, we can adjust for the covariates to further improve estimation efficiency. Assume that each unit has covariates X_{ij} , and we can compute the within-pair differences in covariates $\hat{\tau}_{X,i}$ and their average $\hat{\tau}_X$ in the same way as the outcome. We can show that

$$E(\hat{\tau}_{X,i}) = 0, \quad E(\hat{\tau}_X) = 0,$$

and

$$\text{cov}(\hat{\tau}_X) = n^{-2} \sum_{i=1}^n \hat{\tau}_{X,i} \hat{\tau}_{X,i}^\top.$$

In a realized MPE, $\text{cov}(\hat{\tau}_X)$ is not zero unless all the $\hat{\tau}_{X,i}$'s are zero. With an unlucky draw of (Z_1, \dots, Z_n) , it is possible that $\hat{\tau}_X$ differs substantially from zero. Similar to the discussion in the CRE, adjusting for the imbalance of the covariate means is likely to improve estimation efficiency.

Consider a class of estimators indexed by γ :

$$\hat{\tau}(\gamma) = \hat{\tau} - \gamma^\top \hat{\tau}_X$$

which has mean 0 for any fixed γ . We want to choose γ to minimize the variance of $\hat{\tau}(\gamma)$. Its variance is a quadratic function of γ :

$$\text{var}\{\hat{\tau}(\gamma)\} = \text{var}(\hat{\tau} - \gamma^\top \hat{\tau}_X) = \text{var}(\hat{\tau}) + \gamma^\top \text{cov}(\hat{\tau}_X) \gamma - 2\gamma^\top \text{cov}(\hat{\tau}_X, \hat{\tau}),$$

which is minimized at

$$\tilde{\gamma} = \text{cov}(\hat{\tau}_X)^{-1} \text{cov}(\hat{\tau}_X, \hat{\tau}).$$

We have obtained the formula of $\text{cov}(\hat{\tau}_X)$ in the above, which can also be written as

$$\text{cov}(\hat{\tau}_X) = n^{-2} \sum_{i=1}^n |\hat{\tau}_{X,i}| |\hat{\tau}_{X,i}|^\top,$$

where $|\cdot|$ denotes component-wise absolute value of a vector. So $\text{cov}(\hat{\tau}_X)$ is fixed and known from the observed data. However, $\text{cov}(\hat{\tau}_X, \hat{\tau})$ depends on unknown potential outcomes. Fortunately, we can obtain an unbiased estimator for it, as shown in Theorem 7.2 below.

Theorem 7.2 *An unbiased estimator for $\text{cov}(\hat{\tau}_X, \hat{\tau})$ is*

$$\hat{\theta} = \{n(n-1)\}^{-1} \sum_{i=1}^n (\hat{\tau}_{X,i} - \hat{\tau}_X)(\hat{\tau}_i - \hat{\tau}).$$

The proof of Theorem 7.2 is similar to that of Theorem 7.1. I leave it to Problem 7.2.

Therefore, we can estimate the optimal coefficient $\tilde{\gamma}$ by

$$\begin{aligned}\hat{\gamma} &= \left(n^{-2} \sum_{i=1}^n \hat{\tau}_{X,i} \hat{\tau}_{X,i}^\top \right)^{-1} \left\{ \{n(n-1)\}^{-1} \sum_{i=1}^n (\hat{\tau}_{X,i} - \hat{\tau}_X)(\hat{\tau}_i - \hat{\tau}) \right\} \\ &\approx \left(\sum_{i=1}^n (\hat{\tau}_{X,i} - \hat{\tau}_X)(\hat{\tau}_{X,i} - \hat{\tau}_X)^\top \right)^{-1} \sum_{i=1}^n (\hat{\tau}_{X,i} - \hat{\tau}_X)(\hat{\tau}_i - \hat{\tau}),\end{aligned}$$

which is approximately the coefficient of the $\hat{\tau}_{X,i}$'s in the OLS fit of the $\hat{\tau}_i$'s on the $\hat{\tau}_{X,i}$'s with an intercept. The final estimator is

$$\hat{\tau}_{\text{adj}} = \hat{\tau}(\hat{\gamma}) = \hat{\tau} - \hat{\gamma}^\top \hat{\tau}_X,$$

which, by the property of OLS, is approximately the intercept in the OLS fit of the $\hat{\tau}_i$'s on the $\hat{\tau}_{X,i}$'s with an intercept.

A conservative variance estimator for $\hat{\tau}_{\text{adj}}$ is then

$$\hat{V}_{\text{adj}} = \hat{V} + \hat{\gamma}^\top \text{cov}(\hat{\tau}_X) \hat{\gamma} - 2\hat{\gamma}^\top \hat{\theta} = \hat{V} - \hat{\theta}^\top \text{cov}(\hat{\tau}_X)^{-1} \hat{\theta}.$$

A subtle technical issue is whether $\hat{\tau}(\hat{\gamma})$ has the same optimality as $\hat{\tau}(\tilde{\gamma})$. With large samples, we can show $\hat{\tau}(\hat{\gamma}) - \hat{\tau}(\tilde{\gamma}) = -(\hat{\gamma} - \tilde{\gamma})^\top \hat{\tau}_X$ is of higher order since it is the product of two “small” terms $\hat{\gamma} - \tilde{\gamma}$ and $\hat{\tau}_X$. I omit the tedious details for asymptotic analysis, but hope the result makes some intuitive sense to the readers.

Moreover, Fogarty (2018b) discussed the asymptotically equivalent regression formulation of the above covariate-adjusted procedure, and gave a rigorous proof for associated CLT. I summarize the regression formulation below without giving the regularity conditions.

Proposition 7.2 *Under the MPE, the covariate-adjusted estimator $\hat{\tau}_{\text{adj}}$ and the associated variance estimator \hat{V}_{adj} can be conveniently approximated by the intercept and the associated variance estimator from the OLS fit of the vector of the $\hat{\tau}_i$'s on the 1's and the matrix of the $\hat{\tau}_{X,i}$'s.*

I leave the proof of Proposition 7.2 as Problem 7.3. Interestingly, neither Proposition 7.1 nor 7.2 requires the EHW correction of the variance estimator. Because we reduce the data from the MRE to the within-pair differences, it is unnecessary to center the covariates unlike in Lin (2013)'s estimator for the CRE.

7.5 Examples

7.5.1 Darwin's data comparing cross-fertilizing and self-fertilizing on the height of corns

This is a classical example from Fisher (1935). It contains 15 pairs of corns with either cross-fertilizing or self-fertilizing, with the height being the outcome. The R package `HistData` provides the original data, where `cross` and `self` are the heights under cross-fertilizing and self-fertilizing, respectively, and `diff` denotes their difference.

```
> library("HistData")
> ZeaMays
   pair  pot  cross    self    diff
1      1    1 23.500 17.375  6.125
2      2    1 12.000 20.375 -8.375
3      3    1 21.000 20.000  1.000
4      4    2 22.000 20.000  2.000
5      5    2 19.125 18.375  0.750
6      6    2 21.500 18.625  2.875
7      7    3 22.125 18.625  3.500
8      8    3 20.375 15.250  5.125
9      9    3 18.250 16.500  1.750
10     10   3 21.625 18.000  3.625
11     11   3 23.250 16.250  7.000
12     12   4 21.000 18.000  3.000
13     13   4 22.125 12.750  9.375
14     14   4 23.000 15.500  7.500
15     15   4 12.000 18.000 -6.000
```

In total, the MPE has $2^{15} = 32768$ possible treatment assignment which is a tractable number in R. The following function can enumerate all possible treatment assignment for the MPE:

```
MP_enumerate = function(i, n.pairs)
{
  if(i > 2^n.pairs) print("i is too large.")
  a = 2^((n.pairs-1):0)
  b = 2*a
  2*sapply(i-1,
            function(x)
              as.integer((x %% b)>=a)) - 1
}
```

So we enumerate all the treatment assignments, and calculate the corresponding $\hat{\tau}$'s and the one-sided exact p -value.

```
> difference = ZeaMays$diff
> n.pairs    = length(difference)
```

```

> abs.diff = abs(difference)
> t.obs = mean(difference)
> t.ran = sapply(1:2^15,
+                 function(x){
+                   sum(MP_enumerate(x, 15)*abs.diff)
+                 })/n.pairs
> pvalue = mean(t.ran>=t.obs)
> pvalue
[1] 0.02633667

```

Figure 7.1 shows the exact randomization of $\hat{\tau}$.

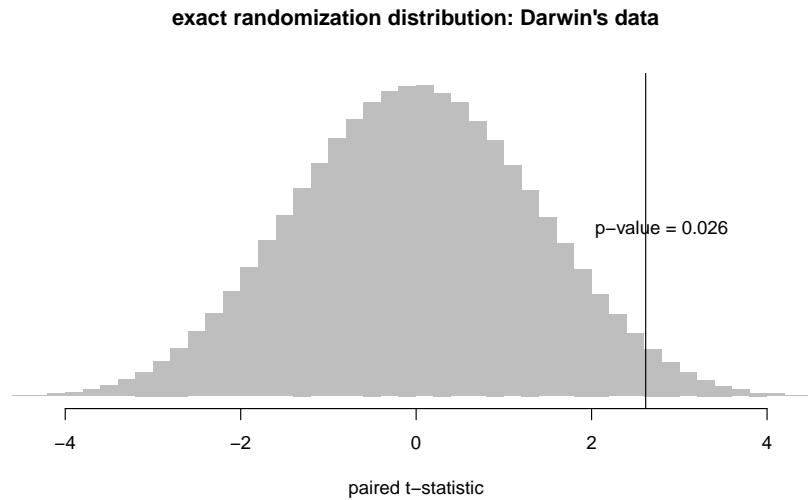


FIGURE 7.1: Randomization distribution of $\hat{\tau}$ using Darwin's data

7.5.2 Children's television workshop experiment data

I also re-analyze the data from Ball et al. (1973) which was also analyzed by Imbens and Rubin (2015). It contains 8 pairs, and the following table summarizes the within-pair covariate and outcome, as well as their differences:

	x.control	x.treatment	y.control	y.treatment	diffx	diffy
1	12.9	12.0	54.6	60.6	-0.9	6.0
2	15.1	12.3	56.5	55.5	-2.8	-1.0
3	16.8	17.2	75.2	84.8	0.4	9.6
4	15.8	18.9	75.6	101.9	3.1	26.3
5	13.9	15.3	55.3	70.6	1.4	15.3
6	14.5	16.6	59.3	78.4	2.1	19.1

7	17.0	16.0	87.0	84.2	-1.0	-2.8
8	15.8	20.1	73.7	108.6	4.3	34.9

We can use the OLS to obtain the point estimators and standard errors: without adjusting for covariates, we have

```
> unadj = summary(lm(diffy ~ 1, data = dataxy))$coef
> round(unadj, 3)
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.425     4.636   2.896   0.023
```

with adjusting for covariates, we have

```
> adj = summary(lm(diffy ~ diffx, data = dataxy))$coef
> round(adj, 3)
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.994     1.410   6.381   0.001
diffx       5.371     0.599   8.964   0.000
```

The above results assume large n , and p -values are justified if we believe the large- n approximation. However, $n = 8$ is not large. In total, we have $2^8 = 256$ possible treatment assignments, so the smallest possible p -value is $1/256 = 0.0039$, which is much larger than the p -value based on the Normal approximation of the covariate-adjusted estimator. In this example, it will be more reasonable to use the FRT with the studentized statistic (i. e., the `t value` from the `lm` function) to calculate exact p -values. Figure 7.2 shows the exact distributions of the two studentized statistic, as well as the two-sided p -values. The figure highlights the fact that the randomization distribution of the test statistics are discrete, taking at most 256 possible values. The Normal approximations are unlikely to be accurate especially at the tails. We should report the p -values based on the FRT.

7.6 Comparing the MPE and CRE

Imai (2008b) compared the MPE and CRE. Heuristically, the conclusion is that the MPE gives more precise estimators if the matching is well done and the covariates are predictive to the outcome. However, without the outcome data in the design stage, it is hard to decide whether this holds. In the FRT, if covariates are predictive to the outcome, the MPE usually gives more powerful tests compared to the CRE. Greevy et al. (2004) illustrated this using simulation based on the Wilcoxon sign rank statistic. However, this can be a subtle issue with finite samples. Consider an experiment with $2n$ units, with n units receiving the treatment and n units receiving the control. If we test the sharp null hypothesis at level 0.05, then in the MPE, we need at least $2 \times 5 = 10$ units since the smallest p -value is $1/2^5 = 1/32 < 0.05$ but $1/2^4 = 1/16 > 0.05$, but in the CRE, we need at least $2 \times 4 = 8$ units since the smallest p -value is

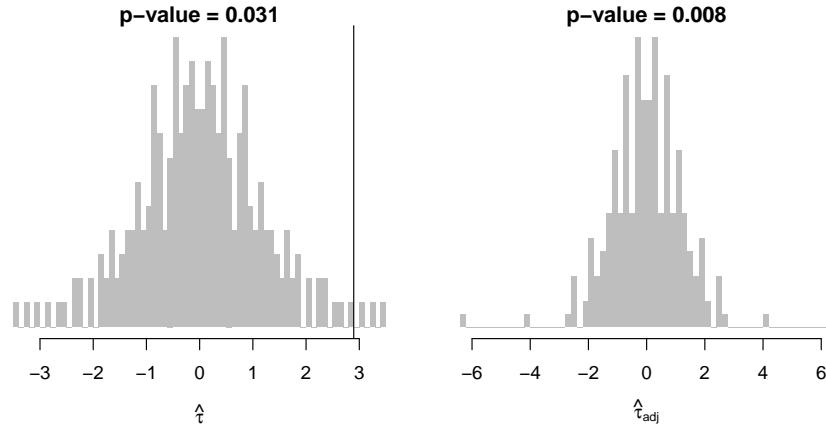


FIGURE 7.2: Randomization distributions of the studentized statistics in Section 7.5.2

$1/(8) = 1/70 < 0.05$ but $1/(6) = 1/20 = 0.05$. So with 8 units, it is impossible to reject the sharp null hypothesis in the MPE but it is possible in the CRE. Even if the covariates are perfect predictors of the outcome, the MPE is not superior to the CRE based on the FRT.

7.7 Extension to the general matched experiment

It is straightforward to extend the MPE to the general matched experiment with varying numbers of control units. Assume that we have n matched sets indexed by $i = 1, \dots, n$. For matched set i , we have $1 + M_i$ units. The M_i 's can vary. The total number of experimental units is $N = n + \sum_{i=1}^n M_i$. Let ij index the unit j within matched set i ($i = 1, \dots, n$ and $j = 1, \dots, M_i + 1$). Unit ij has potential outcomes $Y_{ij}(1)$ and $Y_{ij}(0)$ under the treatment and control, respectively.

Within matched set i ($i = 1, \dots, n$), the experimenter randomly selects exactly one unit to receive the treatment with the rest M_i units receiving the control. This general matched experiment is also a special case of the SRE with n strata of size $1 + M_i$ ($i = 1, \dots, n$). Let Z_{ij} be the treatment indicator for unit ij , which reveals one of the potential outcomes as

$$Y_{ij} = Z_{ij} Y_{ij}(1) + (1 - Z_{ij}) Y_{ij}(0).$$

The average causal effect within matched set i equals

$$\tau_i = (M_i + 1)^{-1} \sum_{j=1}^{M_i+1} \{Y_{ij}(1) - Y_{ij}(0)\}.$$

Since it is a SRE, an unbiased estimator of τ_i is

$$\hat{\tau}_i = \sum_{j=1}^{M_i+1} Z_{ij} Y_{ij} - M_i^{-1} \sum_{i=1}^n (1 - Z_{ij}) Y_{ij}$$

which is the difference in means of the outcomes within matched set i .

Below we discuss the statistical inference with the general matched experiment.

7.7.1 FRT

As usual, we can always use the FRT to test the sharp null hypothesis

$$H_{0F} : Y_{ij}(1) = Y_{ij}(0) \text{ for all } i = 1, \dots, n; j = 1, \dots, M_i + 1.$$

Because the general matched experiment is a special case of the SRE with many small strata, we can use the test statistics defined in Examples 5.4, 5.5, 7.2, 7.3, 7.4, as well as the estimators and the corresponding t -statistics from the following two subsections.

7.7.2 Estimating the average of the within-strata effects

We first focus on estimating the average of the within-strata effects:

$$\tau = n^{-1} \sum_{i=1}^n \tau_i.$$

It has an unbiased estimator

$$\hat{\tau} = n^{-1} \sum_{i=1}^n \hat{\tau}_i.$$

Interestingly, we can show that Theorem 7.1 holds for the general matched experiment, so are other results for the MPE. In particular, we can use the OLS fit of the $\hat{\tau}_i$'s on the intercept to obtain the point and variance estimators for τ . With covariates, we can use the OLS fit of the $\hat{\tau}_i$'s on the intercept and the $\hat{\tau}_{X,i}$'s, where

$$\hat{\tau}_{X,i} = \sum_{j=1}^{M_i+1} Z_{ij} X_{ij} - M_i^{-1} \sum_{i=1}^n (1 - Z_{ij}) X_{ij}$$

is the corresponding difference in means of the covariates within matched set i .

7.7.3 A more general causal estimand

Importantly, the τ above is the average of the τ_i 's, which does not equal the average causal effect for the N units in the experiment when the M_i 's vary. The average causal effect equals

$$\tau' = N^{-1} \sum_{i=1}^n \sum_{j=1}^{1+M_i} \{Y_{ij}(1) - Y_{ij}(0)\} = \sum_{i=1}^n \frac{1+M_i}{N} \tau_i.$$

To unify the discussion, I consider the weighted causal effect

$$\tau_w = \sum_{i=1}^n w_i \tau_i$$

with $\sum_{i=1}^n w_i = 1$, which includes τ as a special case with $w_i = n^{-1}$ and τ' as a special case with $w_i = (1 + M_i)/N$ for $i = 1, \dots, n$. It is straightforward to obtain an unbiased estimator

$$\hat{\tau}_w = \sum_{i=1}^n w_i \hat{\tau}_i,$$

and calculate its variance

$$\text{var}(\hat{\tau}_w) = \sum_{i=1}^n w_i^2 \text{var}(\hat{\tau}_i).$$

However, estimating the variance of this estimator is quite tricky because the $\hat{\tau}_i$'s are independent random variable without any replicates. This is a famous problem in theoretical statistics studied by Hartley et al. (1969) and Rao (1970). Fogarty (2018a) also discussed this problem without recognizing these previous works. I will give the final form of the variance estimator without detailing the motivation:

$$\hat{V}_w = \sum_{i=1}^n c_i (\hat{\tau}_i - \hat{\tau}_w)^2$$

where

$$c_i = \frac{\frac{w_i^2}{1-2w_i}}{1 + \sum_{i=1}^n \frac{w_i^2}{1-2w_i}}.$$

As a sanity check, c_i reduces to $\{n(n-1)\}^{-1}$ in the MPE with $M_i = 1$ and $w_i = n^{-1}$. For simplicity, we focus on the case with $w_i < 1/2$ for all i 's, that is, there is no matched set containing more than half of the total weights. The following theorem extends Theorem 7.1.

Theorem 7.3 Under the general matched experiment with varying M_i 's, we have

$$E(\hat{V}_w) - \text{var}(\hat{\tau}_w) = \sum_{i=1}^n c_i(\tau_i - \tau_w)^2 \geq \text{var}(\hat{\tau}_w) \geq 0$$

with equality holding if the τ_i 's are constant.

Although the theoretical motivation for \hat{V}_w is quite complicated, it is not too difficult to verify Theorem 7.3 directly. I relegate the proof to Problem 7.9.

7.8 Homework Problems

7.1 The true variance of $\hat{\tau}$ in the MPE

Express $\text{var}(\hat{\tau})$ in terms of the first two finite-population moments potential outcomes.

7.2 A covariance estimator

Prove Theorem 7.2.

7.3 Variance estimators via OLS

Prove Propositions 7.1 and 7.2.

7.4 Point and variance estimator with binary outcome

This problem extends Example 7.5 to Neymanian inference.

Express $\hat{\tau}$ and \hat{V} in terms of the counts in Table 7.1.

7.5 Minimum sample size for the FRT

Extend the discussion in Section 7.6. Consider an experiment with $2n$ units, with n units receiving the treatment and n units receiving the control, and test the sharp null hypothesis at level 0.001. What is the minimum value of n for an MPE so that the smallest p -value does not exceed than 0.001, and what is the correponding minimum value of n for a CRE.

7.6 Re-analyzing Darwin's data

In `MPE_FRT_darwin.R`, I analyze Darwin's data using the FRT based on the test statistic $\hat{\tau}$.

Re-analyze this dataset using the FRT with the Wilcoxon signed rank sum statistic.

Re-analyze this dataset based on the Neymanian inference: unbiased point estimator, conservative variance estimator, 95% confidence interval.

7.7 Re-analyzing children's television workshop experiment data

In `MPE_Neyman_star.R`, I analyze the data from based on Neymanian inference.

Re-analyze this dataset using the FRT with different test statistics.

Re-analyze this dataset using the FRT with covariate adjustment, e.g., you can define test statistics based on residuals from the OLS fit of the observed outcome on covariates. Will the conclusion change if you do not include an intercept in your OLS fit?

7.8 Re-analyzing Angrist and Lavy (2009)'s data

The original analysis was quite complicated. For this problem, please focus only on Table A1 of the original paper viewing the schools as experimental units. Angrist and Lavy (2009) essentially conducted an MPE on the schools. Dropping pair 6 and all the pairs with noncompliance results in 14 complete pairs, with data shown below and also in `AL2009.csv`:

	pair	z	pr99	pr00	pr01	pr02
1	1	0	0.046	0.000	0.091	0.185
2	1	1	0.036	0.051	0.000	0.047
3	2	0	0.054	0.094	0.184	0.034
4	2	1	0.050	0.108	0.110	0.095
5	3	0	0.114	0.000	0.056	0.075
6	3	1	0.098	0.054	0.030	0.068
7	4	0	0.148	0.162	0.082	0.075
8	4	1	0.134	0.390	0.339	0.458
9	5	0	0.152	0.105	0.083	0.129
10	5	1	0.145	0.077	0.579	0.167
11	6	0	0.188	0.214	0.375	0.545
12	6	1	0.179	0.165	0.483	0.444
13	7	0	0.193	0.771	0.328	0.583
14	7	1	0.189	0.186	0.168	0.368
15	8	0	0.197	0.350	0.000	0.383
16	8	1	0.200	0.071	0.667	0.429
17	9	0	0.213	0.176	0.164	0.172
18	9	1	0.209	0.165	0.092	0.151
19	10	0	0.211	0.667	0.250	0.617
20	10	1	0.219	0.250	0.500	0.350
21	11	0	0.219	0.153	0.185	0.219
22	11	1	0.224	0.363	0.372	0.342
23	12	0	0.255	0.226	0.213	0.327
24	12	1	0.257	0.098	0.107	0.095
25	13	0	0.261	0.071	0.000	NA
26	13	1	0.263	0.441	0.448	0.435
27	14	0	0.286	0.161	0.126	0.181
28	14	1	0.285	0.389	0.353	0.309

The outcomes are the Bagrut passing rates in years 2001 and 2002, with the Bagrut passing rates in 1999 and 2000 as pretreatment covariates. Re-analyze the data based on the Neymanian inference with and without covariates. In particular, how do you deal with the missing outcome in pair 25?

7.9 Variance estimation in the general matched experiment

This problem contains more details for Section 7.7.

First, prove Theorem 7.1 for the general matched experiment.

Second, prove Theorem 7.3.

Hint: For the second part, we need to first verify that $\hat{\tau}_i - \hat{\tau}_w$ has mean $\tau_i - \tau_w$ and variance

$$\text{var}(\hat{\tau}_i - \hat{\tau}_w) = \text{var}(\hat{\tau}_w) + (1 - 2w_i)\text{var}(\hat{\tau}_i).$$

7.10 Recommended readings

Greevy et al. (2004) provided an algorithm to form matched pairs based on covariates. Imai (2008b) discussed estimation of the average causal effect without covariates, and Fogarty (2018b) discussed covariate adjustment in MPEs.

8

Unification of the Fisherian and Neymanian Inferences in Randomized Experiments

Previous chapters cover both the Fisherian and Neymanian inferences for different types of experiments. The Fisherian perspective focuses on the finite-sample exact p -value for testing the strong null hypothesis of no causal effects for any units whatsoever, and the Neymanian perspective focuses on unbiased estimation with a conservative large-sample confidence interval for the average causal effect. Both of them are justified by the physical randomization of the experiments. They are the two important forms of design-based or randomization-based inference for causal effects. They are related but also have distinct features.

In 1935, Neyman presented his seminal paper on randomization-based inference to the Royal Statistical Society. His paper (Neyman, 1935) was attacked by Fisher in the discussion session. Sabbaghi and Rubin (2014) reviewed this famous Neyman–Fisher controversy and presented some new results for this old problem. Instead of going to philosophical issues, this chapter provides a unified discussion.

8.1 Testing strong and weak null hypotheses in the CRE

Let us revisit the treatment-control CRE. The Fisherian perspective focuses on testing the strong null hypothesis

$$H_{0F} : Y_i(1) = Y_i(0) \text{ for all units } i = 1, \dots, n.$$

The FRT delivers a finite-sample exact p_{FRT} .

By duality of the confidence interval and hypothesis testing, the Neymanian perspective gives a test for the weak null hypothesis

$$H_{0N} : \tau = 0 \iff H_{0N} : \bar{Y}(1) = \bar{Y}(0)$$

based on

$$t = \frac{\hat{\tau}}{\sqrt{\hat{V}}} = \sqrt{\frac{\text{var}(\hat{\tau})}{\hat{V}}} \times \frac{\hat{\tau}}{\sqrt{\text{var}(\hat{\tau})}} \xrightarrow{d} C \times N(0, 1),$$

with $C \leq 1$. Using $N(0, 1)$ quantiles for the studentized statistic t , we have a conservative large-sample test for H_{0N} .

Furthermore, Ding and Dasgupta (2017) show that the FRT with the studentized statistic t has the dual guarantees:

1. the associate p_{FRT} is finite-sample exact under H_{0F} ;
2. it is asymptotically conservative under H_{0N} .

Importantly, this is a feature of the studentized statistic t . Ding and Dasgupta (2017) showed that the FRT with other test statistics may not have the dual guarantee. In particular, the FRT with $\hat{\tau}$ may be asymptotically anti-conservative under H_{0N} . I give some heuristics below to illustrate the importance of studentization in the FRT.

Under H_{0N} , we have

$$\hat{\tau} \sim N\left(0, \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n}\right).$$

The FRT pretends that the Science Table is $(Y_i, Y_i)_{i=1}^n$, so the permutation distribution of $\hat{\tau}$ is

$$(\hat{\tau})^\pi \sim N\left(0, \frac{s^2}{n_1} + \frac{s^2}{n_0}\right),$$

where $(\cdot)^\pi$ denotes the permutation distribution and s^2 is the sample variance of the observed outcomes. Based on (3.7) in Chapter 3, we can approximate the asymptotic variance of $(\hat{\tau})^\pi$ under H_{0F} as

$$\begin{aligned} \frac{s^2}{n_1} + \frac{s^2}{n_0} &= \frac{n}{n_1 n_0} \left\{ \frac{n_1 - 1}{n - 1} \hat{S}^2(1) + \frac{n_0 - 1}{n - 1} \hat{S}^2(0) + \frac{n_1 n_0}{n(n - 1)} \hat{\tau}^2 \right\} \\ &\approx \frac{\hat{S}^2(1)}{n_0} + \frac{\hat{S}^2(0)}{n_1} \\ &\approx \frac{S^2(1)}{n_0} + \frac{S^2(0)}{n_1}, \end{aligned}$$

which does not match the asymptotic variance of $\hat{\tau}$. Ideally, we should compute the p -value under H_{0N} based the true distribution of $\hat{\tau}$, which, however, depends on the unknown potential outcomes. In contrast, we use the FRT to compute the p_{FRT} based on the permutation distribution $(\hat{\tau})^\pi$, which does not match the true distribution of $\hat{\tau}$ under H_{0N} even with large samples. Therefore, the FRT with $\hat{\tau}$ may not control the type one error rate under H_{0N} even with large samples.

Fortunately, the undesired property of the FRT with $\hat{\tau}$ goes away if we replace the test statistic $\hat{\tau}$ with the studentized version t . Under H_{0N} , we have

$$t \sim N(0, C^2)$$

where $C^2 \leq 1$ with equality holding if $Y_i(1) - Y_i(0) = \tau$ for all units $i = 1, \dots, n$. The FRT generates the permutation distribution

$$t^\pi \sim N(0, 1)$$

where the variance equals 1 because the Science Table used by the FRT has zero individual causal effects. Under H_{0N} , because the true distribution of t is more dispersed than the corresponding permutation distribution, the p_{FRT} based on t is asymptotically conservative.

8.2 Covariate-adjusted FRTs in the CRE

Extending the discussion in Section 8.1 to the case with covariates, Zhao and Ding (2021a) recommend using the FRT with the studentized Lin (2013)'s estimator:

$$t_L = \frac{\hat{\tau}_L}{\sqrt{\hat{V}_L}},$$

which is the robust t -statistic for the coefficient of Z_i in the OLS fit of Y_i on $1, Z_i, X_i$ and $Z_i X_i$. They show that the FRT with t_L has multiple guarantees:

1. the associate p_{FRT} is finite-sample exact under H_{0F} ;
2. it is asymptotically conservative under H_{0N} ;
3. it is asymptotically more powerful than the FRT with t when H_{0N} does not hold and the covariates are predictive to the outcomes;
4. the above properties holds even if the linear outcome model is misspecified.

Similarly, this is a feature of the studentized statistic t_L . Zhao and Ding (2021a) show that other covariate-adjusted FRTs reviewed in Section 6.2.1 may be either anti-conservative under H_{0N} or less powerful than the FRT with t_L when H_{0N} does not hold.

8.3 General recommendations

The recommendations for the SRE parallel those for the CRE if both the strong and weak null hypotheses are of interest. Without additional covariates, Zhao and Ding (2021a) recommend using the FRT with

$$t_S = \frac{\hat{\tau}_S}{\sqrt{\hat{V}_S}};$$

with additional covariates, they recommend using the FRT with

$$t_{L,S} = \frac{\hat{\tau}_{L,S}}{\sqrt{\hat{V}_{L,S}}}.$$

The analysis of ReM is trickier. Zhao and Ding (2021a) show that the FRT with t does not have the dual guarantees in Section 8.1, but the FRT with t_L still has the guarantees in Section 8.2. This highlights the importance of both covariate adjustment and studentization in ReM.

Similar results hold for the MPE. Without covariates, we recommend using the FRT with the t -statistic for the intercept in the OLS fit of $\hat{\tau}_i$ on 1; with covariates, we recommend using the FRT with the t -statistic for the intercept in the OLS fit of $\hat{\tau}_i$ on 1 and $\hat{\tau}_{x,i}$. Figure 7.2 in Chapter 7 are based on these recommended FRTs.

Overall, the FRTs with studentized statistics are safer choices. When the large-sample Normal approximations to the studentized statistics are accurate, the FRTs give p_{FRT} 's that are almost identical to those based on Normal approximations. When the large-sample approximations are inaccurate, the FRTs at least guarantees valid p -values under the strong null hypotheses. This is the recommendation of this book.

8.4 A case study

Chong et al. (2016) conducted a randomized experiment on 219 students of a rural secondary school in the Cajamarca district of Peru during the 2009 school year. They first provided the village clinic with iron supplements and trained the local staff to distribute one free iron pill to any adolescent who requested one in person. They then randomly assign students to three arms with three different types of videos: in the first video, a popular soccer player was encouraging the use of iron supplements to maximize energy (“soccer” arm); in the second video, a physician was encouraging the use of iron supplements to improve overall health (“physician” arm); the third video did not mention iron at all (“control” arm). The experiment was stratified on the class level (1–5). The treatment and control group sizes within classes are shown below:

	class 1	class 2	class 3	class 4	class 5
soccer	16	19	15	10	10
physician	17	20	15	11	10
control	15	19	16	12	10

One outcome of interest is the average grades in the third and fourth quarters of 2009, and an important background covariate was the anemia status at baseline. We make pairwise comparisons of the “soccer” arm versus

the “control” arm and the “physician” arm versus the “control” arm. We also compare the FRTs with and without using the covariate indicating the baseline anemia status. We use their dataset to illustrate the FRTs in complete randomization and stratified randomization. The ten subgroup analyses within the same class levels use the FRTs with t and t_L for the CRE and the two overall analyses averaging over all class levels use the FRTs with t_S and $t_{L,S}$ for the SRE.

Table 8.1 shows the point estimators, standard errors, the p -value based on the Normal approximation of the robust t -statistics, and the p -value based on the FRTs. In most strata, covariate adjustment decreases the standard error since the baseline anemia status is predictive to the outcome. Table 8.1 also exhibits two exceptions: within class 2, covariate adjustment increases the standard error when comparing “soccer” and “control”; in class 4, covariate adjustment increases the standard error when comparing “physician” and “control”. This is due to the small group sizes within these strata, causing the asymptotic approximation dubious. Nevertheless, in these two scenarios, the differences in the standard error are in the third digit. The p -values from the Normal approximation and the FRT are close with the latter being slightly larger in most cases. Based on the theory, the p -values based on the FRT should be trusted since it has an additional guarantee of being finite-sample exact under the sharp null hypothesis. This becomes important in this example since the groups sizes are quite small within strata.

We echo Bind and Rubin (2020)’s suggestion that when conducting the FRTs, not only the p -values but also the randomization distributions of the test statistics should be reported. Figure 8.1 compares the histograms of the randomization distributions of the robust t -statistics with the asymptotic approximations. In the subgroup analysis, we can observe discrepancy between the randomization distributions and $N(0, 1)$; average over all class levels, the discrepancy becomes unnoticeable. Overall, in this application, the p -values based on the Normal approximation do not differ substantially from those based on the FRTs. Two approaches yield coherent conclusions: the video with a physician telling the benefits of iron supplements improved the academic performance and the effect was most significant among student in class 3; in contrast, the video with a famous soccer player telling the benefits of the iron supplements did not have any significant effect.

8.5 Homework Problems

8.1 Re-analyzing Angrist and Lavy (2009)’s data

This is the Fisherian counterpart of Problem 7.8. Report the p_{FRT} ’s from the FRTs with studentized statistics.

TABLE 8.1: Re-analysis of Chong's data. N corresponds to the unadjusted estimators and tests, and L corresponds to the covariate-adjusted estimators and tests.

(a) soccer versus control				
	est	s.e.	p_{normal}	p_{frt}
class 1				
N	0.051	0.502	0.919	0.924
L	0.050	0.489	0.919	0.929
class 2				
N	-0.158	0.451	0.726	0.722
L	-0.176	0.452	0.698	0.700
class 3				
N	0.005	0.403	0.990	0.989
L	-0.096	0.385	0.803	0.806
class 4				
N	-0.492	0.447	0.271	0.288
L	-0.511	0.447	0.253	0.283
class 5				
N	0.390	0.369	0.291	0.314
L	0.443	0.318	0.164	0.186
all				
N	-0.051	0.204	0.802	0.800
L	-0.074	0.200	0.712	0.712
(b) physician versus control				
	est	s.e.	p_{normal}	p_{frt}
class 1				
N	0.567	0.426	0.183	0.192
L	0.588	0.418	0.160	0.174
class 2				
N	0.193	0.438	0.659	0.666
L	0.265	0.409	0.517	0.523
class 3				
N	1.305	0.494	0.008	0.012
L	1.501	0.462	0.001	0.003
class 4				
N	-0.273	0.413	0.508	0.515
L	-0.313	0.417	0.454	0.462
class 5				
N	-0.050	0.379	0.895	0.912
L	-0.067	0.279	0.811	0.816
all				
N	0.406	0.202	0.045	0.047
L	0.463	0.190	0.015	0.017

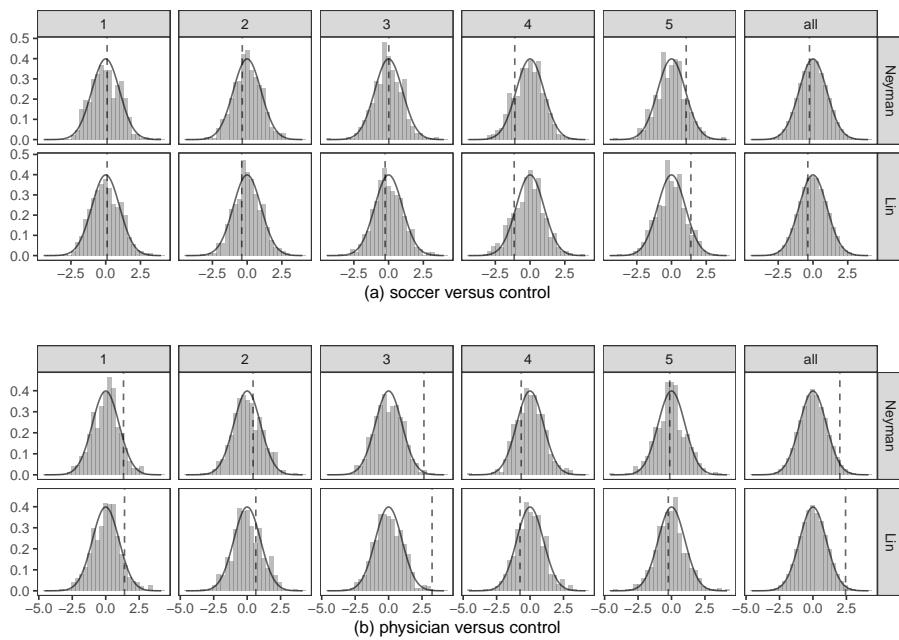


FIGURE 8.1: Re-analyzing Chong et al. (2016)'s data: randomization distributions with 5×10^4 Monte Carlo draws and the $N(0, 1)$ approximations

8.2 Replication of Zhao and Ding (2021a)'s Figure 1

Zhao and Ding (2021a) use simulation to evaluate the finite-sample properties of the p_{FRT} 's from the FRTs with various test statistics. Based on their Figure 1, they recommend using the FRT with $t_{\text{L,S}}$ to analyze the SRE. Replicate their Figure 1.

8.3 Recommended reading

Zhao and Ding (2021a).

9

Bridging Finite and Super Population Causal Inference

We have focused on the finite population perspective in randomized experiment. It treats all the potential outcomes as fixed numbers or conditions on them if they are realizations of some random variables. The advantage of this perspective is that it focuses on the design of the experiments and requires minimal assumptions on the data generating process of the outcomes. However, it is often criticized for having only *internal validity* but not necessarily *external validity*. Obviously, all experimenters care about not only the internal validity but also the external validity of their experiments. Since all statistical properties are conditional on the potential outcomes for the units we have, the results are only about the observed units. Then a natural question arises: do the finite population results generalize to a bigger population?

This is a fair critique on the finite population framework conditional on the potential outcomes. However, this can be a philosophical question. What we observed is a finite population, so any experimental design and analysis directly give us information about this finite population. Randomization only ensures internal validity given the potential outcomes of these units. The external validity of the results depend on the sampling process of the units. If the finite population is a representative sample of a larger population we are interested in, then of course the experimental results also have external validity. Otherwise, the results based on randomization inference may not generalize. Pearl and Bareinboim (2014) discussed this *transportability* problem from a different perspective.

For some statisticians, this is just a technical problem. We can change the statistical framework, assuming that the units are sampled from a super population. Then all the statements are about the population of interest. This is a convenient framework, although it does not really solve the problem mentioned above. Below, I will introduce this framework for two purposes: first, it gives a different perspective for randomized experiments; second, it serves as a bridge between Parts II and III of this book. The latter purpose is more important, since the super population framework allows us to derive more fruitful results for observational studies in which the treatment is not randomly assigned.

9.1 CRE

Assume

$$\{Z_i, Y_i(1), Y_i(0), X_i\}_{i=1}^n \stackrel{\text{IID}}{\sim} \{Z, Y(1), Y(0), X\}$$

from a super population. With a little abuse of notation, we define the population average causal effect as

$$\tau = E\{Y(1) - Y(0)\} = E\{Y(1)\} - E\{Y(0)\}.$$

Under the super population framework, we can formulate the CRE as below.

Definition 9.1 (CRE under the super population framework) $Z \perp\!\!\!\perp \{Y(1), Y(0), X\}$

Under Definition 9.1, the average causal effect can be written as

$$\begin{aligned}\tau &= E\{Y(1) | Z = 1\} - E\{Y(0) | Z = 0\} \\ &= E(Y | Z = 1) - E(Y | Z = 0),\end{aligned}\tag{9.1}$$

which equals the difference in expectations of the outcomes. Since τ can be expressed as a function of the distributions of the observables, it is *nonparametrically identifiable*¹. The identification formula (9.1) immediately suggests a moment estimator $\hat{\tau}$, which is the difference in means of the outcomes defined before. Conditioning on Z , this is then a standard two-sample problem comparing the means of two independent samples. We have

$$E(\hat{\tau} | Z) = \tau, \quad \text{var}(\hat{\tau} | Z) = \frac{\text{var}\{Y(1)\}}{n_1} + \frac{\text{var}\{Y(0)\}}{n_0}.$$

Under IID sampling, the sample variances are unbiased for the population variances, so Neyman (1923)'s variance estimator is unbiased for $\text{var}(\hat{\tau} | Z)$. The conservativeness problem goes away under this super population framework.

We can also discuss covariate adjustment. Based on the OLS decompositions (see Chapter A2)

$$Y(1) = \gamma_1 + \beta_1^\top X + \varepsilon(1),\tag{9.2}$$

$$Y(0) = \gamma_0 + \beta_0^\top X + \varepsilon(0),\tag{9.3}$$

we have

$$\tau = E\{Y(1) - Y(0)\} = \gamma_1 - \gamma_0 + (\beta_1 - \beta_0)^\top E(X),$$

since the residuals $\varepsilon(1)$ and $\varepsilon(0)$ have mean zero due to the inclusion of the

¹In causal inference, we say that a parameter is nonparametrically identifiable if it can be determined by the distribution of the observed variables without imposing further parametric assumptions.

intercepts. We can use the OLS with the treated and control data to estimate the coefficients in (9.2) and (9.3), respectively. The sample versions of the coefficients are $\hat{\gamma}_1, \hat{\beta}_1, \hat{\gamma}_0, \hat{\beta}_0$, so a covariate-adjusted estimator for τ is

$$\hat{\tau}_{\text{adj}} = \hat{\gamma}_1 - \hat{\gamma}_0 + (\hat{\beta}_1 - \hat{\beta}_0)^T \bar{X}.$$

If we center covariates with $\bar{X} = 0$, the above estimator reduces to Lin (2013)'s estimator

$$\hat{\tau}_L = \hat{\gamma}_1 - \hat{\gamma}_0,$$

which equals the coefficient of Z in the pooled regression with treatment-covariates interactions.

Unfortunately, the EHW variance estimator does not work for $\hat{\tau}_L$ because of the additional uncertainty \bar{X} under the super population framework. Berk et al. (2013), Negi and Wooldridge (2021) and Zhao and Ding (2021a) proposed a correction of the EHW variance estimator by adding an additional term

$$(\hat{\beta}_1 - \hat{\beta}_0)^T S_X^2 (\hat{\beta}_1 - \hat{\beta}_0)/n.$$

A conceptually simpler yet computationally intensive approach is to use the bootstrap to estimate the variance; see Chapter A1.5.

9.2 SRE

We can extend the discussion in Section 9.1 to the SRE since it is equivalent to independent CREs within strata. The notation below will be slightly different from that in Chapter 5.

Assume that

$$\{Z_i, Y_i(1), Y_i(0), X_i\} \stackrel{\text{IID}}{\sim} \{Z, Y(1), Y(0), X\}.$$

With a discrete covariate $X_i \in \{1, \dots, K\}$, we can formulate the SRE as below.

Definition 9.2 (SRE under the super population framework) $Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$.

Under Definition 9.2, the conditional average causal effect can be rewritten as

$$\tau_{[k]} = E\{Y(1) - Y(0) \mid X = k\} = E(Y \mid Z = 1, X = k) - E(Y \mid Z = 0, X = k),$$

so the average causal effect can be rewritten as

$$\tau = E\{Y(1) - Y(0)\} = \sum_{k=1}^K \text{pr}(X=k) E\{Y(1) - Y(0) \mid X = k\} = \sum_{k=1}^K \text{pr}(X=k) \tau_{[k]}.$$

The discussion in Section 9.1 holds with all strata, so we can derive the super population analog for the SRE. When there are more than two treatment and control units within each strata, we can use \hat{V}_S as an unbiased variance estimator for $\text{var}(\hat{\tau}_S)$.

9.3 Homework Problems

9.1 OLS decomposition of the observed outcome under the CRE

Based on (9.2) and (9.3), show that the OLS decomposition of the observed outcome on the treatment, covariates and their interaction is

$$Y = \alpha_0 + \alpha_Z Z + \alpha_X^\top X + \alpha_{ZX}^\top XZ + \varepsilon$$

where

$$\alpha_0 = \gamma_0, \quad \alpha_Z = \gamma_1 - \gamma_0, \quad \alpha_X = \beta_0, \quad \alpha_{ZX} = \beta_1 - \beta_0, \quad \varepsilon = Z\varepsilon(1) + (1-Z)\varepsilon(0).$$

That is,

$$(\alpha_0, \alpha_Z, \alpha_X, \alpha_{ZX}) = \arg \min_{a_0, a_Z, a_X, a_{ZX}} E(Y - a_0 - a_Z Z - a_X^\top X - a_{ZX}^\top XZ)^2.$$

9.2 Recommended reading

Ding et al. (2017a) provide a unified discussion of the finite-population and super-population inferences for the average causal effect.

Part III

Observational studies



10

Observational Studies, Selection Bias, and Nonparametric Identification of Causal Effects

Cochran (1965) summarized two common characteristics of observational studies:

1. the objective is to elucidate cause-and-effect relationships;
2. it is not feasible to use controlled experimentation.

The first characteristic is identical to that of randomized experiments discussed in Part II, but the second differs fundamentally from randomized experiments.

Dorn (1953) suggested that the planner of an observational study should always ask himself the following question:

How would the study be conducted if it were possible to do it by controlled experimentation?

It is always helpful to follow Dorn (1953)'s suggestion because the potential outcomes framework has an intrinsic link to an experiment, either a real experiment or a thought experiment. Part III of this book will discuss causal inference with observational studies. It will clarify the fundamental differences between observational studies and randomized experiments. Nevertheless, many ideas of causal inference with observational studies are deeply connected to those with randomized experiments.

10.1 Motivating Examples

Example 10.1 (job training program) *LaLonde (1986) was interested in the causal effect of a job training program on earnings. His compared the results based on a randomized experiment to the results based on observational studies. We have used the experimental data before, which is the `lalonde` dataset in the `Matching` package; we have also used an observational counterpart `cps1re74.csv` in Problem 1.3. LaLonde (1986) found that many traditional*

econometric methods for observational studies gave quite different estimates compared to the estimates based on the experimental data. Dehejia and Wahba (1999) re-analyzed the data using methods motivated by causal inference, and found that those methods can recover the experimental gold standard. Since then, this became a canonical example in causal inference with observational studies.

Example 10.2 (smoking and homocysteine) Bazzano et al. (2003) compared the homocysteine levels in daily smokers and never smokers based on the data from the National Health and Nutrition Examination Survey (NHANES) 2005–2006. Rosenbaum (2018) documented the data as `homocyst` in the package `senstrat`. The dataset has the following important covariates:

<code>female</code>	1=female, 0=male
<code>age3</code>	three age categories: 20–39, 40–50, ≥ 60
<code>ed3</code>	three education categories: < High School, High School, some College
<code>bmi3</code>	three BMI categories: <30, [30, 35), ≥ 35
<code>pov2</code>	TRUE=income at least twice the poverty level, FALSE otherwise

Example 10.3 (school meal program and body mass index) Chan et al. (2016) used a subsample of the data from NHANES 2007–2008 to study whether participation in school meal programs lead to an increase in BMI for school children. They documented the data as `nhanes_bmi` in the package `ATE`. The dataset has the following important covariates:

<code>age</code>	<code>age</code>
<code>ChildSex</code>	gender (1: Male, 0: Female)
<code>black</code>	race (1: Black, 0: otherwise)
<code>mexam</code>	race (1: Hispanic: 0 otherwise)
<code>pir200_plus</code>	Family above 200% of the federal poverty level
<code>WIC</code>	Participation in the special supplemental nutrition program
<code>Food_Stamp</code>	Participation in food stamp program
<code>fsdchbi</code>	Childhood food security
<code>AnyIns</code>	Any insurance
<code>RefSex</code>	Gender of the adult respondent (1: Male, 0: Female)
<code>RefAge</code>	Age of the adult respondent

10.2 Causal effects and selection bias under the potential outcomes framework

For unit i ($i = 1, \dots, n$), we have pretreatment covariates X_i , a binary treatment indicator Z_i , and an observed outcome Y_i with two potential outcomes $Y_i(1)$ and $Y_i(0)$ under treatment and control, respectively. For simplicity, we assume

$$\{X_i, Z_i, Y_i(1), Y_i(0)\}_{i=1}^n \stackrel{\text{IID}}{\sim} \{X, Z, Y(1), Y(0)\}.$$

So we can drop the subscript i for quantities depending on this population. The causal effects of interest are the average causal effect

$$\tau = E\{Y(1) - Y(0)\},$$

the average causal effect on the treated units

$$\tau_T = E\{Y(1) - Y(0) \mid Z = 1\},$$

and the average causal effect on the control units:

$$\tau_C = E\{Y(1) - Y(0) \mid Z = 0\}.$$

By the linearity of the expectation, we have

$$\begin{aligned}\tau_T &= E\{Y(1) \mid Z = 1\} - E\{Y(0) \mid Z = 1\} \\ &= E(Y \mid Z = 1) - E\{Y(0) \mid Z = 1\}\end{aligned}$$

and

$$\begin{aligned}\tau_C &= E\{Y(1) \mid Z = 0\} - E\{Y(0) \mid Z = 0\} \\ &= E\{Y(1) \mid Z = 0\} - E(Y \mid Z = 0).\end{aligned}$$

In the above two formulas of τ_T and τ_C , the quantities $E(Y \mid Z = 1)$ and $E(Y \mid Z = 0)$ are directly observable from the data, but the quantities $E\{Y(0) \mid Z = 1\}$ and $E\{Y(1) \mid Z = 0\}$ are not. The latter two are *counterfactuals* because they are the means of the potential outcomes corresponding to the treatment level that is the opposite of the actual received treatment.

The simple difference in means, also known as the *prima facie* causal effect,

$$\begin{aligned}\tau_{PF} &= E(Y \mid Z = 1) - E(Y \mid Z = 0) \\ &= E\{Y(1) \mid Z = 1\} - E\{Y(0) \mid Z = 0\}\end{aligned}$$

is generally biased for the causal effects defined above. For example,

$$\tau_{PF} - \tau_T = E\{Y(0) \mid Z = 1\} - E\{Y(0) \mid Z = 0\}$$

and

$$\tau_{PF} - \tau_C = E\{Y(1) \mid Z = 1\} - E\{Y(1) \mid Z = 0\}$$

are not zero in general, and they quantifies the *selection bias*. They measure the differences in the means of the potential outcomes across the treatment and control groups.

Why randomization is so important? Rubin (1978) first used potential outcomes to quantify the benefit of randomization. We have used the fact in Chapter 9 that

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \tag{10.1}$$

in the CRE, which implies that the selection bias terms are both zero:

$$\tau_{\text{PF}} - \tau_{\text{T}} = E\{Y(0) | Z = 1\} - E\{Y(0) | Z = 0\} = 0$$

and

$$\tau_{\text{PF}} - \tau_{\text{C}} = E\{Y(1) | Z = 1\} - E\{Y(1) | Z = 0\} = 0.$$

So under complete randomization (10.1),

$$\tau = \tau_{\text{T}} = \tau_{\text{C}} = \tau_{\text{PF}}.$$

From the above discussion, the fundamental benefit of randomization is to balance the distributions of the potential outcomes across the treatment and control groups, which is more important than to balance the distributions of the observed covariates.

Without randomization, the selection bias terms can be arbitrarily large especially for unbounded outcomes. This highlights the fundamental difficulty of causal inference with observational studies.

10.3 Sufficient conditions for nonparametric identification

10.3.1 Identification

Causal inference with observational studies is challenging. It relies on strong assumptions. A strategy is to use the information of the pretreatment covariates and assume that conditioning on the observed covariates X , the selection bias terms are zero, that is,

$$E\{Y(0) | Z = 1, X\} = E\{Y(0) | Z = 0, X\}, \quad (10.2)$$

$$E\{Y(1) | Z = 1, X\} = E\{Y(1) | Z = 0, X\}. \quad (10.3)$$

The assumptions in (10.2) and (10.3) state that the differences in the means of the potential outcomes across the treatment and control groups are entirely due to the difference in the observed covariates. So given the same value of the covariates, the potential outcomes have the same means across the treatment and control groups. Mathematically, (10.2) and (10.3) ensure that the conditional versions of the effects are identical:

$$\tau(X) = \tau_{\text{T}}(X) = \tau_{\text{C}}(X) = \tau_{\text{PF}}(X),$$

where

$$\begin{aligned} \tau(X) &= E\{Y(1) - Y(0) | X\}, \\ \tau_{\text{T}}(X) &= E\{Y(1) - Y(0) | Z = 1, X\}, \\ \tau_{\text{C}}(X) &= E\{Y(1) - Y(0) | Z = 0, X\}, \\ \tau_{\text{PF}}(X) &= E(Y | Z = 1, X) - E(Y | Z = 0, X). \end{aligned}$$

In particular, $\tau(X)$ is often called the *conditional average causal effect*.

A key result in this chapter is that the average causal effect τ is *nonparametrically identifiable* under (10.2) and (10.3). The notion of nonparametrically identifiability does not appear frequently in classic statistics, but it is key to causal inference with observational studies.

Definition 10.1 (identification) *A parameter θ is identifiable if it can be written as a function of the distribution of the observed data under certain model assumptions. A parameter θ is nonparametrically identifiable if it can be written as a function of the distribution of the observed data without any parametric model assumptions.*

Definition 10.1 is too abstract at the moment. I will use more concrete examples in later chapters to illustrate its meaning. It is often neglected in standard statistics problems. For instance, the mean $\theta = E(Y)$ is nonparametrically identifiable if we have IID draws of Y_i 's; the Pearson correlation coefficient $\theta = \text{corr}(X, Y)$ is nonparametrically identifiable if we have IID draws of the pairs (X_i, Y_i) 's. In those examples, the parameters are nonparametrically identifiable automatically. However, Definition 10.1 is fundamental in causal inference with observational studies. In particular, the parameter of interest $\tau = E\{Y(1) - Y(0)\}$ depends on some unobserved random variables, so it is unclear whether it is nonparametrically identifiable based on observed data. Under the assumptions in (10.2) and (10.3), it is nonparametrically identifiable, with detailed below.

Because $\tau_{\text{PF}}(X)$ depends only on the observables, it is nonparametrically identified by definition. Moreover, (10.2) and (10.3) ensure that the three causal effects are the same as $\tau_{\text{PF}}(X)$, so $\tau(X)$, $\tau_{\text{T}}(X)$ and $\tau_{\text{C}}(X)$ are all nonparametrically identified. Consequently, the unconditional versions are also nonparametrically identified under (10.2) and (10.3) due to the law of total expectation:

$$\tau = E\{\tau(X)\}, \quad \tau_{\text{T}} = E\{\tau_{\text{T}}(X) | Z = 1\}, \quad \tau_{\text{C}} = E\{\tau_{\text{C}}(X) | Z = 0\}.$$

From now on, we focus on τ unless stated otherwise. The following theorem summarized the identification formulas of τ .

Theorem 10.1 *Under (10.2) and (10.3), the average causal effect τ is identified by*

$$\tau = E\{\tau(X)\} \tag{10.4}$$

$$= E\{E(Y | Z = 1, X) - E(Y | Z = 0, X)\} \tag{10.5}$$

$$= \int \{E(Y | Z = 1, X = x) - E(Y | Z = 0, X = x)\} F(dx). \tag{10.6}$$

The formula (10.5) was formally established by Rosenbaum and Rubin (1983b), which is also called the g-formula by Robins (see Hernán and Robins, 2020).

With a discrete covariate, we can write the identification formula in Theorem 10.1 as

$$\begin{aligned}\tau &= \sum_x E(Y | Z = 1, X = x) \text{pr}(X = x) \\ &\quad - \sum_x E(Y | Z = 0, X = x) \text{pr}(X = x),\end{aligned}\tag{10.7}$$

and also the simple difference in means as

$$\begin{aligned}\tau_{\text{PF}} &= \sum_x E(Y | Z = 1, X = x) \text{pr}(X = x | Z = 1) \\ &\quad - \sum_x E(Y | Z = 0, X = x) \text{pr}(X = x | Z = 0)\end{aligned}\tag{10.8}$$

by the law of total probability. Comparing (10.7) and (10.8), we can see that although both formulas compare the conditional expectations $E(Y | Z = 1, X = x)$ and $E(Y | Z = 0, X = x)$, they average over different distribution of the covariates. The causal parameter τ averages the conditional expectations over the common distribution of the covariate, but the difference in means τ_{PF} averages the conditional expectations over two different distributions of covariate in the treated and control groups.

Usually, we impose a stronger assumption:

$$Y(z) \perp\!\!\!\perp Z | X \quad (z = 0, 1).\tag{10.9}$$

This assumption has many names:

1. *ignorability* due to Rubin (1978);
2. *unconfoundedness* which is popular among epidemiologists;
3. *selection on observables* which is popular among social scientists;
4. *conditional independence* which is merely a description of the notation $\perp\!\!\!\perp$ in the assumption.

Sometimes, we impose an even stronger assumption

$$\{Y(1), Y(0)\} \perp\!\!\!\perp Z | X\tag{10.10}$$

which is called *strong ignorability* (Rosenbaum and Rubin, 1983b). If the parameter of interest is τ , then the stronger assumptions (10.9) and (10.10) are just imposed for notational simplicity. They are not necessary in this case. However, they cannot be relaxed if the parameter of interest is the causal effects on other scales (for example, distribution, quantile, or some transformation of the outcome). The *strong ignorability* assumption requires that the potential outcomes vector be independent of the treatment given covariates, but the *ignorability* assumption only requires each potential outcome be independent of the treatment given covariates. The former is stronger than the

latter. However, their difference is rather technical and of pure probability interests; see Problem 10.4. In most reasonable statistical models, they are identical; see Section 10.3.2 below. We will not distinguish them in this book and will simply use *ignorability* to refer to both.

10.3.2 Plausibility of the assumption

A fundamental problem of causal inference with observational studies is the plausibility of the ignorability assumption. The above discussion may seem too mathematical in the sense that the ignorability assumption serves as a sufficient condition to ensure the nonparametric identification of the average causal effect. What is its scientific meaning? Intuitively, it rules out all unmeasured covariates that affect treatment and outcome simultaneously. Those “common causes” of the treatment and outcomes are called confounders. That is why the ignorability assumption is also called the unconfoundedness assumption. More mathematically, we can interpret the ignorability assumption based on the outcome data generating process. If

$$\begin{aligned} Y(1) &= f_1(X, V_1), \\ Y(0) &= f_0(X, V_0), \\ Z &= 1\{g(X, V) \geq 0\} \end{aligned}$$

with $(V_1, V_0) \perp\!\!\!\perp V$, then (10.9) and (10.10) hold. In the above data generating process, the “common causes” X of the treatment and the outcome are all observed, the remaining random components are independent. If the data generating process changes to

$$\begin{aligned} Y(1) &= f_1(X, U, V_1), \\ Y(0) &= f_0(X, U, V_0), \\ Z &= 1\{g(X, U, V) \geq 0\} \end{aligned}$$

with $(V_1, V_0) \perp\!\!\!\perp V$, then (10.9) or (10.10) does not hold in general. The unmeasured “common cause” U induces dependence between the treatment and potential outcomes even conditioning on the observed covariates X . If we do not have access to U and analyze the data based only on (Z, X, Y) , the final estimator will be biased for the causal parameter in general. This type of bias is called the *omitted variable bias* in econometrics.

The ignorability assumption can be reasonable if we observe a rich set of covariates X that affect the treatment and the outcome simultaneously. I start with this assumption, discussing identification and estimation strategies in Part III of this book. However, it is fundamentally untestable. We may justify it based on the scientific background knowledge, but we are often not sure whether it holds or not. Parts IV and V of this book will discuss other strategies when this assumption is not plausible.

10.4 Two simple estimation strategies and their limitations

10.4.1 Stratification or standardization based on discrete covariates

If the covariate $X_i \in \{1, \dots, K\}$ is discrete, then ignorability (10.9) reads as

$$Y(z) \perp\!\!\!\perp Z \mid X = k \quad (z = 0, 1; k = 1, \dots, K),$$

which essentially assumes that the observational study is a SRE. Therefore, we can use the estimator

$$\hat{\tau} = \sum_{k=1}^K \pi_{[k]} \left\{ \hat{Y}_{[k]}(1) - \hat{Y}_{[k]}(0) \right\},$$

which is identical to the stratified or post-stratified estimator discussed in Chapter 5.

This method is still widely used in practice. Example 10.2 contains discrete covariates, and I relegate the analysis to Problem 10.3. However, there are several obvious difficulties in implementing this method. First, it works well for the case with small K . For large K , it is very likely that many strata have $n_{[k]1} = 0$ or $n_{[k]0} = 0$, leading to the ill-defined $\hat{\tau}_{[k]}$'s for those strata. This is related to the issue of overlap which will be discussed in Chapter 20. Second, it is not obvious how to apply this stratification method to multidimensional continuous or mixed covariates X . A standard method is to create strata based on the initial covariates and then apply the stratification method. This may result in arbitrariness in the analysis.

10.4.2 Outcome regression

The most commonly-used method based on the outcome regression is to run the OLS with an additive model of the observed outcome on the treatment indicator and covariates, which assumes

$$E(Y \mid Z, X) = \beta_0 + \beta_z Z + \beta_x^\top X.$$

If the above linear model is correct, then we have

$$\begin{aligned} \tau(X) &= E(Y \mid Z = 1, X) - E(Y \mid Z = 0, X) \\ &= (\beta_0 + \beta_z + \beta_x^\top X) - (\beta_0 + \beta_x^\top X) \\ &= \beta_z, \end{aligned}$$

which implies that the treatment effect is homogeneous with respect to the covariates. This, coupled with ignorability, implies that

$$\tau = E\{\tau(X)\} = \beta_z.$$

Therefore, if ignorability holds and the outcome model is linear, then the average causal effect equals the coefficient of Z . This is one of the most important applications of the linear model. However, the causal interpretation of the coefficient of Z is valid only under two strong assumptions: ignorability and the linear model.

We have discussed in Chapter 6, the above procedure is suboptimal even in randomized experiments, because it ignores the treatment effect heterogeneity induced by the covariates. If we assume

$$E(Y | Z, X) = \beta_0 + \beta_z Z + \beta_x^T X + \beta_{zx}^T X Z,$$

we have

$$\begin{aligned}\tau(X) &= E(Y | Z = 1, X) - E(Y | Z = 0, X) \\ &= (\beta_0 + \beta_z + \beta_x^T X + \beta_{zx}^T X) - (\beta_0 + \beta_x^T X) \\ &= \beta_z + \beta_{zx}^T X,\end{aligned}$$

which, coupled with ignorability, implies that

$$\tau = E\{\tau(X)\} = E(\beta_z + \beta_{zx}^T X) = \beta_z + \beta_{zx}^T E(X).$$

The estimator for τ is then $\hat{\beta}_z + \hat{\beta}_{zx}^T \bar{X}$, where $\hat{\beta}_z$ is the regression coefficient and \bar{X} is the sample mean of X . If we center the covariates to ensure $\bar{X} = 0$, then the estimator is simply the regression coefficient of Z . To simplify the procedure, we usually center the covariates at the beginning; also recall Lin (2013)'s estimator introduced in Chapter 6. Rosenbaum and Rubin (1983b) and Hirano and Imbens (2001) discussed this estimator.

In general, we can use other more complex models to estimate the causal effects. For example, if we build two predictors $\hat{\mu}_1(X)$ and $\hat{\mu}_0(X)$ based on the treated and control data, respectively, then we have an estimator for the conditional average causal effect

$$\hat{\tau}(X) = \hat{\mu}_1(X) - \hat{\mu}_0(X)$$

and an estimator for the average causal effect:

$$\hat{\tau} = n^{-1} \sum_{i=1}^n \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}.$$

The estimator $\hat{\tau}$ above has the same form as the projective estimator discussed in Chapter 6. It is sometimes called the *outcome imputation* estimator. For example, we may model a binary outcome using a logistic model

$$E(Y | Z, X) = \text{pr}(Y = 1 | Z, X) = \frac{e^{\beta_0 + \beta_z Z + \beta_x^T X}}{1 + e^{\beta_0 + \beta_z Z + \beta_x^T X}},$$

then based on the estimators of the coefficients $\hat{\beta}_0, \hat{\beta}_z, \hat{\beta}_x$, we have the following estimator for the average causal effect:

$$\hat{\tau} = n^{-1} \sum_{i=1}^n \left\{ \frac{e^{\hat{\beta}_0 + \hat{\beta}_z + \hat{\beta}_x^\top X_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_z + \hat{\beta}_x^\top X_i}} - \frac{e^{\hat{\beta}_0 + \hat{\beta}_x^\top X_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_x^\top X_i}} \right\}.$$

This estimator is not simply the coefficient of the treatment in the logistic model.¹ It is a nonlinear function of all the coefficients as well as the empirical distribution of the covariates. In econometrics, this estimator is called the *average partial effect* or *average marginal effect* of the treatment in the logistic model. Many econometric software packages can report this estimator associated with the standard error. Similarly, we can also derive the corresponding estimator based on a fully interacted logistic model; see Problem 10.2.

For all the estimators discussed above, we can use the nonparametric bootstrap to estimate the standard errors. See Chapter A1.5.

The above predictors for the conditional means of the outcome can also be other machine learning tools. In particular, Hill (2011) championed the use of tree methods for estimating τ , and Wager and Athey (2018) proposed to use them also for estimating $\hat{\tau}(X)$. Wager and Athey (2018) also combined the tree methods with the ideas in the next chapter. Since then, machine learning and causal inference has been an active research area (e.g., Hahn et al., 2020; Künzel et al., 2019).

The biggest problem of the above approach based on outcome regressions is its sensitivity to the specification of the outcome model. Problem 1.3 gave such an example. Depending on the incentive of empirical research and publications, people sometimes reported their favorable causal effects estimates after searching over a wide set of candidate models, without confessing this searching process. This is a major source of *p*-hacking in causal inference.

10.5 Homework Problems

10.1 Nonparametric identification of other causal effects

Under ignorability and overlap, show that

1. the distributional causal effect

$$\text{DCE}_y = \text{pr}\{Y(1) > y\} - \text{pr}\{Y(0) > y\}$$

¹If the logistic outcome model is correct, then $\hat{\beta}_z$ estimates the conditional odds ratio of the treatment on the outcome given covariates, which does not equal τ . Freedman (2008c) gave a warning of using the logistic regression coefficient to estimate τ in CREs. See Chapter A2 for more details of the logistic regression.

- is nonparametrically identifiable for all y ;
2. the quantile causal effect

$$\text{QCE}_q = \text{quantile}_q\{Y(1)\} - \text{quantile}_q\{Y(0)\},$$

is nonparametrically identifiable for all q , where $\text{quantile}_q\{\cdot\}$ is the q th quantile of a random variable.

Remark: In probability theory, $\text{pr}\{Y(z) \leq y\}$ is the cumulative distribution function and $\text{pr}\{Y(z) > y\}$ is the survival function of the potential outcome $Y(z)$. The distributional causal effect compares the survival functions of the potential outcomes under treatment and control.

10.2 Outcome imputation estimator in the fully interacted logistic model

Assume that a binary outcome follows a logistic model

$$E(Y | Z, X) = \text{pr}(Y = 1 | Z, X) = \frac{e^{\beta_0 + \beta_z Z + \beta_x^\top X + \beta_{xz}^\top XZ}}{1 + e^{\beta_0 + \beta_z Z + \beta_x^\top X + \beta_{xz}^\top XZ}}.$$

What is the corresponding outcome regression estimator for the average causal effect?

10.3 Data analysis: stratification and regression

Use the dataset `homocyst` in the package `senstrat`. The outcome is `homocysteine`, the homocysteine level, and the treatment is `z`, where $z = 1$ for a daily smoker and $z = 0$ for a never smoker. Covariates are `female`, `age3`, `ed3`, `bmi3`, `pov2` with detailed explanations in the package, and `st` is a stratum indicator, defined by all the combinations of the discrete covariates.

1. How many strata have only treated or control units? What is the proportion of the units in these strata? Drop these strata and perform a stratified analysis of the observational study. Report the point estimator, variance estimator and 95% confidence interval for the average causal effect.
2. Run the OLS of the outcome on the treatment indicator and covariates without interactions. Report the coefficient of the treatment and the robust standard error.
Drop the strata with only treated or control units. Re-run the OLS and report the result.
3. Apply Lin (2013)'s estimator of the average causal effect. Report the coefficient of the treatment and the robust standard error.
If you do not drop the strata with only treated or control units, what will happen?
4. Compare the results in the above three analyses. Which one is more credible?

10.4 Ignorability versus strong ignorability

Given an example such that the ignorability holds but the strong ignorability does not hold.

Remark: This is related to a classic probability problem of finding three random variables A, B, C such that

$$A \perp\!\!\!\perp C \text{ and } B \perp\!\!\!\perp C \text{ but } (A, B) \not\perp\!\!\!\perp C.$$

10.5 Recommended reading

Cochran (1965) is a classic reference on observational studies. It contains many useful insights but does not use the formal potential outcomes framework.

11

The Central Role of the Propensity Score in Observational Studies for Causal Effects

Rosenbaum and Rubin (1983b) proposed the key concept *propensity score* and discussed its role in causal inference with observational studies. It is one of the most cited papers in statistics, and Titterington (2013) listed it as the second most cited paper published in *Biometrika* during the past 100 years. Its citations are growing very fast during the recent years.

Under the IID sampling assumption, we have four random variables associated with each unit: $\{X, Z, Y(1), Y(0)\}$. Following the basic probability rule, we can factorize the joint distribution as

$$\begin{aligned} & \text{pr}\{X, Z, Y(1), Y(0)\} \\ = & \text{pr}(X) \times \text{pr}\{Y(1), Y(0) | X\} \times \text{pr}\{Z | X, Y(1), Y(0)\}, \end{aligned}$$

where $\text{pr}(X)$ is the covariate distribution, $\text{pr}\{Y(1), Y(0) | X\}$ is the outcome model, and $\text{pr}\{Z | X, Y(1), Y(0)\}$ is the treatment assignment mechanism. Usually, we do not want to model the covariates because they are background information happening before the treatment and outcome. If we want to move beyond the outcome model, then we must focus on the treatment assignment mechanism, which leads to the definition of the propensity score.

Definition 11.1 (propensity score) Define

$$e(X, Y(1), Y(0)) = \text{pr}\{Z = 1 | X, Y(1), Y(0)\}$$

as the propensity score. Under strong ignorability, we have

$$e(X, Y(1), Y(0)) = \text{pr}\{Z = 1 | X, Y(1), Y(0)\} = \text{pr}(Z = 1 | X),$$

so the propensity score reduces to

$$e(X) = \text{pr}(Z = 1 | X),$$

the conditional probability of the receiving the treatment given the observed covariates.

Rosenbaum and Rubin (1983b) used $e(X) = \text{pr}(Z = 1 | X)$ as the definition of the propensity score because they focused on observational studies

under ignorability. It is sometimes helpful to view $e(X, Y(1), Y(0)) = \text{pr}\{Z = 1 | X, Y(1), Y(0)\}$ as the general definition of the propensity score even when ignorability fails. See Problem 11.1 for more details.

Following Rosenbaum and Rubin (1983b), this chapter will demonstrate that $e(X)$ is a key quantity in causal inference with observational studies under ignorability.

11.1 The propensity score as a dimension reduction tool

11.1.1 Theory

Theorem 11.1 *If $Z \perp\!\!\!\perp \{Y(1), Y(0)\} | X$, then $Z \perp\!\!\!\perp \{Y(1), Y(0)\} | e(X)$.*

Theorem 11.1 states that if strong ignorability holds conditional on covariates X , then it also holds conditional on the scalar propensity score $e(X)$. The ignorability requires conditioning on many background characteristics Z of the units, but Theorem 11.1 implies that controlling for the propensity score $e(X)$ removes all confounding induced by covariates X . The original covariates X can be general and have many dimensions, but the propensity score $e(X)$ is a one-dimensional scalar variable bounded between 0 and 1. Therefore, the propensity score reduces the dimension of the original covariates but still maintain the ignorability. As a technical statistical terminology, we can view the propensity score as a *dimensional reduction* tool. We will first prove Theorem 11.1 below and then given an application of the dimension reduction property of the propensity score.

Proof of Theorem 11.1: By the definition of conditional independence, we need to show that

$$\text{pr}\{Z = 1 | Y(1), Y(0), e(X)\} = \text{pr}\{Z = 1 | e(X)\}. \quad (11.1)$$

The left-hand side of (11.1) equals

$$\begin{aligned} & \text{pr}\{Z = 1 | Y(1), Y(0), e(X)\} \\ &= E\{Z | Y(1), Y(0), e(X)\} \\ &= E\left[E\{Z | Y(1), Y(0), e(X), X\} | Y(1), Y(0), e(X)\right] \\ &\quad (\text{tower property; see Section A1.1.1}) \\ &= E\left[E\{Z | Y(1), Y(0), X\} | Y(1), Y(0), e(X)\right] \\ &= E\left\{E(Z | X) | Y(1), Y(0), e(X)\right\} \quad (\text{strong ignorability}) \\ &= E\left\{e(X) | Y(1), Y(0), e(X)\right\} \\ &= e(X). \end{aligned}$$

The right-hand side of (11.1) equals

$$\begin{aligned}
 & \text{pr}\{Z = 1 \mid e(X)\} \\
 &= E\{Z \mid e(X)\} \\
 &= E\left[E\{Z \mid e(X), X\} \mid e(X)\right] \quad (\text{tower property}) \\
 &= E\left\{E(Z \mid X) \mid e(X)\right\} \\
 &= E\left\{e(X) \mid e(X)\right\} \\
 &= e(X).
 \end{aligned}$$

So the left-hand side of (11.1) equals the right-hand side of (11.1). \square

11.1.2 Propensity score stratification

Theorem 11.1 motivates a simple method for estimating causal effects: propensity score stratification. Starting from the simple case, we assume that the propensity score is known and only takes K possible values $\{e_1, \dots, e_K\}$ with K being much smaller than the sample size n . Theorem 11.1 reduces to

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid e(X) = e_k \quad (k = 1, \dots, K).$$

Therefore, we have a stratified randomized experiment (SRE), that is, we have K independent CREs within strata of the propensity score. We can analyze the observational data in the same way as the SRE stratified on $e(X)$.

In general, the propensity score is not known and is not discrete. We often fit a statistical model for $\text{pr}(Z = 1 \mid X)$ (for example, a logistic model) to obtain the estimated propensity score $\hat{e}(X)$. This estimated propensity score can take as many values as the sample size, but we can discretize it to approximate the simple case above. For example, we can discretize the estimated propensity score by its K quantiles to obtain $\hat{e}'(X)$: $\hat{e}'(X_i) = e_k$, the k/K -th quantile of $\hat{e}(X)$, if $\hat{e}(X_i)$ is between the $(k-1)/K$ -th and k/K -th quantiles of $\hat{e}(X)$. Then we have

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid \hat{e}'(X) = e_k \quad (k = 1, \dots, K).$$

approximately. So we can analyze the observational data in the same way as the SRE stratified on $\hat{e}'(X)$. The ignorability holds only approximately given $\hat{e}'(X)$. We can further use regression adjustment based on covariate to remove bias and improve efficiency. To be more specific, we can obtain Lin (2013)'s estimator within each stratum and construct the final estimator by a weighted average.

With unknown propensity score, we need to fit a statistical model to obtain the estimated propensity score $\hat{e}(X)$. This makes the final estimator dependent on the model specification. However, the propensity score stratification estimator only requires the correct ordering of the estimated propensity scores

rather than their exact values, which makes it relatively robust compared to other methods. This robustness property of propensity score stratification appeared in many numerical examples but its rigorous quantification is still missing in the literature.

An important practical question is how to choose K ? If K is too small, then the strong ignorability does not hold even approximately given $\hat{e}'(X)$. If K is too large, then we do not have enough units within each stratum of the estimated propensity score and many strata have only treated or control units. Therefore, we face a trade-off in practice. Following Cochran (1968)'s heuristics, Rosenbaum and Rubin (1983b) and Rosenbaum and Rubin (1984) suggested $K = 5$ which removes a large amount of bias in many settings. However, with extremely large dataset, propensity score stratification leads to biased estimators with a fixed K (Lunceford and Davidian, 2004). It is thus reasonable to increase K as long as each stratum has enough treated and control units. Wang et al. (2020) suggested an aggressive choice of K , which is the maximum number of strata such that the stratified estimator is well defined. But the rigorous theory for this procedure is not fully established.

Another important practical question is how to compute the standard errors of the estimators based on propensity score stratification? Some researcher conditioned on the discretized propensity scores $\hat{e}'(X)$ and reported standard errors based on the SRE. This effectively ignored the uncertainty in the estimated propensity scores. Other researchers bootstrapped the whole procedure to account for full uncertainty. However, the theory for the bootstrap is still unclear due to the discreteness of this estimator.

11.1.3 Application

To illustrate the propensity score stratification method, I revisited Example 10.3. Figure 11.1 shows the histograms of the estimated propensity scores with different numbers of bins ($K = 5, 10, 30$).

Based on propensity score stratification, we can calculate the point estimators and the standard errors for difference choice of $K \in \{5, 10, 20, 50, 80\}$ as follows (with the function `Neyman_SRE` defined in Chapter 5 for analyzing the SRE):

```
> pscore = glm(z ~ x, family = binomial)$fitted.values
> n.strata = c(5, 10, 20, 50, 80)
> strat.res = sapply(n.strata, FUN = function(nn){
+   q.pscore = quantile(pscore, (1:(nn-1))/nn)
+   ps.strata = cut(pscore, breaks = c(0,q.pscore,1),
+                   labels = 1:nn)
+   Neyman_SRE(z, y, ps.strata)})
```

5 10 20 50 80

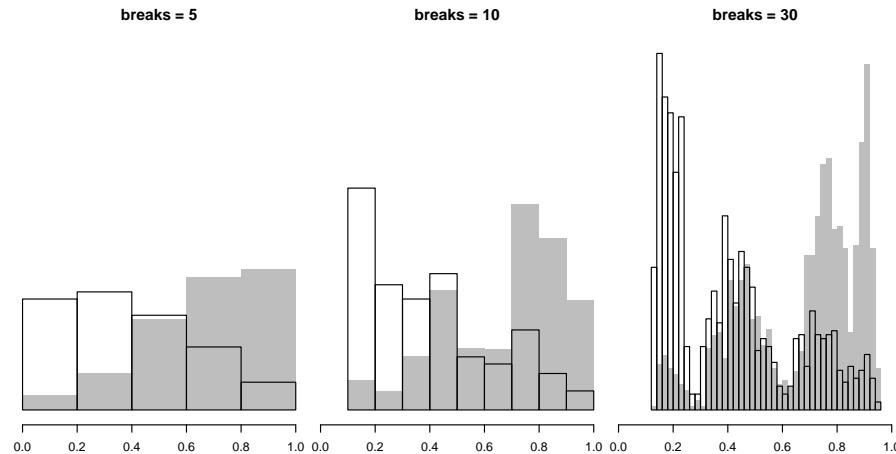


FIGURE 11.1: Histograms of the estimated propensity scores based on the `nhanes_bmi` data: white for the control group and grey for the treatment group

```
est -0.116 -0.178 -0.200 -0.265 -0.204
se  0.283  0.282  0.279  0.272      NA
```

Increasing K from 5 to 50 reduces the standard error. However, we cannot go as extreme as $K = 80$ because the standard error is not well-defined in some strata with only one treated or control unit. The above estimators show negative but insignificant effect of the meal program on the BMI.

We can also compare the above estimator with the three simple regression estimators: the one without adjusting for any covariates and Fisher and Lin's estimators.

```
naive fisher lin
est 0.534 0.061 -0.017
se  0.225 0.227  0.226
```

The naive difference in means differ greatly from other methods. Although the point estimates are different, two regression estimators and the propensity score stratification estimators give qualitatively the same results. The propensity score stratification estimators are stable across different choices of K .

11.2 Propensity score weighting

11.2.1 Theory

Theorem 11.2 *If $Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$ and $0 < e(X) < 1$, then*

$$E\{Y(1)\} = E\left\{\frac{ZY}{e(X)}\right\}, \quad E\{Y(0)\} = E\left\{\frac{(1-Z)Y}{1-e(X)}\right\},$$

and

$$\tau = E\{Y(1) - Y(0)\} = E\left\{\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right\}.$$

Before proving Theorem 11.2, it is important to note the additional assumption $0 < e(X) < 1$. It is called the *overlap* or *positivity* condition. The formulas in Theorem 11.2 become infinity if $e(X) = 0$ or 1 for some values of X . It is not a restriction due to the identification formulas based on propensity score weighting. Although it was not stated explicitly in Theorem 10.1, the conditional expectations $E(Y \mid Z = 1, X)$ and $E(Y \mid Z = 0, X)$ in the identification formula of τ in (10.5) is well defined only if $0 < e(X) < 1$. The overlap condition can be viewed as a technical condition to ensure that the formulas in Theorems 10.1 and 11.2 are well defined. It can also cause some philosophical issues for causal inference with observational studies. When unit i has $e(X_i) = 1$, we always observe its potential outcome under the treatment, $Y_i(1)$, but can never observe its potential outcome under the control, $Y_i(0)$. In this case, the potential outcome $Y_i(0)$ may not even be well defined, making the definition of the causal effect ambiguous for unit i . King and Zeng (2006) called $Y_i(0)$ an *extreme counterfactual* when $e(X_i) = 1$, and discussed their dangers in causal inference. A similar problem arises if unit i has $e(X_i) = 0$.

In sum, $Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$ requires adequate covariates to ensure the conditional independence of the treatment and potential outcomes, and $0 < e(X) < 1$ requires residual randomness in the treatment conditional on the covariates. In fact, Rosenbaum and Rubin (1983b)'s definition of strong ignorability includes both of these conditions. In the modern literature, they are often stated separately.

Proof of Theorem 11.2: I only prove the result for $E\{Y(1)\}$ because the

proof of the result for $E\{Y(0)\}$ is similar. We have

$$\begin{aligned}
 & E \left\{ \frac{ZY}{e(X)} \right\} \\
 = & E \left\{ \frac{ZY(1)}{e(X)} \right\} \\
 = & E \left[E \left\{ \frac{ZY(1)}{e(X)} \mid X \right\} \right] \quad (\text{tower property}) \\
 = & E \left[\frac{1}{e(X)} E \{ZY(1) \mid X\} \right] \\
 = & E \left[\frac{1}{e(X)} E(Z \mid X) E \{Y(1) \mid X\} \right] \quad (\text{strong ignorability}) \\
 = & E \left[\frac{1}{e(X)} e(X) E \{Y(1) \mid X\} \right] \\
 = & E [E \{Y(1) \mid X\}] \\
 = & E\{Y(1)\}.
 \end{aligned}$$

□

11.2.2 Inverse propensity score weighting estimators

Theorem 11.2 implies the following moment estimator for the average causal effect:

$$\hat{\tau}^{\text{ht}} = \frac{1}{n} \sum_{i=1}^n \frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - Z_i) Y_i}{1 - \hat{e}(X_i)},$$

where $\hat{e}(X_i)$ is the estimated propensity score. This is the inverse propensity score weighting (IPW) estimator, which is also called the Horvitz–Thompson (HT) estimator. Horvitz and Thompson (1952) proposed it in survey sampling and Rosenbaum (1987a) used in causal inference with observational studies.

However, the estimator $\hat{\tau}^{\text{ht}}$ has many problems. In particular, it is not invariant to location transformation of the outcome. For example, if we change Y_i to $Y_i + c$ with a constant c , then it becomes $\hat{\tau}^{\text{ht}} + c(\hat{1}_T - \hat{1}_C)$, where

$$\hat{1}_T = \frac{1}{n} \sum_{i=1}^n \frac{Z_i}{\hat{e}(X_i)}, \quad \hat{1}_C = \frac{1}{n} \sum_{i=1}^n \frac{(1 - Z_i)}{1 - \hat{e}(X_i)}$$

are two different estimates of the constant 1. I use the funny notation $\hat{1}_T$ and $\hat{1}_C$ because with the true propensity score these two terms both have expectation 1; see Problem 11.3. In general, $\hat{1}_T - \hat{1}_C$ is not zero in finite sample. Since adding a constant to every outcome should not change the average causal effect, this estimator is not reasonable because of its dependence

on c . A simple fix to the problem is to normalize the weights by $\hat{1}_T$ and $\hat{1}_C$ respectively, resulting in the following estimator

$$\hat{\tau}^{\text{hajek}} = \frac{\sum_{i=1}^n \frac{Z_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^n \frac{Z_i}{\hat{e}(X_i)}} - \frac{\sum_{i=1}^n \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}}{\sum_{i=1}^n \frac{1-Z_i}{1-\hat{e}(X_i)}}.$$

This is the Hajek estimator due to Hájek (1971). We can verify that the Hajek estimator is invariant to the location transformation, that is, if we replace Y_i by $Y_i + c$, then $\hat{\tau}^{\text{hajek}}$ remains the same. Moreover, many numerical studies have found that $\hat{\tau}^{\text{hajek}}$ is much more stable than $\hat{\tau}^{\text{ht}}$ in finite samples.

11.2.3 A problem of weighting and a fundamental problem of causal inference

In many asymptotic analysis, we require a *strong overlap* condition

$$0 < \alpha_L \leq e(X) \leq \alpha_U < 1,$$

that is, the true propensity score is bounded away from 0 and 1. However, D'Amour et al. (2021) pointed out that this is a rather strong assumption especially with many covariates. Chapter 20 will discuss this problem in detail.

Even if the strong overlap condition holds for the true propensity score, the estimated propensity scores can be close to 0 or 1. When this happens, the weighting estimators blow up to infinity resulting in extremely unstable behaviors in finite samples. We can either truncate the estimated propensity score by changing it to

$$\max \left[\alpha_L, \min \{ \hat{e}(X_i), \alpha_U \} \right],$$

or trim the observations by dropping units with $\hat{e}(X_i)$ outside the interval $[\alpha_L, \alpha_U]$. Crump et al. (2009) suggested $\alpha_L = 0.1$ and $\alpha_U = 0.9$, and Kurth et al. (2005) suggested $\alpha_L = 0.05$ and $\alpha_U = 0.95$. Yang and Ding (2018) established some asymptotic theory for trimming.

11.2.4 Application

Revisiting Example 10.3, we can obtain the weighting estimators based on different truncations of the the estimated propensity scores. The following results are the two weighting estimators with the bootstrap standard errors, with truncations at $(0, 1)$, $(0.01, 0.99)$, $(0.05, 0.95)$, and $(0.1, 0.9)$:

```
$trunc0
      HT   Hajek
est  -1.516 -0.156
se    0.495  0.238
```

```
$trunc.01
      HT  Hajek
est -1.516 -0.156
se   0.464  0.231

$trunc.05
      HT  Hajek
est -1.499 -0.152
se   0.472  0.248

$trunc.1
      HT  Hajek
est -0.713 -0.054
se   0.435  0.229
```

The HT estimator gives results far away from all other estimators we discussed so far. The point estimates seem too large and they are negatively significant unless we truncate the estimated propensity scores at (0.1, 0.9). This is an example showing the instability of the HT estimator.

11.3 The propensity score as a balancing score

11.3.1 Theory

Theorem 11.3 *The propensity score satisfies*

$$Z \perp\!\!\!\perp X \mid e(X).$$

Moreover, for any function $h(\cdot)$, we have

$$E \left\{ \frac{Zh(X)}{e(X)} \right\} = E \left\{ \frac{(1-Z)h(X)}{1-e(X)} \right\} \quad (11.2)$$

provided the existence of the moments on both sides of (11.2).

Rosenbaum and Rubin (1983b) also introduced the notion of *balancing score* $b(X)$, which satisfies $Z \perp\!\!\!\perp X \mid b(X)$. By Theorem 11.3, the propensity score is a balancing score. Theorem 11.3 also states that any function $h(X)$ of the covariates has the same mean across the treatment and control groups, if weighted by the inverse of the propensity score.

Moreover, Rosenbaum and Rubin (1983b) showed that the propensity score $e(X)$ is the coarsest balancing score, that is, the propensity score $e(X)$ is a function of any balancing score. Problem 11.5 gives more details.

Proof of Theorem 11.3: First, we show $Z \perp\!\!\!\perp X \mid e(X)$, that is,

$$\Pr\{Z = 1 \mid X, e(X)\} = \Pr\{Z = 1 \mid e(X)\}. \quad (11.3)$$

Following similar steps as the proof of Theorem 11.1, we can show that the left-hand side of (11.3) equals

$$\text{pr}\{Z = 1 \mid X, e(X)\} = \text{pr}(Z = 1 \mid X) = e(X),$$

and the right-hand side of (11.3) equals

$$\begin{aligned} \text{pr}\{Z = 1 \mid e(X)\} &= E\{Z \mid e(X)\} \\ &= E\left[E\{Z \mid X, e(X)\} \mid e(X)\right] \\ &= E\left[E\{Z \mid X\} \mid e(X)\right] \\ &= E\left[e(X) \mid e(X)\right] \\ &= e(X). \end{aligned}$$

Therefore, (11.3) holds.

Second, we show (11.2). We can use similar steps as the proof of Theorem 11.1. But given Theorem 11.1, we have a simpler proof. If we view $h(X)$ as an outcome, then its two potential outcomes are identical and the strong ignorability holds: $Z \perp\!\!\!\perp h(X) \mid X$. The difference between the left-hand and right-hand sides of (11.2) is the average causal effect of Z on $h(X)$, which is zero. \square

11.3.2 Covariate balance check

The proof of Theorem 11.3 is simple. But Theorem 11.3 has useful implications for the statistical analysis. Before getting access to the outcome data, we can check whether the propensity score model is specified well enough to ensure covariate balance in the data. Rubin (2007) viewed this as the design stage of the observational study, and Rubin (2008) argued that this can result in more objective causal inference because the design stage does not involve the values of the outcomes. While this is a useful recommendation in practice, it is not entirely clear how to quantify the objectiveness.

In propensity score stratification, we have the discretized estimated propensity score $\hat{e}'(X)$ and approximately

$$Z \perp\!\!\!\perp X \mid \hat{e}'(X) = e_k \quad (k = 1, \dots, K).$$

Therefore, we can check whether the covariate distributions are the same across the treatment and control groups within each stratum of the discretized estimated propensity score.

In propensity score weighting, we can view $h(X)$ as a pseudo outcome and estimate the average causal effect on $h(X)$. Because the true average causal effect on $h(X)$ is 0, the estimate should not be significantly different from 0. A canonical choice of $h(X)$ is X .

Let us revisit Example 10.3 again. Based on propensity score stratification

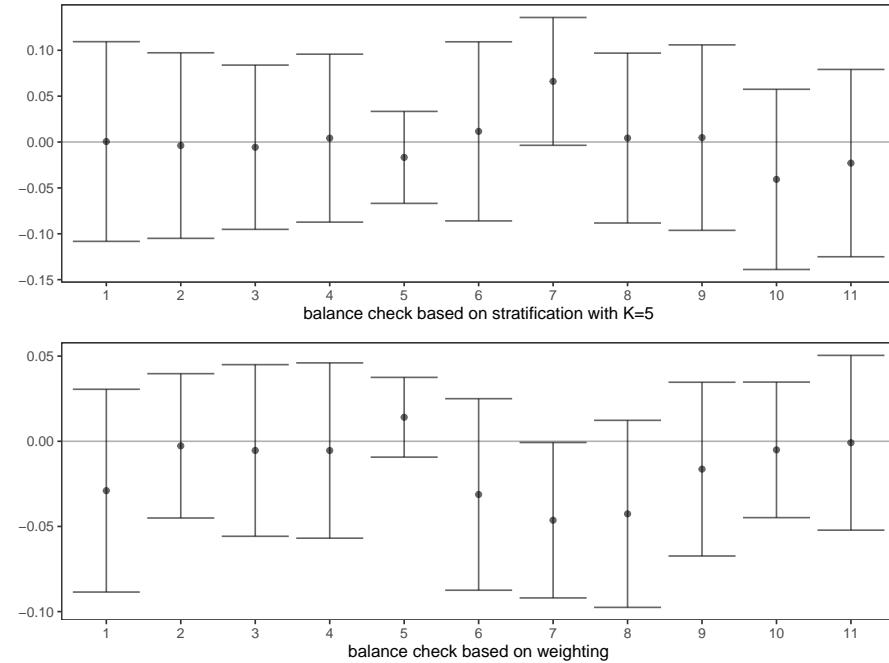


FIGURE 11.2: Balance check: point estimates and 95% confidence intervals of the average causal effect on covariates

with $K = 5$, all the covariates except `Food_Stamp` are well balanced across the treatment and control groups. Similar result holds for the Hajek estimator. Figure 11.2 shows the balance checking results.

11.4 Homework Problems

11.1 Another version of Theorem 11.1

Prove that

$$Z \perp\!\!\!\perp \{Y(1), Y(0), X\} \mid e(X, Y(1), Y(0)).$$

Remark: This result implies that

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid \{X, e(X, Y(1), Y(0))\}.$$

Rosenbaum (2020) and Rosenbaum and Rubin (2023) pointed out this result and called $e(X, Y(1), Y(0))$ the *principal unobserved covariate*.

11.2 Another version of Theorem 11.1

If $Z \perp\!\!\!\perp Y(z) \mid X$ for $z = 0, 1$, then $Z \perp\!\!\!\perp Y(z) \mid e(X)$ for $z = 0, 1$. That is, if ignorability holds conditional on covariates X , then it also holds conditional on the scalar propensity score $e(X)$. Prove this theorem.

11.3 More results on the IPW estimators

This is related to the discussion of the IPW estimators in Section 11.2.2.

Prove

$$E \left\{ \frac{1}{n} \sum_{i=1}^n \frac{Z_i}{e(X_i)} \right\} = 1, \quad E \left\{ \frac{1}{n} \sum_{i=1}^n \frac{(1-Z_i)}{1-e(X_i)} \right\} = 1.$$

11.4 Re-analysis of Rosenbaum and Rubin (1983a)

Use Table 1 of Rosenbaum and Rubin (1983a). If you are interested, you can read the whole paper. It is a canonical paper. But for this problem, you only need Table 1.

Rosenbaum and Rubin (1983a) fitted a logistic regression model for the propensity score and stratified the data into 5 subclasses. Because the treatment (Surgical versus Medical) is binary and the outcome is also binary (improved or not), they represented the data by a table.

Based on this table, estimate the average causal effect, and report the 95% confidence interval.

11.5 Balancing score and propensity score: more theoretical results

Rosenbaum and Rubin (1983b) defined $b(X)$ as a balancing score if $Z \perp\!\!\!\perp X \mid b(X)$. Here, $b(X)$ can be a scalar or a vector. An obvious balancing score is $b(X) = X$, but it is not a useful one without any simplification of the original covariates. By Theorem 11.3, the propensity score is a special balancing score. More interestingly, Rosenbaum and Rubin (1983b) showed that the propensity score is the coarsest balancing score, as in Theorem 11.4 below which includes Theorem 11.3 as a special case.

Theorem 11.4 $b(X)$ is a balancing score if and only if $b(X)$ is finer than $e(X)$ in the sense that $e(X) = f(b(X))$ for some function $f(\cdot)$.

Theorem 11.4 is relevant in subgroup analysis. In particular, we may be interested in not only the average causal effect τ but also the subgroup effects for boys and girls. Without loss of generality, assume the first component of X is the indicator for girls, and we can interested in estimating

$$\tau(x_1) = E\{Y(1) - Y(0) \mid X_1 = x_1\}, \quad (x_1 = 1, 0).$$

Theorem 11.4 implies that under ignorability,

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid e(X), X_1 \tag{11.4}$$

because $b(X) = \{e(X), X_1\}$ is finer than $e(X)$ and thus a balancing score. The conditional independence in (11.4) ensures ignorability holds given the propensity score, within each level of X_1 . Therefore, we can perform the same analysis based on the propensity score, within each level of X_1 , yielding estimates for two subgroup effects.

With the above motivation in mind, now prove Theorem 11.4.

11.6 Some basics of subgroup effects

This problem is related to Problem 11.5, but you can work on it independently.

Consider a standard observational study with covariates $X = (X_1, X_2)$, where X_1 denotes a binary subgroup indicator (e.g., statistics major or not statistics major) and X_2 contains the rest covariates. The parameter of interest is the subgroup causal effect

$$\tau(x_1) = E\{Y(1) - Y(0) \mid X_1 = x_1\}, \quad (x_1 = 1, 0).$$

Show that

$$\tau(x_1) = E \left\{ \frac{1(X_1 = x_1)ZY}{e(X)} - \frac{1(X_1 = x_1)(1-Z)Y}{1-e(X)} \right\} / \text{pr}(X_1 = x_1)$$

and give the corresponding Horvitz–Thompson and Hajek estimators for $\tau(x_1)$.

11.7 Recommended reading

The title of this chapter is the same as the title of the classic paper by Rosenbaum and Rubin (1983b). Most results in this chapter are directly drawn from their original paper.

Rubin (2007) and Rubin (2008) highlighted the importance of the design stage of observational studies for more objective causal inference



12

The Doubly Robust or the Augmented Inverse Propensity Score Weighting Estimator for the Average Causal Effect

Under unconfoundedness $Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$ and overlap $0 < e(X) < 1$, Chapter 11 has shown two identification formulas of the average causal effect $\tau = E\{Y(1) - Y(0)\}$. First, the outcome imputation formula is

$$\tau = E\{\mu_1(X)\} - E\{\mu_0(X)\} \quad (12.1)$$

where

$$\begin{aligned}\mu_1(X) &= E\{Y(1) \mid X\} = E(Y \mid Z = 1, X), \\ \mu_0(X) &= E\{Y(0) \mid X\} = E(Y \mid Z = 0, X)\end{aligned}$$

are the two conditional mean functions of the outcome given covariates. Second, the inverse propensity score weighting (IPW) formula is

$$\tau = E\left\{\frac{ZY}{e(X)}\right\} - E\left\{\frac{(1-Z)Y}{1-e(X)}\right\} \quad (12.2)$$

where

$$e(X) = \text{pr}(Z = 1 \mid X)$$

is the propensity score introduced in Chapter 11.

The outcome imputation estimator requires fitting a model for the outcome given the treatment and covariates. It is consistent if the outcome model is correctly specified. The IPW estimator requires fitting a model for the treatment given covariates. It is consistent if the propensity score model is correctly specified.

Mathematically, we have many combinations of (12.1) and (12.2) that lead to different identification formulas of the average causal effect. Below I will discuss a particular combination that has appealing theoretical properties. This combination motivates an estimator that is consistent if either the propensity score or the outcome model is correctly specified. It is called the *doubly robust* estimator, championed by James Robins (Scharfstein et al., 1999; Bang and Robins, 2005).

12.1 The doubly robust estimator

12.1.1 Population version

We posit a *working model* for the conditional means of the outcome $\mu_1(X, \beta_1)$ and $\mu_0(X, \beta_0)$, indexed by the parameters β_1 and β_0 . For example, if the conditional means are linear or logistic under the working model, then the parameters are just the regression coefficients. If the outcome model is correctly specified, then $\mu_1(X, \beta_1) = \mu_1(X)$ and $\mu_0(X, \beta_0) = \mu_0(X)$. We posit a working model for the propensity score $e(X, \alpha)$, indexed by the parameter α . For example, if the working model is logistic, then α is the regression coefficient. If the propensity score model is correctly specified, then $e(X, \alpha) = e(X)$. In practice, both models may be misspecified.

Define

$$\tilde{\mu}_1^{\text{dr}} = E \left[\frac{Z\{Y - \mu_1(X, \beta_1)\}}{e(X, \alpha)} + \mu_1(X, \beta_1) \right], \quad (12.3)$$

$$\tilde{\mu}_0^{\text{dr}} = E \left[\frac{(1-Z)\{Y - \mu_0(X, \beta_0)\}}{1 - e(X, \alpha)} + \mu_0(X, \beta_0) \right], \quad (12.4)$$

which can also be written as

$$\tilde{\mu}_1^{\text{dr}} = E \left[\frac{ZY}{e(X, \alpha)} - \frac{Z - e(X, \alpha)}{e(X, \alpha)} \mu_1(X, \beta_1) \right], \quad (12.5)$$

$$\tilde{\mu}_0^{\text{dr}} = E \left[\frac{(1-Z)Y}{1 - e(X, \alpha)} - \frac{e(X, \alpha) - Z}{1 - e(X, \alpha)} \mu_0(X, \beta_0) \right]. \quad (12.6)$$

The formulas in (12.3) and (12.4) augment the outcome imputation estimator by inverse propensity score weighting terms of the residuals. The formulas in (12.5) and (12.6) augment the IPW estimator by the imputed outcomes. For this reason, the doubly robust estimator is also called the *augmented inverse propensity score weighting* (AIPW) estimator.

The augmentation strengthens the theoretical properties in the following sense.

Theorem 12.1 *Assume unconfoundedness $Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$ and overlap $1 < e(X) < 1$.*

1. *If either $e(X, \alpha) = e(X)$ or $\mu_1(X, \beta_1) = \mu_1(X)$, then $\tilde{\mu}_1^{\text{dr}} = E\{Y(1)\}$.*
2. *If either $e(X, \alpha) = e(X)$ or $\mu_0(X, \beta_0) = \mu_0(X)$, then $\tilde{\mu}_0^{\text{dr}} = E\{Y(0)\}$.*
3. *If either $e(X, \alpha) = e(X)$ or $\{\mu_1(X, \beta_1) = \mu_1(X), \mu_0(X, \beta_0) = \mu_0(X)\}$, then $\tilde{\mu}_1^{\text{dr}} - \tilde{\mu}_0^{\text{dr}} = \tau$.*

By Theorem 12.1, $\tilde{\mu}_1^{\text{dr}} - \tilde{\mu}_0^{\text{dr}}$ equals τ if either the propensity score model or the outcome model is correctly specified. That's why it is called the doubly robust estimator.

Proof of Theorem 12.1: I only prove the result for $\mu_1 = E\{Y(1)\}$. The proof for the result for $\mu_0 = E\{Y(0)\}$ is similar. We have the decomposition

$$\begin{aligned}\tilde{\mu}_1^{\text{dr}} - E\{Y(1)\} &= E\left[\frac{Z\{Y(1) - \mu_1(X, \beta_1)\}}{e(X, \alpha)} - \{Y(1) - \mu_1(X, \beta_1)\}\right] \\ &= E\left[\frac{Z - e(X, \alpha)}{e(X, \alpha)}\{Y(1) - \mu_1(X, \beta_1)\}\right] \\ &= E\left(E\left[\frac{Z - e(X, \alpha)}{e(X, \alpha)}\{Y(1) - \mu_1(X, \beta_1)\} \mid X\right]\right) \\ &= E\left[E\left\{\frac{Z - e(X, \alpha)}{e(X, \alpha)} \mid X\right\} \times E\{Y(1) - \mu_1(X, \beta_1) \mid X\}\right] \\ &= E\left[\frac{e(X) - e(X, \alpha)}{e(X, \alpha)} \times \{\mu_1(X) - \mu_1(X, \beta_1)\}\right].\end{aligned}$$

Therefore, $\tilde{\mu}_1^{\text{dr}} - E\{Y(1)\} = 0$ if either $e(X, \alpha) = e(X)$ or $\mu_1(X, \beta_1) = \mu_1(X)$. \square

12.1.2 Sample version

From the population versions of $\tilde{\mu}_1^{\text{dr}}$ and $\tilde{\mu}_0^{\text{dr}}$, we can construct the sample versions by the following steps:

1. obtain the fitted values of the propensity scores: $e(X, \hat{\alpha})$;
2. obtain the fitted values of the outcome means: $\mu_1(X, \hat{\beta}_1)$ and $\mu_0(X, \hat{\beta}_0)$;
3. construct the doubly robust estimator: $\hat{\tau}^{\text{dr}} = \hat{\mu}_1^{\text{dr}} - \hat{\mu}_0^{\text{dr}}$, where

$$\hat{\mu}_1^{\text{dr}} = \frac{1}{n} \sum_{i=1}^n \left[\frac{Z_i\{Y_i - \mu_1(X_i, \hat{\beta}_1)\}}{e(X_i, \hat{\alpha})} + \mu_1(X_i, \hat{\beta}_1) \right]$$

and

$$\hat{\mu}_0^{\text{dr}} = \frac{1}{n} \sum_{i=1}^n \left[\frac{(1 - Z_i)\{Y_i - \mu_0(X_i, \hat{\beta}_0)\}}{1 - e(X_i, \hat{\alpha})} + \mu_0(X_i, \hat{\beta}_0) \right];$$

4. approximate the variance of $\hat{\tau}^{\text{dr}}$ via the nonparametric bootstrap by resampling from $(Z_i, X_i, Y_i)_{i=1}^n$ (Funk et al., 2011).

Analogous to (12.5) and (12.6), we can also rewrite $\hat{\mu}_1^{\text{dr}}$ and $\hat{\mu}_0^{\text{dr}}$ as

$$\begin{aligned}\hat{\mu}_1^{\text{dr}} &= \frac{1}{n} \sum_{i=1}^n \left[\frac{Z_i Y_i}{e(X_i, \hat{\alpha})} - \frac{Z_i - e(X_i, \hat{\alpha})}{e(X_i, \hat{\alpha})} \mu_1(X_i, \hat{\beta}_1) \right], \\ \hat{\mu}_0^{\text{dr}} &= \frac{1}{n} \sum_{i=1}^n \left[\frac{(1 - Z_i) Y_i}{1 - e(X_i, \hat{\alpha})} - \frac{e(X_i, \hat{\alpha}) - Z_i}{1 - e(X_i, \hat{\alpha})} \mu_0(X_i, \hat{\beta}_0) \right].\end{aligned}$$

12.2 More intuition and theory for the doubly robust estimator

Although the beginning of this chapter claims that the basic identification formulas based on outcome regression and inverse propensity score weight immediately yield infinitely many other identification formulas, the particular forms of the double robust estimators in (12.3) and (12.4) are not obvious to come up with. The original motivation for (12.3) and (12.4) was quite theoretical, which relies on the *semiparametric efficiency theory* in advanced mathematical statistics (Bickel et al., 1993). It is beyond the level of this book. Below I will give two more intuitive perspectives to construct (12.3) and (12.4). Both Sections 12.2.1 and 12.2.2 below focus on the estimation of $E\{Y(1)\}$ since the estimation of $E\{Y(0)\}$ is similar by symmetry.

12.2.1 Reducing the variance of the IPW estimator

The IPW estimator for μ_1 based on

$$\mu_1 = E \left\{ \frac{ZY}{e(X)} \right\}$$

completely ignores the outcome model of Y . It has the advantages of being consistent without assuming any outcome model. However, if the covariates are predictive to the outcome, the residual based on a working outcome model usually has smaller variance than the outcome even if this working outcome model is wrong. With a possibly mis-specified outcome model $\mu_1(X, \beta_1)$, a trivial decomposition holds:

$$\mu_1 = E\{Y(1)\} = E\{Y(1) - \mu_1(X, \beta_1)\} + E\{\mu_1(X, \beta_1)\}.$$

If we apply the IPW formula to the first term in the above formula viewing $Y(1) - \mu_1(X, \beta_1)$ as a pseudo potential outcome under the treatment, we can rewrite the above formula as

$$\mu_1 = E \left\{ \frac{Z\{Y - \mu_1(X, \beta_1)\}}{e(X)} \right\} + E\{\mu_1(X, \beta_1)\} \quad (12.7)$$

$$= E \left\{ \frac{Z\{Y - \mu_1(X, \beta_1)\}}{e(X)} + \mu_1(X, \beta_1) \right\}, \quad (12.8)$$

which holds if the propensity score model is correct without assuming that the outcome model is correct. Using a working model to improve efficiency is an old idea from survey sampling. Little and An (2004) and Lumley et al. (2011) pointed out its connection with the doubly robust estimator.

12.2.2 Reducing the bias of the outcome regression estimator

The discussion in Section 12.2.1 starts with the IPW estimator and improves its efficiency based on a working outcome model. Alternatively, we can also start with an outcome regression estimator based on

$$\tilde{\mu}_1 = E\{\mu_1(X, \beta_1)\}$$

which may not be the same as μ_1 since the outcome may be wrong. The bias of this estimator is $E\{\mu_1(X, \beta_1) - Y(1)\}$, which can be estimated by an IPW estimator

$$B = E \left\{ \frac{Z\{\mu_1(X, \beta_1) - Y\}}{e(X)} \right\}$$

if the propensity score model is correct. So a de-biased estimator is $\tilde{\mu}_1 - B$, which is identical to (12.8).

12.3 Examples

12.3.1 Summary of some canonical estimators for τ

The following R implements the outcome imputation, Hovitz–Thompson, Hajek, and doubly robust estimators for τ . These estimators can be conveniently implemented based on the fitted values of the `glm` function. The default choice for the propensity score model is the logistic model, and the default choice for the outcome model is the linear model with `out.family = gaussian`¹. For binary outcomes, we can also specify `out.family = binomial` to fit the logistic model.

```
OS_est = function(z, y, x, out.family = gaussian,
                  truncpscore = c(0, 1))
{
  ## fitted propensity score
  pscore    = glm(z ~ x, family = binomial)$fitted.values
  pscore    = pmax(truncpscore[1], pmin(truncpscore[2], pscore))
```

¹The `glm` function is more general than the `lm` function. With `out.family = gaussian`, `glm` is identical to `lm`.

```

## fitted potential outcomes
outcome1 = glm(y ~ x, weights = z,
               family = out.family)$fitted.values
outcome0 = glm(y ~ x, weights = (1 - z),
               family = out.family)$fitted.values

## regression imputation estimator
ace.reg = mean(outcome1 - outcome0)
## IPW estimators
ace.ipw0 = mean(z*y/pscore - (1 - z)*y/(1 - pscore))
ace.ipw = mean(z*y/pscore)/mean(z/pscore) -
            mean((1 - z)*y/(1 - pscore))/mean((1 - z)/(1 - pscore))
## doubly robust estimator
res1 = y - outcome1
res0 = y - outcome0
ace.dr = ace.reg + mean(z*res1/pscore - (1 - z)*res0/(1 - pscore))

return(c(ace.reg, ace.ipw0, ace.ipw, ace.dr))
}

```

It is tedious to calculate the analytic formulas for the variances of the above estimators. The bootstrap provides convenient approximations for the variances based on resampling from $\{Z_i, X_i, Y_i\}_{i=1}^n$. Building upon `OS_est`, the following function returns point estimators as well as the bootstrap standard errors.

```

OS_ATE = function(z, y, x, n.boot = 2*10^2,
                  out.family = gaussian, truncpscore = c(0, 1))
{
  point.est = OS_est(z, y, x, out.family, truncpscore)

  ## nonparametric bootstrap
  n.sample = length(z)
  x = as.matrix(x)
  boot.est = replicate(n.boot,
    {id.boot = sample(1:n.sample, n.sample, replace = TRUE)
     OS_est(z[id.boot], y[id.boot], x[id.boot, ],
            out.family, truncpscore)})

  boot.se = apply(boot.est, 1, sd)

  res = rbind(point.est, boot.se)
  rownames(res) = c("est", "se")
  colnames(res) = c("reg", "HT", "Hajek", "DR")

  return(res)
}

```

12.3.2 Simulation

I will use simulation to evaluate the finite-sample properties of the estimators under four scenarios:

1. both the propensity score and outcome models are correct;
2. the propensity score model is wrong but the outcome model is correct;
3. the propensity score model is correct but the outcome model is wrong;
4. both the propensity score and outcome models are wrong.

I will report the average bias, the true standard error, and the average estimated standard error of the estimators over simulation.

In case 1, the data generating process is

```
x      = matrix(rnorm(n*2), n, 2)
x1    = cbind(1, x)
beta.z = c(0, 1, 1)
pscore = 1/(1 + exp(- as.vector(x1%*%beta.z)))
z     = rbinom(n, 1, pscore)
beta.y1 = c(1, 2, 1)
beta.y0 = c(1, 2, 1)
y1    = rnorm(n, x1%*%beta.y1)
y0    = rnorm(n, x1%*%beta.y0)
y     = z*y1 + (1 - z)*y0
```

In case 2, I modify the propensity score model to be nonlinear:

```
x1      = cbind(1, x, exp(x))
beta.z = c(-1, 0, 0, 1, -1)
pscore = 1/(1 + exp(- as.vector(x1%*%beta.z)))
```

In case 3, I modify the outcome model to be nonlinear:

```
beta.y1 = c(1, 0, 0, 0.2, -0.1)
beta.y0 = c(1, 0, 0, -0.2, 0.1)
y1     = rnorm(n, x1%*%beta.y1)
y0     = rnorm(n, x1%*%beta.y0)
```

In case 4, I modify both the propensity score and the outcome model.

We set the sample size to be $n = 500$ and generate 500 independent data sets according to the data generating processes above. In case 1,

	reg	HT	Hajek	DR
ave.bias	0.00	0.02	0.03	0.01
true.se	0.11	0.28	0.26	0.13
est.se	0.10	0.25	0.23	0.12

All estimators are nearly unbiased. The two weighting estimators have larger variances. In case 2,

	reg	HT	Hajek	DR
ave.bias	0.00	-0.76	-0.75	-0.01
true.se	0.12	0.59	0.47	0.18
est.se	0.13	0.50	0.38	0.18

The two weighting estimators are severely biased due to the misspecification of the propensity score model. The regression imputation and doubly robust estimators are nearly unbiased. In case 3,

	reg	HT	Hajek	DR
ave.bias	-0.05	0.00	-0.01	0.00
true.se	0.11	0.15	0.14	0.14
est.se	0.11	0.14	0.13	0.14

The regression imputation estimator has larger bias than the other three estimators due to the misspecification of the outcome model. The weighting and doubly robust estimators are nearly unbiased. In case 4,

	reg	HT	Hajek	DR
ave.bias	-0.08	0.11	-0.07	0.16
true.se	0.13	0.32	0.20	0.41
est.se	0.13	0.25	0.16	0.26

All estimators are biased because both the propensity score and outcome models are wrong. The Horvitz–Thompson and doubly robust estimator has the largest bias. When both models are wrong, the doubly robust estimator appears to be doubly fragile.

In all the cases above, the bootstrap standard errors are close to the true ones when the estimators are nearly unbiased for the true average causal effect.

12.3.3 Applications

Revisiting Example 10.3, we obtain the following estimators and bootstrap standard errors:

	reg	HT	Hajek	DR
est	-0.017	-1.516	-0.156	-0.019
se	0.230	0.492	0.246	0.233

The two weighting estimators are much larger than the other two estimators. Truncating the estimated propensity score at [0.1, 0.9], we obtain the following estimators and bootstrap standard errors:

	reg	HT	Hajek	DR
est	-0.017	-0.713	-0.054	-0.043
se	0.223	0.422	0.235	0.231

The Hajek estimator becomes much close to the regression imputation and doubly robust estimators, while the Horvitz–Thompson estimator is still an outlier.

12.4 Some further discussion

Recall the proof of Theorem 12.1, the key for the double robustness property is the product structure in

$$\tilde{\mu}_1^{\text{dr}} - E\{Y(1)\} = E \left[\frac{e(X) - e(X, \alpha)}{e(X, \alpha)} \times \{\mu_1(X) - \mu_1(X, \beta_1)\} \right],$$

which ensures that the estimation error is zero if either $e(X) = e(X, \alpha)$ or $\mu_1(X) = \mu_1(X, \beta_1)$. This delicate structure renders the doubly robust estimator possibly doubly fragile when both the propensity score and the outcome models are misspecified. The product of two errors multiply to yield potentially much larger errors. Kang and Schafer (2007) criticized the doubly robust estimator based on extensive simulation studies. They found that the finite-sample performance of the doubly robust estimator can be even more wild than the simple regression imputation and IPW estimators.

Despite the critique from Kang and Schafer (2007), the doubly robust estimator has been a standard strategy in causal since the seminal work of Scharfstein et al. (1999). Recently, it resurrected in the theoretical statistics and econometrics literature with a fancier name “double machine learning” (Chernozhukov et al., 2018). The basic idea is to replace the working models for the propensity score and outcome by machine learning tools which can be viewed as more flexible models than the traditional parametric models.

12.5 Homework problems

12.1 A sanity check

Consider the case in which the covariate is discrete $X \in \{1, \dots, K\}$ and the parameter of interest is μ_1 . Without imposing any model assumptions, the estimated propensity score $\hat{e}(X)$ is the proportion of units receiving the treatment and the estimated outcome mean is the sample mean of the outcome $\hat{Y}_{[k]1} = \hat{E}(Y \mid Z = 1, X = k)$ under treatment, within stratum $X = k$ ($k = 1, \dots, K$). Show that the stratified estimator, outcome regression estimator, IPW estimator, and the doubly robust estimator are all the same.

12.2 An alternative form of the doubly robust estimator for τ

Motivated by (12.7), we have an alternative form of doubly robust estimator for μ_1 :

$$\tilde{\mu}_1^{\text{dr2}} = \frac{E \left[\frac{Z\{Y - \mu_1(X, \beta_1)\}}{e(X, \alpha)} \right]}{E \left[\frac{Z}{e(X, \alpha)} \right]} + E\{\mu_1(X, \beta_1)\}.$$

Show that $\tilde{\mu}_1^{\text{dr2}} = \mu_1$ if either $e(X, \alpha) = e(X)$ or $\mu_1(X, \beta_1) = \mu_1(X)$. Give the analogous formula for estimating μ_0 . Give the sample analogue of the doubly robust estimator for τ based on these formulas. Note that this form of doubly robust estimator appeared in Robins et al. (2007).

12.3 Data analysis of Example 10.1

Analyze the dataset `cps1re74.csv` using the methods discussed so far.

12.4 Recommended reading

Lunceford and Davidian (2004) gave a nice review and comparison of many methods discussed in Chapters 11 and 12.

13

The Average Causal Effect on the Treated Units and Other Estimands

Chapters 10–12 focused on the identification and estimation of the average causal effect $\tau = E\{Y(1) - Y(0)\}$ under the unconfoundedness and overlap assumptions. Conceptually, it is straightforward to extend the discussion to the average causal effects on the treated and control units:

$$\begin{aligned}\tau_T &= E\{Y(1) - Y(0) \mid Z = 1\}, \\ \tau_C &= E\{Y(1) - Y(0) \mid Z = 0\}.\end{aligned}$$

Because of the symmetry, this chapter focuses on τ_T and also included extensions to other estimands.

13.1 Nonparametric identification of τ_T

The average causal effect on the treated units equals

$$\tau_T = E(Y \mid Z = 1) - E\{Y(0) \mid Z = 1\},$$

where the first term $E(Y \mid Z = 1)$ is directly identifiable from the data and the second term $E\{Y(0) \mid Z = 1\}$ is counterfactual. The key assumption to identify the second term is the following unconfoundedness and overlap assumptions.

Assumption 13.1 $Z \perp\!\!\!\perp Y(0) \mid X$ and $e(X) < 1$.

Because the key is to identify $E\{Y(0) \mid Z = 1\}$, we only need the “one-sided” unconfoundedness and overlap assumptions. Under Assumption 13.1, we have the following identification result for τ_T .

Theorem 13.1 *Under Assumption 13.1, we have*

$$\begin{aligned}E\{Y(0) \mid Z = 1\} &= E\{E(Y \mid Z = 0, X) \mid Z = 1\} \\ &= \int E(Y \mid Z = 0, X = x)F(dx \mid Z = 1).\end{aligned}$$

Theorem 13.1 implies that τ_T is nonparametrically identified by

$$\tau_T = E(Y | Z = 1) - E\{E(Y | Z = 0, X) | Z = 1\} \quad (13.1)$$

Proof of Theorem 13.1: We have

$$\begin{aligned} E\{Y(0) | Z = 1\} &= E[E\{Y(0) | Z = 1, X\} | Z = 1] \\ &= E[E\{Y(0) | Z = 0, X\} | Z = 1] \\ &= E\{E(Y | Z = 0, X) | Z = 1\} \\ &= \int E(Y | Z = 0, X = x)F(dx | Z = 1). \end{aligned}$$

□

With a discrete X , the identification formula in Theorem 13.1 reduces to

$$E\{Y(0) | Z = 1\} = \sum_{k=1}^K E(Y | Z = 0, X = k)\text{pr}(X = k | Z = 1),$$

motivating the following stratified estimator for τ_T :

$$\hat{\tau}_T = \hat{Y}(1) - \sum_{k=1}^K \hat{\pi}_{[k]1} \hat{Y}_{[k]}(0),$$

where $\hat{\pi}_{[k]1} = n_{[k]1}/n_1$ is the proportion of category k of X among the treated units.

For continuous X , we need to fit an outcome model for $E(Y | Z = 0, X)$ using the control units. If the fitted values for the control potential outcomes are $\hat{\mu}_0(X_i)$, then the outcome regression estimator is

$$\hat{\tau}_T = \hat{Y}(1) - n_1^{-1} \sum_{i=1}^n Z_i \hat{\mu}_0(X_i) = n_1^{-1} \sum_{i=1}^n Z_i \{Y_i - \hat{\mu}_0(X_i)\}.$$

Example 13.1 If we specify a linear model for all units

$$E(Y | Z, X) = \beta_0 + \beta_z Z + \beta_x^T X,$$

then

$$\begin{aligned} \tau_T &= E(Y | Z = 1) - E(\beta_0 + \beta_x^T X | Z = 1) \\ &= E(Y | Z = 1) - \beta_0 - \beta_x^T E(X | Z = 1). \end{aligned}$$

If we run OLS to obtain $(\hat{\beta}_0, \hat{\beta}_z, \hat{\beta}_x)$, then the estimator is

$$\hat{\tau}_T = \hat{Y}(1) - \hat{\beta}_0 - \hat{\beta}_x^T \hat{X}(1).$$

Using the property of the OLS (see A2.3), we have

$$\sum_{i=1}^n Z_i(Y_i - \hat{\beta}_0 - \hat{\beta}_z Z_i - \hat{\beta}_x^\top X_i) = 0 \implies \hat{Y}(1) - \hat{\beta}_0 - \hat{\beta}_z - \hat{\beta}_x^\top \hat{X}(1) = 0.$$

Therefore, the above estimator reduces to $\hat{\tau}_T = \hat{\beta}_z$, the OLS coefficient of Z .

By the property of the OLS, we can also write $\hat{\beta}_z$ as the difference in means of the adjusted outcome $\hat{Y}_i - \hat{\beta}_x^\top X_i$, resulting in

$$\begin{aligned}\hat{\tau}_T &= \left\{ \hat{Y}(1) - \hat{\beta}_x^\top \hat{X}(1) \right\} - \left\{ \hat{Y}(0) - \hat{\beta}_x^\top \hat{X}(0) \right\} \\ &= \left\{ \hat{Y}(1) - \hat{Y}(0) \right\} - \hat{\beta}_x^\top \left\{ \hat{X}(1) - \hat{X}(0) \right\}.\end{aligned}\quad (13.2)$$

Therefore, $\hat{\tau}_T$ equals the simple difference in means of the outcome, adjusted by the imbalance of the covariates in the treatment and control groups.

Section 10.4.2 shows that $\hat{\beta}_z$ is an estimator for τ , and this example further shows that $\hat{\beta}_z$ is an estimator for τ_T . This is not surprising because the linear model assumes constant causal effects across units.

Example 13.2 The identification formula depends only on $E(Y | Z = 0, X)$, so we need only to specify a model for the control units. When this model is linear,

$$E(Y | Z = 0, X) = \beta_{0|0} + \beta_{x|0}^\top X,$$

we have

$$\begin{aligned}\tau_T &= E(Y | Z = 1) - E(\beta_{0|0} + \beta_{x|0}^\top X | Z = 1) \\ &= E(Y | Z = 1) - \beta_{0|0} - \beta_{x|0}^\top E(X | Z = 1).\end{aligned}$$

If we run OLS with only the control units to obtain $(\hat{\beta}_{0|0}, \hat{\beta}_{x|0})$, then the estimator is

$$\hat{\tau}_T = \hat{Y}(1) - \hat{\beta}_{0|0} - \hat{\beta}_{x|0}^\top \hat{X}(1).$$

Using the property of the OLS (see A2.3), we have

$$\hat{Y}(0) = \hat{\beta}_{0|0} + \hat{\beta}_{x|0}^\top \hat{X}(0).$$

Therefore, the above estimator reduces to

$$\hat{\tau}_T = \left\{ \hat{Y}(1) - \hat{Y}(0) \right\} - \hat{\beta}_{x|0}^\top \left\{ \hat{X}(1) - \hat{X}(0) \right\},$$

which is similar to (13.2) with a different coefficient for the difference in means of the covariates.

As an algebraic fact, we can show that this estimator equals the coefficient of Z in the OLS fit of the outcome on the treatment, covariates, and their interactions, with the covariates centered by $\hat{X}(1)$. See Problem 13.1 for more details.

13.2 Inverse propensity score weighting and doubly robust estimation of τ_T

Theorem 13.2 Under Assumption 13.1, we have

$$E\{Y(0) | Z = 1\} = E \left\{ \frac{e(X)}{e} \frac{1 - Z}{1 - e(X)} Y \right\} \quad (13.3)$$

and

$$\tau_T = E(Y | Z = 1) - E \left\{ \frac{e(X)}{e} \frac{1 - Z}{1 - e(X)} Y \right\}, \quad (13.4)$$

where $e = \text{pr}(Z = 1)$ is the marginal probability of the treatment.

Proof of Theorem 13.2: The left-hand side of (13.3) equals

$$\begin{aligned} E\{Y(0) | Z = 1\} &= E\{ZY(0)\}/e \\ &= E[E(Z | X)E\{Y(0) | X\}]/e \\ &= E[e(X)E\{Y(0) | X\}]/e. \end{aligned}$$

The right-hand side of (13.3) equals

$$\begin{aligned} E \left\{ \frac{e(X)}{e} \frac{1 - Z}{1 - e(X)} Y \right\} &= E \left[E \left\{ \frac{e(X)}{e} \frac{1 - Z}{1 - e(X)} Y(0) | X \right\} \right] \\ &= E \left[\frac{e(X)}{e\{1 - e(X)\}} E \{(1 - Z)Y(0) | X\} \right] \\ &= E \left[\frac{e(X)}{e\{1 - e(X)\}} E(1 - Z | X) E\{Y(0) | X\} \right] \\ &= E[e(X)E\{Y(0) | X\}]/e. \end{aligned}$$

So (13.3) holds. \square

We have two inverse propensity score weighting estimators

$$\hat{\tau}_T^{\text{ht}} = \hat{Y}(1) - n_1^{-1} \sum_{i=1}^n \hat{o}(X_i)(1 - Z_i)Y_i$$

and

$$\hat{\tau}_T^{\text{hajek}} = \hat{Y}(1) - \frac{\sum_{i=1}^n \hat{o}(X_i)(1 - Z_i)Y_i}{\sum_{i=1}^n \hat{o}(X_i)(1 - Z_i)},$$

where $\hat{o}(X_i) = \hat{e}(X_i)/\{1 - \hat{e}(X_i)\}$ is the fitted odds of the treatment given covariates.

The estimation of $E(Y | Z = 1)$ is simple. We have a doubly robust

estimator for $E\{Y(0) | Z = 1\}$ which combines the propensity score and the outcome model. Define

$$\tilde{\mu}_{0T}^{\text{dr}} = E[o(X, \alpha)(1 - Z)\{Y - \mu_0(X, \beta_0)\} + Z\mu_0(X, \beta_0)] / e, \quad (13.5)$$

where $o(X, \alpha) = e(X, \alpha) / \{1 - e(X, \alpha)\}$.

Theorem 13.3 *Under Assumption 13.1, if either $e(X, \alpha) = e(X)$ or $\mu_0(X, \beta_0) = \mu_0(X)$, then $\mu_{0T}^{\text{dr}} = E\{Y(0) | Z = 1\}$.*

Proof of Theorem 13.3: We have the decomposition

$$\begin{aligned} & e[\tilde{\mu}_{0T}^{\text{dr}} - E\{Y(0) | Z = 1\}] \\ &= E[o(X, \alpha)(1 - Z)\{Y(0) - \mu_0(X, \beta_0)\} + Z\mu_0(X, \beta_0)] - E\{ZY(0)\} \\ &= E[o(X, \alpha)(1 - Z)\{Y(0) - \mu_0(X, \beta_0)\} - Z\{Y(0) - \mu_0(X, \beta_0)\}] \\ &= E[\{o(X, \alpha)(1 - Z) - Z\}\{Y(0) - \mu_0(X, \beta_0)\}] \\ &= E\left[\frac{e(X, \alpha) - Z}{1 - e(X, \alpha)}\{Y(0) - \mu_0(X, \beta_0)\}\right] \\ &= E\left[E\left\{\frac{e(X, \alpha) - Z}{1 - e(X, \alpha)} | X\right\} \times E\{Y(0) - \mu_0(X, \beta_0) | X\}\right] \\ &= E\left[\frac{e(X, \alpha) - e(X)}{1 - e(X, \alpha)} \times \{\mu_0(X) - \mu_0(X, \beta_0)\}\right]. \end{aligned}$$

Therefore, $\tilde{\mu}_{0T}^{\text{dr}} - E\{Y(0) | Z = 1\} = 0$ if either $e(X, \alpha) = e(X)$ or $\mu_0(X, \beta_0) = \mu_0(X)$. \square

From the population versions of $\tilde{\mu}_{0T}^{\text{dr}}$, we can construct the sample version by the following steps:

1. obtain the fitted values of the propensity scores $e(X, \hat{\alpha})$;
2. obtain the fitted values of the outcome mean under control $\mu_0(X, \hat{\beta}_0)$;
3. construct the doubly robust estimator: $\hat{\tau}_T^{\text{dr}} = \hat{Y}(1) - \hat{\mu}_{0T}^{\text{dr}}$, where

$$\hat{\mu}_{0T}^{\text{dr}} = \frac{1}{n_1} \sum_{i=1}^n \left[e(X_i, \hat{\alpha}) \frac{(1 - Z_i)\{Y_i - \mu_0(X_i, \hat{\beta}_0)\}}{1 - e(X_i, \hat{\alpha})} + Z_i \mu_0(X_i, \hat{\beta}_0) \right];$$

4. estimate the variance of τ_T via the bootstrap by resampling from $(Z_i, X_i, Y_i)_{i=1}^n$.

Hahn (1998), Mercatanti and Li (2014), Shinozaki and Matsuyama (2015) and Yang and Ding (2018) are references discussing the estimation of τ_T .

13.3 An example

The following R code implements two outcome regression estimators, two IPW estimators, and the doubly robust estimator for τ_T , as well as the bootstrap variance estimators. To avoid extreme estimated propensity scores, we can also truncated them from the above.

```

ATT.est = function(z, y, x, out.family = gaussian, Utruncpscore = 1)
{
  ## sample size
  nn = length(z)
  nn1 = sum(z)

  ## fitted propensity score
  pscore = glm(z ~ x, family = binomial)$fitted.values
  pscore = pmin(Utruncpscore, pscore)
  odds.pscore = pscore/(1 - pscore)

  ## fitted potential outcomes
  outcome0 = glm(y ~ x, weights = (1 - z),
                  family = out.family)$fitted.values

  ## regression imputation estimator
  ace.reg0 = lm(y ~ z + x)$coef[2]
  ace.reg = mean(y[z==1]) - mean(outcome0[z==1])
  ## propensity score weighting estimator
  ace.ipw0 = mean(y[z==1]) -
    mean(odds.pscore*(1 - z)*y)*nn/nn1
  ace.ipw = mean(y[z==1]) -
    mean(odds.pscore*(1 - z)*y)/mean(odds.pscore*(1 - z))
  ## doubly robust estimator
  res0 = y - outcome0
  ace.dr = ace.reg - mean(odds.pscore*(1 - z)*res0)*nn/nn1

  return(c(ace.reg0, ace.reg, ace.ipw0, ace.ipw, ace.dr))
}

OS_ATT = function(z, y, x, n.boot = 10^2,
                   out.family = gaussian, Utruncpscore = 1)
{
  point.est = ATT.est(z, y, x, out.family, Utruncpscore)

  ## nonparametric bootstrap
  n.sample = length(z)
  x = as.matrix(x)
  boot.est = replicate(n.boot,
    {id.boot = sample(1:n.sample, n.sample, replace = TRUE)}
  )
}
```

```

ATT.est(z[id.boot], y[id.boot], x[id.boot, ],
        out.family, Utruncpscore}))}

boot.se      = apply(boot.est, 1, sd)

res          = rbind(point.est, boot.se)
rownames(res) = c("est", "se")
colnames(res) = c("reg0", "reg", "HT", "Hajek", "DR")

return(res)
}

```

Now we re-analyze the data in Example 10.3 to estimate τ_T . We obtain

	reg0	reg	HT	Hajek	DR
est	0.061	-0.351	-1.992	-0.351	-0.187
se	0.227	0.258	0.705	0.328	0.287

without truncating the estimated propensity scores, and

	reg0	reg	HT	Hajek	DR
est	0.061	-0.351	-0.597	-0.192	-0.230
se	0.223	0.255	0.579	0.302	0.276

by truncating the estimated propensity scores from the above at 0.9. The HT estimator is sensitive to the truncation as expected. The regression estimator in Example 13.1 is quite different from other estimators. It imposes an unnecessary assumption that the regression functions in the treatment and control group share the same coefficient of X . The regression estimator in Example 13.2 is much close to the Hajek and doubly robust estimators. The estimates above are slightly different from those in Section 12.3.3, suggesting some treatment effect heterogeneity across τ_T and τ .

13.4 Other estimands

Li et al. (2018a) gave a unified discussion of the causal estimands in observational studies. Starting from the conditional average causal effect $\tau(X)$, they proposed a general class of estimands

$$\tau^h = \frac{E\{h(X)\tau(X)\}}{E\{h(X)\}}$$

indexed by a weighting function $h(X)$ with $E\{h(X)\} \neq 0$. The normalization in the denominator is to ensure that a constant causal effect $\tau(X) = \tau$ averages to the same τ .

Under the unconfoundedness assumption,

$$\tau^h = \frac{E[h(X)\{\mu_1(X) - \mu_0(X)\}]}{E\{h(X)\}}$$

which motivates the outcome regression estimator

$$\hat{\tau}^h = \frac{\sum_{i=1}^n h(X_i)\{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}}{\sum_{i=1}^n h(X_i)}.$$

Moreover, we can show that τ^h has the following weighting form:

Theorem 13.4 *Under ignorability and overlap, we have*

$$\tau^h = E \left\{ \frac{ZYh(X)}{e(X)} - \frac{(1-Z)Yh(X)}{1-e(X)} \right\} / E\{h(X)\}.$$

The proof of Theorem 13.4 is similar to those of Theorems 11.2 and 13.2 which is relegated to Problem 13.8. Based on Theorem 13.4, we can construct the corresponding IPW estimator.

By Theorem 13.4, each unit is associated with the weight due to the definition of the estimand as well as the weight due to the inverse of the propensity score. Finally, the treated units are weighted by $h(X)/e(X)$ and the control units are weighted by $h(X)/\{1-e(X)\}$. Li et al. (2018a, Table 1) summarized several estimands, and I present a part of it below:

population	$h(X)$	estimand	weights
combined	1	τ	$1/e(X)$ and $1/\{1-e(X)\}$
treated	$e(X)$	τ_T	1 and $e(X)/\{1-e(X)\}$
control	$1-e(X)$	τ_C	$\{1-e(X)\}/e(X)$ and 1
overlap	$e(X)\{1-e(X)\}$	τ_O	$1-e(X)$ and $e(X)$

The overlap population and the corresponding estimand

$$\tau_O = \frac{E[e(X)\{1-e(X)\}\tau(X)]}{E[e(X)\{1-e(X)\}]}$$

is new to us. This estimand has the largest weight for units with $e(X) = 1/2$ and downweights the units with extreme propensity scores. A nice feature of this estimand is that its IPW estimator is rather stable without the possibly extremely small values of $e(X)$ and $1 - e(X)$ in the denominator. If $e(X) \perp\!\!\!\perp \tau(X)$ including the special case of $\tau(X) = \tau$, the parameter τ_O reduces to τ . In general, however, the estimand τ_O may cause controversy because it changes the initial population and depends on the propensity score which may be misspecified in practice. Li et al. (2018a) and Li et al. (2019) gave some justifications and numerical evidence. This estimand will appear again in Chapter 14.

We can also construct the doubly robust estimator for τ^h . I relegate the details to Problem 13.9.

13.5 Homework Problems

13.1 An algebraic fact about a regression estimator for τ_T

This problem provides more details for Example 13.2.

Show that if we center the covariates by $X_i - \hat{X}(1)$ for all units, then $\hat{\tau}_T$ equals the coefficient of Z in the OLS fit of the outcome on the treatment, covariates, and their interactions.

13.2 Simulation for the average causal effect on the treated units

In `OS_ATE.R` in Chapter 12, I ran some simulation studies for τ . Run similar simulation studies for τ_T with either correct or incorrect propensity score or outcome models.

You can choose different model parameters, larger numbers of simulation and bootstrap replicates. Report your findings, including at least the bias, variance, and variance estimator via the bootstrap. You can also report other properties of the estimators, for example, the asymptotic Normality and the coverage rates of the confidence intervals.

13.3 An alternative form of the doubly robust estimator for τ_T

Motivated by (13.5), we have an alternative form of doubly robust estimator for $E\{Y(0) | Z = 1\}$:

$$\tilde{\mu}_{0T}^{\text{dr2}} = \frac{E[o(X, \alpha)(1 - Z)\{Y - \mu_0(X, \beta_0)\}]}{E[o(X, \alpha)(1 - Z)]} + E\{Z\mu_0(X, \beta_0)\}/e.$$

Show that under Assumption 13.1, $\tilde{\mu}_{0T}^{\text{dr2}} = E\{Y(0) | Z = 1\}$ if either $e(X, \alpha) = e(X)$ or $\mu_0(X, \beta_0) = \mu_0(X)$. Give the sample analogue of the doubly robust estimator for τ_T .

13.4 Average causal effect on the control units

Prove the identification formulas for τ_C , analogous to (13.1) and (13.4). Propose the doubly robust estimator for τ_C .

13.5 Estimating individual effect and conditional average causal effect

Assume that $\{Z_i, X_i, Y_i(1), Y_i(0)\}_{i=1}^n \stackrel{\text{IID}}{\sim} \{Z, X, Y(1), Y(0)\}$. The individual effect is $\tau_i = Y_i(1) - Y_i(0)$ and the conditional average causal effect is $\tau(X_i) = E\{Y_i(1) - Y_i(0) | X_i\}$. Since we will discuss individual effect, we do not drop the subscript i since τ mean the average causal effect, not the population version of $Y(1) - Y(0)$.

1. Under randomization with $Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0)\}$ and $e = \text{pr}(Z_i = 1)$,

17213 *The Average Causal Effect on the Treated Units and Other Estimands*

show that

$$\delta_i = \frac{Z_i Y_i}{e} - \frac{(1 - Z_i) Y_i}{1 - e}$$

is an unbiased predictor of the individual effect in the sense that

$$E(\delta_i - \tau_i) = 0 \quad (i = 1, \dots, n).$$

Further show that $E(\delta_i) = \tau$ for all $i = 1, \dots, n$.

2. Under ignorability with $Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0)\} \mid X_i$ and $e(X_i) = \text{pr}(Z_i = 1 \mid X_i)$, show that

$$\delta_i = \frac{Z_i Y_i}{e(X_i)} - \frac{(1 - Z_i) Y_i}{1 - e(X_i)}$$

is an unbiased predictor of the individual effect and the conditional average causal effect in the sense that

$$E(\delta_i - \tau_i) = 0, \quad E\{\delta_i - \tau(X_i)\} = 0, \quad (i = 1, \dots, n).$$

Further show that $E(\delta_i) = \tau$ for all $i = 1, \dots, n$.

13.6 General estimand and (τ_T, τ_C)

Assume unconfoundedness. Show that $\tau^h = \tau_T$ if $h(X) = e(X)$, and $\tau^h = \tau_C$ if $h(X) = 1 - e(X)$.

13.7 More on τ_O

Show that

$$\tau_O = \frac{E[\{1 - e(X)\}\tau(X) \mid Z = 1]}{E\{1 - e(X) \mid Z = 1\}} = \frac{E\{e(X)\tau(X) \mid Z = 0\}}{E\{e(X) \mid Z = 0\}}.$$

13.8 IPW for the general estimand

Prove Theorem 13.4.

13.9 Doubly robust estimation for general estimand

For a given $h(X)$, we have the following formulas for constructing the doubly robust estimator for τ^h :

$$\begin{aligned} \tilde{\mu}_1^{h,\text{dr}} &= E \left[\frac{Zh(X)\{Y - \mu_1(X, \beta_1)\}}{e(X, \alpha)} + h(X)\mu_1(X, \beta_1) \right], \\ \tilde{\mu}_0^{h,\text{dr}} &= E \left[\frac{(1 - Z)h(X)\{Y - \mu_0(X, \beta_0)\}}{1 - e(X, \alpha)} + h(X)\mu_0(X, \beta_0) \right]. \end{aligned}$$

Show that under ignorability and overlap,

1. if either $e(X, \alpha) = e(X)$ or $\mu_1(X, \beta_1) = \mu_1(X)$, then $\tilde{\mu}_1^{h,\text{dr}} = E\{h(X)Y(1)\}$;
2. if either $e(X, \alpha) = e(X)$ or $\mu_0(X, \beta_0) = \mu_0(X)$, then $\tilde{\mu}_0^{h,\text{dr}} = E\{h(X)Y(0)\}$;
3. if either $e(X, \alpha) = e(X)$ or $\{\mu_1(X, \beta_1) = \mu_1(X), \mu_0(X, \beta_0) = \mu_0(X)\}$, then

$$\frac{\tilde{\mu}_1^{h,\text{dr}} - \tilde{\mu}_0^{h,\text{dr}}}{E\{h(X)\}} = \tau^h.$$

Remark: Tao and Fu (2019) proved the above results. However, they hold only for a given $h(X)$. The most interesting cases of τ_T , τ_C and τ_O all have weight depending on the propensity score $e(X)$, which must be estimated in the first place. The above formulas do not apply to constructing the doubly robust estimators for τ_T and τ_C ; there does not exist a doubly robust estimator for τ_O .

13.10 Recommended reading

Shinozaki and Matsuyama (2015) focused on τ_T , and Li et al. (2018a) discussed general τ^h .



14

Using the Propensity Score in Regressions for Causal Effects

Since Rosenbaum and Rubin (1983b)'s seminal paper, many creative uses of the propensity score have appeared in the literature (e.g., Bang and Robins, 2005; Robins et al., 2007; Van der Laan and Rose, 2011; Vansteelandt and Daniel, 2014). This chapter discusses two simple methods to use the propensity score: including the propensity score as a covariate in regressions and running regressions weighted by the inverse of the propensity score. I choose to focus on these two methods because

1. they are easy to implement, which involve only standard statistical software packages for regressions;
 2. their properties are comparable to many more complex methods;
 3. they can be easily extended to allow for flexible statistical models including machine learning algorithms.
-

14.1 Regressions with the propensity score as a covariate

By Theorem 11.1, if unconfoundedness holds conditioning on X , then it also holds conditioning on $e(X)$:

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid e(X).$$

Analogous to (10.5), τ is also nonparametrically identified by

$$\tau = E\left[E\{Y \mid Z = 1, e(X)\} - E\{Y \mid Z = 0, e(X)\}\right],$$

which motivates methods based on regressions of Y on Z and $e(X)$.

The simplest regression specification is the OLS fit of Y on $\{1, Z, e(X)\}$, with the coefficient of Z as an estimator, denoted by τ_e . For simplicity, I will discuss the population OLS:

$$\arg \min_{a,b,c} E\{Y - a - bZ - ce(X)\}^2$$

with τ_e defined as the coefficient of Z . It is consistent for τ if we have a correct propensity score model and the outcome model is indeed linear in Z and $e(X)$. The more interesting result is that τ_e estimates τ_O if we have a correct propensity score model even if the outcome model is completely misspecified.

Theorem 14.1 *If $Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$, then the coefficient of Z in the OLS fit of Y on $\{1, Z, e(X)\}$ equals*

$$\tau_e = \tau_O = \frac{E\{h_O(X)\tau(X)\}}{E\{h_O(X)\}},$$

recalling that $h_O(X) = e(X)\{1 - e(X)\}$ and $\tau(X) = E\{Y(1) - Y(0) \mid X\}$.

An unusual feature of Theorem 14.1 is that the overlap condition is not needed any more. Even if some units have propensity score $e(X)$ equaling 0 or 1, their associate weight $e(X)\{1 - e(X)\}$ is zero so that they do not contribute anything to the final parameter τ_O .

Proof of Theorem 14.1: Based on the FWL theorem reviewed in Section A2.3, we can obtain τ_e in two steps: first, we obtain the residual \tilde{Z} from the OLS fit of Z on $\{1, e(X)\}$; then, we obtain τ_e from the OLS fit of Y on \tilde{Z} .

The coefficient of $e(X)$ in the OLS fit of Z on $\{1, e(X)\}$ is

$$\begin{aligned} \frac{\text{cov}\{Z, e(X)\}}{\text{var}\{e(X)\}} &= \frac{E[\text{cov}\{Z, e(X) \mid X\}] + \text{cov}\{E(Z \mid X), e(X)\}}{\text{var}\{e(X)\}} \\ &= \frac{0 + \text{var}\{e(X)\}}{\text{var}\{e(X)\}} = 1, \end{aligned}$$

so the intercept is $E(Z) - E\{e(X)\} = 0$ and the residual is $\tilde{Z} = Z - e(X)$. This makes sense since $Z - e(X)$ is uncorrelated with any function of X .

Therefore, we can obtain τ_e from the univariate OLS fit of Y on a centered variable $Z - e(X)$:

$$\tau_e = \frac{\text{cov}\{Z - e(X), Y\}}{\text{var}\{Z - e(X)\}}.$$

The denominator simplifies to

$$\begin{aligned} \text{var}\{Z - e(X)\} &= E\{Z - e(X)\}^2 \\ &= E\{Z + e(X)^2 - 2Ze(X)\} \\ &= e(X) + e(X)^2 - 2e(X)^2 = h_O(X). \end{aligned}$$

The numerator simplifies to

$$\begin{aligned}
 & \text{cov}\{Z - e(X), Y\} \\
 = & E[\{Z - e(X)\}Y] \\
 = & E[\{Z - e(X)\}ZY(1)] + E[\{Z - e(X)\}(1 - Z)Y(0)] \\
 & \quad (\text{since } Y = ZY(1) + (1 - Z)Y(0)) \\
 = & E[\{Z - Ze(X)\}Y(1)] - E[e(X)(1 - Z)Y(0)] \\
 = & E[Z\{1 - e(X)\}Y(1)] - E[e(X)(1 - Z)Y(0)] \\
 = & E[e(X)\{1 - e(X)\}\mu_1(X)] - E[e(X)\{1 - e(X)\}\mu_0(X)] \\
 & \quad (\text{tower property and ignorability}) \\
 = & E\{h_O(X)\tau(X)\}.
 \end{aligned}$$

The conclusion follows. \square

From the proof of Theorem 14.1, we can simply run the OLS of Y on the centered treatment $\tilde{Z} = Z - e(X)$. Lee (2018) proposed this procedure. Moreover, we can also include X in the OLS fit which may improve efficiency in finite sample. However, this does not change the estimand, which is still τ_O . I summarize these two results in the corollary below.

Corollary 14.1 *If $Z \perp\!\!\!\perp \{Y(1), Y(0)\} | X$, then*

- (1) *the coefficient of $Z - e(X)$ in the OLS fit of Y on $Z - e(X)$ or $\{1, Z - e(X)\}$ equals τ_O ;*
- (2) *the coefficient of Z in the OLS fit of Y on $\{1, Z, e(X), X\}$ equals τ_O .*

Proof of Corollary 14.1: (1) The first result is an intermediate step in the proof of Theorem 14.1. The second result holds because regressing Y on $Z - e(X)$ or $\{1, Z - e(X)\}$ does not change the coefficient of $Z - e(X)$ since it has mean zero.

(2) It follows from the fact that

$$Z - e(X) = Z - 0 - 1 \cdot e(X) - 0^T X$$

is the residual of the OLS fit of Z on $\{1, e(X), X\}$, since $Z - e(X)$ is uncorrelated with any functions of X . \square

Theorem 14.1 motivates a two-step estimator for τ_O : first, fit a propensity score model to obtain $\hat{e}(X_i)$; second, run OLS of Y_i on $(1, X_i, \hat{e}(X_i))$ to obtain the coefficient of Z_i . Corollary 14.1 motivates another two-step estimator for τ_O : first, fit a propensity score model to obtain $\hat{e}(X_i)$; second, run OLS of Y_i on $Z_i - \hat{e}(X_i)$ to obtain the coefficient of Z_i . Although OLS is convenient for obtaining point estimators, the corresponding standard errors are incorrect due to the uncertainty in the first step estimation of the propensity score. We can use the bootstrap to approximate the standard errors.

Robins et al. (1992) discussed many OLS estimators based on the propensity score. The above results seem special cases of their general theory although they did not point out the connection with the estimand under the overlap weight, which was resurrected by Li et al. (2018a). Lee (2018) proposed to regress Y on $Z - e(X)$ from a different perspective without making connections to the existing results in Robins et al. (1992) and Li et al. (2018a).

Rosenbaum and Rubin (1983b) proposed to estimate the average causal effect based on the OLS fit of Y on $\{1, Z, e(X), Ze(X)\}$. When this outcome model is correct, their estimator is consistent for the average causal effect. However, when the model is incorrect, the corresponding estimator has a much more complicated interpretation. Little and An (2004) suggested constructing estimators based on the OLS of Y on Z and a flexible function of $e(X)$ and showed it enjoys certain doubly robustness property. Due to the complexity in implementation, I omit the discussion.

14.2 Regressions weighted by the inverse of the propensity score

14.2.1 Average causal effect

We first re-examine the Hajek estimator of τ :

$$\hat{\tau}^{\text{hajek}} = \frac{\sum_{i=1}^n \frac{Z_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^n \frac{Z_i}{\hat{e}(X_i)}} - \frac{\sum_{i=1}^n \frac{(1-Z_i) Y_i}{1-\hat{e}(X_i)}}{\sum_{i=1}^n \frac{1-Z_i}{1-\hat{e}(X_i)}},$$

which equals the difference between the weighted means of the outcomes in the treatment and control groups. Numerically, it is identical to the coefficient of Z_i in the following weighted least squares (WLS) of Y_i on $(1, Z_i)$.

Proposition 14.1 $\hat{\tau}^{\text{hajek}}$ equals $\hat{\beta}$ from the following WLS:

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta} \sum_{i=1}^n w_i (Y_i - \alpha - \beta Z_i)^2$$

with weights

$$w_i = \frac{Z_i}{\hat{e}(X_i)} + \frac{1-Z_i}{1-\hat{e}(X_i)} = \begin{cases} \frac{1}{\hat{e}(X_i)} & \text{if } Z_i = 1; \\ \frac{1}{1-\hat{e}(X_i)} & \text{if } Z_i = 0. \end{cases} \quad (14.1)$$

Imbens (2004) pointed out the result in Proposition 14.1. I leave it as a Problem 14.1. By Proposition 14.1, it is convenient to obtain $\hat{\tau}^{\text{hajek}}$ based on WLS. However, due to the uncertainty in the estimated propensity score, the standard error reported by WLS is incorrect for the true standard error

of $\hat{\tau}^{\text{hajek}}$. The bootstrap provides a convenient approximation to the true standard error.

Why does the WLS give a consistent estimator for τ ? Recall that in the CRE with a constant propensity score, we can simply use the coefficient of Z_i in the OLS fit of Y_i on $(1, Z_i)$ to estimate τ . In observational studies, units have different probabilities of receiving the treatment and control, respectively. If we weight the treated units by $1/e(X_i)$ and the control units by $1/\{1-e(X_i)\}$, then they can represent the whole population and we effectively have a pseudo randomized experiment. Consequently, the difference between the weighted means are consistent for τ . The numerical equivalence of $\hat{\tau}^{\text{hajek}}$ and WLS is not only a fun numerical fact itself but also useful for motivation more complex estimator with covariate adjustment. I give one extension below.

Recall that in the CRE, we can use the coefficient of Z_i in the OLS fit of Y_i on $(1, Z_i, X_i, Z_iX_i)$ to estimate τ , where the covariates are centered with $\bar{X} = 0$. This is Lin (2013)'s estimator which uses covariates to improve efficiency. A natural extension to observational studies is to estimate τ using the coefficient of Z_i in the WLS fit of Y_i on $(1, Z_i, X_i, Z_iX_i)$ with weights defined in (14.1). Hirano and Imbens (2001) used this estimator in an application. The fully interacted linear model is equivalent to two separate linear models for the treated and control groups. If the linear models

$$E(Y | Z = 1, X) = \beta_{10} + \beta_{1x}^T X, \quad E(Y | Z = 0, X) = \beta_{00} + \beta_{0x}^T X,$$

are correctly specified, then both OLS and WLS give consistent estimators for the coefficients and the estimators of the coefficient of Z is consistent for τ . More interestingly, the estimator of the coefficient of Z based on WLS is also consistent for τ if the propensity score model is correct and the outcome model is incorrect. That is, the estimator based on WLS is doubly robust. Robins et al. (2007) discussed this property and attributed this result to M. Joffe's unpublished paper. I will give more details below.

Let $\hat{e}(X_i)$ be the fitted propensity score and $(\mu_1(X_i, \hat{\beta}_1), \mu_0(X_i, \hat{\beta}_0))$ be the fitted values of the outcome means based on the WLS. The outcome regression estimator is

$$\hat{\tau}_{\text{wls}}^{\text{reg}} = \frac{1}{n} \sum_{i=1}^n \mu_1(X_i, \hat{\beta}_1) - \frac{1}{n} \sum_{i=1}^n \mu_0(X_i, \hat{\beta}_0)$$

and the doubly robust estimator for τ is

$$\hat{\tau}_{\text{wls}}^{\text{dr}} = \hat{\tau}_{\text{wls}}^{\text{reg}} + \frac{1}{n} \sum_{i=1}^n \frac{Z_i \{Y_i - \mu_1(X_i, \hat{\beta}_1)\}}{\hat{e}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - Z_i) \{Y_i - \mu_0(X_i, \hat{\beta}_0)\}}{1 - \hat{e}(X_i)}.$$

An interesting result is that this doubly robust estimator equals the outcome regression estimator, which reduces to the coefficient of Z_i in the WLS fit of Y_i on $(1, Z_i, X_i, Z_iX_i)$ if we use weights (14.1).

Theorem 14.2 *If $\bar{X} = 0$ and $(\mu_1(X_i, \hat{\beta}_1), \mu_0(X_i, \hat{\beta}_0)) = (\hat{\beta}_{10} + \hat{\beta}_{1x}^T X_i, \hat{\beta}_{00} +$*

$\hat{\beta}_{0x}^T X_i)$ based on the WLS fit of Y_i on $(1, Z_i, X_i, Z_i X_i)$ with weights (14.1), then

$$\hat{\tau}_{wls}^{dr} = \hat{\tau}_{wls}^{reg} = \hat{\beta}_{10} - \hat{\beta}_{00},$$

which is the coefficient of Z_i in the WLS fit.

Proof of Theorem 14.2: The WLS fit of Y_i on $(1, Z_i, X_i, Z_i X_i)$ is equivalent to two WLS fits based on the treated and control data. Both WLS fits include intercepts, so the first order conditions must satisfy

$$\sum_{i=1}^n \frac{Z_i(Y_i - \hat{\beta}_{10} - \hat{\beta}_{1x}^T X_i)}{\hat{e}(X_i)} = 0$$

and

$$\sum_{i=1}^n \frac{(1 - Z_i)(Y_i - \hat{\beta}_{00} - \hat{\beta}_{0x}^T X_i)}{1 - \hat{e}(X_i)} = 0.$$

So the difference between $\hat{\tau}^{dr}$ and $\hat{\tau}^{reg}$ is exactly zero. Both reduces to

$$\frac{1}{n} \sum_{i=1}^n (\hat{\beta}_{10} + \hat{\beta}_{1x}^T X_i) - \frac{1}{n} \sum_{i=1}^n (\hat{\beta}_{00} + \hat{\beta}_{0x}^T X_i) = \hat{\beta}_{10} - \hat{\beta}_{00} + (\hat{\beta}_{1x} - \hat{\beta}_{0x})^T \bar{X} = \hat{\beta}_{10} - \hat{\beta}_{00}$$

with centered covariates. So they both equal the coefficient of Z_i in the WLS fit of Y_i on $(1, Z_i, X_i, Z_i X_i)$. \square

Freedman and Berk (2008) discouraged the use of the WLS estimator above based on some simulation studies. They showed that when the outcome model is correct, the WLS estimator is worse than the OLS estimator since the WLS estimator has large variability in their simulation setting with homoskedastic outcomes. This may not be true in general. When the errors have variance proportional to the inverse of the propensity scores, the WLS estimator will be more efficient than the OLS estimator. They also showed that the estimated standard error based on the WLS fit is not consistent for the true standard error because it ignores the uncertainty in the estimated propensity score. This can be easily fixed by using the bootstrap to approximate the variance of the WLS estimator. Nevertheless, they found that “weighting may help under some circumstances” because when the outcome model is incorrect, the WLS estimator is still consistent if the propensity score model is correct.

I end this section with Table 14.1 summarizing the regression estimators for causal effects in both randomized experiments and observational studies.

14.2.2 Average causal effect on the treated units

The results for τ_T parallel those for τ . First, the Hajek estimator for τ_T

$$\hat{\tau}_T^{\text{hajek}} = \hat{Y}(1) - \frac{\sum_{i=1}^n \hat{o}(X_i)(1 - Z_i)Y_i}{\sum_{i=1}^n \hat{o}(X_i)(1 - Z_i)},$$

with $\hat{o}(X_i) = \hat{e}(X_i)/\{1 - \hat{e}(X_i)\}$, equals the coefficient of Z_i in the following WLS fit Y_i on $(1, Z_i)$.

TABLE 14.1: Regression estimators in CREs and unconfounded observational studies. The weights w_i 's are defined in (14.1) .

	CRE	unconfounded observational studies
without X	$Y_i \sim Z_i$	$Y_i \sim Z_i$ with weights w_i
with X	$Y_i \sim (Z_i, X_i, Z_i X_i)$	$Y_i \sim (Z_i, X_i, Z_i X_i)$ with weights w_i

Proposition 14.2 $\hat{\tau}_T^{hajek}$ is numerically identical to $\hat{\beta}$ in the following WLS:

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta} \sum_{i=1}^n w_{Ti} (Y_i - \alpha - \beta Z_i)^2$$

with weights

$$w_{Ti} = Z_i + (1 - Z_i) \hat{o}(X_i) = \begin{cases} 1 & \text{if } Z_i = 1; \\ \hat{o}(X_i) & \text{if } Z_i = 0. \end{cases} \quad (14.2)$$

Similar to Proposition 14.1, Proposition 14.2 is a pure linear algebra result. I relegate its proof as Problem 14.1.

Second, if we center covariates with $\hat{X}(1) = 0$, then we can estimate τ_T using the coefficient of Z_i in the WLS fit of Y_i on $(1, Z_i, X_i, Z_i X_i)$ with weights defined in (14.2). Similarly, this estimator equals the regression estimator

$$\hat{\tau}_{T,wls}^{\text{reg}} = \hat{Y}(1) - \frac{1}{n_1} \sum_{i=1}^n Z_i \mu_0(X_i, \hat{\beta}_0),$$

which also equals the doubly robust estimator

$$\hat{\tau}_{T,wls}^{\text{dr}} = \hat{\tau}_{T,wls}^{\text{reg}} - \frac{1}{n_1} \sum_{i=1}^n \hat{o}(X_i)(1 - Z_i)\{Y_i - \mu_0(X_i, \hat{\beta}_0)\}.$$

Theorem 14.3 If $\hat{X}(1) = 0$ and $\mu_0(X_i, \hat{\beta}_0) = \hat{\beta}_{00} + \hat{\beta}_{0x}^T X_i$ based on the WLS fit of Y_i on $(1, Z_i, X_i, Z_i X_i)$ with weights (14.2), then

$$\hat{\tau}_{T,wls}^{\text{dr}} = \hat{\tau}_{T,wls}^{\text{reg}} = \hat{\beta}_{10} - \hat{\beta}_{00},$$

which is the coefficient of Z_i in the WLS fit.

Proof of Theorem 14.3: Based on the WLS fits in the treatment and control groups, we have

$$\sum_{i=1}^n Z_i(Y_i - \hat{\beta}_{10} - \hat{\beta}_{1x}^T X_i) = 0, \quad (14.3)$$

$$\sum_{i=1}^n \hat{o}(X_i)(1 - Z_i)(Y_i - \hat{\beta}_{00} - \hat{\beta}_{0x}^T X_i) = 0. \quad (14.4)$$

The second result (14.4) ensures that $\hat{\tau}_{T,wls}^{dr} = \hat{\tau}_{T,wls}^{reg}$. Both reduces to

$$\hat{Y}(1) - \frac{1}{n_1} \sum_{i=1}^n Z_i (\hat{\beta}_{00} + \hat{\beta}_{0x}^T X_i) = \frac{1}{n_1} \sum_{i=1}^n Z_i (Y_i - \hat{\beta}_{00} - \hat{\beta}_{0x}^T X_i).$$

With covariates centered with $\hat{X}(1) = 0$, the first result (14.3) implies that $\hat{Y}(1) = \hat{\beta}_{10}$ which further simplifies the estimators to $\hat{\beta}_{10} - \hat{\beta}_{00}$. \square

14.3 Homework problems

14.1 Hajek estimators as WLS estimators

Prove Propositions 14.1 and 14.2.

Hint: These are special cases of Problem A2.2 on the univariate WLS.

14.2 Predictive estimator and doubly robust estimator

Another outcome regression estimator is the predictive estimator

$$\hat{\tau}^{pred} = \hat{\mu}_1^{pred} - \hat{\mu}_0^{pred}$$

where

$$\hat{\mu}_1^{pred} = \frac{1}{n} \sum_{i=1}^n \left\{ Z_i Y_i + (1 - Z_i) \mu_1(X_i, \hat{\beta}_1) \right\}$$

and

$$\hat{\mu}_0^{pred} = \frac{1}{n} \sum_{i=1}^n \left\{ Z_i \mu_0(X_i, \hat{\beta}_1) + (1 - Z_i) Y_i \right\}.$$

It differs from the outcome regression estimator discussed before in that it only predicts the counterfactual outcomes but not the observed outcomes.

Show that the doubly robust estimator equals $\hat{\tau}^{pred}$ if $(\mu_1(X_i, \hat{\beta}_1), \mu_0(X_i, \hat{\beta}_1)) = (\hat{\beta}_{10} + \hat{\beta}_{1x}^T X_i, \hat{\beta}_{00} + \hat{\beta}_{0x}^T X_i)$ are from the WLS fits of Y_i on $(1, X_i)$ based on the treated and control data, respectively, with weights

$$w_i = Z_i / \hat{o}(X_i) + (1 - Z_i) \hat{o}(X_i) = \begin{cases} \frac{1}{\hat{o}(X_i)} = \frac{1 - \hat{e}(X_i)}{\hat{e}(X_i)} & \text{if } Z_i = 1; \\ \hat{o}(X_i) = \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} & \text{if } Z_i = 0. \end{cases} \quad (14.5)$$

Remark: Cao et al. (2009) and Vermeulen and Vansteelandt (2015) motivated the weights in (14.5) from other more theoretical perspectives.

14.3 Weighted logistic regression with a binary outcome

With a binary outcome, we can replace linear outcome models by the logistic outcome models. Show that with weights in the logistic regressions, the doubly robust estimators equals the outcome regression estimator. The result holds for both τ and τ_T .

14.4 Causal inference with a misspecified linear regression

Define the population OLS of Y on Z, X as

$$(\beta_0, \beta_1, \beta_2) = \arg \min_{b_0, b_1, b_2} E(Y - b_0 - b_1 Z - b_2^T X)^2.$$

Recall that $e(X) = \text{pr}(Z = 1 | X)$ is the propensity score, and define $\tilde{e}(X) = \gamma_0 + \gamma_1^T X$ as the OLS projection of A on X with

$$(\gamma_0, \gamma_1) = \arg \min_{c_0, c_1} E(A - c_0 - c_1^T X)^2.$$

1. Show that

$$\beta_1 = \frac{E[\tilde{w}(X)\{\mu_1(X) - \mu_0(X)\}]}{E\{\tilde{w}(X)\}} + \frac{E\{e(X) - \tilde{e}(X)\}\mu_0(X)}{E\{\tilde{w}(X)\}}$$

where $\tilde{w}(X) = e(X)\{1 - \tilde{e}(X)\}$.

2. When X contains the dummy variables for a discrete covariate, show that

$$\beta_1 = \frac{E[w(X)\{\mu_1(X) - \mu_0(X)\}]}{E\{w(X)\}}$$

where $w(X) = e(X)\{1 - e(X)\}$ is the overlap weight.

Remark: Vansteelandt and Dukes (2022) gave the formula in the first part without a detailed proof. The result in part 2 was derived many times in the literature (e.g., Angrist, 1998; Ding, 2021).

14.5 Data re-analysis

Re-analyze the dataset in `karolinska.txt` and the dataset `nhanes_bmi` in the `ATE` package.

14.6 Recommended reading

Kang and Schafer (2007) gave a critical review of the doubly robust estimator, using simulation to compare it with many other estimators. Robins et al. (2007) gave a very insightful comment on Kang and Schafer (2007).

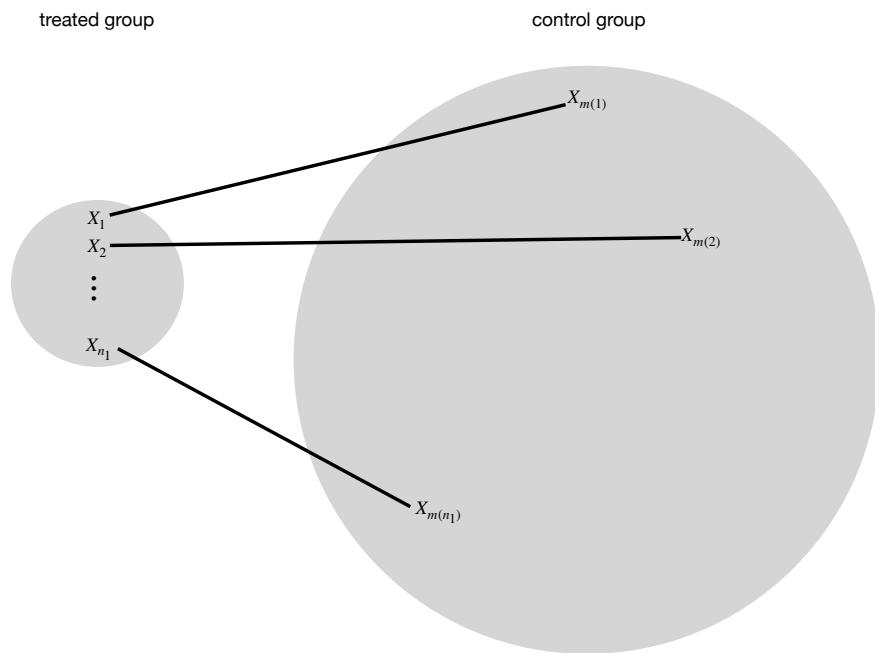


15

Matching in Observational Studies

Matching has a long history in empirical research. W. Cochran and D. Rubin popularized it in statistical causal inference. Cochran and Rubin (1973) is an early review paper. Rubin (2006b) collects Rubin's contributions to this topic. This chapter also discusses modern contributions by Abadie and Imbens (2006, 2008, 2011).

15.1 A simple starting point: many more control units



Consider a simple case with the number of control units n_0 being much larger than the number of treated units n_1 . For unit $i = 1, \dots, n_1$ in the treated

group, we find a unit $m(i)$ in the control group such that $X_i = X_{m(i)}$. In the ideal case, we have exact matches. Therefore, the units within a matched pair have the same propensity score $e(X_i) = e(X_{m(i)})$. Consequently, conditioning on the event that one unit receives the treatment and the other receives the control, the probability of unit i receiving the treatment and unit $m(i)$ receives the control is

$$\begin{aligned} & \text{pr}(Z_i = 1, Z_{m(i)} = 0 \mid Z_i + Z_{m(i)} = 1, X_i, X_{m(i)}) \\ &= \frac{\text{pr}(Z_i = 1, Z_{m(i)} = 0 \mid X_i, X_{m(i)})}{\text{pr}(Z_i = 1, Z_{m(i)} = 0 \mid X_i, X_{m(i)}) + \text{pr}(Z_i = 0, Z_{m(i)} = 1 \mid X_i, X_{m(i)})} \\ &= \frac{e(X_i)\{1 - e(X_{m(i)})\}}{e(X_i)\{1 - e(X_{m(i)})\} + \{1 - e(X_i)\}e(X_{m(i)})} \\ &= \frac{1}{2}. \end{aligned}$$

That is, the treatment assignment is identical to the MPE conditioning on the covariates and the event that each pair has a treated and control units. So we can analyze the exactly matched observational study as if it is a MPE, using either the FRT or the Neymanian approach in Chapter 7. This gives us inference on the causal effect on the treated units.

We can also find multiple control units for each treated unit. In general, we can find M_i matched control units for the treated unit i . When the M_i 's vary, it is called the *variable-ratio matching* (Ming and Rosenbaum, 2000, 2001; Pimentel et al., 2015). With perfect matching, the treatment assignment mechanism is identical to the general matched experiment discussed in Section 7.7. We can use the analytic results in that section to analyzed the matched observational study.

15.2 A more complicated but realistic scenario

Even if the control group is large, we often do not have exact matches. What we can achieve is that $X_i \approx X_{m(i)}$ or $X_i - X_{m(i)}$ is small under some distance metric. So we have only approximate matches. For example, we define

$$m(i) = \arg \min_{k: Z_k=0} d(X_i, X_k),$$

where $d(X_i, X_k)$ measures the distance between X_i and X_k . Some canonical choices of the distance are the Euclidean distance

$$d(X_i, X_k) = \|X_i - X_k\|_2^2,$$

and the Mahalanobis distance¹

$$d(X_i, X_k) = (X_i - X_k)^T \Omega^{-1} (X_i - X_k)$$

with Ω being the sample covariance matrix of the X_i 's from the whole population or only the control group.

I review some subtle issues about matching below. See Stuart (2010) for a review paper.

1. (one-to-one or one-to- M matching) The above discussion focused on one-to-one matching
2. I focus on matching with replacement but some practitioners prefer matching without replacement. If the pool of control units is large, these two methods will not matter too much for the final result. Matching with replacement is computationally more convenient, but matching without replacement involves computationally intensive discrete optimization. Matching with replacement usually gives matches of higher quality but it introduces dependence by using the same units multiple times. In contrast, the advantage of matching without replacement is the independence of matched units and the simplicity in the subsequent data analysis.
3. Because of the residual covariate imbalance within matched pairs, it is crucial to use covariate adjustment when analyzing the data. In this case, covariate adjustment is not only for efficiency gain but also for bias correction.
4. If X is “high dimensional”, it is likely that $d(X_i, X_k)$ is too large for some unit i in the treated group and for all choices of the units in the control group. In this case, we may have to drop some units that are hard to find matches. By doing this, we effectively change the study population of interest.
5. It is hard to avoid the above problem. For example, if $X_i \sim N(0, I_p)$, $X_k \sim N(0, I_p)$, and $X_i \perp\!\!\!\perp X_k$, then

$$\|X_i - X_k\|_2^2 \sim \|N(0, 2I_p)\|_2^2 = 2\chi_p^2$$

which has mean $2p$ and variance $8p$. Theory shows that with large p , imperfect matching causes large bias in causal effect estimation. This suggests that if p is large, we must have some dimension reduction before matching. Rosenbaum and Rubin (1983b) proposed to match based on the propensity score. With the estimated propensity score, we find pairs of units $\{i, m(i)\}$ with small values of $|\hat{e}(X_i) - \hat{e}(X_{m(i)})|$ or $|\text{logit}\{\hat{e}(X_i)\} - \text{logit}\{\hat{e}(X_{m(i)})\}|$, i.e., we have a one dimensional matching problem.

¹We define $\|v\|_2^2 = \sum_{j=1}^p v_j^2$ for a vector $v = (v_1, \dots, v_p)^T$. It denotes the squared length of the vector v .

15.3 Matching estimator for the average causal effect

In a sequence of papers, Abadie and Imbens (AI) rigorously characterized the repeated sampling properties of the matching estimator and proposed the corresponding large-sample confidence intervals for the average causal effect. They chose the standard setup for observational studies with $\{X_i, Z_i, Y_i(1), Y_i(0)\}_{i=1}^n \stackrel{\text{IID}}{\sim} \{X, Z, Y(1), Y(0)\}$.

15.3.1 Point estimation and bias correction

AI focused on 1 to M matching with replacement. For a treated unit i , we can simply impute the potential outcome under treatment as $\hat{Y}_i(1) = Y_i$, and impute the potential outcome under control as

$$\hat{Y}_i(0) = M^{-1} \sum_{k \in J_i} Y_k,$$

where J_i is the set of matched units from the control group for unit i . For example, we can compute $d(X_i, X_k)$ for all k in the control group, and then define J_i as the indices of k with the M smallest values of $d(X_i, X_k)$.

For a control unit i , we simply impute the potential outcome under control as $\hat{Y}_i(0) = Y_i$, and impute the potential outcome under treatment as

$$\hat{Y}_i(1) = M^{-1} \sum_{k \in J_i} Y_k,$$

where J_i is the set of matched units from the treatment group for unit i .

The matching estimator is

$$\hat{\tau}^m = n^{-1} \sum_{i=1}^n \{\hat{Y}_i(1) - \hat{Y}_i(0)\}.$$

AI showed that $\hat{\tau}^m$ has non-negligible bias especially when X is multidimensional and the number of control units is comparable to the number of treated units. Through some technical derivations, they proposed the following estimator for the bias:

$$\hat{B} = n^{-1} \sum_{i=1}^n \hat{B}_i$$

where

$$\hat{B}_i = (2Z_i - 1)M^{-1} \sum_{k \in J_i} \{\hat{\mu}_{1-Z_i}(X_i) - \hat{\mu}_{1-Z_i}(X_k)\}$$

with $\{\hat{\mu}_1(X_i), \hat{\mu}_0(X_i)\}$ being the predicted outcomes by, for example, from OLS fits. For a treated unit with $Z_i = 1$, the estimated bias is

$$\hat{B}_i = M^{-1} \sum_{k \in J_i} \{\hat{\mu}_0(X_i) - \hat{\mu}_0(X_k)\}$$

which corrects the discrepancy in predicted control potential outcomes due to the mis-match in covariates; for a control unit with $Z_i = 0$, the estimates bias is

$$\hat{B}_i = -M^{-1} \sum_{k \in J_i} \{\hat{\mu}_1(X_i) - \hat{\mu}_1(X_k)\}$$

which corrects the discrepancy in predicted treated potential outcomes due to the mis-match in covariates.

The final bias corrected matching estimator is

$$\hat{\tau}^{\text{mbc}} = \hat{\tau}^{\text{m}} - \hat{B},$$

which has the following linear expansion.

Proposition 15.1 *We have*

$$\hat{\tau}^{\text{mbc}} = n^{-1} \sum_{i=1}^n \hat{\psi}_i \quad (15.1)$$

where

$$\hat{\psi}_i = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + (2Z_i - 1)(1 + K_i/M)\{Y_i - \hat{\mu}_{Z_i}(X_i)\}$$

with K_i being the times that unit i is used as a match.

The linear expansion in Proposition 15.1 follows from simple but tedious algebra. I leave its proof as Problem 15.1. The linear expansion motivates a simple variance estimator

$$\hat{V}^{\text{mbc}} = \frac{1}{n^2} \sum_{i=1}^n (\hat{\psi}_i - \hat{\tau}^{\text{mbc}})^2,$$

by viewing $\hat{\tau}^{\text{mbc}}$ as sample averages of the $\hat{\psi}_i$'s. In the literature, Abadie and Imbens (2008) first showed that the simple bootstrap by resampling the original data does not work for estimating the variance of the matching estimators, but their proposed variance estimation procedure is not easy to implement. Otsu and Rai (2017) proposed to bootstrap the $\hat{\psi}_i$'s in the linear expansion, which yields the variance estimator \hat{V}^{mbc} .

15.3.2 Connection with the doubly robust estimators

The bias-corrected matching estimators and the doubly robust estimators are closely related. They both equal the outcome regression estimator with some modifications based on the residuals

$$\hat{R}_i = \begin{cases} Y_i - \hat{\mu}_1(X_i) & \text{if } Z_i = 1; \\ Y_i - \hat{\mu}_0(X_i) & \text{if } Z_i = 0. \end{cases}$$

For the average causal effect τ , recall the outcome regression estimator

$$\hat{\tau}^{\text{reg}} = n^{-1} \sum_{i=1}^n \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}$$

and the doubly robust estimator

$$\hat{\tau}^{\text{dr}} = \hat{\tau}^{\text{reg}} + n^{-1} \sum_{i=1}^n \left\{ \frac{Z_i \hat{R}_i}{\hat{e}(X_i)} - \frac{(1 - Z_i) \hat{R}_i}{1 - \hat{e}(X_i)} \right\}.$$

Furthermore, we can verify that $\hat{\tau}^{\text{mbc}}$ has a form very similar to $\hat{\tau}^{\text{dr}}$.

Proposition 15.2 *The bias-corrected matching estimator for τ equals*

$$\hat{\tau}^{\text{mbc}} = \hat{\tau}^{\text{reg}} + n^{-1} \sum_{i=1}^n \left\{ \left(1 + \frac{K_i}{M} \right) Z_i \hat{R}_i - \left(1 + \frac{K_i}{M} \right) (1 - Z_i) \hat{R}_i \right\}.$$

I leave the proof of Proposition 15.2 as Problem 15.2. From Proposition 15.2, we can view matching as a nonparametric method to estimate the propensity score, and the resulting bias-corrected matching estimator as a doubly robust estimator. For instance, $1 + K_i/M$ should be similar to $1/\hat{e}(X_i)$. When a treated unit has a small $e(X_i)$, the resulting weight $1/\hat{e}(X_i)$ will be large, and at the same time, it will be matched with many control units, resulting in large K_i and thus large $1 + K_i/M$. However, this connection also raised an obvious question regarding matching. With a fixed M , the estimator $1 + K_i/M$ for $1/e(X_i)$ will be very noisy. Allowing M to grow with the sampling size is likely to improve the matching-based nonparametric estimator for the propensity score and thus improve the asymptotic properties of the matching and bias-corrected matching estimators. Lin et al. (2023) provided a formal theory.

15.4 Matching estimator for the average causal effect on the treated

For the average causal effect on the treated

$$\tau_T = E(Y | Z = 1) - E\{Y(0) | Z = 1\},$$

we only need to impute the missing potential outcomes under control for all the treated units, resulting the following estimator

$$\hat{\tau}_T^m = n_1^{-1} \sum_{i=1}^n Z_i \{Y_i - \hat{Y}_i(0)\}.$$

Again it is biased with multidimensional X . Otsu and Rai (2017) propose to estimate its bias by

$$\hat{B}_T = n_1^{-1} \sum_{i=1}^n Z_i \hat{B}_{T,i}$$

where

$$\hat{B}_{T,i} = M^{-1} \sum_{k \in J_i} \{\hat{\mu}_0(X_i) - \hat{\mu}_0(X_k)\}$$

corrects the bias due to the mis-match of covariates for a treated unit with $Z_i = 1$.

The final bias-corrected estimator is

$$\hat{\tau}_T^{mbc} = \hat{\tau}_T^m - \hat{B}_T,$$

which has the following linear expansion.

Proposition 15.3 *We have*

$$\hat{\tau}_T^{mbc} = n_1^{-1} \sum_{i=1}^n \hat{\psi}_{T,i}, \quad (15.2)$$

where

$$\hat{\psi}_{T,i} = Z_i \{Y_i - \hat{\mu}_0(X_i)\} - (1 - Z_i) K_i / M \{Y_i - \hat{\mu}_0(X_i)\}.$$

I leave the proof as Problem 15.1. Motivated by Otsu and Rai (2017), we can view $\hat{\tau}_T^{mbc}$ as n/n_1 multiplied by the sample average of the $\hat{\psi}_{T,i}$'s, so an intuitive variance estimator is

$$\hat{V}_T^{mbc} = \left(\frac{n}{n_1} \right)^2 \frac{1}{n^2} \sum_{i=1}^n (\hat{\psi}_{T,i} - \hat{\tau}_T^{mbc} n_1/n)^2 = \frac{1}{n_1^2} \sum_{i=1}^n (\hat{\psi}_{T,i} - \hat{\tau}_T^{mbc} n_1/n)^2.$$

Similar to the discussion in Section 15.3.2, we can compare the doubly robust and bias-corrected matching estimators with the outcome regression estimator. For the average causal effect on the treated units τ_T , recall the outcome regression estimator

$$\hat{\tau}_T^{\text{reg}} = n_1^{-1} \sum_{i=1}^n Z_i \{Y_i - \hat{\mu}_0(X_i)\},$$

and the doubly robust estimator

$$\hat{\tau}_T^{\text{dr}} = \hat{\tau}_T^{\text{reg}} - n_1^{-1} \sum_{i=1}^n \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} (1 - Z_i) \hat{R}_i.$$

Furthermore, we can verify that $\hat{\tau}_T^{mbc}$ has a form very similar to $\hat{\tau}_T^{\text{dr}}$.

Proposition 15.4 *The bias correction matching estimator for τ_T equals*

$$\hat{\tau}_T^{\text{mbc}} = \hat{\tau}_T^{\text{reg}} - n_1^{-1} \sum_{i=1}^n \frac{K_i}{M} (1 - Z_i) \hat{R}_i.$$

I leave the proof of Proposition 15.4 as Problem 15.3. Proposition 15.4 suggests that matching essentially uses K_i/M to estimate the odds of the treatment given covariates.

15.5 A case study

15.5.1 Experimental data

Now I revisit the LaLonde data using Sekhon (2011)'s `Matching` package. We have used this package several times for the dataset `lalonde`, and now we will use its key function `Match`. The experimental part gives us the following results:

```
> library("car")
> library("Matching")
> y = lalonde$re78
> z = lalonde$treat
> x = as.matrix(lalonde[, c("age", "educ", "black",
+ "hisp", "married", "nodegr",
+ "re74", "re75")])
>
> ## analysis the randomized experiment
> neymanols = lm(y ~ z)
> neymanols$coef[2]
      z
1794.343
> sqrt(hccm(neymanols, type = "hc2"))[2, 2])
[1] 670.9967
>
> xc = scale(x)
> linols = lm(y ~ z*xc)
> linols$coef[2]
      z
1621.584
> sqrt(hccm(linols, type = "hc2"))[2, 2])
[1] 694.7217
```

Both the unadjusted and adjusted estimators shows positive significant results on the job training program. We can analyze the data as if it is an observational study, yielding the following results:

```

> matchest.adj = Match(Y = y, Tr = z, X = x, BiasAdjust = TRUE)
> summary(matchest.adj)

Estimate... 2119.7
AI SE..... 876.42
T-stat..... 2.4185
p.val..... 0.015583

Original number of observations..... 445
Original number of treated obs..... 185
Matched number of observations..... 185
Matched number of observations (unweighted). 268

```

Both the point estimator and standard error increase, but qualitatively, the conclusion remains the same.

15.5.2 Observational data

Then I revisit the observational counterpart of the data:

```

> dat <- read.table("cps1re74.csv", header=T)
> dat$u74 <- as.numeric(dat$re74==0)
> dat$u75 <- as.numeric(dat$re75==0)
> y = dat$re78
> z = dat$treat
> x = as.matrix(dat[, c("age", "educ", "black",
+ "hispan", "married", "nodegree",
+ "re74", "re75", "u74", "u75")])

```

If we use simple OLS estimators, we obtain results that are far from the experimental benchmark:

```

> neymanols = lm(y ~ z)
> neymanols$coef [2]
      z
-8506.495
> sqrt(hccm(neymanols, type = "hc2"))[2, 2]
[1] 583.4426
>
> xc = scale(x)
> linols = lm(y ~ z*xc)
> linols$coef [2]
      z
-4265.801
> sqrt(hccm(linols, type = "hc2"))[2, 2]
[1] 3211.772

```

However, if we use matching, the results almost recovers those based on the experimental data:

```
> matchest = Match(Y = y, Tr = z, X = x, BiasAdjust = TRUE)
```

```
> summary(matchest)

Estimate... 1747.8
AI SE..... 916.59
T-stat..... 1.9068
p.val..... 0.056543

Original number of observations..... 16177
Original number of treated obs..... 185
Matched number of observations..... 185
Matched number of observations (unweighted). 248
```

Ignoring the ties in the matched data, we can also use the matched-pairs analysis, which again yields results similar to those based on the experimental data:

```
> diff = y[matchest$index.treated] -
+       y[matchest$index.control]
> round(summary(lm(diff ~ 1))$coef[1, ], 2)
  Estimate Std. Error t value Pr(>|t|)
  1581.44    558.55     2.83   0.01
>
> diff.x = x[matchest$index.treated, ] -
+       x[matchest$index.control, ]
> round(summary(lm(diff ~ diff.x))$coef[1, ], 2)
  Estimate Std. Error t value Pr(>|t|)
  1842.06    578.37     3.18   0.00
```

15.5.3 Covariate balance checks

Moreover, we can use simple OLS to check covariate balance. Before matching, the covariates are highly imbalanced, signified by many stars associated with the coefficients.

```
> lm.before = lm(z ~ x)
> summary(lm.before)

Call:
lm(formula = z ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.18508 -0.01057  0.00303  0.01018  1.01355 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.404e-03 6.326e-03   0.222   0.8243  
xage        -4.043e-04 8.512e-05  -4.750  2.05e-06 *** 
xeduc       3.220e-04 4.073e-04   0.790   0.4293  

```

```

xblack      1.070e-01  2.902e-03  36.871  < 2e-16 ***
xhispan    6.377e-03  3.103e-03   2.055   0.0399 *
xmarried   -1.525e-02 2.023e-03  -7.537  5.06e-14 ***
xnodegree  1.345e-02  2.523e-03   5.331  9.89e-08 ***
xre74      7.601e-07  1.806e-07   4.208  2.59e-05 ***
xre75      -1.231e-07 1.829e-07  -0.673   0.5011
xu74       4.224e-02  3.271e-03  12.914  < 2e-16 ***
xu75       2.424e-02  3.399e-03   7.133  1.02e-12 ***

```

```

Residual standard error: 0.09935 on 16166 degrees of freedom
Multiple R-squared:  0.1274,   Adjusted R-squared:  0.1269
F-statistic: 236.1 on 10 and 16166 DF,  p-value: < 2.2e-16

```

But after matching, the covariates are well balanced, signified by the absence of stars for all coefficients.

```

> lm.after = lm(z ~ x,
+                 subset = c(matchest$index.treated,
+                               matchest$index.control))
> summary(lm.after)

Call:
lm(formula = z ~ x, subset = c(matchest$index.treated, matchest$index.control))

Residuals:
    Min      1Q      Median      3Q      Max 
-0.66864 -0.49161 -0.03679  0.50378  0.65122 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.003e-01  2.427e-01   2.474   0.0137 *  
xage        3.199e-03  3.427e-03   0.933   0.3511    
xeduc      -1.501e-02 1.634e-02  -0.918   0.3590    
xblack      6.141e-05  7.408e-02   0.001   0.9993    
xhispan    1.391e-02  1.208e-01   0.115   0.9084    
xmarried   -1.328e-02 6.729e-02  -0.197   0.8437    
xnodegree -3.023e-02 7.144e-02  -0.423   0.6723    
xre74      6.754e-06  9.864e-06   0.685   0.4939    
xre75      -9.848e-06 1.279e-05  -0.770   0.4417    
xu74       2.179e-02  1.027e-01   0.212   0.8321    
xu75      -2.642e-02  8.327e-02  -0.317   0.7512    

Residual standard error: 0.5043 on 485 degrees of freedom
Multiple R-squared:  0.005101,   Adjusted R-squared:  -0.01541
F-statistic: 0.2487 on 10 and 485 DF,  p-value: 0.9909

```

15.6 Discussion

With many covariates, matching based on the original covariates may suffer from the curse of dimensionality. Rosenbaum and Rubin (1983b) suggested to use matching based on the estimated propensity score. Abadie and Imbens (2016) provided a formal theory for this strategy.

15.7 Homework Problems

15.1 Linear expansions of the bias-corrected estimators

Prove Propositions 15.1 and 15.3.

15.2 Doubly robust form of the bias-corrected matching estimator for τ

Prove Proposition 15.2.

15.3 Doubly robust form of the bias-corrected matching estimator for τ_T

Prove Proposition 15.4.

15.4 Data re-analyses

In `OS_ATE.R`, I analyze two datasets using regression imputation, two IPW and the doubly robust estimators. Reanalyze them using the propensity score stratification estimator and the Abadie–Imbens matching estimator. Compare these estimators.

Note that you should choose different number of strata for the propensity score stratification estimator, and check covariate balance. You should also choose different number of matches for the matching estimator. You can even apply various estimators to the matched data. Are your results sensitive to your choices?

15.5 Data re-analyses

In `Matching.R`, I analyzed the LaLonde observational study using matching. Matching performs well because it gives an estimator that is close to the experimental gold standard. Reanalyze the data using the regression imputation, propensity score stratification, two IPW and the doubly robust estimators. Compare the results to the matching estimator and to the estimator from the experimental gold standard.

Note that you have many choices. For example, the number of strata for stratification and the threshold to trim to data based on the estimated propen-

sity scores. You may consider fitting different propensity score and outcome models, e.g., including some quadratic terms of the basic covariates. You can even apply these estimators to the matched data.

This is a classic dataset and hundreds of papers have used it. You can read some references (Dehejia and Wahba, 1999; Hainmueller, 2012) and you can also be creative in your own data analysis.

15.6 Data re-analyses

Ho et al. (2007) is an influential paper in political science, based on which the authors have developed an R package `MatchIt` (Ho et al., 2011). Ho et al. (2007) analyzed two datasets, both of which are available from the Harvard Dataverse.

Reanalyze these two datasets using the methods discussed so far. You can also try other methods as long as you can justify them.

15.7 Recommended reading

The literature of matching estimators is massive, and three excellent review papers are Sekhon (2009), Stuart (2010) and Imbens (2015).



Part IV

Difficulties and challenges of observational studies



16

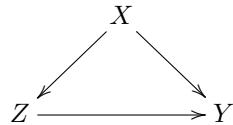
Difficulties of Unconfoundedness in Observational Studies for Causal Effects

Part III of this book discusses causal inference with observational studies under two assumptions: unconfoundedness and overlap. Both are strong assumptions and likely to be violated in practice. This chapter will discuss the difficulties of the unconfoundedness assumption. Chapters 17–19 will discuss various strategies for sensitivity analysis in observational studies with unmeasured confounding. Chapter 20 will discuss the difficulties of the overlap assumption.

16.1 Some basics of the causal diagram

Pearl (1995) introduced the causal diagram as a powerful tool for causal inference in empirical research. Pearl (2000) is a textbook on the causal diagram. Here I introduce the causal diagram as an intuitive tool for illustrating the causal relationships among variables.

For example, if we have the causal diagram

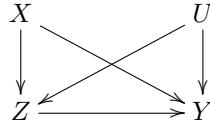


and focus on the causal effect of Z on Y , we can read it as

$$\begin{cases} X \sim F_X(x), \\ Z = f_Z(X, \varepsilon_Z), \\ Y(z) = f_Y(X, z, \varepsilon_Y(z)), \end{cases}$$

where $\varepsilon_Z \perp\!\!\!\perp \varepsilon_Y(z)$ for both $z = 0, 1$. In the above, covariates X are generated from a distribution $F_X(x)$, the treatment assignment is a function of X with a random error term ε_Z , and the potential outcome $Y(z)$ is a function of X , z and a random error term $\varepsilon_Y(z)$. We can easily read from the equations that $Z \perp\!\!\!\perp Y(z) | X$, i.e., the unconfoundedness assumption holds.

If we have a causal diagram



we can read it as

$$\left\{ \begin{array}{l} X \sim F_X(x), \\ U \sim F_U(u), \\ Z = f_Z(X, U, \varepsilon_Z), \\ Y(z) = f_Y(X, U, z, \varepsilon_Y(z)), \end{array} \right.$$

where $\varepsilon_Z \perp\!\!\!\perp \varepsilon_Y(z)$ for both $z = 0, 1$. We can easily read from the equations that $Z \perp\!\!\!\perp Y(z) \mid (X, U)$ but $Z \not\perp\!\!\!\perp Y(z) \mid X$, i.e., the unconfoundedness assumption holds conditioning on (X, U) but does not hold conditioning on X only. In this case, U is an unmeasured confounder. In this diagram, U is called an unmeasured confounder.

16.2 Assessing ignorability

The weak ignorability

$$Z \perp\!\!\!\perp Y(1) \mid X, \quad Z \perp\!\!\!\perp Y(0) \mid X$$

implies that

$$\begin{aligned} \text{pr}\{Y(1) \mid Z = 1, X\} &= \text{pr}\{Y(1) \mid Z = 0, X\}, \\ \text{pr}\{Y(0) \mid Z = 1, X\} &= \text{pr}\{Y(0) \mid Z = 0, X\}. \end{aligned}$$

So the ignorability assumption basically requires that the counterfactual distribution $\text{pr}\{Y(1) \mid Z = 0, X\}$ equals the observed distribution $\text{pr}\{Y(1) \mid Z = 1, X\}$, and the counterfactual distribution $\text{pr}\{Y(0) \mid Z = 1, X\}$ equals the observed distribution $\text{pr}\{Y(0) \mid Z = 0, X\}$. Because the counterfactual distributions are not directly identifiable from the data, the ignorability assumption is fundamentally untestable without additional assumptions. I will discuss two strategies to assess ignorability. Here, “assess” is a weaker notion than “test”. The former is referred to as supplementary analysis that support or undermine the initial analysis, but the latter is referred to formal statistical testing.

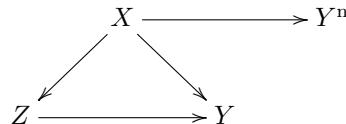
16.2.1 Using negative outcomes

Assume that Y^n is an outcome similar to Y and ideally, shares the same confounding structure as Y . If we believe $Z \perp\!\!\!\perp Y(z) \mid X$, then we also tend to

believe $Z \perp\!\!\!\perp Y^n(z) | X$. Moreover, we know, a priori, the effect of Z on Y^n :

$$\tau(Z \rightarrow Y^n) = E\{Y^n(1) - Y^n(0)\}.$$

An important example is that $\tau(Z \rightarrow Y^n) = 0$. A causal diagram satisfying these requirements is below:



Example 16.1 Cornfield et al. (1959) studied the causal role of cigarette smoking on lung cancer based on observational studies. They controlled for many important background variables but it is still possible to have some unmeasured confounders biasing the observed effects. To strengthen the evidence, they also reported the effect of cigarette smoking on car accident which was close to zero, the anticipated effect based on biology. So even if they could not rule out unmeasured confounding in the analysis, this supplementary analysis based on a negative outcome makes the evidence of the causal effect of cigarette smoking on lung cancer stronger.

Example 16.2 Imbens and Rubin (2015) suggested using the lagged outcome as a negative outcome. In most cases, it is reasonable to believe that the lagged outcome and the outcome have similar confounding structure. Since the lagged outcome happens before the treatment, the average causal effect on it must be 0. However, their suggestion should be used with caution since in most studies we simply treat lagged outcomes as an observed confounder.

In some sense, the covariate balance check in Chapter 11 is a special case of using negative controls. Similar to the problem of using lagged outcomes as negative controls, those covariates are usually a part of the ignorability assumption. Therefore, the failure of covariate balance check does not really falsify the ignorability assumption but rather the model specification of the propensity score.

Example 16.3 Observational studies in elderly persons have shown that vaccination against influenza remarkably reduces one's risk of pneumonia/influenza hospitalization and all-cause mortality in the following season, after adjustment for measured covariates. Jackson et al. (2006) were skeptical about the large magnitude and thus conducted supplementary analysis on negative outcomes. Vaccination often begins in autumn, but influenza transmission is often minimal until winter. Based on biology, the effect of vaccination should be most prominent during influenza season. But Jackson et al. (2006) found greater effect before the influenza season, suggesting that the observed effect is due to unmeasured confounding.

Jackson et al. (2006) seems the most convincing one since the influenza-related outcomes before and during the influenza season should have similar confounding patterns. Cornfield et al. (1959)'s additional evidence seems weaker since car accident and lung cancer have very different causal mechanisms with respect to cigarette smoking. In fact, Fisher (1957)'s critique was that the relationship between cigarette smoking on lung cancer may be due to an unobserved genetic factor. Such a genetic factor might affect cigarette smoking and lung cancer simultaneously, but it seems unlikely that it also affects car accident.

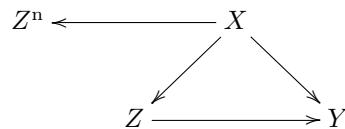
Lipsitch et al. (2010) is a recent article on negative outcomes. Rosenbaum (1989) discussed the role of known effects in causal inference.

16.2.2 Using negative exposures

Negative exposures are duals of negative outcomes. Assume Z^n is a treatment variable similar to Z and shares the same confounding structure as Z . If we believe $Z \perp\!\!\!\perp Y(z) | X$, then we tend to believe $Z^n \perp\!\!\!\perp Y(z) | X$. Moreover, we know, a priori, the effect of Z^n on Y

$$\tau(Z^n \rightarrow Y) = E\{Y(1^n) - Y(0^n)\}.$$

An important example is that $\tau(Z^n \rightarrow Y) = 0$. A causal diagram satisfying these requirements is below:



Example 16.4 Sanderson et al. (2017) give many examples of negative exposures in determining the effect of intrauterine exposure on later outcomes by comparing the association of a maternal exposure during pregnancy with the outcome of interest, with the association of the paternal exposure with the same outcome. They review studies on the effect of maternal and paternal smoking on offspring outcomes, the effect of maternal and paternal BMI on later offspring BMI and autism spectrum disorder. In these examples, we expect the the association of the maternal exposure with the outcome is larger than that of the paternal exposure with the outcome.

16.2.3 Summary

The unconfoundedness assumption is fundamentally untestable without additional assumptions. Although negative outcomes and negative controls in observational studies cannot prove or disprove unconfoundedness, using them in supplementary analyses can strengthen the evidence for causation. However, it is often non-trivial to conduct this type of supplementary analyses

because it involves more data and more importantly, deeper understanding of the causal problems in order to find convincing negative outcomes and negative controls.

16.3 Problems of over adjustment

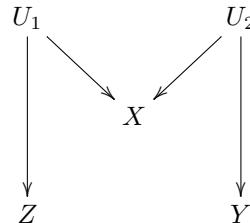
We have discussed many methods for estimating causal effects under ignorability:

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X.$$

This is an assumption conditioning on X . It is crucial to select the right set of X that ensure the conditional independence. Rosenbaum (2002b) wrote that “there is no reason to avoid adjustment for a variable describing subjects before treatment.” Similarly, Rubin (2007) wrote that “typically, the more conditional an assumption, the more acceptable it is.” Both argued that we should control for all observed pretreatment covariate. VanderWeele and Shpitser (2011) called it the *pretreatment criterion*. Pearl disagreed with this recommendation and gave two counterexamples below.

16.3.1 M-bias

M-bias appears in the following causal diagram with an M-structure:



We can read from the diagram the data generating process:

$$\begin{cases} U_1 \perp\!\!\!\perp U_2, \\ X = f_X(U_1, U_2, \varepsilon_X), \\ Z = f_Z(U_1, \varepsilon_Z), \\ Y = Y(z) = f_Y(U_2, \varepsilon_Y), \end{cases}$$

where $(\varepsilon_X, \varepsilon_Z, \varepsilon_Y)$ are independent random error terms. In the above causal diagram, X is observed, but U_1 and U_2 are unobserved. If we change the value of Z , the value of Y will not change at all. So the true causal effect of Z on Y must be 0. From the data-generating equations, we can easily read that $Z \perp\!\!\!\perp Y$,

20616 *Difficulties of Unconfoundedness in Observational Studies for Causal Effects*

so the association between Z and Y is 0, and, in particular,

$$\tau_{\text{PF}} = E(Y | Z = 1) - E(Y | Z = 0) = 0.$$

This means that without adjusting for the covariate X , the simple estimator is unbiased for the true parameter.

However, if we condition on X , then $U_1 \not\perp\!\!\!\perp U_2 | X$, and consequently, $Z \not\perp\!\!\!\perp Y | X$ and

$$\int \{E(Y | Z = 1, X = x) - E(Y | Z = 0, X = x)\} F(dx) \neq 0$$

in general. To gain intuition, we consider the case with Gaussian linear models¹:

$$\begin{cases} X = aU_1 + bU_2 + \varepsilon_X, \\ Z = cU_1 + \varepsilon_Z, \\ Y = Y(z) = dU_2 + \varepsilon_Y, \end{cases}$$

where $(U_1, U_2, \varepsilon_X, \varepsilon_Z, \varepsilon_Y) \stackrel{\text{IID}}{\sim} N(0, 1)$. We have

$$\text{cov}(Z, Y) = \text{cov}(cU_1 + \varepsilon_Z, dU_2 + \varepsilon_Y) = 0,$$

but by the result in Problem 1.2, the partial correlation coefficient between Z and Y given X is

$$\rho_{ZY|X} = \frac{\rho_{ZY} - \rho_{ZX}\rho_{YX}}{\sqrt{1 - \rho_{ZX}^2}\sqrt{1 - \rho_{YX}^2}} \propto -\rho_{ZX}\rho_{YX} \propto -\text{cov}(Z, X)\text{cov}(Y, X) = -abcd,$$

the product of the coefficients on the path from Z to Y . So the unadjusted estimator is unbiased but the adjusted estimator has bias proportional to $abcd$.

The following simple example illustrates M-bias.

```
> n = 10^6
>
> ## M bias
> U1 = rnorm(n)
> U2 = rnorm(n)
> X = U1 + U2 + rnorm(n)
> Z = U1 + rnorm(n)
> Y = U2 + rnorm(n)
>
> round(summary(lm(Y ~ Z))$coef[2, 1], 3)
[1] -0.001
> round(summary(lm(Y ~ Z + X))$coef[2, 1], 3)
[1] -0.201
>
```

¹It is not ideal for our discussion of binary Z , but it simplifies the derivations. Ding and Miratrix (2015) gave detailed discussion with more natural models for binary Z .

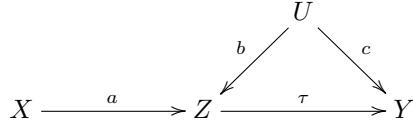
```

> Z = (Z >= 0)
> round(summary(lm(Y ~ Z))$coef[2, 1], 3)
[1] -0.002
> round(summary(lm(Y ~ Z + X))$coef[2, 1], 3)
[1] -0.421

```

16.3.2 Z-bias

Consider the following causal diagram:



with the data generating process²

$$\begin{cases} Z = aX + bU + \varepsilon_Z, \\ Y = \tau z + cU + \varepsilon_Y, \end{cases}$$

where $(U, X, \varepsilon_Z, \varepsilon_Y) \stackrel{\text{IID}}{\sim} N(0, 1)$. In this data generating process, we have $X \perp\!\!\!\perp U$, $X \not\perp\!\!\!\perp Z$, and X affects Y only through Z .

The unadjusted estimator is

$$\tau_{\text{unadj}} = \frac{\text{cov}(Z, Y)}{\text{var}(Z)} = \frac{\text{cov}(Z, \tau Z + cU)}{\text{var}(Z)} = \tau + \frac{ccov(aX + bU, U)}{\text{var}(Z)} = \tau + \frac{cb}{a^2 + b^2 + 1},$$

which has bias $bc/(a^2 + b^2 + 1)$. The adjusted estimator from the OLS of Y on (Z, X) satisfies

$$\begin{cases} E\{Z(Y - \tau_{\text{adj}}Z - \alpha X)\} = 0, \\ E\{X(Y - \tau_{\text{adj}}Z - \alpha X)\} = 0, \end{cases}$$

which is equivalent to

$$\begin{cases} E(ZY) = \tau_{\text{adj}}\text{var}(Z) + \alpha E(XZ), \\ E(XY) = \tau_{\text{adj}}E(XZ) + \alpha \text{var}(X). \end{cases}$$

We need to solve for $(\tau_{\text{adj}}, \alpha)$ from the above two linear equations:

$$\begin{aligned} \tau_{\text{adj}} &= \frac{\begin{vmatrix} E(ZY) & E(XZ) \\ E(XY) & \text{var}(X) \end{vmatrix}}{\begin{vmatrix} \text{var}(Z) & E(XZ) \\ E(XZ) & \text{var}(X) \end{vmatrix}} = \frac{E(ZY)\text{var}(X) - E(XZ)E(XY)}{\text{var}(Z)\text{var}(X) - E(XZ)^2} \\ &= \frac{\tau(a^2 + b^2 + 1) + bc - a\tau a}{(a^2 + b^2 + 1) - a^2} = \frac{\tau(b^2 + 1) + bc}{b^2 + 1} = \tau + \frac{bc}{b^2 + 1}, \end{aligned}$$

²Again, we generate continuous Z from a linear model to simplify the derivations. Ding et al. (2017b) extended the theory to more general causal models, especially for binary Z .

20816 Difficulties of Unconfoundedness in Observational Studies for Causal Effects

which has bias $bc/(b^2 + 1)$.

So the unadjusted estimator has smaller bias than the adjusted estimator. More interestingly, the stronger the association between X and Z is (measured by a), the larger the bias of the adjusted estimator is.

The mathematical derivation is not extremely hard. But this type of bias seems rather mysterious. Here is the intuition. The treatment is a function of X , U , and other random errors. If we condition on X , it is merely a function of U and other random errors. Therefore, conditioning makes Z less random, and more critically, makes the unmeasured confounder U play a more important role in Z . Consequently, the confounding bias due to U is amplified by conditioning on X . This idealized example illustrates the danger of over adjusting for some covariates.

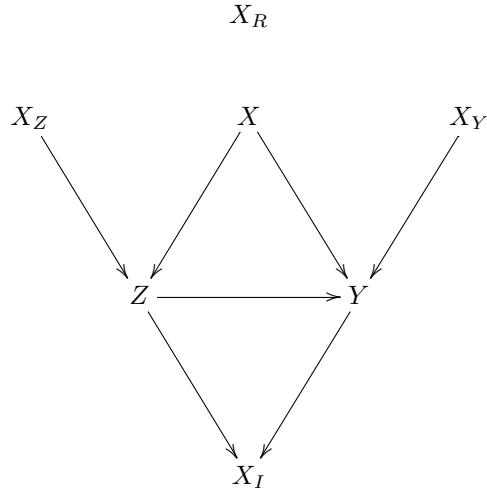
Heckman and Navarro-Lozano (2004) observed the phenomenon in simulation studies, and Wooldridge (2016, technical report in 2006) verified it in linear models. Pearl (2010, 2011) explained it using causal diagrams. This type of bias is called Z-bias because in Pearl's original papers, he used the symbol Z for our variable X . Throughout the book, however, Z is used for the treatment variable. In Part V of this book, we will call Z an instrumental variable if it satisfies the causal diagram presented in this subsection. This justifies the *instrumental variable bias* as another name of this type of bias.

The following simple example illustrates Z-bias.

```
> X = rnorm(n)
> U = rnorm(n)
> Z = X + U + rnorm(n)
> Y = U + rnorm(n)
>
> round(summary(lm(Y ~ Z))$coef[2, 1], 3)
[1] 0.334
> round(summary(lm(Y ~ Z + X))$coef[2, 1], 3)
[1] 0.501
>
> Z = 2*X + U + rnorm(n)
> round(summary(lm(Y ~ Z))$coef[2, 1], 3)
[1] 0.167
> round(summary(lm(Y ~ Z + X))$coef[2, 1], 3)
[1] 0.5
>
> Z = 10*X + U + rnorm(n)
> round(summary(lm(Y ~ Z))$coef[2, 1], 3)
[1] 0.01
> round(summary(lm(Y ~ Z + X))$coef[2, 1], 3)
[1] 0.5
```

16.3.3 What covariates should we adjust for in observational studies?

We never know the true underlying data generating process which can be quite complicated. However, the following causal diagram helps to clarify many ideas. It already rules out the possibility of M -bias discussed in Section 16.3.1.



The covariates above have different features:

1. X affects both the treatment and the outcome. Conditioning on X ensures ignorability, so we should control for X .
2. X_R is pure random noise not affecting either the treatment or the outcome. Including it in analysis does not bias the estimate but it introduces unnecessary variability in finite sample.
3. X_Z is an instrumental variable that affects the outcome only through the treatment. In the diagram above, including it in analysis does not bias the estimate although it increases variability. However, with unmeasured confounding, including it in analysis amplifies the bias as shown in Section 16.3.1.
4. X_Y affects the outcome only but not the treatment. Without conditioning on it, the ignorability still holds. Since they are predictive to the outcome, including them in analysis often improves precision.
5. X_I is affected by the treatment and outcome. It is a post-treatment variable, not a pretreatment covariate. We should not include it if the goal is to infer the effect of the treatment on the outcome. We will discuss issues with post-treatment variables in causal inference in Part VI of this book.

If we believe the above causal diagram, we should adjust for at least X to remove bias and more ideally, further adjust for X_Y to reduce variance.

16.4 Homework Problems

16.1 Cochran's formula or the omitted variable bias formula

Sir David Cox calls the following result Cochran's formula (Cochran, 1938; Cox, 2007) and econometricians call it the omitted variable bias formula (Angrist and Pischke, 2008). A special case appeared in Fisher (1925). It is also a sister of the Frisch–Waugh–Lovell Theorem in Chapter A2.3.

The formula has two versions. All vectors below are column vectors.

1. (Population version) Assume $(y_i, x_{i1}, x_{i2})_{i=1}^n$ are iid, where y_i is a scalar, x_{i1} has dimension K , and x_{i2} has dimension L .

We have the following OLS decompositions of random variables

$$y_i = \beta_1^\top x_{i1} + \beta_2^\top x_{i2} + \varepsilon_i, \quad (16.1)$$

$$y_i = \gamma^\top x_{i1} + e_i, \quad (16.2)$$

$$x_{i2} = \delta^\top x_{i1} + v_i. \quad (16.3)$$

Equation (16.1) is called the long regression, and Equation (16.2) is called the short regression. In Equation (16.3), δ is a matrix because it is a regression of a vector on a vector. You can view (16.3) as regression of each component of x_{i2} on x_{i1} .

Show that $\gamma = \beta_1 + \delta\beta_2$.

2. (Sample version) We have an $n \times 1$ vector Y , an $n \times K$ matrix X_1 , and an $n \times L$ matrix X_2 . We do not assume any randomness. All results below are purely linear algebra.

We can obtain the following OLS fits:

$$Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{\varepsilon},$$

$$Y = X_1 \hat{\gamma} + \hat{e},$$

$$X_2 = X_1 \hat{\delta} + \hat{v},$$

where $\hat{\varepsilon}, \hat{e}, \hat{v}$ are the residuals. Again, the last OLS fit means the OLS fit of each column of X_2 on X_1 , and therefore the residual \hat{v} is an $n \times L$ matrix.

Show that $\hat{\gamma} = \hat{\beta}_1 + \hat{\delta}\hat{\beta}_2$.

Remark: The product terms $\delta\beta_2$ and $\hat{\delta}\hat{\beta}_2$ are often referred to as the omitted-variable bias at the population level and sample level, respectively.

16.2 Recommended reading

Imbens (2020) reviews and compares the roles of potential outcomes and causal diagrams for causal inference.

17

E-Value: Evidence for Causation in Observational Studies with Unmeasured Confounding

All the methods discussed in Part III rely crucially on the ignorability assumption. They require controlling for all confounding between the treatment and outcome. However, we cannot use the data to validate the ignorability assumption. Observational studies are often criticized due to the possibility of unmeasured confounding. The famous Yule–Simpson Paradox demonstrates that an unmeasured binary confounder can completely overturn an observed association between the treatment and outcome. However, to overturn a larger observed association, this unmeasured confounder must have stronger association with the treatment and the outcome. In other words, not all observational studies are created equal. Some provide stronger evidence for causation than others.

The following three chapters will discuss various sensitivity analysis techniques that can quantify the evidence of causation based on observational studies in the presence of unmeasured confounding. This chapter starts with the E-value, introduced by VanderWeele and Ding (2017) based on the theory in Ding and VanderWeele (2016). It is more useful for observational studies using logistic regressions. Chapter 18 discusses sensitivity analysis for the average causal effect based on inverse probability weighting, outcome regression, and doubly robust estimation. Chapter 19 discusses Rosenbaum’s framework for sensitivity analysis for matched observational studies.

17.1 Cornfield-type sensitivity analysis

Although we do not assume ignorability given X :

$$Z \not\perp\!\!\!\perp \{Y(1), Y(0)\} \mid X,$$

we still assume latent ignorability given X and another unmeasured confounder U :

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid (X, U).$$

The technique in this chapter works the best for a binary outcome Y although it can be extended to other non-negative outcomes (Ding and VanderWeele, 2016). Focus on binary Y now. The true conditional causal effect on the risk ratio scale is defined as

$$\text{RR}_{ZY|x}^{\text{true}} = \frac{\text{pr}\{Y(1) = 1 | X = x\}}{\text{pr}\{Y(0) = 1 | X = x\}},$$

and the observed conditional risk ratio equals

$$\text{RR}_{ZY|x}^{\text{obs}} = \frac{\text{pr}(Y = 1 | Z = 1, X = x)}{\text{pr}(Y = 1 | Z = 0, X = x)}.$$

In general, with an unmeasured confounder U ,

$$\text{RR}_{ZY|x}^{\text{true}} \neq \text{RR}_{ZY|x}^{\text{obs}}$$

because

$$\text{RR}_{ZY|x}^{\text{true}} = \frac{\int \text{pr}(Y = 1 | Z = 1, X = x, U = u)F(du | X = x)}{\int \text{pr}(Y = 1 | Z = 0, X = x, U = u)F(du | X = x)}$$

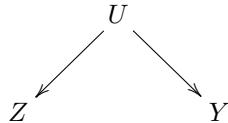
and

$$\text{RR}_{ZY|x}^{\text{obs}} = \frac{\int \text{pr}(Y = 1 | Z = 1, X = x, U = u)F(du | Z = 1, X = x)}{\int \text{pr}(Y = 1 | Z = 0, X = x, U = u)F(du | Z = 0, X = x)}$$

are averaged over different distributions of U .

Doll and Hill (1950) found that the risk ratio of cigarette smoking on lung cancer was 9 even after adjusting for many observed covariates X ¹. Fisher (1957) criticized their result to be noncausal because it is possible that a hidden gene simultaneously causes cigarette smoking and lung cancer although the true causal effect of cigarette smoking on lung cancer is absent. This is the *common cause* hypothesis, also discussed by Reichenbach (1957). Cornfield et al. (1959) took a more constructive perspective and asked: how strong this unmeasured confounder must be in order to explain away the observed association between cigarette smoking and lung cancer? Below we will use Ding and VanderWeele (2016)'s general formulation of the problem.

Consider the following causal diagram:



which conditions on X . So $Z \perp\!\!\!\perp Y | (X, U)$. Conditioning on X and U , we

¹Their original analysis was based on a case-control study and estimated the odds ratio of cigarette smoking on lung cancer. But the risk ratio is close to the odds ratio since lung cancer is a rare outcome.

observe no association between Z and Y ; but conditioning on only X , we observe association between Z and Y . Although we can allow U to be general as Ding and VanderWeele (2016), we assume that U is binary to simplify the presentation.

Define two sensitivity parameters:

$$\text{RR}_{ZU|x} = \frac{\text{pr}(U = 1 | Z = 1, X = x)}{\text{pr}(U = 1 | Z = 0, X = x)} \equiv \frac{f_{1,x}}{f_{0,x}}$$

measures the treatment-confounder association, and

$$\text{RR}_{UY|x} = \frac{\text{pr}(Y = 1 | U = 1, X = x)}{\text{pr}(Y = 1 | U = 0, X = x)},$$

measures the confounder-outcome association, conditional on covariates $X = x$. Without loss of generality, we assume that $\text{RR}_x^{\text{obs}} > 1$, $\text{RR}_{ZU|x} > 1$, and $\text{RR}_{UY|x} > 1$. We can show the main result below.

Theorem 17.1 *Under $Z \perp\!\!\!\perp Y | (X, U)$, we have*

$$\text{RR}_{ZY|x}^{\text{obs}} \leq \frac{\text{RR}_{ZU|x} \text{RR}_{UY|x}}{\text{RR}_{ZU|x} + \text{RR}_{UY|x} - 1}.$$

Theorem 17.1 shows the upper bound of the observed risk ratio of the treatment on the outcome if the conditional independence $Z \perp\!\!\!\perp Y | (X, U)$ holds. Under this conditional independence assumption, the association between the treatment and the outcome is purely due to the association between the treatment, $\text{RR}_{ZU|x}$, and the confounder and the association between the confounder and the outcome, $\text{RR}_{UY|x}$. The upper bound equals $\text{RR}_{ZU|x} \text{RR}_{UY|x} / (\text{RR}_{ZU|x} + \text{RR}_{UY|x} - 1)$. A similar inequality appeared in Lee (2011). It is also related to Cochran's formula or the omitted-variable bias formula for linear models, which was reviewed in Problem 16.1.

Reversely, to generate a certain value of the observed risk ratio RR_x^{obs} , the two confounding measures $\text{RR}_{ZU|x}$ and $\text{RR}_{UY|x}$ cannot be arbitrary. Their function $\text{RR}_{ZU|x} \text{RR}_{UY|x} / (\text{RR}_{ZU|x} + \text{RR}_{UY|x} - 1)$ must be at least at large as RR_x^{obs} .

I will give the proof of Theorem 17.1 below.

Proof of Theorem 17.1: We can decompose $\text{RR}_{ZY|x}^{\text{obs}}$ as

$$\begin{aligned}
 & \text{RR}_{ZY|x}^{\text{obs}} \\
 = & \frac{\text{pr}(Y = 1 | Z = 1, X = x)}{\text{pr}(Y = 1 | Z = 0, X = x)} \\
 = & \frac{\left[\text{pr}(U = 1 | Z = 1, X = x) \text{pr}(Y = 1 | Z = 1, U = 1, X = x) \right.}{\left. + \text{pr}(U = 0 | Z = 1, X = x) \text{pr}(Y = 1 | Z = 1, U = 0, X = x) \right]} \\
 = & \frac{\left[\text{pr}(U = 1 | Z = 0, X = x) \text{pr}(Y = 1 | Z = 0, U = 1, X = x) \right.}{\left. + \text{pr}(U = 0 | Z = 0, X = x) \text{pr}(Y = 1 | Z = 0, U = 0, X = x) \right]} \\
 = & \frac{\left[\text{pr}(U = 1 | Z = 1, X = x) \text{pr}(Y = 1 | U = 1, X = x) \right.}{\left. + \text{pr}(U = 0 | Z = 1, X = x) \text{pr}(Y = 1 | U = 0, X = x) \right]} \\
 = & \frac{\left[\text{pr}(U = 1 | Z = 0, X = x) \text{pr}(Y = 1 | U = 1, X = x) \right.}{\left. + \text{pr}(U = 0 | Z = 0, X = x) \text{pr}(Y = 1 | U = 0, X = x) \right]} \\
 = & \frac{f_{1,x} \text{RR}_{UY|x} + 1 - f_{1,x}}{f_{0,x} \text{RR}_{UY|x} + 1 - f_{0,x}} \\
 = & \frac{(\text{RR}_{UY|x} - 1)f_{1,x} + 1}{\frac{\text{RR}_{UY|x} - 1}{\text{RR}_{ZU|x}} f_{1,x} + 1}.
 \end{aligned}$$

We can verify that $\text{RR}_{ZY|x}^{\text{obs}}$ is increasing in $f_{1,x}$. So letting $f_{1,x} = 1$, we have

$$\text{RR}_{ZY|x}^{\text{obs}} \leq \frac{(\text{RR}_{UY|x} - 1) + 1}{\frac{\text{RR}_{UY|x} - 1}{\text{RR}_{ZU|x}} + 1} = \frac{\text{RR}_{ZU|x} \text{RR}_{UY|x}}{\text{RR}_{ZU|x} + \text{RR}_{UY|x} - 1}.$$

□

In the proof of Theorem 17.1, we have obtain an identity

$$\text{RR}_{ZY|x}^{\text{obs}} = \frac{(\text{RR}_{UY|x} - 1)f_{1,x} + 1}{\frac{\text{RR}_{UY|x} - 1}{\text{RR}_{ZU|x}} f_{1,x} + 1}.$$

But this identity involves three parameters

$$\{f_{1,x}, \text{RR}_{ZU|x}, \text{RR}_{UY|x}\};$$

see Problem 17.2 for a related formula. In contrast, the upper bound in Theorem 17.1 involves only two parameters

$$\{\text{RR}_{ZU|x}, \text{RR}_{UY|x}\}$$

which measure the strength of the confounder.

17.2 E-value

Lemma 17.1 below is useful for deriving interesting corollaries of Theorem 17.1.

Lemma 17.1 Define $\beta(w_1, w_2) = w_1 w_2 / (w_1 + w_2 - 1)$ for $w_1 > 1$ and $w_2 > 1$.

1. $\beta(w_1, w_2)$ is symmetric in w_1 and w_2 ;
2. $\beta(w_1, w_2)$ increasing in both w_1 and w_2 ;
3. $\beta(w_1, w_2) \leq w_1$ and $\beta(w_1, w_2) \leq w_2$;
4. $\beta(w_1, w_2) \leq w^2 / (2w - 1)$, where $w = \max(w_1, w_2)$.

Using Theorem 17.1 and Lemma 17.1(3), we have

$$\text{RR}_{ZU|x} \geq \text{RR}_{ZY|x}^{\text{obs}}, \quad \text{RR}_{UY|x} \geq \text{RR}_{ZY|x}^{\text{obs}},$$

or, equivalently,

$$\min(\text{RR}_{ZU|x}, \text{RR}_{UY|x}) \geq \text{RR}_{ZY|x}^{\text{obs}}.$$

Therefore, to explain away the observed relative risk, both confounding measures $\text{RR}_{ZU|x}$ and $\text{RR}_{UY|x}$ must be at least as large as $\text{RR}_{ZY|x}^{\text{obs}}$. Cornfield et al. (1959) first derived the inequality $\text{RR}_{ZU|x} \geq \text{RR}_{ZY|x}^{\text{obs}}$, also called the *Cornfield inequality* (Gastwirth et al., 1998). Schlesselman (1978) derived the inequality $\text{RR}_{UY|x} \geq \text{RR}_{ZY|x}^{\text{obs}}$. These are related to the *data processing inequality* in information theory².

If we define $w = \max(\text{RR}_{ZU|x}, \text{RR}_{UY|x})$, then we can use Theorem 17.1 and Lemma 17.1(4) to obtain

$$\begin{aligned} w^2 / (2w - 1) &\geq \beta(\text{RR}_{ZU|x}, \text{RR}_{UY|x}) \geq \text{RR}_x^{\text{obs}} \\ \implies w^2 - 2\text{RR}_x^{\text{obs}}w + \text{RR}_x^{\text{obs}} &\geq 0, \end{aligned}$$

which is a quadratic inequality. One root $\text{RR}_{ZY|x}^{\text{obs}} - \sqrt{\text{RR}_{ZY|x}^{\text{obs}}(\text{RR}_{ZY|x}^{\text{obs}} - 1)}$ is always smaller than or equal to 1, so we have

$$w = \max(\text{RR}_{ZU|x}, \text{RR}_{UY|x}) \geq \text{RR}_{ZY|x}^{\text{obs}} + \sqrt{\text{RR}_{ZY|x}^{\text{obs}}(\text{RR}_{ZY|x}^{\text{obs}} - 1)}.$$

Therefore, to explain away the observed relative risk, the maximum of the confounding measures $\text{RR}_{ZU|x}$ and $\text{RR}_{UY|x}$ must be at least as large as $\text{RR}_{ZY|x}^{\text{obs}} + \sqrt{\text{RR}_{ZY|x}^{\text{obs}}(\text{RR}_{ZY|x}^{\text{obs}} - 1)}$. Based on this result, VanderWeele and Ding (2017) introduced the following notion of E-value for measuring the *evidence* of causation with observational studies.

²In information theory, the *mutual information*

$$I(A, B) = \iint p(a, b) \log_2 \frac{p(a, b)}{p(a)p(b)} da db$$

measures the dependence between two random variables A and B , where $p(\cdot)$ denotes the joint or marginal density of (A, B) . The *data processing inequality* is a famous result: if $Z \perp\!\!\!\perp Y | U$, then $I(Z, Y) \geq I(Z, U)$ and $I(Z, Y) \geq I(U, Y)$. Lihua Lei and Bin Yu pointed out to me the connection between Cornfield's inequality and the data processing inequality.

Definition 17.1 (E-Value) With the observed conditional risk ratio $\text{RR}_{ZY|x}^{\text{obs}}$, define the E-Value as

$$\text{RR}_{ZY|x}^{\text{obs}} + \sqrt{\text{RR}_{ZY|x}^{\text{obs}}(\text{RR}_{ZY|x}^{\text{obs}} - 1)}$$

The E-value is defined for the parameter $\text{RR}_{ZY|x}^{\text{obs}}$. In practice, $\text{RR}_{ZY|x}^{\text{obs}}$ is estimated with sampling error. We can calculate the E-value based on the estimated $\text{RR}_{ZY|x}^{\text{obs}}$, as well as the corresponding E-values for the lower and upper confidence limits of $\text{RR}_{ZY|x}^{\text{obs}}$.

Fisher's p -value measures the evidence for causal effects in randomized experiment. We have discussed the p -value based on the FRT in Part II of this book. However, in observational studies with large sample sizes, p -values can be a poor measure of evidence for causal effects. Even if the true causal effects are 0, a tiny amount of unmeasured confounding can bias the estimate, which can result in extremely small p -values given the small sampling uncertainty. The sampling uncertainty is usually secondary in observational studies with large sample sizes, but the uncertainty due to unmeasured confounding is often the first order problem that does not diminish with increased sample sizes. VanderWeele and Ding (2017) argued that the E-value is a better measure of the evidence for causal effects in observational studies.

17.3 A classic example

I revisit a classic example below.

Example 17.1 Hammond and Horn (1958) used the U.S. population to study the cigarette smoking and lung cancer relationship. Ignoring covariates, their data can be represented by a 2×2 table:

	Lung cancer	No lung cancer
Smoker	397	78557
Non-smoker	51	108778

Based on the data, they obtained an estimate of the risk ratio 10.73 with a 95% confidence interval [8.02, 14.36]. To explain away the point estimate, the E-value is

$$10.73 + \sqrt{10.73 \times (10.73 - 1)} = 20.95;$$

to explain away the lower confidence limit, the E-value is

$$8.02 + \sqrt{8.02 \times (8.02 - 1)} = 15.52.$$

Figure 17.1 shows the joint values of the two confounding measures to

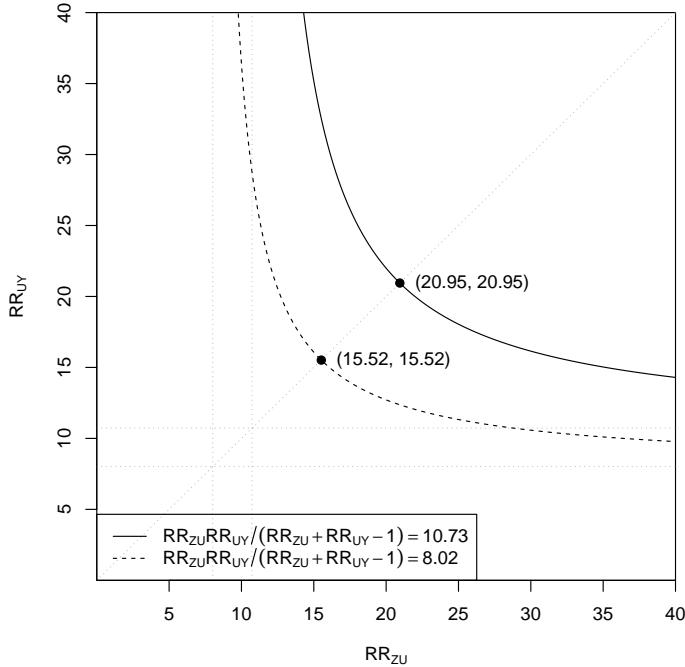


FIGURE 17.1: Magnitude of confounding to explain away the observed risk ratio in Hammond and Horn (1958)

explain away the point estimate and lower confidence limit of the risk ratio. In particular, to explain way the point estimate, they must lie in the area above the solid curve; to explain away the lower confidence limit, they must lie in the area above the dashed curve.

17.4 Extensions

17.4.1 E-value and Bradford Hill's criteria for causation

The E-value provides evidence for causation. But evidence is not a proof. With a larger E-value, we need a stronger unmeasured confounder to explain away the observed risk ratio; the evidence for causation is stronger. With a smaller E-value, we need a weaker unmeasured confounder to explain away

the observed risk ratio; the evidence for causation is weaker. Coupled with the discussion in Section 17.5.1, a larger observed risk ratio have stronger evidence for causation. This is closely related to Sir Bradford Hill's first criterion for causation: *strength* of the association (Bradford Hill, 1965). Theorem 17.1 provides a mathematical quantification of his heuristic argument.

In a famous paper, Bradford Hill (1965) proposed a set of nine criteria to provide evidence for causation between a presumed cause and outcome. His criteria are

1. strength;
2. consistency;
3. specificity;
4. temporality;
5. biological gradient;
6. plausibility;
7. coherence;
8. experiment;
9. analogy.

The E-value is a way to justify his first criterion. That is, stronger association often provides stronger evidence for causation because to explain a stronger association, we need stronger confounding measures. We have discussed randomized experiments in Part II, which corroborates his eighth criterion. Due to the space limit, I omit the detailed discussion of his other criteria and encourage the readers to read (Bradford Hill, 1965). Recently, this paper is re-printed as Bradford Hill (2020) with insightful comments from many leading researchers in causal inference.

17.4.2 E-value after logistic regression

With a binary outcome, it is common for epidemiologists to use a logistic regression of the outcome Y_i on the treatment indicator Z_i and covariates X_i :

$$\text{pr}(Y_i = 1 \mid Z_i, X_i) = \frac{e^{\beta_0 + \beta_1 Z_i + \beta_2^\top X_i}}{1 + e^{\beta_0 + \beta_1 Z_i + \beta_2^\top X_i}}.$$

In the logistic model above, the coefficient of Z_i is the log of the conditional odds ratio between the treatment and the outcome given the covariates:

$$\beta_1 = \log \frac{\text{pr}(Y_i = 1 \mid Z_i = 1, X_i = x)/\text{pr}(Y_i = 0 \mid Z_i = 1, X_i = x)}{\text{pr}(Y_i = 1 \mid Z_i = 0, X_i = x)/\text{pr}(Y_i = 0 \mid Z_i = 0, X_i = x)}.$$

Importantly, the logistic model assumes a common odds ratio across all values of the covariates. Moreover, when the outcome is rare in that $\text{pr}(Y_i = 1 \mid Z_i =$

$1, X_i = x)$ and $\text{pr}(Y_i = 1 | Z_i = 0, X_i = x)$ are close to 0, the conditional odds ratio approximates the conditional risk ratio (see Proposition 1.1(3)):

$$\beta_1 \approx \log \frac{\text{pr}(Y_i = 1 | Z_i = 1, X_i = x)}{\text{pr}(Y_i = 1 | Z_i = 0, X_i = x)} = \log \text{RR}_{ZY|x}^{\text{obs}}.$$

Therefore, based on the estimated logistic regression coefficient and the corresponding confidence limits, we can calculate the E-value immediately. This is the leading application of the E-value.

Example 17.2 The *NCHS2003.txt* contains the National Center for Health Statistics birth certificate data, with the following binary indicator variables useful for us:

<i>PTbirth</i>	pre-term birth
<i>preeclampia</i>	pre-eclampsia ³
<i>ageabove35</i>	an older mother with age ≥ 35 (the treatment)
<i>somecollege</i>	college education
<i>mar</i>	marital status
<i>smoking</i>	smoking status
<i>drinking</i>	drinking status
<i>hispanic</i>	mother's ethnicity
<i>black</i>	mother's ethnicity
<i>nativeamerican</i>	mother's ethnicity
<i>asian</i>	mother's ethnicity

This version of the data is from Valeri and Vanderweele (2014). This example focuses on the outcome *PTbirth* and Problem 17.3. The following R code computes the E-values after fitting a logistic regression. Based on the E-values, we conclude that to explain away the point estimate, the maximum confounding measure must be larger than 1.94, and to explain away the lower confidence limit, the maximum confounding measure must be larger than 1.91. Although these confounding measures are not as strong as those in Section 17.3, they appear to be fairly large in epidemiologic studies.

```
> eval.evalue = function(rr)
+ {
+   rr + sqrt(rr*(rr - 1))
+ }
>
> NCHS2003 = read.table("NCHS2003.txt", header = TRUE, sep = "\t")
>
> ## outcome: PTbirth
> y_logit = glm(PTbirth ~ ageabove35 +
+               mar + smoking + drinking + somecollege +
+               hispanic + black + nativeamerican + asian,
+               data = NCHS2003,
+               family = binomial)
> log.or    = summary(y_logit)$coef[2, 1:2]
```

```

> est      = exp(log_or[1])
> lower.ci = exp(log_or[1] - 1.96*log_or[2])
> est
Estimate
1.305982
> evalute(est)
Estimate
1.938127
>
> lower.ci
Estimate
1.294619
> evalute(lower.ci)
Estimate
1.912211

```

17.4.3 Non-zero true causal effect

Theorem 17.1 assumes no true causal effect of the treatment on the outcome. Ding and VanderWeele (2016) proved a general theorem allowing for non-zero true causal effect.

Theorem 17.2 *Modify the definition of $\text{RR}_{UY|x}$ as*

$$\text{RR}_{UY|x} = \max_{z=0,1} \frac{\text{pr}(Y = 1 | Z = z, U = 1, X = x)}{\text{pr}(Y = 1 | Z = z, U = 0, X = x)}.$$

We have

$$\text{RR}_{ZY|x}^{\text{true}} \geq \text{RR}_{ZY|x}^{\text{obs}} / \frac{\text{RR}_{ZU|x} \text{RR}_{UY|x}}{\text{RR}_{ZU|x} + \text{RR}_{UY|x} - 1}.$$

Theorem 17.1 is a special case of Theorem 17.2 with $\text{RR}_{ZY|x}^{\text{true}} = 1$. See the original paper of Ding and VanderWeele (2016) for the proof of Theorem 17.2. Without assuming any additional assumptions, Theorem 17.2 states a lower bound of the true risk ratio $\text{RR}_{ZY|x}^{\text{true}}$ given the observed risk ratio $\text{RR}_{ZY|x}^{\text{obs}}$ and the two sensitivity parameters $\text{RR}_{ZU|x}$ and $\text{RR}_{UY|x}$.

When the treatment is apparently preventive to the outcome, the observed risk ratio is smaller than 1. In this case, Theorems 17.1 and 17.2 are not directly useful, and we must re-label the treatment levels and calculate the E-value based on $1/\text{RR}_{ZY|x}^{\text{obs}}$.

17.5 Critiques and responses

Since the original paper was published, E-value has been a standard number reported by many epidemiologic studies. Nevertheless, it also attracted

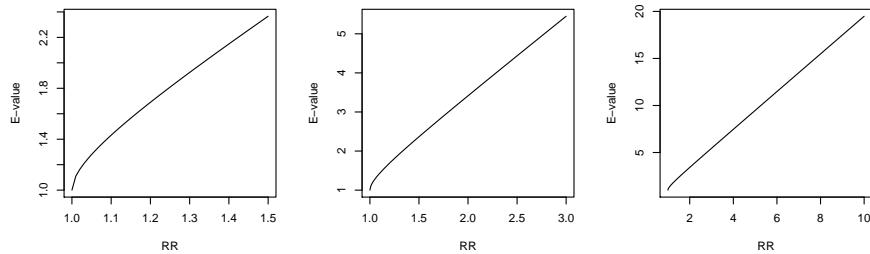


FIGURE 17.2: E-value as a monotone transformation of the risk ratio: three figures have different ranges of the risk ratio.

critiques (Ioannidis et al., 2019). I will review some limitations of E-values below.

17.5.1 E-value is just a monotone transformation of the risk ratio

From Figure 17.2, we can see that if the risk ratio is large, then the E-value $\text{RR}_{ZY|x}^{\text{obs}} + \sqrt{\text{RR}_{ZY|x}^{\text{obs}}(\text{RR}_{ZY|x}^{\text{obs}} - 1)}$ is nearly $2\text{RR}_{ZY|x}^{\text{obs}}$ which is linear in the risk ratio. For small risk ratio, the E-value is more nonlinear. Critics often say that the E-value is merely a monotone transformation of the point estimator or the confidence limits of the risk ratio. So it does not provide any additional information.

This is partially true. Indeed, the E-value is entirely based on the point estimator or the confidence limits of the risk ratio. It has a meaningful interpretation based on Theorem 17.1: to explain away the observed risk ratio, the maximum of the confounding measures must be at least as large as the E-value.

17.5.2 Calibration of the E-value

The E-value equals the maximum value of the association between the confounder and the treatment and that between the confounder and the outcome to completely explain away an observed association. An obvious problem is that this confounder is fundamentally latent. So it is not trivial to decide whether a certain E-value is large or small. Another related problem is that the E-value depends on how many observed covariates X we have controlled for since it quantifies the strength of the residual confounding given X . Therefore, E-values across studies are not directly comparable. The E-value provides evidence for causation but this evidence should be assessed carefully based on background knowledge of the problem of interest.

The following leave-one-covariate-out approach is an intuitive approach to calibrate the E-value. With $X = (X_1, \dots, X_p)$, we can pretend that the component X_j were not observed and compute the Z - X_j and X_j - Y risk ratios given other observed covariates ($j = 1, \dots, p$). These risk ratios provide the range for the confounding measures due to U if we believe that the unmeasured U is not as strong as all of the observed covariates. However, I am not aware of any formal justification of this approach.

17.5.3 It works the best for a binary outcome and the risk ratio

Theorem 17.1 works well for a binary outcome and the risk ratio. Ding and VanderWeele (2016) also proposed sensitivity analysis methods for other causal parameters, but they are not as elegant as the E-value for binary outcome based on the risk ratio. The next chapter will propose a simple sensitivity analysis method for the average causal effect that include several methods in Part III as special cases.

17.6 Homework Problems

17.1 Lemma 17.1

Prove Lemma 17.1.

17.2 Schlesselman (1978)'s formula

For simplicity, we condition on X implicitly in the following discussion. With binary treatment Z , outcome Y , and unmeasured confounder U , show that

$$\frac{\text{RR}_{ZY}^{\text{obs}}}{\text{RR}_{ZY}^{\text{true}}} = \frac{1 + (\gamma - 1)\text{pr}(U = 1 \mid Z = 1)}{1 + (\gamma - 1)\text{pr}(U = 1 \mid Z = 0)}$$

assuming a common risk ratio of the treatment on the outcome within both $U = 0$ and $U = 1$:

$$\text{RR}_{ZY|U=0} = \text{RR}_{ZY|U=1},$$

and also a common risk ratio of the confounder on the outcome within both $Z = 0$ and $Z = 1$:

$$\text{RR}_{UY|Z=0} = \text{RR}_{UY|Z=1}, \text{ denoted by } \gamma.$$

Hint: First verify that if $\text{RR}_{ZY|U=0} = \text{RR}_{ZY|U=1}$ then

$$\text{RR}_{ZY}^{\text{true}} = \text{RR}_{ZY|U=0} = \text{RR}_{ZY|U=1}.$$

This identity shows the *collapsibility* of the risk ratio. In epidemiology, the risk ratio is a *collapsible* measure of association.

Remark: Schlesselman (1978)'s formula does not assume conditional independence $Z \perp\!\!\!\perp Y | U$, but assumes homogeneity of the $Z-Y$ and $U-Y$ risk ratios. It is a classic formula for sensitivity analysis. It is an identity that is simple to implement with pre-specified

$$\{\gamma, \text{pr}(U = 1 | Z = 1), \text{pr}(U = 1 | Z = 0)\}.$$

However, it involves more sensitivity parameters than Theorem 17.1. Even though Theorem 17.1 only gives an inequality, it is not a loose inequality compared to Schlesselman (1978)'s formula under stronger assumptions. With Theorem 17.1, Schlesselman (1978)'s formula is only of historical interest.

17.3 E-value after logistic regression: data analysis

This problem uses the same dataset as Example 17.2.

Report the E-value for the outcome **preeclampsia**.

17.4 Cornfield-type inequalities for the risk difference

Consider binary Z, Y, U , and condition on X implicitly. Assume latent ignorability given U . Show that under $Z \perp\!\!\!\perp Y | U$, we have

$$\text{RD}_{ZY}^{\text{obs}} = \text{RD}_{ZU} \times \text{RD}_{UY} \quad (17.1)$$

where $\text{RD}_{ZY}^{\text{obs}}$ is the observed risk difference of Z on Y , and RD_{ZU} and RD_{UY} are the treatment-confounder and confounder-outcome risk differences, respectively (recall the definition of the risk difference in Chapter 1.2.2).

Remark: Without loss of generality, assume that $\text{RD}_{ZY}^{\text{obs}}, \text{RD}_{ZU}, \text{RD}_{UY}$ are all positive. Then (17.1) implies that

$$\min(\text{RD}_{ZU}, \text{RD}_{UY}) \geq \text{RD}_{ZY}^{\text{obs}}$$

and

$$\max(\text{RD}_{ZU}, \text{RD}_{UY}) \geq \sqrt{\text{RD}_{ZY}^{\text{obs}}}.$$

These are the Cornfield inequalities for the risk difference with a binary confounder. They show that for an unmeasured confounder to explain away an observed risk difference $\text{RD}_{ZY}^{\text{obs}}$, the treatment-confounder and confounder-outcome risk differences must both be larger than $\text{RD}_{ZY}^{\text{obs}}$, and the maximum of them must be larger than the square root of $\text{RD}_{ZY}^{\text{obs}}$.

Cornfield et al. (1959) obtained, but did not appreciate the significance of (17.1). Gastwirth et al. (1998) and Poole (2010) discussed the first Cornfield condition for the risk difference, and Ding and VanderWeele (2014) discussed the second.

Ding and VanderWeele (2014) also derived more general results without assuming a binary U . Unfortunately, the results for a general U are weaker

than those above for a binary U , that is, the inequalities become looser with more levels of U . This motivated Ding and VanderWeele (2016) to focus on the Cornfield inequalities for the risk ratio, which do not deteriorate with more levels of U .

17.5 Recommended reading

Ding and VanderWeele (2016) extended and unified the Cornfield-type sensitivity analysis, which is the theoretical basis for the notion of E-value.

18

Sensitivity Analysis for the Average Causal Effect with Unmeasured Confounding

Cornfield-type sensitivity analysis works the best for binary outcomes on the risk ratio scale, conditioning on the observed covariates. Although Ding and VanderWeele (2016) also proposed Cornfield-type sensitivity analysis methods for the average causal effect, they are not general enough and are not convenient to apply. Below I give a more direct approach to sensitivity analysis based on the conditional expectations of the potential outcomes. The idea appeared in early work of Robins (1999) and Scharfstein et al. (1999). This chapter is based on Lu and Ding (2023)'s recent formulation.

The approach is closely related to the idea of deriving worse-case bounds on the average potential outcomes. I will first review the simpler idea of bounds, and then discuss the approach to sensitivity analysis.

18.1 Introduction

Recall the canonical setup of an observational study with $\{Z_i, X_i, Y_i(1), Y_i(0)\}_{i=1}^n \stackrel{\text{IID}}{\sim} \{Z, X, Y(1), Y(0)\}$ and focus on the average causal effect

$$\tau = E\{Y(1) - Y(0)\}.$$

It decomposes to

$$\begin{aligned}\tau &= [E(Y | Z = 1)\text{pr}(Z = 1) + E\{Y(1) | Z = 0\}\text{pr}(Z = 0)] \\ &\quad - [E\{Y(0) | Z = 1\}\text{pr}(Z = 1) + E(Y | Z = 0)\text{pr}(Z = 0)].\end{aligned}$$

So the fundamental difficulty is to estimate the counterfactual means

$$E\{Y(1) | Z = 0\}, \quad E\{Y(0) | Z = 1\}.$$

There are in general two extreme strategies to estimate them.

We have discussed the first strategy in Part III, which relies on ignorability. Assuming

$$\begin{aligned}E\{Y(1) | Z = 1, X\} &= E\{Y(1) | Z = 0, X\}, \\ E\{Y(0) | Z = 1, X\} &= E\{Y(0) | Z = 0, X\},\end{aligned}$$

TABLE 18.1: Science Table with bounded outcome $[\ell, u]$, where ℓ and u are two constants

Z	$Y(1)$	$Y(0)$	Lower $Y(1)$	Upper $Y(1)$	Lower $Y(0)$	Upper $Y(0)$
1	$Y_1(1)$?	$Y_1(1)$	$Y_1(1)$	ℓ	u
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
1	$Y_{n_1}(1)$?	$Y_{n_1}(1)$	$Y_{n_1}(1)$	ℓ	u
0	?	$Y_{n_1+1}(0)$	ℓ	u	$Y_{n_1+1}(0)$	$Y_{n_1+1}(0)$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
0	?	$Y_n(0)$	ℓ	u	$Y_n(0)$	$Y_n(0)$

we can identify the counterfactual means by the observables:

$$E\{Y(1) | Z = 0\} = E\{E(Y | Z = 1, X) | Z = 0\}$$

and, similarly,

$$E\{Y(0) | Z = 1\} = E\{E(Y | Z = 0, X) | Z = 1\}.$$

The second strategy in the next section assumes nothing except that the outcomes are bounded between ℓ and u . This is natural for binary outcomes with $\ell = 0$ and $u = 1$. With this assumption, the two counterfactual means are also bounded between ℓ and u , which implies the worse-case bounds on τ . I will review this strategy below.

18.2 Manski-type worse-case bounds on the average causal effect without assumptions

Assume that the outcome is bounded between ℓ and u . From the decomposition

$$E\{Y(1)\} = E\{Y(1) | Z = 1\}\text{pr}(Z = 1) + E\{Y(1) | Z = 0\}\text{pr}(Z = 0),$$

we can derive that $E\{Y(1)\}$ has lower bound

$$E\{Y | Z = 1\}\text{pr}(Z = 1) + \ell\text{pr}(Z = 0)$$

and upper bound

$$E\{Y | Z = 1\}\text{pr}(Z = 1) + u\text{pr}(Z = 0).$$

Similarly, from the decomposition

$$E\{Y(0)\} = E\{Y(0) | Z = 1\}\text{pr}(Z = 1) + E\{Y(0) | Z = 0\}\text{pr}(Z = 0),$$

we can derive that $E\{Y(0)\}$ has lower bound

$$\ell \text{pr}(Z = 1) + E\{Y | Z = 0\} \text{pr}(Z = 0)$$

and upper bound

$$\text{upr}(Z = 1) + E\{Y | Z = 0\} \text{pr}(Z = 0).$$

Combining these bounds, we can derive that the average causal effect $\tau = E\{Y(1)\} - E\{Y(0)\}$ has lower bound

$$E\{Y | Z = 1\} \text{pr}(Z = 1) + \ell \text{pr}(Z = 0) - \text{upr}(Z = 1) - E\{Y | Z = 0\} \text{pr}(Z = 0)$$

and upper bound

$$E\{Y | Z = 1\} \text{pr}(Z = 1) + \text{upr}(Z = 0) - \ell \text{pr}(Z = 1) - E\{Y | Z = 0\} \text{pr}(Z = 0).$$

The length of the bounds is $u - \ell$, which is not informative but is better than the a priori bounds $[\ell - u, u - \ell]$ with length $2(u - \ell)$. Without further assumptions, the observed data distribution does not uniquely determine τ . In this case, we say that τ is *partially identified*, with the formal definition below.

Definition 18.1 (partial identification) *A parameter θ is partially identified if the observed data distribution is compatible with multiple values of θ .*

Compare Definitions 10.1 and 18.1. If the parameter θ is uniquely determined by the observed data distribution, then it is identifiable; otherwise, it is partially identifiable. Therefore, τ is identifiable with the ignorability assumption, but only partially identifiable without the ignorability assumption.

Cochran (1953) used the idea of worse-case bounds in surveys with missing data, but abandoned the idea because it often gives very conservative results. Similarly, the worst-case bounds above are often uninteresting from a practical perspective because they often cover 0. Moreover, this strategy is not applicable to the settings with unbounded outcomes.

Manski applied the idea to causal inference (Manski, 1990) and many other econometric models (Manski, 2003). This idea of bounding causal parameters with minimal assumptions is powerful when coupled with other qualitative assumptions. Manski (2003) surveyed many strategies. For instance, we may believe that the treatment does not harm any units, so the monotonicity assumption holds: $Y(1) \geq Y(0)$. Then the lower bound on τ is zero but the upper bound is unchanged. Another type of assumption is $Z = I\{Y(1) \geq Y(0)\}$, that is, the treatment selection is based on the difference between the latent potential outcomes. This assumption can also improve the bounds on τ .

18.3 Sensitivity analysis for the average causal effect

The first strategy is optimistic which assumes that the potential outcomes do not differ across treatment and control groups, conditioning on the observed covariates. The second strategy is pessimistic which does not infer the counterfactual means based on the observed data at all. The following strategy is in-between.

18.3.1 Identification formulas

Define

$$\frac{E\{Y(1) | Z = 1, X\}}{E\{Y(1) | Z = 0, X\}} = \varepsilon_1(X),$$

$$\frac{E\{Y(0) | Z = 1, X\}}{E\{Y(0) | Z = 0, X\}} = \varepsilon_0(X),$$

which are the sensitivity parameters. For simplicity, we can further assume that they are constant independent of X . In practice, we need to fix them or vary them in a pre-specified range. Recall that $\mu_1(X) = E(Y | Z = 1, X)$ and $\mu_0(X) = E(Y | Z = 0, X)$. We can identify the two counterfactual means and the average causal effect as follows.

Theorem 18.1 *With known $\varepsilon_1(X)$ and $\varepsilon_0(X)$, we have*

$$E\{Y(1) | Z = 0\} = E\{\mu_1(X)/\varepsilon_1(X) | Z = 0\},$$

$$E\{Y(0) | Z = 1\} = E\{\mu_0(X)\varepsilon_0(X) | Z = 1\}$$

and therefore

$$\begin{aligned} \tau &= E\{ZY + (1 - Z)\mu_1(X)/\varepsilon_1(X)\} \\ &\quad - E\{Z\mu_0(X)\varepsilon_0(X) + (1 - Z)Y\} \end{aligned} \tag{18.1}$$

$$\begin{aligned} &= E\{Z\mu_1(X) + (1 - Z)\mu_1(X)/\varepsilon_1(X)\} \\ &\quad - E\{Z\mu_0(X)\varepsilon_0(X) + (1 - Z)\mu_0(X)\}. \end{aligned} \tag{18.2}$$

I leave the proof of Theorem 18.1 to Problem 18.1. With the fitted outcome model, (18.1) and (18.2) motivate the following predictive and projective estimators for τ :

$$\begin{aligned} \hat{\tau}^{\text{pred}} &= \left\{ n^{-1} \sum_{i=1}^n Z_i Y_i + n^{-1} \sum_{i=1}^n (1 - Z_i) \hat{\mu}_1(X_i) / \varepsilon_1(X_i) \right\} \\ &\quad - \left\{ n^{-1} \sum_{i=1}^n Z_i \hat{\mu}_0(X_i) \varepsilon_0(X_i) + n^{-1} \sum_{i=1}^n (1 - Z_i) Y_i \right\}, \end{aligned}$$

and

$$\begin{aligned}\hat{\tau}^{\text{proj}} &= \left\{ n^{-1} \sum_{i=1}^n Z_i \hat{\mu}_1(X_i) + n^{-1} \sum_{i=1}^n (1 - Z_i) \hat{\mu}_1(X_i) / \varepsilon_1(X_i) \right\} \\ &\quad - \left\{ n^{-1} \sum_{i=1}^n Z_i \hat{\mu}_0(X_i) \varepsilon_0(X_i) + n^{-1} \sum_{i=1}^n (1 - Z_i) \hat{\mu}_0(X_i) \right\}.\end{aligned}$$

The terminologies “predictive” and “projective” are from the survey sampling literature (Firth and Bennett, 1998; Ding and Li, 2018). The estimators $\hat{\tau}^{\text{pred}}$ and $\hat{\tau}^{\text{proj}}$ differ slightly: the former uses the observed outcomes when available; in contrast, the latter replaces the observed outcomes with the fitted values.

More interesting, we can also identify τ by an inverse probability weighting formula.

Theorem 18.2 *With known $\varepsilon_1(X)$ and $\varepsilon_0(X)$, we have*

$$E\{Y(1)\} = E\left\{w_1(X) \frac{Z}{e(X)} Y\right\}, \quad E\{Y(0)\} = E\left\{w_0(X) \frac{1-Z}{1-e(X)} Y\right\},$$

where

$$w_1(X) = e(X) + \{1 - e(X)\}/\varepsilon_1(X), \quad w_0(X) = e(X)\varepsilon_0(X) + 1 - e(X).$$

I leave the proof of Theorem 18.2 to Problem 18.2. Theorem 18.2 modifies the classic inverse probability weighting formulas with two extra factors $w_1(X)$ and $w_0(X)$ depending on both the propensity score and the sensitivity parameters. With the fitted propensity score model, Theorem 18.2 motivates the following estimators for τ :

$$\begin{aligned}\hat{\tau}^{\text{ht}} &= n^{-1} \sum_{i=1}^n \frac{\{\hat{e}(X_i)\varepsilon_1(X_i) + 1 - \hat{e}(X_i)\}Z_i Y_i}{\varepsilon_1(X_i)\hat{e}(X_i)} \\ &\quad - n^{-1} \sum_{i=1}^n \frac{\{\hat{e}(X_i)\varepsilon_0(X_i) + 1 - \hat{e}(X_i)\}(1 - Z_i) Y_i}{1 - \hat{e}(X_i)}\end{aligned}$$

and

$$\begin{aligned}\hat{\tau}^{\text{haj}} &= \sum_{i=1}^n \frac{\{\hat{e}(X_i)\varepsilon_1(X_i) + 1 - \hat{e}(X_i)\}Z_i Y_i}{\varepsilon_1(X_i)\hat{e}(X_i)} / \sum_{i=1}^n \frac{Z_i}{\hat{e}(X_i)} \\ &\quad - n^{-1} \sum_{i=1}^n \frac{\{\hat{e}(X_i)\varepsilon_0(X_i) + 1 - \hat{e}(X_i)\}(1 - Z_i) Y_i}{1 - \hat{e}(X_i)} / \sum_{i=1}^n \frac{1 - Z_i}{1 - \hat{e}(X_i)}.\end{aligned}$$

More interestingly, with fitted propensity score and outcome models, the following estimator for τ is doubly robust:

$$\hat{\tau}^{\text{ht}} = \hat{\tau}^{\text{ipw}} - n^{-1} \sum_{i=1}^n \{Z_i - \hat{e}(X_i)\} \left\{ \frac{\hat{\mu}_1(X_i)}{\hat{e}(X_i)\varepsilon_1(X_i)} + \frac{\hat{\mu}_0(X_i)\varepsilon_0(X_i)}{1 - \hat{e}(X_i)} \right\}.$$

That is, with known $\varepsilon_1(X_i)$ and $\varepsilon_0(X_i)$, the estimator $\hat{\tau}^{\text{dr}}$ is consistent for τ if either the propensity score model or the outcome model is correctly specified. We can use the bootstrap to approximate the variance of the above estimators. See Lu and Ding (2023) for technical details.

When $\varepsilon_1(X_i) = \varepsilon_0(X_i) = 1$, the above estimators reduce to the predictive estimator, inverse probability weighting estimator, and the doubly robust estimators introduced in Part III.

18.4 Example

We revisit Example 10.3. With

$$\varepsilon_1(X) = \varepsilon_0(X) \in \{1/2, 1/1.7, 1/1.5, 1/1.3, 1, 1.3, 1.5, 1.7, 2\},$$

we obtain an array of doubly robust estimates of τ .

	1/2	1/1.7	1/1.5	1/1.3	1	1.3	1.5	1.7	
2	11.62	10.44	9.40	8.03	4.96	0.97	-1.69	-4.35	-8.34
1/2	9.22	8.05	7.00	5.64	2.57	-1.42	-4.08	-6.75	-10.74
1/1.5	7.63	6.45	5.41	4.05	0.97	-3.02	-5.68	-8.34	-12.33
1/1.3	6.03	4.86	3.81	2.45	-0.62	-4.61	-7.27	-9.94	-13.93
1	3.64	2.47	1.42	0.06	-3.01	-7.01	-9.67	-12.33	-16.32
1.3	1.80	0.63	-0.42	-1.78	-4.85	-8.85	-11.51	-14.17	-18.16
1.5	0.98	-0.19	-1.24	-2.60	-5.67	-9.66	-12.33	-14.99	-18.98
1.7	0.36	-0.82	-1.86	-3.23	-6.30	-10.29	-12.95	-15.61	-19.60
2	-0.35	-1.52	-2.57	-3.93	-7.00	-10.99	-13.65	-16.32	-20.31

The signs of the estimates are not sensitive to sensitivity parameters larger than 1, but they are quite sensitive to sensitivity parameters smaller than 1. When the participants of the meal plan tend to have higher BMI, the average causal effect of the meal plan on BMI is negative. However, this conclusion can be quite sensitive if the participants of the meal plan tend to have lower BMI.

18.5 Homework Problems

18.1 Proof of Theorem 18.1

Prove Theorem 18.1.

18.2 Proof of Theorem 18.2

Prove Theorem 18.2.

18.3 Sensitivity analysis for the average causal effect on the treated units τ_T

This problem extends Chapter 13 to allow for unmeasured confounding for estimating

$$\tau_T = E\{Y(1) - Y(0) \mid Z = 1\} = E(Y \mid Z = 1) - E\{Y(0) \mid Z = 1\}.$$

We can easily estimate $E(Y \mid Z = 1)$ by the sample moment. The only counterfactual term is $E\{Y(0) \mid Z = 1\}$. Therefore, we only need the sensitivity parameter $\varepsilon_0(X)$. We have the following two identification formulas with a known $\varepsilon_0(X)$.

Theorem 18.3 *With known $\varepsilon_0(X)$, we have*

$$\begin{aligned} E\{Y(0) \mid Z = 1\} &= E\{Z\mu_0(X)\varepsilon_0(X)\}/e \\ &= E\left\{e(X)\varepsilon_0(X)\frac{1-Z}{1-e(X)}Y\right\}/e, \end{aligned}$$

where $e = \text{pr}(Z = 1)$

Prove Theorem 18.3.

Remark: Theorem 18.3 motivates using $\hat{\tau}_T^* = \hat{\mu}_{T1} - \hat{\mu}_{T0}^*$ to estimate τ_T , where $\hat{\mu}_{T1} = \sum_{i=1}^n Z_i Y_i / \sum_{i=1}^n Z_i$ and

$$\begin{aligned} \hat{\mu}_{T0}^{\text{reg}} &= n_1^{-1} \sum_{i=1}^n Z_i \varepsilon_0(X_i) \hat{\mu}_0(X_i), \\ \hat{\mu}_{T0}^{\text{ht}} &= n_1^{-1} \sum_{i=1}^n \varepsilon_0(X_i) \hat{o}(X_i) (1 - Z_i) Y_i, \\ \hat{\mu}_{T0}^{\text{haj}} &= \sum_{i=1}^n \varepsilon_0(X_i) \hat{o}(X_i) (1 - Z_i) Y_i / \sum_{i=1}^n \hat{o}(X_i) (1 - Z_i), \end{aligned}$$

with $\hat{o}(X_i) = \hat{e}(X_i) / \{1 - \hat{e}(X_i)\}$ being the estimated conditional odds of the treatment. Moreover, we can construct the doubly robust estimator $\hat{\tau}_T^{\text{dr}} = \hat{\mu}_{T1} - \hat{\mu}_{T0}^{\text{dr}}$ for τ_T , where

$$\hat{\mu}_{T0}^{\text{dr}} = \hat{\mu}_{T0}^{\text{ht}} - n_1^{-1} \sum_{i=1}^n \varepsilon_0(X_i) \frac{\hat{e}(X_i) - Z}{1 - \hat{e}(X_i)} \hat{\mu}_0(X_i).$$

Lu and Ding (2023) provide more details and also propose a doubly robust estimator for τ_T .

18.4 R code

Implement the estimators in Problem 18.3.

18.5 Recommended reading

Rosenbaum and Rubin (1983a) and Imbens (2003) are two classic papers on sensitivity analysis which, however, involve more complicated procedures.

19

Rosenbaum-Style p -Values for Matched Observational Studies with Unmeasured Confounding

Rosenbaum (1987b) introduced a sensitivity analysis technique for matched observational studies. Although it works for general settings (Rosenbaum, 2002b), the theory is most elegant for one-to-one matching. Different from Chapters 17 and 18, Rosenbaum-type sensitivity analysis works the best for matched observational studies for testing the sharp null hypothesis of no individual treatment effect.

19.1 The model for sensitivity analysis with matched data

Consider exactly matched pairs from an observational study, with (i, j) indexing unit j in pair i ($i = 1, \dots, n; j = 1, 2$). Assume iid sampling, and define the propensity score as

$$e_{ij} = \text{pr}\{Z_{ij} = 1 \mid X_i, Y_{ij}(1), Y_{ij}(0)\}.$$

Let $\mathbb{S}_i = \{Y_{i1}(1), Y_{i1}(0), Y_{i2}(1), Y_{i2}(0)\}$ denote the set of all potential outcomes within pair i . Conditioning on the event that $Z_{i1} + Z_{i2} = 1$, we have

$$\begin{aligned} \pi_{i1} &= \text{pr}\{Z_{i1} = 1 \mid X_i, \mathbb{S}_i, Z_{i1} + Z_{i2} = 1\} \\ &= \frac{\text{pr}\{Z_{i1} = 1, Z_{i2} = 0 \mid X_i, \mathbb{S}_i\}}{\text{pr}\{Z_{i1} + Z_{i2} = 1 \mid X_i, \mathbb{S}_i\}} \\ &= \frac{\text{pr}\{Z_{i1} = 1, Z_{i2} = 0 \mid X_i, \mathbb{S}_i\}}{\text{pr}\{Z_{i1} = 1, Z_{i2} = 0 \mid X_i, \mathbb{S}_i\} + \text{pr}\{Z_{i1} = 0, Z_{i2} = 1 \mid X_i, \mathbb{S}_i\}} \\ &= \frac{e_{i1}(1 - e_{i2})}{e_{i1}(1 - e_{i2}) + (1 - e_{i1})e_{i2}} \end{aligned}$$

Define $o_{ij} = e_{ij}/(1 - e_{ij})$ as the odds of the treatment for unit (i, j) , and we have

$$\pi_{i1} = \frac{o_{i1}}{o_{i1} + o_{i2}}.$$

Under ignorability, e_{ij} is only a function of X_i , and therefore, $e_{i1} = e_{i2}$ and $\pi_{i1} = 1/2$. Thus the treatment assignment mechanism conditioning on covariates and potential outcomes is equivalent to that from an MPE with equal treatment and control probabilities. This is a strategy to analyze matched observational studies we discussed in Chapter 15.1.

In general, e_{ij} is also a function of the unobserved potential outcomes, and it can range from 0 to 1. Rosenbaum (1987b)'s model for sensitivity analysis imposes bounds on the odds ratio o_{i1}/o_{i2} .

Assumption 19.1 (Rosenbaum's sensitivity model) *The odds ratios are bounded by*

$$o_{i1}/o_{i2} \leq \Gamma, \quad o_{i2}/o_{i1} \leq \Gamma,$$

for some pre-specified $\Gamma \geq 1$. Equivalently,

$$\frac{1}{1 + \Gamma} \leq \pi_{i1} \leq \frac{\Gamma}{1 + \Gamma}$$

for some pre-specified $\Gamma \geq 1$.

Under Assumption 19.1, we have a biased MPE with unequal and varying treatment and control probabilities across pairs. When $\Gamma = 1$, we have π_{i1} and thus a standard MPE. Therefore, $\Gamma > 1$ measures the deviation from the ideal MPE due to the omitted variables in matching.

19.2 Worst-case p -values under Rosenbaum's sensitivity model

Consider testing the sharp null hypothesis

$$H_{0F} : Y_{ij}(1) = Y_{ij}(0) \text{ for } i = 1, \dots, n \text{ and } j = 1, 2$$

based on the within-pair differences $\hat{\tau}_i = (2Z_{i1} - 1)(Y_{i1} - Y_{i2})$ ($i = 1, \dots, n$). Under H_{0F} , $|\hat{\tau}_i|$ is fixed but $S_i = I(\hat{\tau}_i > 0)$ is random if $\hat{\tau}_i \neq 0$. Consider the following class of test statistics:

$$T = \sum_{i=1}^n S_i q_i,$$

where $q_i \geq 0$ is a function of $(|\hat{\tau}_1|, \dots, |\hat{\tau}_n|)$. Special cases include the sign statistic, the pair t statistic (up to some constant shift), and the Wilcoxon sign rank statistic:

$$T = \sum_{i=1}^n S_i, \quad T = \sum_{i=1}^n S_i |\hat{\tau}_i|, \quad T = \sum_{i=1}^n S_i R_i,$$

where (R_1, \dots, R_n) are the ranks of $(|\hat{\tau}_1|, \dots, |\hat{\tau}_n|)$.

What is the null distribution of the test statistic with general Γ ? It can be quite complicated because we do not fully specify the exact values of the π_{i1} 's. Fortunately, we know that the worse case distribution correspond to

$$S_i \stackrel{\text{IID}}{\sim} \text{Bernoulli} \left(\frac{\Gamma}{1 + \Gamma} \right).$$

Here, the FRT with T has the largest p -value under the “the worst case” distribution. The corresponding distribution has mean

$$E_\Gamma(T) = \frac{\Gamma}{1 + \Gamma} \sum_{i=1}^n q_i,$$

and variance

$$\text{var}_\Gamma(T) = \frac{\Gamma}{(1 + \Gamma)^2} \sum_{i=1}^n q_i^2,$$

with a Normal approximation

$$\frac{T - \frac{\Gamma}{1+\Gamma} \sum_{i=1}^n q_i}{\sqrt{\frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^n q_i^2}} \xrightarrow{\text{d}} N(0, 1).$$

In practice, we can report a sequence of p -values as a function of Γ .

19.3 Revisiting the LaLonde data

We conduct Rosenbaum-style sensitivity analysis in the matched LaLonde data. We consider using the test statistic $T = \sum_{i=1}^n S_i |\hat{\tau}_i|$. Under the ideal matched pair experiment with $\Gamma = 1$, we can simulate the distribution of T and obtain the p -value 0.002, as shown in the first subfigure in Figure 19.1. With a slightly larger $\Gamma = 1.1$, the distribution of T shifts to the right and the p -value increases to 0.011. If we further increase Γ to 1.3, then the distribution of T shifts further and the p -value exceeds 0.05. Figure 19.2 shows the histogram of the $\hat{\tau}_i$'s and the p -value as a function of Γ ; $\Gamma = 1.233$ measures the maximum confounding that we can still reject the null hypothesis at level 0.05.

We can also use the `senmw` function in the `sensitivitymw` package to obtain a sequence of p -values against Γ , as shown in Figure 19.2.

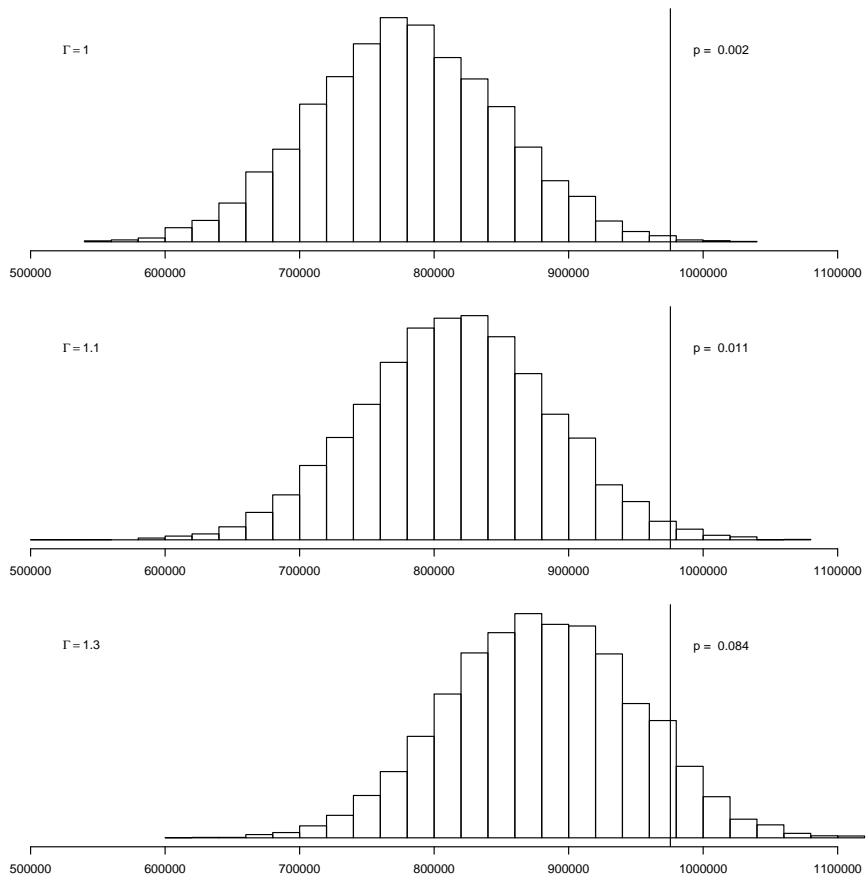
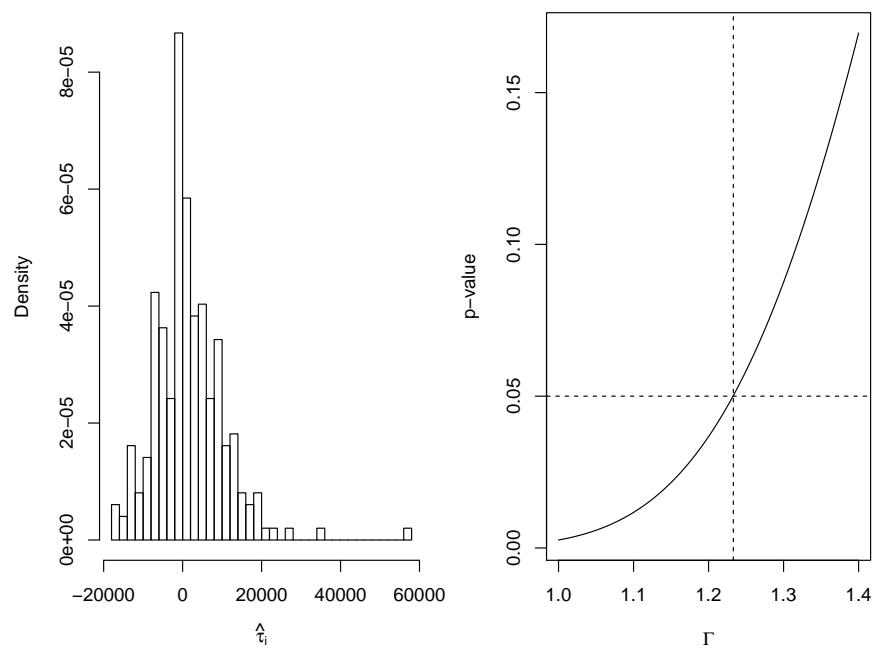


FIGURE 19.1: Distributions of $T = \sum_{i=1}^n S_i |\hat{\tau}_i|$ with S_i iid Bernoulli($\Gamma/(1 + \Gamma)$), based on the LaLonde data.

FIGURE 19.2: p -value as a function of Γ , based on the LaLonde data.

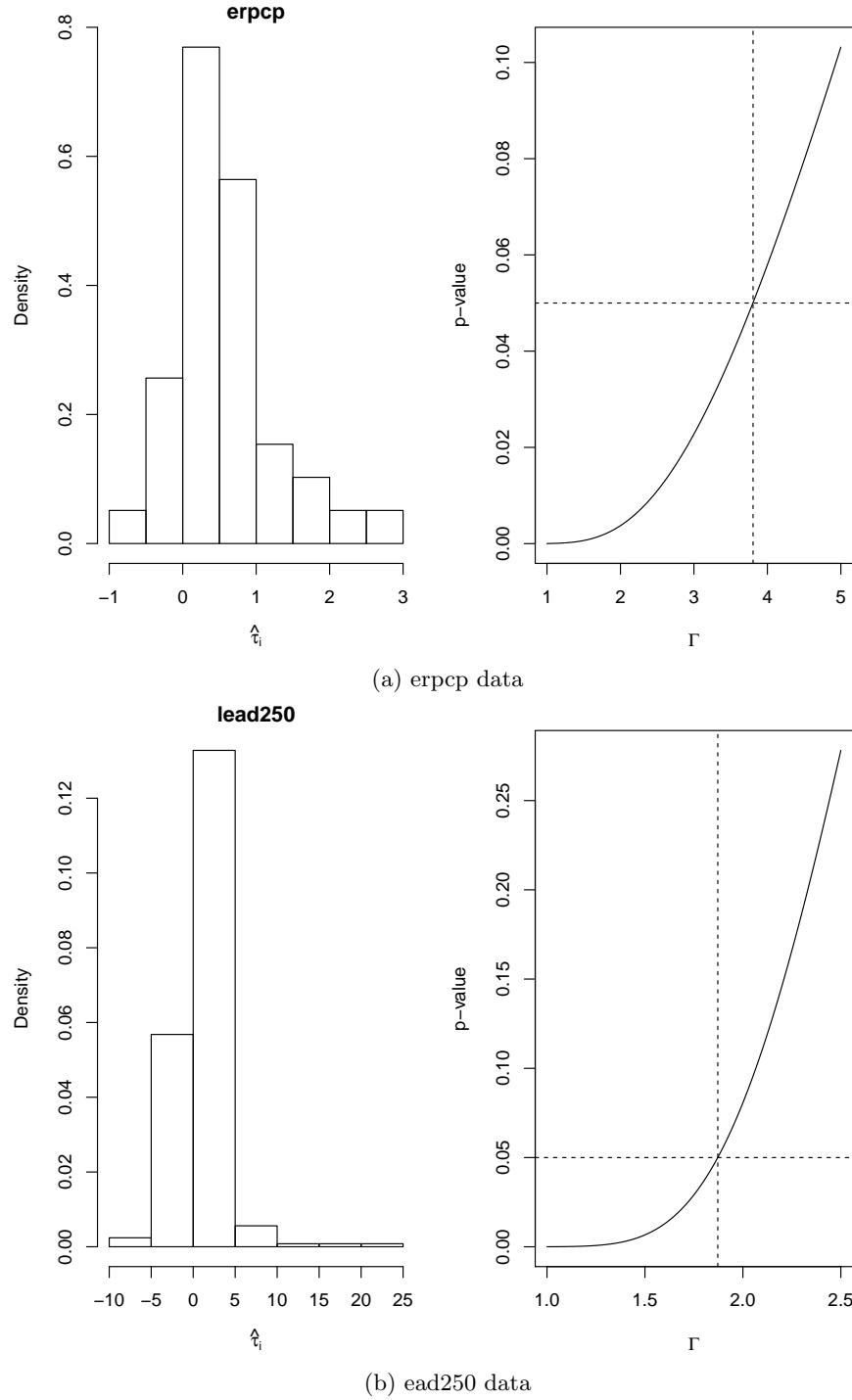


FIGURE 19.3: Two examples

19.4 Homework Problems*19.1 Application of Rosenbaum's approach*

Re-analyze Example 10.3 using Rosenbaum's approach.

19.2 Recommended reading

Rosenbaum (2015) provides a tutorial for his two R packages for sensitivity analysis with matched observational studies.



20

Overlap in Observational Studies: Difficulties and Opportunities

20.1 Implications of overlap

In Part III of this book, causal inference with observational studies relies on two critical assumptions: unconfoundedness

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$$

and overlap

$$0 < e(X) < 1.$$

D'Amour et al. (2021) pointed out the tension between these two assumptions: typically, more covariates make the unconfoundedness assumption more plausible (ignoring M-bias discussed in Chapter 16.3.1), but more covariates make the overlap assumption less plausible because the treatment becomes more predictable.

If some units has $e(X) = 0$ or $e(X) = 1$, then we have philosophic difficulty of thinking about the counterfactual potential outcomes (King and Zeng, 2006). In particular, if a unit deterministically receives the treatment, then it may not be meaningful to conceive its potential outcome under control; vice versa. Even if the true propensity score is not exactly 0 or 1, the estimated propensity score can be very close to 0 or 1 in finite sample, which makes the estimators based on inverse probability weighting numerically unstable. Many statistical analyses in fact require a strict version of overlap:

Assumption 20.1 (strict overlap) $\eta \leq e(X) \leq 1 - \eta$ for some $\eta \in (0, 1/2)$.

However, D'Amour et al. (2021, Corollary 1) showed that Assumption 20.1 has very strong implications. For simplicity, I present only one of their results. Let X_k ($k = 1, \dots, p$) be the k th component of the covariate $X = (X_1, \dots, X_p)$, and $e = \text{pr}(Z = 1)$ be the proportion of the treated units.

Theorem 20.1 *Assumption 20.1 implies that $\eta \leq e \leq 1 - \eta$ and*

$$\begin{aligned} & p^{-1} \sum_{k=1}^p |E(X_k \mid Z = 1) - E(X_k \mid Z = 0)| \\ & \leq p^{-1/2} C^{1/2} \left\{ e \lambda_1^{1/2} + (1 - e) \lambda_0^{1/2} \right\}, \end{aligned} \tag{20.1}$$

where

$$C = \frac{(e - \eta)(1 - e - \eta)}{e^2(1 - e)^2\eta(1 - \eta)}$$

is a positive constant depending only on (e, η) , and λ_1 and λ_0 are the maximum eigenvalues of the covariance matrices $\text{cov}(X | Z = 1)$ and $\text{cov}(X | Z = 0)$, respectively.

What is the order of the maximum eigenvalues in Theorem 20.1? D'Amour et al. (2021) showed that it is usually smaller than $O(p)$ unless the components of X are highly correlated. If the components of X are highly correlated, then some components are redundant after including other components. If the components of X are not highly correlated, then the right-hand side converges to zero. So the average difference in means of the covariates is close to zero, that is, the treatment and control groups are nearly balanced in means averaged over all dimensions of the covariates. Mathematically, the left-hand side of (20.1) converging to zero rules out the possibility that all dimensions of X have non-vanishing difference in means across treatment and control groups. It is a strong requirement in observational studies with many covariates.

20.1.1 Trimming in the presence of limited overlap

When Assumption 20.1 does not hold, it is common to trim the units based on the estimated propensity scores (Crump et al., 2009; Yang and Ding, 2018). Trimming drops units within regions of little overlap, which changes the population and estimand. The restrictive implications of overlap in Assumption 20.1 suggest that trimming must be employed more often and one may need to trim a large proportion of units to achieve desirable overlap in high dimensions.

20.1.2 Outcome modeling in the presence of limited overlap

The somewhat negative results in D'Amour et al. (2021) also highlight the limitation of focusing only on the propensity score in the presence of limited overlap. With high dimensional covariates, outcome modeling becomes more important. In particular, if the outcome means only depend on a function of the original covariates in that

$$E\{Y(z) | X\} = f_z(r(X)), \quad (z = 0, 1)$$

then it suffices to control for $r(X)$, a lower dimensional summary of the original covariates. Due to the dimension reduction, the strict overlap condition on $r(X)$ can be much weaker than the strict overlap condition on X . This is conceptually straightforward, but the corresponding theory and methods are missing.

20.2 Causal inference with no overlap: regression discontinuity

Starting from the simple case with a univariate X . An extreme treatment assignment mechanism is a deterministic one:

$$Z = 1(X \geq x_0),$$

where x_0 is a predetermined threshold. An interesting consequence of this assignment is that the unconfoundedness assumption holds automatically:

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$$

because Z is a deterministic function of X . However, the overlap assumption is violated by definition:

$$e(X) = \text{pr}(Z = 1 \mid X) = 1(X \geq x_0) = \begin{cases} 1 & \text{if } X \geq x_0, \\ 0 & \text{if } X < x_0. \end{cases}$$

So our analytic strategies discussed in Part IV are no longer applicable here. We must change our perspective.

The discussion here seems contrived, with a deterministic treatment assignment. Interestingly, it has many applications in practice, and is called *regression discontinuity*. Below, I first review some canonical examples and then give a mathematical formulation of this type of studies.

20.2.1 Examples and graphical diagnostics

Example 20.1 *Thistlethwaite and Campbell (1960) first proposed the idea of regression-discontinuity analysis. Their motivating example was to study the effect of students' winning Certificated of Merit on later career plans, where the Certificated of Merit was determined by whether the Scholarship Qualifying Test score was above a certain threshold. Their initial analysis was mainly graphical. Figure 20.1 shows one of their graphs.*

Example 20.2 *Bor et al. (2014) used regression discontinuity to study the effect of when to start HIV patients with antiretroviral on their mortality, where the treatment is determined by whether the patients' CD4 counts were below 200 cells/ μL .¹*

Example 20.3 *Carpenter and Dobkin (2009) studied the effect of alcohol consumption on mortality, which leverages the minimum legal drinking age as a discontinuity for alcohol consumption. They derived mortality*

¹CD4 cells are white blood cells that fight infection.

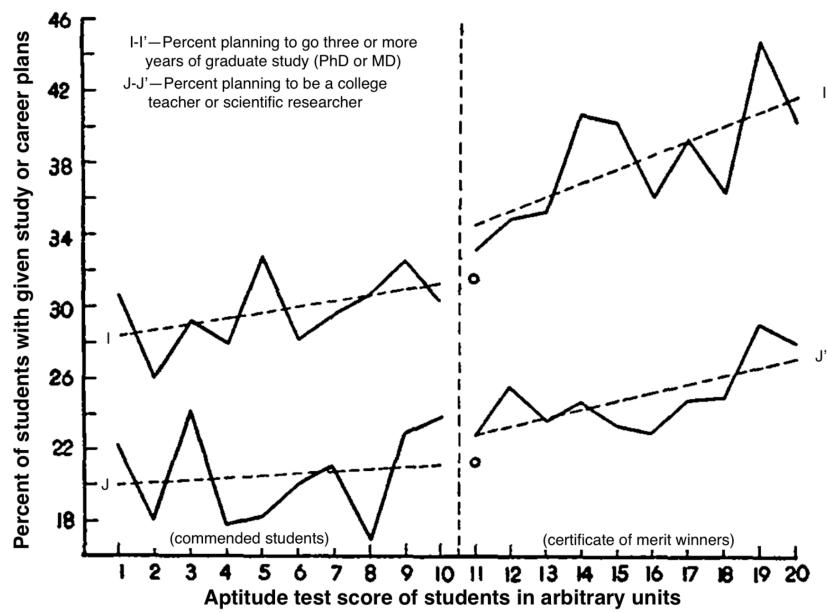


FIGURE 20.1: A graph from Thistlethwaite and Campbell (1960) with minor modifications of the unclear text in the original paper

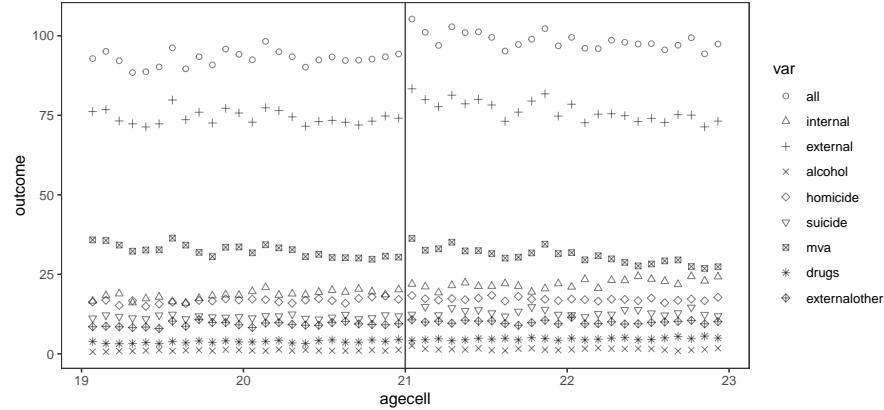


FIGURE 20.2: Minimum legal drinking age example

data from the National Center for Health Statistics, including the decedent's date of birth and date of death. They computed age profile of deaths per 100,000 person years with outcomes measured by the following nine variables:

<i>all</i>	all deaths, the sum of <i>internal</i> and <i>external</i>
<i>internal</i>	deaths due to internal causes
<i>external</i>	deaths due to external causes, the sum of the rest
<i>homicide</i>	homicides
<i>suicide</i>	suicides
<i>mva</i>	motor vehicle accidents
<i>alcohol</i>	deaths with a mention of alcohol
<i>drugs</i>	deaths with a mention of drug use
<i>externalother</i>	deaths due to other external causes

Figure 20.2 plots the number of deaths per 100,000 person years for nine measures based on the data used by Angrist and Pischke (2014). From the jumps at age 21, it seems obvious that there is an increase of mortality at age 21, primarily due to motor vehicle accidents. I leave the formal analysis as Problem 20.3.

20.2.2 A mathematical formulation of regression discontinuity

The technical term for the variable X that determines the treatment is the *running variable*. Intuitively, regression discontinuity can identify a *local average causal effect* at the cutoff point x_0 :

$$\tau(x_0) = E\{Y(1) - Y(0) \mid X = x_0\}.$$

In particular, for the potential outcome under treatment, we have

$$E\{Y(1) | X = x_0\} = \lim_{\varepsilon \rightarrow 0+} E\{Y(1) | X = x_0 + \varepsilon\} \quad (20.2)$$

$$= \lim_{\varepsilon \rightarrow 0+} E\{Y(1) | Z = 1, X = x_0 + \varepsilon\} \quad (20.3)$$

$$= \lim_{\varepsilon \rightarrow 0+} E(Y | Z = 1, X = x_0 + \varepsilon), \quad (20.4)$$

where (20.2) holds if $E\{Y(1) | X = x\}$ is continuous from the right at x_0 and (20.3) follows by the definition of Z . Similarly, for the potential outcome under control, we have

$$E\{Y(0) | X = x_0\} = \lim_{\varepsilon \rightarrow 0+} E(Y | Z = 0, X = x_0 - \varepsilon)$$

if $E\{Y(0) | X = x\}$ is continuous from the left at x_0 . So the local average causal effect at x_0 can be identified by the difference of the two limits. I summarize the key identification result below.

Theorem 20.2 *Assume that $E\{Y(1) | X = x\}$ is continuous from the right at x_0 and $E\{Y(0) | X = x\}$ is continuous from the left at x_0 . Then the local average treatment effect at $X = x_0$ is identified by*

$$\tau(x_0) = \lim_{\varepsilon \rightarrow 0+} E(Y | Z = 1, X = x_0 + \varepsilon) - \lim_{\varepsilon \rightarrow 0+} E(Y | Z = 0, X = x_0 - \varepsilon).$$

Since the right-hand side of the above equation only involves observables, the parameter $\tau(x_0)$ is nonparametrically identified. However, the form of the identification formula is totally different from what we derived before. In particular, the identification formula involves limits of two conditional expectation functions.

20.2.3 Regressions near the boundary

If we are lucky, graphical diagnostic sometimes can clearly show the causal effect at the cutoff point. However, many outcomes are noisy so graphical diagnostic is not enough. Figure 20.3 shows two examples with obvious jumps at the cutoff point and two examples without obvious jumps, although the underlying data generating processes all have discontinuities.

Assume that $E(Y | Z = 1, X = x) = \gamma_1 + \beta_1 x$ and $E(Y | Z = 0, X = x) = \gamma_0 + \beta_0 x$ are linear in x . We can run OLS based on the treated and control data to obtain the fitted lines $\hat{\gamma}_1 + \hat{\beta}_1 x$ and $\hat{\gamma}_0 + \hat{\beta}_0 x$, respectively. We can then estimate the average causal effect at the point $X = x_0$ as

$$\hat{\tau}(x_0) = (\hat{\gamma}_1 - \hat{\gamma}_0) + (\hat{\beta}_1 - \hat{\beta}_0)x_0.$$

Numerically, $\hat{\tau}(x_0)$ is identical to the coefficient of Z_i in the OLS

$$Y_i \sim \{1, Z_i, X_i - x_0, Z_i(X_i - x_0)\}, \quad (20.5)$$

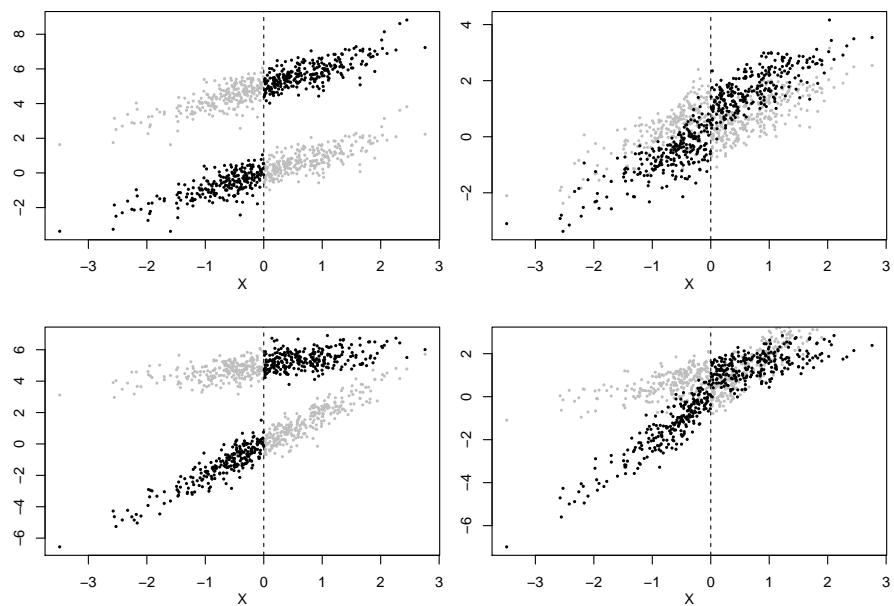


FIGURE 20.3: Examples of regression discontinuity. In the first column, the data generating processes result in visible jumps at the cutoff points; in the second column, the jumps are not so visible. In the first row, the data generating processes have constant $\tau(x)$; in the second row, $\tau(x)$ varies with x .

and it is also identical to the coefficient of Z_i in the OLS

$$Y_i \sim \{1, Z_i, R_i, L_i\}, \quad (20.6)$$

where

$$R_i = \max(X_i - x_0, 0), \quad L_i = \min(X_i - x_0, 0)$$

indicate the right and left parts of $X_i - x_0$, respectively. I leave the algebraic details to Problem 20.1.

However, this approach may be sensitive to the violation of the linear model. Theory suggests running regression using only the local observations near the cutoff point². However, the rule for choosing the “local points” are quite involved. Fortunately, the `rdrobust` function in the `rdrobust` package in R implements various choices of “local points.” Since choosing the “local points” is the key in regression discontinuity, it seems more sensible to report estimates and confidence intervals based on various choices of the “local points.”

20.2.4 An example

Lee (2008) gave a famous example of using regression discontinuity to study the incumbency advantage in the U.S. House. He wrote that “incumbents are, by definition, those politicians who were successful in the previous election. If what makes them successful is somewhat persistent over time, they should be expected to be somewhat more successful when running for re-election.” Therefore, this is a fundamentally challenging causal inference problem. The regression discontinuity is a clever study design to study this problem.

The running variable is the lagged vote in the previous election centered at 0, and the outcome is the vote in current election, with units being the congressional districts. The treatment is the binary indicator for being the current incumbent party in a district, determined by the lagged vote. Figure 20.4 show the raw data.

The `rdrobust` function gives three sets of the point estimate and confidence intervals. They all suggest positive incumbency advantage.

```
> library(rdrobust)
> library(rddtools)
> data(house)
> RDDest = rdrobust(house$y, house$x)
[1] "Mass points detected in the running variable."
> cbind(RDDest$coef, RDDest$ci)
   Coeff    CI Lower    CI Upper
Conventional  0.06372533 0.04224798 0.08520269
Bias-Corrected 0.05937028 0.03789292 0.08084763
Robust        0.05937028 0.03481238 0.08392818
```

²This is called *local linear regression* in nonparametric statistics, which belongs to a broader class of *local polynomial regression* (Fan and Gijbels, 1996).

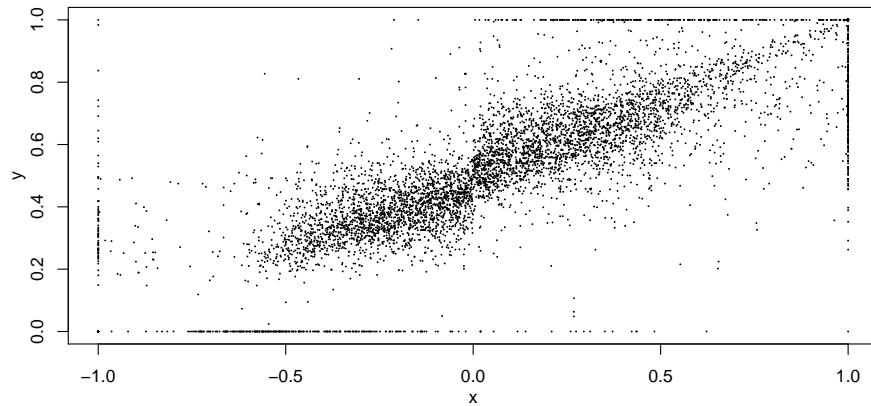


FIGURE 20.4: Raw data of Lee (2008)

Figure 20.5 shows the point estimates and the confidence intervals based on OLS with different choices of the local points defined by $|X| < x_0$. While the point estimates and the confidence intervals are sensitive to the choice of x_0 , the qualitative result remains the same as above.

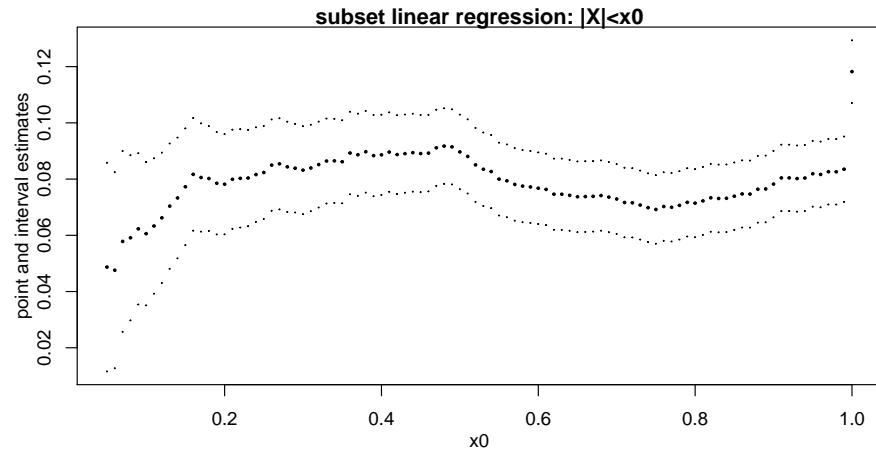


FIGURE 20.5: Estimates based on local linear regressions

20.2.5 Problems of regression discontinuity

What can go wrong with the regression discontinuity analysis? The technical challenge is to specify the neighborhood near the cutoff point. We have discussed this issue above.

In addition, Theorem 20.2 holds under a continuity condition. It may be

violated in practice. For instance, if the mortality rate jumps at the age of 21, then the jumps in Figure 20.2 may not be due to the change of drinking behavior due to the legal drink age. However, it is hard to check the violation of the continuity condition empirically.

McCrary (2008) proposed an indirect test for the validity of the regression discontinuity. He suggested checking the density of the running variable at the cutoff point. The discontinuity in the density of the running variable at the cutoff point may suggest that some units were able to manipulate their treatment status perfectly.

20.3 Homework Problems

20.1 Linear potential outcome models

This problem gives more details for the numerical equivalence in Section 20.2.3.

Show that $\hat{\tau}(x_0)$ equals the coefficients of Z_i in OLS fits (20.5) and (20.5).

Hint: It is helpful to start with the figures of $Z_i(X_i - x_0)$, L_i , and R_i with X_i on the x-axis. The conclusion holds by reparametrizing the OLS regressions.

20.2 Simulation for regression discontinuity

`RDD_numerical.R` simulates potential outcomes from linear models and generates Figure 20.3. Change them to nonlinear models, and compare different point estimators and confidence intervals, including the biases and variances of the point estimators, and the coverage properties of confidence intervals.

20.3 Re-analysis of the data on the minimum legal drink age

Analyze the data `mlda.csv` of Carpenter and Dobkin (2009).

20.4 Recommended reading

D'Amour et al. (2021) discussed the implications of overlap with high dimensional covariates.

Thistlethwaite and Campbell (1960)'s original paper on regression discontinuity was re-printed as Thistlethwaite and Campbell (2016) with many insightful comments. Coincidentally, Thistlethwaite and Campbell (1960) and Rubin (1974) were both published in the *Journal of Educational Psychology*.

Part V

Instrumental variables



21

An Experimental Perspective

The *instrumental variable method* has been a powerful tool in econometrics. It identifies causal effects in studies without unconfoundedness between the treatment and the outcome. It relies on an additional variable, called the *instrumental variable* (IV), that satisfies certain conditions. These conditions may not be easy to digest when you read for the first time. In some sense, IV is a magic. This chapter presents a not-so-magic perspective based on the encouragement design. This again echos Dorn (1953)'s suggestion that the planner of an observational study should always ask himself the following question:

How would the study be conducted if it were possible to do it by controlled experimentation?

The experimental analog of the IV method is the encouragement design (Zelen, 1979; Powers and Swinton, 1984; Holland, 1986).

21.1 Encouragement Design and Noncompliance

Consider an experiment with units indexed by $i = 1, \dots, n$. Let Z_i be the treatment assigned, with 1 for the treatment and 0 for the control. Let D_i be the treatment received, with 1 for the treatment and 0 for the control. When $Z_i \neq D_i$ for some unit i , the noncompliance problem arises. Noncompliance is a very common problem especially in encouragement designs involving human beings as experimental units. In these cases, the experimenters cannot force the units to take the treatment but rather only encourage them to do so. Let Y_i be the outcome of interest.

Consider complete randomization of Z and ignore covariates X now. We have the potential values for the treatment received $\{D_i(1), D_i(0)\}$ and the potential values for the outcome $\{Y_i(1), Y_i(0)\}$, all with respect to the treatment assignment levels 1 and 0. Their observed values are $D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)$ and $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$, respectively. For notational simplicity, we assume $\{Z_i, D_i(1), D_i(0), Y_i(1), Y_i(0)\}_{i=1}^n \stackrel{\text{IID}}{\sim} \{Z, D(1), D(0), Y(1), Y(0)\}$ and sometimes drop the subscript i without causing confusions.

We start with completely randomized experiments.

Assumption 21.1 (randomization) $Z \perp\!\!\!\perp \{D(1), D(0), Y(1), Y(0)\}$.

Randomization allows for identification of the average causal effects on D and Y :

$$\tau_D = E\{D(1) - D(0)\} = E(D | Z = 1) - E(D | Z = 0)$$

and

$$\tau_Y = E\{Y(1) - Y(0)\} = E(Y | Z = 1) - E(Y | Z = 0).$$

We can use simple difference-in-means estimators $\hat{\tau}_D$ and $\hat{\tau}_Y$ to estimate τ_D and τ_Y , respectively.

Reporting the estimate $\hat{\tau}_Y$ with the associated standard error is called the intention-to-treat (ITT) analysis. It estimates the effect of the treatment assignment on the outcome, and complete randomization in Assumption 21.1 justifies this analysis. However, it may not answer the scientific question, that is, the causal effect of the treatment received on the outcome.

21.2 Latent Compliance Status and Effects

21.2.1 Nonparametric identification

Following Imbens and Angrist (1994) and Angrist et al. (1996), we stratify the population based on the joint potential values of $\{D_i(1), D_i(0)\}$. Because D is binary, we have four possible combinations:

$$U_i = \begin{cases} a, & \text{if } D_i(1) = 1 \text{ and } D_i(0) = 1; \\ c, & \text{if } D_i(1) = 1 \text{ and } D_i(0) = 0; \\ d, & \text{if } D_i(1) = 0 \text{ and } D_i(0) = 1; \\ n, & \text{if } D_i(1) = 0 \text{ and } D_i(0) = 0, \end{cases}$$

where “a” is for “always taker,” “c” is for “complier,” “d” is for “defier,” and “n” is for “never taker.” Because we cannot observe $D_i(1)$ and $D_i(0)$ simultaneously, U_i is a latent variable for the compliance behavior of unit i .

Based on U , we can use the law of total probability to decompose the average causal effect on Y into four terms:

$$\begin{aligned} \tau_Y = & E\{Y(1) - Y(0) | U = a\} \text{pr}(U = a) \\ & + E\{Y(1) - Y(0) | U = c\} \text{pr}(U = c) \\ & + E\{Y(1) - Y(0) | U = d\} \text{pr}(U = d) \\ & + E\{Y(1) - Y(0) | U = n\} \text{pr}(U = n). \end{aligned} \quad (21.1)$$

Therefore, τ_Y is a weighted average of four latent subgroup effects. We will look into more details of the latent groups below.

Assumption 21.2 below restricts the third term in (21.1) to be zero.

Assumption 21.2 (monotonicity) $\text{pr}(U = d) = 0$ or $D_i(1) \geq D_i(0)$, that is, there are no defiers.

Assumption 21.2 holds automatically with *one-sided noncompliance* when the units assigned to the control arm have no access to the treatment, i.e., $D_i(0) = 0$ for all units. Under randomization, Assumption 21.2 has a testable implication that

$$\text{pr}(D = 1 | Z = 1) \geq \text{pr}(D = 1 | Z = 0).$$

But Assumption 21.2 is much stronger than the inequality above. The former restricts $D_i(1)$ and $D_i(0)$ at the individual level and the latter restricts them only on average. Nevertheless, when this testable implication holds, we cannot use the observed data to refute Assumption 21.2.

Assumption 21.3 below restricts the first and last terms in (21.1) to be zero based on the mechanism of the treatment assignment on the outcome through only the treatment received.

Assumption 21.3 (exclusion restriction) $Y_i(1) = Y_i(0)$ for always takers with $U_i = a$ and never takers with $U_i = n$.

Assumption 21.3 requires that the treatment assignment affects the outcome only if it affects the treatment received. In double-blind clinical trial¹, it is biologically plausible because the outcome only depends on the actual treatment received. That is, if the treatment assignment does not change the treatment received, it does not change the outcome either. It can be violated if the treatment assignment has *direct effects* on the outcome not through the treatment received. For example, some randomized controlled trials are not double blinded, and the treatment assignment can have some unknown pathways to the outcome.

Under Assumptions 21.2 and 21.3, the decomposition (21.1) only has the second term :

$$\tau_Y = E\{Y(1) - Y(0) | U = c\}\text{pr}(U = c). \quad (21.2)$$

Similarly, we can decompose the average causal effect on D into four terms:

$$\begin{aligned} \tau_D &= E\{D(1) - D(0) | U = a\}\text{pr}(U = a) \\ &\quad + E\{D(1) - D(0) | U = c\}\text{pr}(U = c) \\ &\quad + E\{D(1) - D(0) | U = d\}\text{pr}(U = d) \\ &\quad + E\{D(1) - D(0) | U = n\}\text{pr}(U = n) \\ &= 0 \times \text{pr}(U = a) + 1 \times \text{pr}(U = c) + (-1) \times \text{pr}(U = d) + 0 \times \text{pr}(U = n), \end{aligned}$$

¹In general, it is better to blind the experiment to avoid various biases arising from placebo effects, patients' expectation, etc. In double blind trials, both doctors and patients do not know the treatment; in single blind trials, the patients do not know the treatment but the doctors know. Sometimes, it is impossible to conduct double or even single blind trials. Those trials are called open trials.

which, under Assumption 21.2, reduces to

$$\tau_D = \text{pr}(U = c). \quad (21.3)$$

This is an interesting fact that the proportion of the compliers π_c equals the average causal effect of the treatment assigned on D , an identifiable quantity under complete randomization. Although we do not know all the compliers based on the observed data, we can identify their proportion in the whole population based on (21.3). Combining (21.2) and (21.3), we have the following result.

Theorem 21.1 *Under Assumptions 21.2–21.3, we have*

$$E\{Y(1) - Y(0) \mid U = c\} = \frac{\tau_Y}{\tau_D}$$

if $\tau_D \neq 0$.

Following Imbens and Angrist (1994) and Angrist et al. (1996), we define a new causal effect below.

Definition 21.1 (CACE or LATE) Define

$$\tau_c \equiv E\{Y(1) - Y(0) \mid U = c\}$$

as the “complier average causal effect (CACE)” or the “local average treatment effect (LATE)”. It has alternative forms:

$$\begin{aligned} \tau_c &= E\{Y(1) - Y(0) \mid D(1) = 1, D(0) = 0\} \\ &= E\{Y(1) - Y(0) \mid D(1) > D(0)\}. \end{aligned}$$

Based on Definition 21.1, we can rewrite Theorem 21.1 as

$$\tau_c = \frac{\tau_Y}{\tau_D},$$

that is, the CACE or LATE equals the ratio of the average causal effects on Y over that on D . Under Assumption 21.1, we further identify the CACE below.

Corollary 21.1 *Under Assumptions 21.1–21.3, we have*

$$\tau_c = \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)}.$$

Therefore, under randomization, monotonicity, and exclusion restriction, we can nonparametrically identify the CACE as the ratio of the difference in means of the outcome over the difference in means of the treatment received.

21.2.2 Estimation

Based on Corollary 21.1, we can estimate τ_c by a simple ratio

$$\hat{\tau}_c = \frac{\hat{\tau}_Y}{\hat{\tau}_D},$$

which is called the Wald estimator (Wald, 1940) or the IV estimator. In the above discussion, Z acts as the IV for D .

We can obtain the variance estimator based on the following heuristics (see Example A1.3):

$$\hat{\tau}_c - \tau_c = (\hat{\tau}_Y - \tau_c \hat{\tau}_D)/\hat{\tau}_D \approx (\hat{\tau}_Y - \tau_c \hat{\tau}_D)/\tau_D = \hat{\tau}_A/\tau_D,$$

where $\hat{\tau}_A$ is the difference-in-means of the adjusted outcome $A_i = Y_i - \tau_c D_i$. So the asymptotic variance of $\hat{\tau}_c$ is close to the variance of $\hat{\tau}_A$ divided by τ_D^2 . The variance estimation proceeds in the following steps:

1. obtain the adjusted outcomes $\hat{A}_i = Y_i - \hat{\tau}_c D_i$ ($i = 1, \dots, n$);
2. obtain the Neyman-type variance estimate based on the adjusted outcomes:

$$\hat{V}_{\hat{A}} = \frac{\hat{S}_{\hat{A}}^2(1)}{n_1} + \frac{\hat{S}_{\hat{A}}^2(0)}{n_0},$$

where $\hat{S}_{\hat{A}}^2(1)$ and $\hat{S}_{\hat{A}}^2(0)$ are the sample variances of the \hat{A}_i 's under treatment and control, respectively;

3. obtain the final variance estimator $\hat{V}_{\hat{A}}/\hat{\tau}_D^2$.

Under the null hypothesis that $\tau_c = 0$, we can simply approximate the variance by $\hat{V}_Y/\hat{\tau}_D^2$, where \hat{V}_Y is the Neyman-type variance estimate for the difference in means of Y . This variance estimator is inconsistent if the true τ_c is not zero. Therefore, it works for testing but not for estimation. Nevertheless, it gives interesting insights for the ITT estimator and the Wald estimator. The ITT estimator $\hat{\tau}_Y$ has estimated standard error $\sqrt{\hat{V}_Y}$. The Wald estimator $\hat{\tau}_Y/\hat{\tau}_D$ essentially equals the ITT estimator multiplied by $1/\hat{\tau}_D > 1$, which is larger in magnitude but at the same time its estimated standard error increases by the same factor. The confidence intervals for τ_Y and τ_c are

$$\hat{\tau}_Y \pm z_{1-\alpha/2} \sqrt{\hat{V}_Y}$$

and

$$\hat{\tau}_Y/\hat{\tau}_D \pm z_{1-\alpha/2} \sqrt{\hat{V}_Y/\hat{\tau}_D^2} = \left(\hat{\tau}_Y \pm z_{1-\alpha/2} \sqrt{\hat{V}_Y} \right) / \hat{\tau}_D.$$

These confidence intervals give the same qualitative conclusions since they will both cover zero or not. In some sense, the IV analysis provides the same qualitative information as the ITT analysis of Y although it involves more complicated procedures.

21.3 Covariates

21.3.1 Covariate adjustment in complete randomization

We now consider completely randomized experiments with covariates, and assume $Z \perp\!\!\!\perp \{D(1), D(0), Y(1), Y(0), X\}$. With covariates X , we can obtain Lin (2013)'s estimators $\hat{\tau}_{D,L}$ and $\hat{\tau}_{Y,L}$ for both D and Y , resulting in $\hat{\tau}_{c,L} = \hat{\tau}_{Y,L}/\hat{\tau}_{D,L}$. Recall that

$$\begin{aligned}\hat{\tau}_{D,L} &= \left\{ \hat{D}(1) - \hat{\beta}_{D1}^T \hat{X}(1) \right\} - \left\{ \hat{D}(0) - \hat{\beta}_{D0}^T \hat{X}(0) \right\}, \\ \hat{\tau}_{Y,L} &= \left\{ \hat{Y}(1) - \hat{\beta}_{Y1}^T \hat{X}(1) \right\} - \left\{ \hat{Y}(0) - \hat{\beta}_{Y0}^T \hat{X}(0) \right\},\end{aligned}$$

where $\hat{\beta}_{D1}$ and $\hat{\beta}_{Y1}$ are the coefficients of X in the OLS fits of D and Y in the treated group, and $\hat{\beta}_{D0}$ and $\hat{\beta}_{Y0}$ are the coefficients of X in the OLS fits of D and Y in the control group. We can approximate the standard error of $\hat{\tau}_{c,L}$ based on the following heuristics (again see Example A1.3):

$$\hat{\tau}_{c,L} - \tau_c = (\hat{\tau}_{Y,L} - \tau_c \hat{\tau}_{D,L})/\hat{\tau}_{D,L} \approx (\hat{\tau}_{Y,L} - \tau_c \hat{\tau}_{D,L})/\tau_D = \hat{\tau}_A/\tau_D,$$

where $\hat{\tau}_A$ is the difference-in-means of A , defined as

$$A_i = \begin{cases} (Y_i - \hat{\beta}_{Y1}^T X_i) - \tau_c (D_i - \hat{\beta}_{D1}^T X_i), & \text{if } Z_i = 1, \\ (Y_i - \hat{\beta}_{Y0}^T X_i) - \tau_c (D_i - \hat{\beta}_{D0}^T X_i), & \text{if } Z_i = 0. \end{cases}$$

The variance estimation proceeds in the following steps:

1. obtain the adjusted outcomes \hat{A}_i ($i = 1, \dots, n$) with

$$\hat{A}_i = \begin{cases} (Y_i - \hat{\beta}_{Y1}^T X_i) - \hat{\tau}_{c,L} (D_i - \hat{\beta}_{D1}^T X_i), & \text{if } Z_i = 1, \\ (Y_i - \hat{\beta}_{Y0}^T X_i) - \hat{\tau}_{c,L} (D_i - \hat{\beta}_{D0}^T X_i), & \text{if } Z_i = 0; \end{cases}$$

2. obtain the Neyman-type variance estimate based on the adjusted outcomes:

$$\hat{V}_{\hat{A}} = \frac{\hat{S}_{\hat{A}}^2(1)}{n_1} + \frac{\hat{S}_{\hat{A}}^2(0)}{n_0},$$

where $\hat{S}_{\hat{A}}^2(1)$ and $\hat{S}_{\hat{A}}^2(0)$ are the sample variances of the \hat{A}_i 's under the treatment and control, respectively;

3. obtain the final variance estimator $\hat{V}_{\hat{A}}/\hat{\tau}_{D,L}^2$.

Again under the null with $\tau_c = 0$, we can approximate the estimated standard error for $\hat{\tau}_{c,L}$ by the estimated standard error of $\hat{\tau}_{Y,L}$ (e.g., the EHW standard error in the fully interacted linear model) divided by $\hat{\tau}_{D,L}$.

21.3.2 Covariates in conditional randomization or unconfounded observational studies

If randomization holds conditionally, i.e.,

$$Z \perp\!\!\!\perp \{D(1), D(0), Y(1), Y(0)\} \mid X,$$

then we must adjust for covariates to avoid bias. The analysis is also straightforward since we already have discussed many estimators in Part III for estimating the effects of Z on D and Y , respectively. We can just use them in the ratio formula $\hat{\tau}_c = \hat{\tau}_Y / \hat{\tau}_D$ and use the bootstrap to approximate the asymptotic variance.

21.4 Weak IV

Even $\tau_D > 0$, there is a positive probability that $\hat{\tau}_D$ is zero, so the variance of $\hat{\tau}_c$ is infinity. The variance from the Normal approximation discussed before is not the variance of $\hat{\tau}_c$ but rather the variance of its asymptotic distribution. This is a subtle technical point. When τ_D is close to 0, which is referred to as the weak IV case, the ratio estimator $\hat{\tau}_c = \hat{\tau}_Y / \hat{\tau}_D$ has poor finite-sample properties. Under this scenario, $\hat{\tau}_c$ has finite sample bias and non-Normal asymptotic distribution, and the corresponding Wald-type confidence intervals have poor coverage properties². In the simple case with a binary outcome Y , we know that τ_Y must be bounded between -1 and 1 , but there is no guarantee that $\hat{\tau}_c$ is bounded between -1 and 1 . How do we deal with a weak IV?

From a testing perspective, there is an easy solution. Because $\tau_c = \tau_Y / \tau_D$, so the following two null hypotheses are equivalent:

$$H_0 : \tau_c = 0 \iff H'_0 : \tau_Y = 0.$$

Therefore, we simply test H'_0 , i.e., the average causal effect of Z on Y is zero. This echos our discussion in Section 21.2.2.

From an estimation perspective, we can focus on the confidence interval although the point estimator has poor finite-sample properties. Because $\tau_c = \tau_Y / \tau_D$, this is similar to the classical Fieller–Creasy problem in statistics. Below we discuss a strategy for constructing confidence interval for τ_c motivated by Fieller (1954); see Section A1.4.2. Given the true value τ_c , we have

$$\tau_Y - \tau_c \tau_D = 0.$$

²The theory often assumes that τ_D has the order $n^{-1/2}$. Under this regime, the proportion of compliers goes to 0 as n goes to infinity. The IV method can only identify a subgroup average causal effect with the proportion shrinking to 0. This is a contrived regime for theoretical analysis. It is hard to justify this assumption in practice. The follow discussion does not assume it.

So we can construct a confidence set for τ_c by inverting a sequence of null hypotheses

$$H_0(b) : \tau_c = b$$

This null hypothesis is equivalent to the null hypothesis of zero average causal effect on the outcome $A_i(b) = Y_i - bD_i$:

$$H_0(b) : \tau_{A(b)} = 0.$$

Let $\hat{\tau}_A(b)$ be a generic estimator for $\tau_{A(b)}$ with the associated variance estimator $\hat{V}_A(b)$. In the CRE without covariates, $\hat{\tau}_A(b)$ is the difference in means of the outcome $A_i(b)$ and $\hat{V}_A(b)$ is the Neyman-type variance estimator. In the CRE with covariates, $\hat{\tau}_A(b)$ is Lin (2013)'s estimator for the outcome $A_i(b)$ and $\hat{V}_A(b)$ is the EHW variance estimator in the associated OLS fit of $Y_i - bD_i$ on (Z_i, X_i, Z_iX_i) . In unconfounded observational studies, we can obtain the estimator for the average causal effect on $A_i(b)$ and the associated variance estimator based on many existing strategies in Part III.

Based on $\hat{\tau}_A(b)$ and $\tau_{A(b)}$, we can construct a Wald-type test for $H_0(b)$. Inverting tests, we can construct the following confidence set for τ_c :

$$\left\{ b : \frac{\hat{\tau}_A^2(b)}{\hat{V}_A(b)} \leq z_\alpha^2 \right\}.$$

This is close to the Anderson–Rubin-type confidence interval in econometrics (Anderson and Rubin, 1950). Due to its connection to Fieller (1954), I will call it the Fieller–Anderson–Rubin confidence interval. These weak-IV confidence intervals reduce to the asymptotic confidence intervals when the IV is strong. But they have additional guarantees when the IV is weak. I recommend using them in practice.

Example 21.1 To gain intuition about the Fieller–Anderson–Rubin confidence interval, we look into the simple case of the CRE without covariates. The quadratic inequality in the confidence interval reduces to

$$\begin{aligned} & (\hat{\tau}_Y - b\hat{\tau}_D)^2 \\ & \leq z_\alpha^2 \left[n_1^{-1} \{ \hat{S}_Y^2(1) + b^2 \hat{S}_D^2(1) - 2b\hat{S}_{YD}(1) \} \right. \\ & \quad \left. + n_0^{-1} \{ \hat{S}_Y^2(0) + b^2 \hat{S}_D^2(0) - 2b\hat{S}_{YD}(0) \} \right], \end{aligned}$$

where $\{ \hat{S}_Y^2(1), \hat{S}_D^2(1), \hat{S}_{YD}(1) \}$ and $\{ \hat{S}_Y^2(0), \hat{S}_D^2(0), \hat{S}_{YD}(0) \}$ are the sample variances and covariances of Y and D under treatment and control, respectively. The confidence set can be a close interval, two disconnected intervals, an empty set, or the whole real line. I relegate the detailed discussion to Problem 21.3.

21.5 Application

The `mediation` package contains a dataset `jobs` from Job Search Intervention Study (JOBS II), which was a randomized field experiment that investigates the efficacy of a job training intervention on unemployed workers. The variable `treat` is the indicator for whether a participant was randomly selected for the JOBS II training program, and the variable `comply` is the indicator for whether a participant actually participated in the JOBS II program. An outcome of interest is `job_seek` for measuring the level of job-search self-efficacy with values from 1 to 5. A few standard covariates are `sex`, `age`, `marital`, `nonwhite`, `educ`, and `income`.

Without using covariates, the confidence intervals based on the delta method and the bootstrap are

```
> est  
[1] 0.1087904  
> c(est - 1.96*dse, est + 1.96*dse)  
[1] -0.05002163 0.26760235  
> c(est - 1.96*bse, est + 1.96*bse)  
[1] -0.04657384 0.26415455
```

Adjusting for covariates, the confidence intervals based on the delta method and the bootstrap are

```
> est  
[1] 0.1176332  
> c(est - 1.96*dse, est + 1.96*dse)  
[1] -0.03638421 0.27165070  
> c(est - 1.96*bse, est + 1.96*bse)  
[1] -0.03926737 0.27453386
```

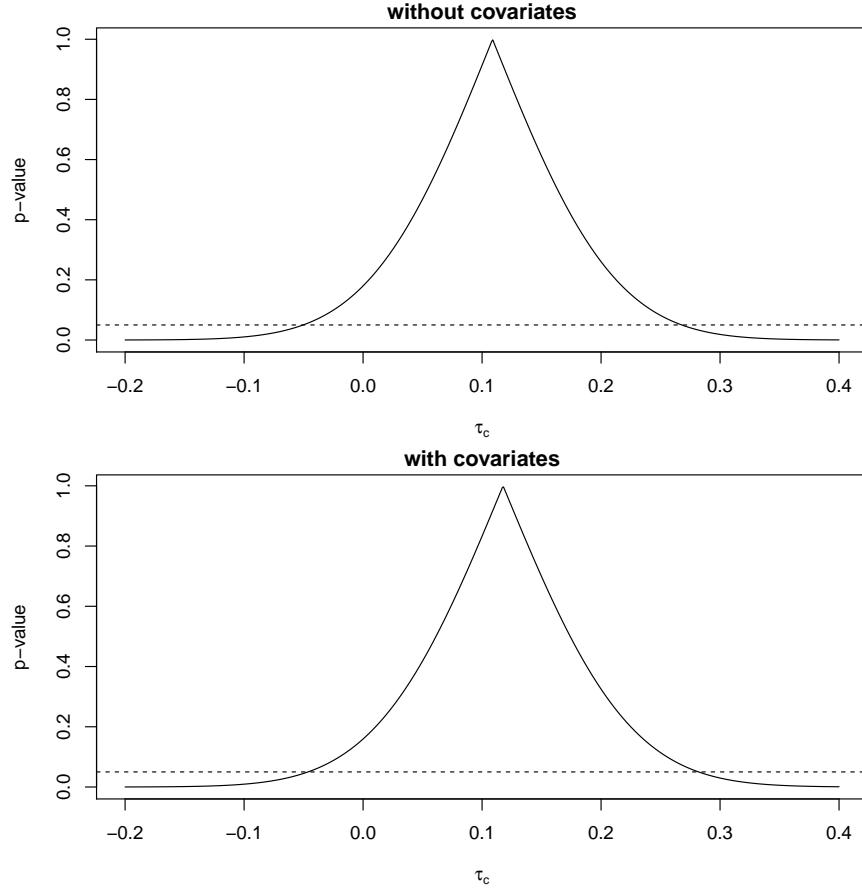
We can also construct confidence interval by inverting tests. Without using covariates, it is

```
> ARCI  
[1] -0.050 0.267
```

adjusting for covariates, it is

```
> ARCI  
[1] -0.046 0.281
```

Figure 21.1 plots the *p*-values for a sequence of tests.

FIGURE 21.1: Confidence interval of τ_c by inverting tests

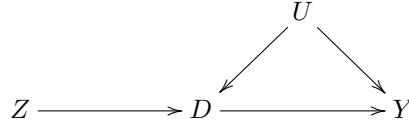
21.6 Interpreting the Complier Average Causal Effect

The notation for potential outcomes $\{D(1), D(0), Y(1), Y(0)\}$ is with respect to the hypothetical intervention of the treatment assigned Z . So τ_c is the average causal effect of the treatment assigned on the outcome for compliers. Fortunately, $D = Z$ for compliers, so we can also interpret τ_c as the average causal effect of the treatment received on the outcome for compliers. This partially answers the scientific question.

Some papers (e.g., Angrist et al., 1996) use different notation. They use $Y_i(z, d)$ for the potential outcome of unit i under a 2×2 factorial experiment with the treatment assigned z and treatment received d . The exclusion restriction assumption has the following form.

Assumption 21.4 (exclusion restriction) $Y_i(z, d) = Y_i(d)$ for all i , that is, the potential outcome is only a function of d .

Based on the causal graph below, Assumption 21.4 rules out the direct arrow from Z to Y . In such case, Z is an IV for D .



Under Assumption 21.4, the augmented notation $Y_i(z, d)$ reduces to $Y_i(d)$, which justifies the name of “exclusion restriction.” Therefore, $Y_i(1, d) = Y_i(0, d)$ for $d = 0, 1$, which, coupled with Assumption 21.2, implies that

$$\begin{aligned} Y_i(z=1) - Y_i(z=0) &= Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \\ &= \begin{cases} 0, & \text{if } U_i = a, \\ 0, & \text{if } U_i = n, \\ Y_i(d=1) - Y_i(d=0), & \text{if } U_i = c. \end{cases} \end{aligned}$$

In the above, we emphasize the potential outcomes are with respect to z, d or both, to avoid confusions. The previous decomposition of τ_Y holds and we have the following result from Imbens and Angrist (1994) and Angrist et al. (1996).

Recall the average causal effect on D , $\tau_D = E\{D(1) - D(0)\}$, define the average causal effect on Y as $\tau_Y = E\{Y(D(1)) - Y(D(0))\}$, and define the complier average causal effect as

$$\tau_c = E\{Y(d=1) - Y(d=0) \mid U = c\}.$$

Theorem 21.2 Under Assumptions 21.2–21.4, we have

$$Y(D(1)) - Y(D(0)) = \{\tau_D\} \times \{Y(d=1) - Y(d=0)\}$$

and $\tau_c = \tau_Y / \tau_D$.

The proof is almost identical to the proof of Theorem 21.1 with modifications of the notation. I leave it as Problem 21.2. From the notation $Y_i(d)$, it is more convenient to interpret τ_c as as the average causal effect of the treatment received on the outcome for compliers.

21.7 Homework problems

21.1 Variance of the Wald estimator

Show that $\text{var}(\hat{\tau}_c) = \infty$.

21.2 Proof of the main theorem of Imbens and Angrist (1994) and Angrist et al. (1996)

Prove Theorem 21.2.

21.3 More on the Fieller–Anderson–Rubin confidence set

The confidence set in Example 21.1 can be a close interval, two disconnected intervals, an empty set, or the whole real line. Find the precise conditions for each case.

21.4 Binary IV and ordinal treatment received

Angrist and Imbens (1995) discussed a more general setting with a binary IV Z , an ordinal treatment received $D \in \{0, 1, \dots, J\}$, and an outcome Y . The ordinal treatment received has potential outcomes $D(1)$ and $D(0)$ with respect to the binary IV, and the outcome has potential outcomes $Y(z, d)$ with respect to both the binary IV and the ordinal treatment received. Extend the discussion in Section 21.6 and the corresponding IV assumptions as below.

Assumption 21.5 We have (1) randomization that $Z \perp\!\!\!\perp \{D(z), Y(z, d) : z = 0, 1; d = 0, 1, \dots, J\}$; (2) monotonicity that $D(1) \geq D(0)$; and (3) exclusion restriction that $Y(z, d) = Y(d)$ for all $z = 0, 1$ and $d = 0, 1, \dots, J$.

They proved Theorem 21.3 below.

Theorem 21.3 Under Assumption 21.5, we have

$$\frac{E(Y | Z = 1) - E(Y | Z = 0)}{E(D | Z = 1) - E(D | Z = 0)} = \sum_{j=1}^J w_j E\{Y(j) - Y(j-1) | D(1) \geq j > D(0)\}$$

where

$$w_j = \frac{\Pr\{D(1) \geq j > D(0)\}}{\sum_{j'=1}^J \Pr\{D(1) \geq j' > D(0)\}}.$$

Prove Theorem 21.3.

Remark: When $J = 1$, Theorem 21.3 reduces to Theorem 21.2. It states that the standard IV formula identifies a weighted average of some latent subgroup effects. The weights are proportional to the probability of the latent groups defined by $D(1) \geq j > D(0)$, and the latent subgroup effects $E\{Y(j) - Y(j-1) | D(1) \geq j > D(0)\}$ compare the adjacent levels of the treatment received. However, this weighted average may not be easy to interpret because the latent groups overlap.

The proof can be tedious. A trick is to write the treatment received and outcome under treatment assignment z as

$$D(z) = \sum_{j=1}^J j 1\{D(z) = j\}, \quad Y(D(z)) = \sum_{j=1}^J Y(j) 1\{D(z) = j\}$$

to obtain

$$D(1) - D(0) = \sum_{j=0}^J j[1\{D(1) = j\} - 1\{D(0) = j\}]$$

and

$$Y(D(1)) - Y(D(0)) = \sum_{j=0}^J Y(j)[1\{D(1) = j\} - 1\{D(0) = j\}].$$

Then use the following *Abel's lemma*, also called *summation by parts*:

$$\sum_{j=0}^J f_j(g_{j+1} - g_j) = f_J g_{J+1} - f_0 g_0 - \sum_{j=1}^J g_j(f_j - f_{j-1})$$

for appropriately specified sequences (f_j) and (g_j) .

21.5 Data analysis: a flu shot encouragement design (McDonald et al., 1992)

The dataset in `fludata.txt` is from a randomized encouragement design of McDonald et al. (1992), which was also re-analyzed by Hirano et al. (2000). It contains the following variables:

<code>assign</code>	binary encouragement to receive the flu shot
<code>receive</code>	binary indicator for receiving the flu shot
<code>outcome</code>	binary outcome for flu related hospitalization
<code>age</code>	age of the patient
<code>sex</code>	sex of the patient
<code>race</code>	race of the patient
<code>copd, dm, heartd, renal, liverd</code>	various disease background covariates

Analyze the data with and without adjusting for the covariates.

21.6 Data analysis: the Karolinska data

Rubin (2008) used the Karolinska data as an example for the IV method. In `karolinska.txt`, whether a patient was diagnosed at large volume hospital can be viewed as an IV for whether a patient was treated at a large volume hospital. This is plausible at least conditioning on other observed covariates. See Rubin (2008)'s analysis for more details.

Reanalyze the data assuming that the IV is randomly assigned conditional on observed covariates.

21.7 Data analysis: a job training program (Schochet et al., 2008)

`jobtraining.rtf` contains the description of the data files `X.csv` and `Y.csv`.

`X.csv` contains the pretreatment covariates; you can view the sampling weight variable `wgt` as a covariate too. It is generally difficult to deal with sampling weights. Many previous analyses made this simplification. Conduct analyses with and without covariates.

`Y.csv` contains the sampling weight, treatment assigned, treatment received, and many post-treatment variables. Therefore, this data contains many outcomes depending on your questions of interest. The data also have many complications. First, some outcomes are missing. Second, unemployed individuals do not have wages or incomes. Third, the outcomes are repeatedly observed over time. When you do the data analysis, please give details about your choice of the questions of interest and estimators.

21.8 Recommended reading

Angrist et al. (1996) bridged the econometric IV perspective and statistical causal inference based on potential outcomes and demonstrated its usefulness with an application.

Some other early references on IV are Permutt and Hebel (1989), Sommer and Zeger (1991), Baker and Lindeman (1994), and Cuzick et al. (1997).

22

Disentangle Mixture Distributions and Instrumental Variable Inequalities

The IV model in Chapter 21 imposes Assumptions 21.1–21.3:

1. $Z \perp\!\!\!\perp \{D(1), D(0), Y(1), Y(0)\}$;
2. $\text{pr}(U = d) = 0$;
3. $Y(1) = Y(0)$ for $U = a$ or n .

Table 22.1 summarizes the observed groups and the corresponding latent groups.

TABLE 22.1: Observed groups and latent groups under Assumption 21.2

$$\begin{array}{llll} Z = 1 & D = 1 & D(1) = 1 & U = c \text{ or } a \\ Z = 1 & D = 0 & D(1) = 0 & U = n \\ Z = 0 & D = 1 & D(0) = 1 & U = a \\ Z = 0 & D = 0 & D(0) = 0 & U = c \text{ or } n \end{array}$$

Interestingly, Assumptions 21.1–21.3 together have some testable implications. Balke and Pearl (1997) called them the *instrumental variable inequalities*. This chapter will give an intuitive derivation of a special case of these inequalities. The proof is a direct consequence of identifying the means of the potential outcomes for all latent groups defined by U .

22.1 Disentangle Mixture Distributions and Instrumental Variable Inequalities

We summarize the main results in Theorem 22.1 below. Recall π_u as the proportion of type $U = u$, and define

$$\mu_{zu} = E\{Y(z) \mid U = u\}, \quad (d = 0, 1; u = a, n, c).$$

Theorem 22.1 Under Assumptions 21.1–21.3, we can identify the proportions of the latent types by

$$\begin{aligned}\pi_n &= \text{pr}(D = 0 | Z = 1), \\ \pi_a &= \text{pr}(D = 1 | Z = 0), \\ \pi_c &= E(D | Z = 1) - E(D | Z = 0),\end{aligned}$$

and the type-specific means of the potential outcomes by

$$\begin{aligned}\mu_{1n} = \mu_{0n} \equiv \mu_n &= E(Y | Z = 1, D = 0), \\ \mu_{1a} = \mu_{0a} \equiv \mu_a &= E(Y | Z = 0, D = 1), \\ \mu_{1c} &= \pi_c^{-1} \{E(DY | Z = 1) - E(DY | Z = 0)\}, \\ \mu_{0c} &= \pi_c^{-1} [E\{(1 - D)Y | Z = 0\} - E\{(1 - D)Y | Z = 1\}].\end{aligned}$$

Proof of Theorem 17.1: Part I: We first identify the proportions of the latent compliance types. We can identify the proportion of the never takers by

$$\begin{aligned}\text{pr}(D = 0 | Z = 1) &= \text{pr}(U = n | Z = 1) \\ &= \text{pr}(U = n) = \pi_n,\end{aligned}$$

and the proportion of the always takes by

$$\begin{aligned}\text{pr}(D = 1 | Z = 0) &= \text{pr}(U = a | Z = 0) \\ &= \text{pr}(U = a) = \pi_a.\end{aligned}$$

Therefore, the proportion of compliers is

$$\begin{aligned}\pi_c &= \text{pr}(U = c) = 1 - \pi_n - \pi_a \\ &= 1 - \text{pr}(D = 0 | Z = 1) - \text{pr}(D = 1 | Z = 0) \\ &= E(D | Z = 1) - E(D | Z = 0) = \tau_D,\end{aligned}$$

which is coherent with our discussion before. Although we do not know individual latent compliance types for all units, we can identify the proportions of never takers, always takers, and compliers.

Part II: We then identify the means of the potential outcomes within latent compliance types. Under Assumption 21.3,

$$\mu_{1a} = \mu_{0a} \equiv \mu_a, \quad \mu_{1n} = \mu_{0n} \equiv \mu_n.$$

The observed group ($Z = 1, D = 0$) only has never takers, so

$$E(Y | Z = 1, D = 0) = E\{Y(1) | Z = 1, U = n\} = E\{Y(1) | U = n\} = \mu_n.$$

The observed group ($Z = 0, D = 1$) only has always takers, so

$$E(Y | Z = 0, D = 1) = E\{Y(0) | Z = 0, U = a\} = E\{Y(0) | U = a\} = \mu_a.$$

The observed group ($Z = 1, D = 1$) has both compliers and always takers, so

$$\begin{aligned} E(Y | Z = 1, D = 1) &= E\{Y(1) | Z = 1, D(1) = 1\} \\ &= E\{Y(1) | D(1) = 1\} \\ &= \text{pr}\{D(0) = 1 | D(1) = 1\}E\{Y(1) | D(1) = 1, D(0) = 1\} \\ &\quad + \text{pr}\{D(0) = 0 | D(1) = 1\}E\{Y(1) | D(1) = 1, D(0) = 0\} \\ &= \frac{\pi_c}{\pi_c + \pi_a}\mu_{1c} + \frac{\pi_a}{\pi_c + \pi_a}\mu_a. \end{aligned}$$

Solve the linear equation above to obtain

$$\begin{aligned} \mu_{1c} &= \pi_c^{-1}\{(\pi_c + \pi_a)E(Y | Z = 1, D = 1) - \pi_aE(Y | Z = 0, D = 1)\} \\ &= \pi_c^{-1}\{\text{pr}(D = 1 | Z = 1)E(Y | Z = 1, D = 1) \\ &\quad - \text{pr}(D = 1 | Z = 0)E(Y | Z = 0, D = 1)\} \\ &= \pi_c^{-1}\{E(DY | Z = 1) - E(DY | Z = 0)\}. \end{aligned}$$

The observed group ($Z = 0, D = 0$) has both compliers and never takers, so we have

$$\begin{aligned} E(Y | Z = 0, D = 0) &= E\{Y(0) | Z = 0, D(0) = 0\} \\ &= E\{Y(0) | D(0) = 0\} \\ &= \text{pr}\{D(1) = 1 | D(0) = 0\}E\{Y(0) | D(1) = 1, D(0) = 0\} \\ &\quad + \text{pr}\{D(1) = 0 | D(0) = 0\}E\{Y(0) | D(1) = 0, D(0) = 0\} \\ &= \frac{\pi_c}{\pi_c + \pi_n}\mu_{0c} + \frac{\pi_n}{\pi_c + \pi_n}\mu_n. \end{aligned}$$

Solve the linear equation above to obtain

$$\begin{aligned} \mu_{0c} &= \pi_c^{-1}\{(\pi_c + \pi_n)E(Y | Z = 0, D = 0) - \pi_nE(Y | Z = 1, D = 0)\} \\ &= \pi_c^{-1}\{\text{pr}(D = 0 | Z = 0)E(Y | Z = 0, D = 0) \\ &\quad - \text{pr}(D = 0 | Z = 1)E(Y | Z = 1, D = 0)\} \\ &= \pi_c^{-1}[E\{(1 - D)Y | Z = 0\} - E\{(1 - D)Y | Z = 1\}]. \end{aligned}$$

□

Based on the formulas of μ_{1c} and μ_{0c} in Theorem 22.1, we have

$$\tau_c = \mu_{1c} - \mu_{0c} = \{E(Y | Z = 1) - E(Y | Z = 0)\}/\pi_c,$$

which is the same as the formula in Theorem 21.1 before.

Theorem 22.1 focuses on identifying the means of the potential outcomes, μ_{zu} . Imbens and Rubin (1997) derived more general identification formulas for the distribution of the potential outcomes; I leave the details to Problem 22.2.

22.2 Testable implications

Is there any additional value of the this detour for deriving the formula of τ_c ? The answer is yes. For binary outcome, the following inequalities must be true:

$$0 \leq \mu_{1c} \leq 1, \quad 0 \leq \mu_{0c} \leq 1,$$

which implies four inequalities

$$\begin{aligned} E(DY | Z = 1) - E(DY | Z = 0) &\geq 0, \\ E(DY | Z = 1) - E(DY | Z = 0) &\leq E(D | Z = 1) - E(D | Z = 0), \\ E\{(1 - D)Y | Z = 0\} - E\{(1 - D)Y | Z = 1\} &\geq 0, \\ E\{(1 - D)Y | Z = 0\} - E\{(1 - D)Y | Z = 1\} &\leq E(D | Z = 1) - E(D | Z = 0). \end{aligned}$$

Rearranging terms, we obtain the following unified inequalities.

Theorem 22.2 (Instrumental Variable Inequalities) *With a binary outcome Y , Assumptions 21.1–21.3 imply*

$$E(Q | Z = 1) - E(Q | Z = 0) \geq 0, \tag{22.1}$$

where $Q = DY, D(1 - Y), (D - 1)Y$ and $D + Y - DY$.

Under the IV assumptions 21.1–21.3, the difference in means for $Q = DY, D(1 - Y), (D - 1)Y$ and $D + Y - DY$ must all be non-negative. Importantly, these implications only involve the distribution of the observed variables. Rejection of the IV inequalities leads to rejection of the IV assumptions.

Balke and Pearl (1997) derived more general IV inequalities without assuming monotonicity. The above proving strategy is due to Jiang and Ding (2020) for a slightly more complex setting. Theorem 22.2 states the testable implications only for a binary outcome. Problem 22.3 gives an equivalent form, and Problem 22.4 gives the result for a general outcome.

22.3 Examples

For a binary outcome, we can estimate all the parameters by the method of moment as below.

```
## function for binary data (Z, D, Y)
## n_{zdy}'s are the counts from 2X2X2 table
IVbinary = function(n111, n110, n101, n100,
                     n011, n010, n001, n000){
```

```

n_tr = n111 + n110 + n101 + n100
n_co = n011 + n010 + n001 + n000
n     = n_tr + n_co

## proportions of the latent strata
pi_n = (n101 + n100)/n_tr
pi_a = (n011 + n010)/n_co
pi_c = 1 - pi_n - pi_a

## four observed means of the outcomes (Z=z,D=d)
mean_y_11 = n111/(n111 + n110)
mean_y_10 = n101/(n101 + n100)
mean_y_01 = n011/(n011 + n010)
mean_y_00 = n001/(n001 + n000)

## means of the outcomes of two strata
mu_n1 = mean_y_10
mu_a0 = mean_y_01
## ER implies the following two means
mu_n0 = mu_n1
mu_a1 = mu_a0
## stratum (Z=1,D=1) is a mixture of c and a
mu_c1 = ((pi_c + pi_a)*mean_y_11 - pi_a*mu_a1)/pi_c
## stratum (Z=0,D=0) is a mixture of c and n
mu_c0 = ((pi_c + pi_n)*mean_y_00 - pi_n*mu_n0)/pi_c

## identifiable quantities from the observed data
list(pi_c = pi_c,
     pi_n = pi_n,
     pi_a = pi_a,
     mu_c1= mu_c1,
     mu_c0= mu_c0,
     mu_n1= mu_n1,
     mu_n0= mu_n0,
     mu_a1= mu_a1,
     mu_a0= mu_a0)
}

```

We then re-visit two canonical examples.

Example 22.1 Investigators *et al.* (2014) assess the effectiveness of the emergency endovascular versus the open surgical repair strategies for patients with a clinical diagnosis of ruptured aortic aneurism. Patients are randomized to either the emergency endovascular or the open repair strategy. The primary outcome is the survival status after 30 days. Let Z be the treatment assigned, with $Z = 1$ for the endovascular strategy and $Z = 0$ for the open repair. Let D be the treatment received. Let Y be the survival status, with $Y = 1$ for dead, and $Y = 0$ for alive. The estimate of τ_c is 0.131 with 95% confidence interval $(-0.036, 0.298)$ including 0. Using the function above, we can obtain

TABLE 22.2: Binary data and IV inequalities

(a) Investigators et al. (2014)'s study

	$Z = 1$		$Z = 0$	
	$D = 1$	$D = 0$	$D = 1$	$D = 0$
$Y = 1$	107	68	24	131
$Y = 0$	42	42	8	79

(b) Hirano et al. (2000)'s study

	$Z = 1$		$Z = 0$	
	$D = 1$	$D = 0$	$D = 1$	$D = 0$
$Y = 1$	31	85	30	99
$Y = 0$	424	944	237	1041

```
$mu_c1
[1] 0.7086064
```

```
$mu_c0
[1] 0.6292042
```

There is no evidence of violating the IV assumptions.

Example 22.2 In Hirano et al. (2000), physicians are randomly selected to receive a letter encouraging them to inoculate patients at risk for flu. The treatment is the actual flu shot, and the outcome is an indicator for flu-related hospital visits. However, some patients do not comply with their assignments. Let Z_i be the indicator of encouragement to receive the flu shot, with $Z = 1$ if the physician receives the encouragement letter, and $Z = 0$ otherwise. Let D be the treatment received. Let Y be the outcome, with $Y = 0$ if for a flu-related hospitalization during the winter, and $Y = 1$ otherwise. The estimate of τ_c is 0.116 with 95% confidence interval $(-0.061, 0.293)$ including 0. Using the function above, we can obtain

```
$mu_c1
[1] -0.004548064
```

```
$mu_c0
[1] 0.1200094
```

Since $\hat{\mu}_{1c} < 0$, there is evidence of violating the IV assumptions.

22.4 Homework problems

22.1 Risk ratio for compliers

With binary outcome, we can define the risk ratio for compliers as

$$RR_c = \frac{\text{pr}\{Y(1) = 1 \mid U = c\}}{\text{pr}\{Y(0) = 1 \mid U = c\}}.$$

Show that under Assumptions 21.1–21.3, we can identify it by

$$RR_c = \frac{E(DY \mid Z = 1) - E(DY \mid Z = 0)}{E\{(D - 1)Y \mid Z = 1\} - E\{(D - 1)Y \mid Z = 0\}}.$$

Remark: Using Theorem 22.1, we can identify any comparisons between $E\{Y(1) \mid U = c\}$ and $E\{Y(0) \mid U = c\}$.

22.2 Disentangle the mixtures: distributional results

This problem extends Theorem 22.1. Define

$$f_{zu}(y) = \text{pr}\{Y(z) = y \mid U = u\}, \quad (d = 0, 1; u = a, n, c)$$

as the density of $Y(z)$ for latent stratum $U = u$, and define

$$g_{zd}(y) = \text{pr}(Y = y \mid Z = z, D = d)$$

as the density of the outcome within the observed group ($Z = z, D = d$). Show Theorem 22.3 below.

Theorem 22.3 *Under Assumptions 21.1–21.3, we can identify the type-specific densities of the potential outcomes by*

$$\begin{aligned} f_{1n}(y) &= f_{0n}(y) \equiv f_n(y) &= g_{10}(y), \\ f_{1a}(y) &= f_{0a}(y) \equiv f_a(y) &= g_{01}(y), \\ f_{1c}(y) &= \pi_c^{-1}\{\text{pr}(D = 1 \mid Z = 1)g_{11}(y) - \text{pr}(D = 1 \mid Z = 0)g_{01}(y)\}, \\ f_{0c}(y) &= \pi_c^{-1}\{\text{pr}(D = 0 \mid Z = 0)g_{00}(y) - \text{pr}(D = 0 \mid Z = 1)g_{10}(y)\}. \end{aligned}$$

22.3 Alternative form of Theorem 22.2

The inequalities in (22.1) can be re-written as

$$\begin{aligned} \text{pr}(D = 1, Y = y \mid Z = 1) &\geq \text{pr}(D = 1, Y = y \mid Z = 0), \\ \text{pr}(D = 0, Y = y \mid Z = 0) &\geq \text{pr}(D = 0, Y = y \mid Z = 1) \end{aligned}$$

for both $y = 0, 1$.

22.4 Instrumental variable inequalities for a general outcome

For a general outcome Y , show that Assumptions 21.1–21.3 imply

$$\begin{aligned}\text{pr}(D = 1, Y \geq y | Z = 1) &\geq \text{pr}(D = 1, Y \geq y | Z = 0), \\ \text{pr}(D = 1, Y < y | Z = 1) &\geq \text{pr}(D = 1, Y < y | Z = 0), \\ \text{pr}(D = 0, Y \geq y | Z = 0) &\geq \text{pr}(D = 0, Y \geq y | Z = 1), \\ \text{pr}(D = 0, Y < y | Z = 0) &\geq \text{pr}(D = 0, Y < y | Z = 1)\end{aligned}$$

for all y .

Remark: Imbens and Rubin (1997) and Kitagawa (2015) discussed similar results. For instance, we can test the first inequality based on an analog of the Kolmogorov–Smirnov statistic:

$$\text{KS}_1 = \max_y \left| \frac{\sum_{i=1}^n Z_i D_i 1(Y_i \leq y)}{\sum_{i=1}^n Z_i D_i} - \frac{\sum_{i=1}^n (1 - Z_i) D_i 1(Y_i \leq y)}{\sum_{i=1}^n (1 - Z_i) D_i} \right|.$$

22.5 Example for the IV inequalities

Give an example in which all the IV inequalities hold and another example in which not all the IV inequalities hold. You need to specify the joint distribution of (Z, D, Y) with binary Z and D .

22.6 Violations of the key assumptions

Theorem 21.1 relies on randomization, monotonicity, and exclusion restriction. The latter two are not testable even in randomized experiments. When they are violated, the IV estimator no longer identifies the complier average causal effect. This problem gives two cases below, which are restatement of Propositions 2 and 3 in Angrist et al. (1996).

Under Assumptions 21.1 and 21.2 without the exclusion restriction, we have

$$\frac{E(Y | Z = 1) - E(Y | Z = 0)}{E(D | Z = 1) - E(D | Z = 0)} - \tau_c = \frac{\pi_a \tau_a + \pi_n \tau_n}{\pi_c}$$

where

$$\tau_u = E\{Y(1) - Y(0) | U = u\}, \quad (U = a, n, c).$$

Under Assumptions 21.1 and 21.3 without the monotonicity, we have

$$\frac{E(Y | Z = 1) - E(Y | Z = 0)}{E(D | Z = 1) - E(D | Z = 0)} - \tau_c = \frac{\pi_d (\tau_c + \tau_d)}{\pi_c - \pi_d}.$$

Prove the above two results.

22.7 Problems of other analyses

In the process of deriving the IV inequalities in Section 22.1, we disentangled the mixture distributions by identifying the proportions of the latent strata

as well as the conditional means of their potential outcomes. These results are helpful for understanding the drawbacks of other seemingly reasonable analyses. I review three estimators below and suppose Assumptions 21.1–21.3 holds.

1. The *as-treated analysis* compares the means of the outcomes among units receiving the treatment and control, yielding

$$\tau_{AT} = E(Y | D = 1) - E(Y | D = 0).$$

Show that

$$\tau_{AT} = \frac{\pi_a \mu_a + \text{pr}(Z = 1)\pi_c \mu_{1c}}{\text{pr}(D = 1)} - \frac{\pi_n \mu_n + \text{pr}(Z = 0)\pi_c \mu_{0c}}{\text{pr}(D = 0)}.$$

2. The *per-protocol analysis* compares the units who comply with the treatment assigned in treatment and control groups, yielding

$$\tau_{PP} = E(Y | Z = 1, D = 1) - E(Y | Z = 0, D = 0).$$

Show that

$$\tau_{PP} = \frac{\pi_a \mu_a + \pi_c \mu_{1c}}{\pi_a + \pi_c} - \frac{\pi_n \mu_n + \pi_c \mu_{0c}}{\pi_n + \pi_c}.$$

3. We may also want to compare the outcomes among units receiving the treatment and control, conditioning on their treatment assignment, yielding

$$\begin{aligned}\tau_{Z=1} &= E(Y | Z = 1, D = 1) - E(Y | Z = 1, D = 0), \\ \tau_{Z=0} &= E(Y | Z = 0, D = 1) - E(Y | Z = 0, D = 0).\end{aligned}$$

Show that they reduce to

$$\tau_{Z=1} = \frac{\pi_a \mu_a + \pi_c \mu_{1c}}{\pi_a + \pi_c} - \mu_n, \quad \tau_{Z=0} = \mu_a - \frac{\pi_n \mu_n + \pi_c \mu_{0c}}{\pi_n + \pi_c}.$$

22.8 Bounds on the average causal effect on the whole population

Extend the discussion in Section 22.1 based on the notation in Section 21.6. With the potential outcome $Y(d)$, define the average causal effect of the treatment received on the outcome as

$$\delta = E\{Y(d = 1) - Y(d = 0)\},$$

and modify the definition of μ_{du} as

$$m_{du} = E\{Y(d) | U = u\}, \quad (z = 0, 1; u = a, n, c).$$

They satisfy

$$\delta = \sum_{u=a,n,c} \pi_u (m_{1u} - m_{0u}).$$

Section 22.1 identifies π_a , π_n , π_c , $m_{1a} = \mu_{1a}$, $m_{0n} = \mu_{0n}$, $m_{1c} = \mu_{1c}$ and $m_{0c} = \mu_{0c}$. But the data do not contain any information about m_{0a} and m_{1n} . Therefore, we cannot identify δ . With a bounded outcome, we can bound δ . Show the following result:

Theorem 22.4 *Under Assumptions 21.2–21.4 with a bounded outcome in $[y, \bar{y}]$, we have $\underline{\delta} \leq \delta \leq \bar{\delta}$, where*

$$\underline{\delta} = \delta' - \bar{y}\text{pr}(D = 1 | Z = 0) + \underline{y}\text{pr}(D = 0 | Z = 1)$$

and

$$\bar{\delta} = \delta' - \underline{y}\text{pr}(D = 1 | Z = 0) + \bar{y}\text{pr}(D = 0 | Z = 1)$$

with $\delta' = E(DY | Z = 1) - E(Y - DY | Z = 0)$.

Remark: In the special case with a binary outcome, the bounds simplify to

$$\underline{\delta} = E(DY | Z = 1) - E(D + Y - DY | Z = 0)$$

and

$$\bar{\delta} = E(DY + 1 - D | Z = 1) - E(Y - DY | Z = 0).$$

22.9 One-sided noncompliance and statistical inference

Consider a randomized encouragement design where the units assigned to the control have no access to the treatment. For unit i , let Z_i be the binary treatment assigned, D_i be the binary treatment received, and Y_i be the outcome of interest. One-sided noncompliance happens when

$$Z_i = 0 \implies D_i = 0 \quad (i = 1, \dots, n).$$

Suppose that Assumption 21.1 holds.

1. Does monotonicity Assumption 21.2 hold in this case? How many latent strata defined by $\{D_i(1), D_i(0)\}$ are there in this problem? How do we identify their proportions by the observed data distribution?
2. State the assumption of exclusion restriction. Under exclusion restriction, show that $E\{Y(z) | U = u\}$ can be identified by the observed data distributions. Give the formulas for all possible values of z and u . How do we identify the complier average causal effect in this case?
3. If we observe pretreatment covariates X_i for all units i , how do we use the covariate information to improve the estimation efficiency of the complier average causal effect?
4. Under Assumption 21.1, the exclusion restriction Assumption 21.3 has testable implications, which are the IV inequalities for one-sided noncompliance. State the IV inequalities.

5. Sommer and Zeger (1991) provided the following dataset:

	$Z = 1$		$Z = 0$	
	$D = 1$	$D = 0$	$D = 1$	$D = 0$
$Y = 1$	9663	2385	0	11514
$Y = 0$	12	34	0	74

Re-analyze it.

Remark: Bloom (1984) first discussed one-sided noncompliance and proposed the IV estimator $\hat{\tau}_c = \hat{\tau}_Y / \hat{\tau}_D$. His notation is different from this chapter.

22.10 One-sided noncompliance with partial adherence

Sanders and Karim (2021, Table 3) reported the following data from a randomized clinical trial aiming to estimate the efficacy of smoking cessation interventions among individuals with psychotic disorders.

group assigned	treatment received	group size	# positive outcomes
Control	None	151	25
Treatment	None	35	7
Treatment	Partial	42	17
Treatment	Full	70	40

Three tiers of treatment received are defined as follows: “full” treatment corresponds to attending all 8 treatment sessions, “partial” corresponds to attending 5 to 7 sessions, and “none” corresponds to < 5 sessions. The outcome is defined as the binary indicator of smoking reduction of 50% or greater relative to baseline, measured at three months.

In this problem, the treatment assignment Z is binary but the treatment received D takes three values 0, 0.5, 1 for “none”, “partial”, and “full.” The three-leveled D causes complications, but it can only be 0 under the control assignment. How many latent strata $U = \{D(1), D(0)\}$ do we have in this problem? Can we identify their proportions?

How do we extend the exclusion restriction to this problem? What can be the causal effects of interest? Can we identify them?

Analyze the data based on the questions above.

22.11 Recommended reading

Balke and Pearl (1997) derived more general IV inequalities.



23

An Econometric Perspective

Chapters 21 and 22 discuss the IV method from the experimental perspective. Figure 23.1 illustrates the intuition behind the discussion.

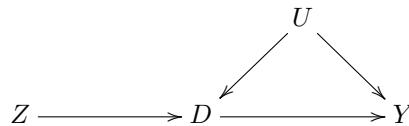


FIGURE 23.1: Causal diagram for IV

In an encouragement design with noncompliance, Z is randomized, so it is independent of the confounder U between the treatment received D and the outcome Y . Importantly, the treatment assignment Z does not have any direct effect on the outcome Y . It acts as an IV for the treatment received D in the sense that it affects the outcome Y only through the treatment received D . This IV is generated by the experimenter.

In many applications, randomization is infeasible. Then how can we draw causal inference in the presence of unmeasured confounding between the treatment and outcome? A clever idea from econometrics is to find *natural experiments* to mimic the setting of encouragement designs. To identify the causal effect of D on Y with unmeasured confounding, we can find another variable Z that satisfies the assumptions of the diagram in Figure 23.1. The variable Z must satisfy the following conditions:

1. it should be close to be randomized so that it is independent of the unmeasured confounding;
2. it should change the distribution of D ;
3. it should not affect the outcome Y directly.

If all these conditions hold, then Z is a valid IV.

This chapter will provide the traditional econometrics perspective on IV. It is based on linear regression. Imbens and Angrist (1994) and Angrist et al. (1996) made a fundamental contribution by clarifying connection between this perspective and the experimental perspective in Chapters 21 and 22. I will start with examples and then give more algebraic details.

23.1 Examples of studies with IVs

Finding IV for causal inference is more an art than a science. The algebraic details in later sections are not the most complicated ones in statistics. However, it is fundamentally challenging to find IV in empirical research. Below are some famous examples.

Example 23.1 *In an encouragement design, Z is the randomly assigned treatment, D is the final treatment received, and Y is the outcome. The IV assumptions encoded by Figure 23.1 is plausible in double-blind trials as discussed in Chapter 21. This is the ideal case for IV.*

Example 23.2 Hearst et al. (1986) reported that men with low lottery number in the Vietnam Era draft lottery had higher mortality rates afterwards. They attributed this to the negative effect of military service. Angrist (1990) further reported that men with low lottery number in the Vietnam Era draft lottery had lower subsequent earnings. He attributed this to the negative effect of military service. These explanations are plausible because the lottery numbers were randomly generated, men with low lottery number were more likely to have military service, and the lottery numbers were unlikely to affect the subsequent mortality or earnings. That is, Figure 23.1 is plausible. Angrist et al. (1996) reanalyzed the data using the IV framework. Here, the lottery number is the IV, military service is the treatment, and mortality or earnings is the outcome.

Example 23.3 Angrist and Krueger (1991) studied the return of schooling in years on earnings, using the quarter of birth as an IV. This IV is plausible because of the pseudo randomization of the quarter of birth. It affected the years of schooling because (1) most states required the students to enter school in the calendar year in which they turned six, and (2) compulsory schooling laws typically required students remain in the school before their sixteenth birthday. More important, it is plausible that the quarter of birth did not affect earnings directly.

Example 23.4 Angrist and Evans (1998) studied the effect of family size on mother's employment and work, using the sibling sex composition as an IV. This IV is plausible because of the pseudo randomization of the sibling sex composition. Moreover, parents in the US with two children of the same sex are more likely to have a third child than those parents with two children of different sex. It is also plausible that the sibling sex composition does not affect mother's employment and work directly.

Example 23.5 Card (1993) studied the effect of schooling on wage, using the geographic variation in college proximity as an IV. In particular, Z contains dummy variables for whether a subject grew up near a two-year college or a

four-year college. Although this study is classic, it might be a poor example for IV because parents' choices of where to live might not be random, and moreover, where a subject grew up might matter for the subsequent wage.

Example 23.6 Voight *et al.* (2012) studied the causal effect of plasma high-density lipoprotein (HDL) cholesterol on the risk of heart attack based on Mendelian randomization. They used some single-nucleotide polymorphisms (SNPs) as genetic IV for HDL, which are random with respect to the unmeasured confounders between HDL and heart attack by Mendel's second law, and affect heart attack only through HDL. I will give more details of Mendelian randomization in Chapter 25.

23.2 Brief Review of the Ordinary Least Squares

Before discussing the econometric view of IV, I will first review the OLS in statistics (see Chapter A2). This is a standard topic in statistics. However, it has different mathematical formulations, and the choice of formulation matters for the interpretation.

The first view is based on projection. Given any pair of random variables (D, Y) with finite second moments, define the population OLS coefficient as

$$\beta = \arg \min_b E(Y - D^\top b)^2 = E(DD^\top)^{-1}E(DY),$$

and then define the residual as $\varepsilon = Y - D^\top \beta$. By definition, Y decomposes into

$$Y = D^\top \beta + \varepsilon, \quad (23.1)$$

which must satisfy

$$E(D\varepsilon) = 0.$$

Based on $(D_i, Y_i)_{i=1}^n \stackrel{\text{IID}}{\sim} (D, Y)$, the OLS estimator of β is

$$\hat{\beta} = \left(\sum_{i=1}^n D_i D_i^\top \right)^{-1} \sum_{i=1}^n D_i Y_i.$$

Because

$$\hat{\beta} = \left(\sum_{i=1}^n D_i D_i^\top \right)^{-1} \sum_{i=1}^n D_i (D_i^\top \beta + \varepsilon_i) = \beta + \left(\sum_{i=1}^n D_i D_i^\top \right)^{-1} \sum_{i=1}^n D_i \varepsilon_i,$$

we can show that $\hat{\beta}$ is consistent for β because of $E(\varepsilon D) = 0$. The classical

EHW robust variance estimator for $\text{cov}(\hat{\beta})$ is

$$\hat{V}_{\text{EHW}} = \left(\sum_{i=1}^n D_i D_i^\top \right)^{-1} \left(\sum_{i=1}^n \hat{\varepsilon}_i^2 D_i D_i^\top \right) \left(\sum_{i=1}^n D_i D_i^\top \right)^{-1}$$

where $\hat{\varepsilon}_i = Y_i - D_i^\top \hat{\beta}$ is the residual.

The second view is to treat

$$Y = D^\top \beta + \varepsilon, \quad (23.2)$$

as a true model for data generating process. That is, given the random variables (D, ε) , we generate Y based on the linear equation (23.2). Importantly, in the data generating process, ε and D may be correlated in that $E(D\varepsilon) \neq 0$. Figure 23.2 gives such an example. This is the fundamental difference compared to the first view where $E(\varepsilon D) = 0$ holds by the definition of the population OLS. Consequently, the OLS estimator can be inconsistent:

$$\hat{\beta} \rightarrow \beta + E(DD^\top)^{-1} E(D\varepsilon) \neq \beta$$

in probability.

I end this section with definitions of *endogenous* and *exogenous* regressors based on (23.2), although their definitions are not unique in econometrics.

Definition 23.1 When $E(\varepsilon D) \neq 0$, the regressor D is called *endogenous*; when $E(\varepsilon D) = 0$, the regressor D is called *exogenous*.

The terminologies in Definition 23.1 are standard in econometrics. When $E(\varepsilon D) \neq 0$, we also say that we have *endogeneity*; when $E(\varepsilon D) = 0$, we also say that we have *exogeneity*.

In first view of OLS, the notions of endogeneity and exogeneity do not play any roles because $E(\varepsilon D) = 0$ by definition. Statisticians holding the first view usually find the notations of endogeneity and exogeneity strange, and consequently, find the idea of IV unnatural. To understand the econometric view of IV, we must switch to the second view of OLS.

23.3 Linear Instrumental Variable Model

When D is endogenous, the OLS estimator is inconsistent. We must use additional information to construct a consistent estimator for β . I will focus on the following linear IV model:

Definition 23.2 (linear IV model) We have

$$Y = D^\top \beta + \varepsilon,$$

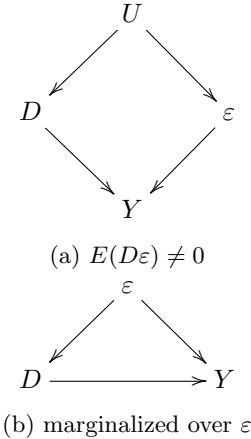
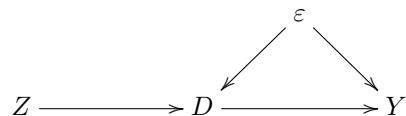


FIGURE 23.2: Different representations of the endogenous regressor

with

$$E(\varepsilon Z) = 0. \quad (23.3)$$

The linear IV model in Definition 23.2 can be illustrated by the following causal graph:



The above linear IV model allows that $E(\varepsilon D) \neq 0$ but requires an alternative moment condition (23.3). With $E(\varepsilon) = 0$ by incorporating the intercept, the new condition states that Z is uncorrelated with the error term ε . But any randomly generated noise is uncorrelated with ε , so an additional condition must hold to ensure that Z is useful for estimating β . Intuitively, the additional condition requires that Z is correlated to D , with more technical details stated below.

The mathematical requirement (23.3) seems simple. However, it is a key challenge in empirical research to find such a variable or variables Z that satisfies (23.3). Since the condition (23.3) involves the unobservable ε , it is generally untestable.

23.4 The Just-Identified Case

We first consider the case in which Z and D have the same dimension and $E(ZD^T)$ has full rank. The condition $E(\varepsilon Z) = 0$ implies that

$$\begin{aligned} E\{Z(Y - D^T \beta)\} &= 0 \implies E(ZY) = E(ZD^T)\beta \\ &\implies \beta = E(ZD^T)^{-1}E(ZY) \end{aligned}$$

if $E(ZD^T)$ is not degenerate. The OLS is a special case if $E(\varepsilon D) = 0$, i.e., D itself acts as an IV for itself. The resulting moment estimator is

$$\hat{\beta}_{IV} = \left(\sum_{i=1}^n Z_i D_i^T \right)^{-1} \sum_{i=1}^n Z_i Y_i. \quad (23.4)$$

In the simple case with an intercept and scalar D and Z , we have

$$\begin{cases} Y = \alpha + \beta D + \varepsilon, \\ E(\varepsilon) = 0, \quad \text{cov}(\varepsilon, Z) = 0, \end{cases}$$

which implies that

$$\text{cov}(Z, Y) = \beta \text{cov}(Z, D) \implies \beta = \frac{\text{cov}(Z, Y)}{\text{cov}(Z, D)}.$$

Standardizing the numerator and denominator by $\text{var}(Z)$, we have

$$\beta = \frac{\text{cov}(Z, Y)/\text{var}(Z)}{\text{cov}(Z, D)/\text{var}(Z)},$$

which equals the ratio between the coefficients of Z in the OLS fits of Y and D on Z . If Z is binary, these coefficients are differences in means and β reduces to

$$\beta = \frac{E(Y | Z = 1) - E(Y | Z = 0)}{E(D | Z = 1) - E(D | Z = 0)}.$$

This is identical to the identification formula in Theorem 21.1. That is, with a binary IV Z and a binary treatment D , the IV estimator recovers the CACE under the potential outcomes framework. This is a key result in Imbens and Angrist (1994) and Angrist et al. (1996).

23.5 The Over-Identified Case

The discussion in Section 23.4 focuses on the *just-identified* case. When Z has lower dimension than X and $E(ZD^T)$ does not have full column rank, the

equation $E(ZY) = E(ZD^T)\beta$ has infinitely many solutions. This is the *under-identified* case in which the coefficient β cannot be uniquely determined even with Z . It is a challenging case beyond the scope of this book. In practice, we need at least as many IVs as the endogenous regressors.

When Z has higher dimension than D and $E(ZD^T)$ has full column rank, we have many ways to determine β from $E(ZY) = E(ZD^T)\beta$. What is more, the sample analog

$$n^{-1} \sum_{i=1}^n Z_i Y_i = n^{-1} \sum_{i=1}^n Z_i D_i^T \beta$$

may not have any solution because the number of equations is larger than the number of unknown parameters.

A computational trick in this case is the *two-stage least squares* (TSLS) estimator (Theil, 1953; Basman, 1957). It is a clever computational trick, which has two steps.

Definition 23.3 (Two-stage least squares) Define the TSLS estimator of the coefficient of D with Z being the IV as follows.

1. Run OLS of D on Z , and obtain the fitted value \hat{D}_i ($i = 1, \dots, n$). If D_i is a vector, then we need to run component-wise OLS to obtain \hat{D}_i . Put the fitted vectors in a matrix \hat{D} with rows \hat{D}_i^T ;
2. Run OLS of Y on \hat{D} , and obtain the coefficient $\hat{\beta}_{\text{TSLS}}$.

To see why TSLS works, we need more algebra. Write it more explicitly as

$$\begin{aligned} \hat{\beta}_{\text{TSLS}} &= \left(\sum_{i=1}^n \hat{D}_i \hat{D}_i^T \right)^{-1} \sum_{i=1}^n \hat{D}_i Y_i \\ &= \left(\sum_{i=1}^n \hat{D}_i \hat{D}_i^T \right)^{-1} \sum_{i=1}^n \hat{D}_i (D_i^T \beta + \varepsilon_i) \\ &= \left(\sum_{i=1}^n \hat{D}_i \hat{D}_i^T \right)^{-1} \sum_{i=1}^n \hat{D}_i D_i^T \beta + \left(\sum_{i=1}^n \hat{D}_i \hat{D}_i^T \right)^{-1} \sum_{i=1}^n \hat{D}_i \varepsilon_i. \end{aligned} \tag{23.5}$$

The first stage OLS fit ensures $D_i = \hat{D}_i + \check{D}_i$ with

$$\sum_{i=1}^n \hat{D}_i \check{D}_i^T = 0 \tag{23.6}$$

being a zero square matrix with the same dimension as D_i . The orthogonality (23.6) implies

$$\sum_{i=1}^n \hat{D}_i D_i^T = \sum_{i=1}^n \hat{D}_i \hat{D}_i^T,$$

which further implies that

$$\hat{\beta}_{\text{TSLS}} = \beta + \left(\sum_{i=1}^n \hat{D}_i \hat{D}_i^\top \right)^{-1} \sum_{i=1}^n \hat{D}_i \varepsilon_i. \quad (23.7)$$

The first stage OLS fit also ensures

$$\hat{D}_i = \hat{\Gamma}^\top Z_i \quad (23.8)$$

which implies that

$$\hat{\beta}_{\text{TSLS}} = \beta + \left\{ \hat{\Gamma}^\top \left(n^{-1} \sum_{i=1}^n Z_i Z_i^\top \right) \hat{\Gamma} \right\}^{-1} \hat{\Gamma}^\top \left(n^{-1} \sum_{i=1}^n Z_i \varepsilon_i \right). \quad (23.9)$$

Based on (23.9), we can see the consistency of the TSLS estimator because the term $n^{-1} \sum_{i=1}^n Z_i \varepsilon_i$ has probability limit $E(Z\varepsilon) = 0$. We can also use (23.9) to show that when Z and D have the same dimension, $\hat{\beta}_{\text{TSLS}}$ is numerically identical to $\hat{\beta}_{\text{IV}}$ defined in Section 23.4, which is left as Problem 23.1.

Based on (23.7), we can obtain the standard error as follows. We first obtain the residual $\hat{\varepsilon}_i = Y_i - \hat{\beta}_{\text{TSLS}}^\top \hat{D}_i$, and then obtain the robust variance estimator as

$$\hat{V}_{\text{TSLS}} = \left(\sum_{i=1}^n \hat{D}_i \hat{D}_i^\top \right)^{-1} \left(\sum_{i=1}^n \hat{\varepsilon}_i^2 \hat{D}_i \hat{D}_i^\top \right) \left(\sum_{i=1}^n \hat{D}_i \hat{D}_i^\top \right)^{-1}.$$

Importantly, the $\hat{\varepsilon}_i$'s are not the residual from the second stage OLS $Y_i - \hat{\beta}_{\text{TSLS}}^\top \hat{D}_i$, so \hat{V}_{TSLS} differs from the robust variance estimator from the second stage OLS.

23.6 A Special Case: A Single IV for a Single Endogenous Treatment

This section focuses on a simple case with a single IV and a single endogenous treatment. It has wide applications. Consider the following *structural equations*:

$$\begin{cases} Y_i = \beta_0 + \beta_1 D_i + \beta_2^\top X_i + \varepsilon_i, \\ D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2^\top X_i + \varepsilon_{2i}, \end{cases} \quad (23.10)$$

where D_i is a scalar endogenous regressor representing the treatment variable of interest (i.e., $E(\varepsilon_i D_i) \neq 0$), Z_i is a scalar IV for D_i (i.e., $E(\varepsilon_i Z_i) = 0$), and X_i contains other exogenous regressors (i.e., $E(\varepsilon_i X_i) = 0$). This is a special case with D replaced by $(1, D, X)$ and Z replaced by $(1, Z, X)$.

23.6.1 Two-stage least squares

The TSLS estimator in Definition 23.3 simplifies to the following form.

Definition 23.4 (TSLS with a single endogenous regressor) *Based on (23.10), the TSLS estimator has the following two steps:*

1. run OLS of D on $(1, Z, X)$, and obtain the fitted value \hat{D}_i ($i = 1, \dots, n$);
2. run OLS of Y on $(1, \hat{D}, X)$, and obtain the coefficient $\hat{\beta}_{\text{TSLS}}$, and in particular, $\hat{\beta}_{1,\text{TSLS}}$, the coefficient of \hat{D} .

23.6.2 Indirect least squares

The structural equation (23.10) implies

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(\gamma_0 + \gamma_1 Z_i + \gamma_2^T X_i + \varepsilon_{2i}) + \beta_2^T X_i + \varepsilon_i \\ &= (\beta_0 + \beta_1 \gamma_0) + \beta_1 \gamma_1 Z_i + (\beta_2 + \beta_1 \gamma_2)^T X_i + (\varepsilon_i + \beta_1 \varepsilon_{2i}). \end{aligned}$$

Define $\Gamma_0 = \beta_0 + \beta_1 \gamma_0$, $\Gamma_1 = \beta_1 \gamma_1$, $\Gamma_2 = \beta_2 + \beta_1 \gamma_2$, and $\varepsilon_{1i} = \varepsilon_i + \beta_1 \varepsilon_{2i}$. We have the following equations

$$\begin{cases} Y_i = \Gamma_0 + \Gamma_1 Z_i + \Gamma_2^T X_i + \varepsilon_{1i}, \\ D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2^T X_i + \varepsilon_{2i}, \end{cases} \quad (23.11)$$

which is called the *reduced form*. The parameter of interest equals the ratio of two coefficients

$$\beta_1 = \Gamma_1 / \gamma_1.$$

In the reduced form, the left-hand side are dependent variables Y and D , and the right-hand side are the exogenous variable Z and X satisfying

$$E(Z\varepsilon_{1i}) = E(Z\varepsilon_{2i}) = 0, \quad E(X\varepsilon_{1i}) = E(X\varepsilon_{2i}) = 0.$$

More importantly, OLS gives consistent estimators for the coefficients in the reduced form.

The reduced form (23.11) suggests that the ratio of two OLS coefficients $\hat{\Gamma}_1$ and $\hat{\gamma}_1$ is a reasonable estimator for β_1 . This is called the *indirect least squares* (ILS) estimator:

$$\hat{\beta}_{1,\text{ILS}} \equiv \hat{\Gamma}_1 / \hat{\gamma}_1.$$

Interestingly, it is numerically identical to the TSLS estimator under (23.10).

Theorem 23.1 *With a single endogenous treatment and a single IV, we have*

$$\hat{\beta}_{1,\text{ILS}} = \hat{\beta}_{1,\text{TSLS}}.$$

Theorem 23.1 is an algebraic fact. Imbens (2014, Section A.3) pointed it out without giving a proof. I relegate its proof to Problem 23.2. The ratio formula makes it clear that the TSLS estimator has poor finite sample properties with a weak instrument variable, i.e., γ_1 is close to zero.

23.6.3 Weak IV

The following inferential procedure is simpler, more transparent, and more robust to weak IV. It is more computationally intensive though. The reduced form (23.11) also implies that

$$Y_i - bD_i = (\Gamma_0 - b\gamma_0) + (\Gamma_1 - b\gamma_1)Z_i + (\Gamma_2 - b\gamma_2)^T X_i + (\varepsilon_{1i} - b\varepsilon_{2i}) \quad (23.12)$$

At the true value $b = \beta_1$, the coefficient of Z_i must be 0. This simple fact suggests a confidence interval for β_1 by inverting tests for $H_0(b) : \beta_1 = b$:

$$\{b : |t_Z(b)| \leq z_\alpha\},$$

where $t_Z(b)$ is the t -statistic for the coefficient of Z based on the OLS fit of (23.12) with the EHW standard error. This confidence interval is more robust than the Wald-type confidence interval based on the TSLS estimator. It is similar to the Fieller–Anderson–Rubin confidence interval discussed in Chapter 21. This procedure makes the TSLS estimator unnecessary, and what is more, we only need to run the OLS fit of Y based on the reduced form if the goal is to test $\beta_1 = 0$ under (23.10).

23.7 Application

Card (1993) used the National Longitudinal Survey of Young Men to estimate the causal effect of education on earnings. The data set contains 3010 men with age between 14 and 24 in the year 1966, and Card (1993) leveraged the geographic variation in college proximity as an IV for education. Here, Z is the indicator of growing up near a four-year college, D measures the years of education, and the outcome Y is the log wage in the year 1976, ranging from 4.6 to 7.8. Additional covariates are ace , age and squared age, a categorical variable indicating living with both parents, single mom, or both parents, and variables summarizing the living areas in the past.

```
> library("car")
>
> ## Card Data
> card.data = read.csv("card1995.csv")
> Y = card.data[, "lwage"]
> D = card.data[, "educ"]
> Z = card.data[, "nearc4"]
> X = card.data[, c("exper", "expersq", "black", "south",
+                   "smsa", "reg661", "reg662", "reg663",
+                   "reg664", "reg665", "reg666",
+                   "reg667", "reg668", "smsa66")]
> X = as.matrix(X)
```

Based on TSLS, the point estimator is 0.132 and the 95% confidence interval is [0.026, 0.237].

```
> Dhat      = lm(D ~ Z + X)$fitted.values
> tslsreg  = lm(Y ~ Dhat + X)
> tslsest  = coef(tslsreg)[2]
> ## correct se by changing the residuals
> res.correct = Y - cbind(1, D, X) %*% coef(tslsreg)
> tslsreg$residuals = as.vector(res.correct)
> tslsse = sqrt(hccm(tslsreg, type = "hc0"))[2, 2]
> res = c(tslsest, tslsest - 1.96*tslsse, tslsest + 1.96*tslsse)
> names(res) = c("est", "l.ci", "u.ci")
> round(res, 3)
  est l.ci u.ci
0.132 0.026 0.237
```

Figure 23.3 shows the p -values for a sequence of tests for the coefficient of D . It also implies the 95% confidence interval for the coefficient of D based on inverting tests, which is [0.028, 0.282].

```
> BetaAR    = seq(-0.1, 0.4, 0.001)
> PvalueAR = sapply(BetaAR,
+                     function(b){
+                         Y_b     = Y - b*D
+                         ARreg   = lm(Y_b ~ Z + X)
+                         coefZ   = coef(ARreg)[2]
+                         seZ     = sqrt(hccm(ARreg)[2, 2])
+                         Tstat   = coefZ/seZ
+                         (1 - pnorm(abs(Tstat)))*2
+                     })
> point.est = BetaAR[which.max(PvalueAR)]
> point.est
[1] 0.132
> ARCI = range(BetaAR[PvalueAR >= 0.05])
> ARCI
[1] 0.028 0.282
```

Comparing the above two methods, the lower confidence limits are very close but the upper confidence limits are slightly different due to the possibly heavy right tail of the distribution of the TSLS estimator.

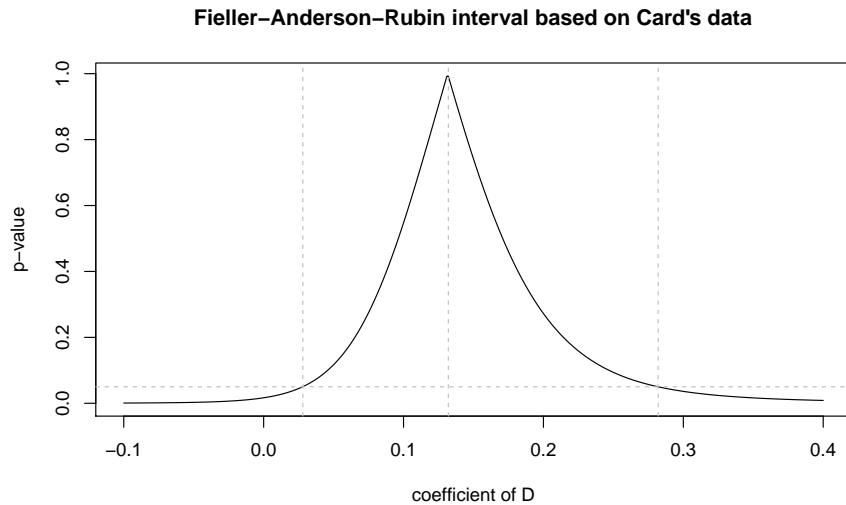


FIGURE 23.3: Re-analysis of Card (1993)'s data

23.8 Homework

23.1 More algebra for TSLS in Section 23.5

1. Show that the $\hat{\Gamma}$ in (23.8) equals

$$\hat{\Gamma} = \left(\sum_{i=1}^n Z_i Z_i^\top \right)^{-1} \sum_{i=1}^n Z_i D_i^\top.$$

2. Show $\hat{\beta}_{\text{TSLS}}$ defined in (23.5) reduces to $\hat{\beta}_{\text{IV}}$ defined in (23.4) if Z and D have the same dimension and

$$n^{-1} \sum_{i=1}^n Z_i Z_i^\top, \quad n^{-1} \sum_{i=1}^n Z_i D_i^\top$$

are both invertible.

23.2 Equivalence between TSLS and ILS

Prove Theorem 23.1.

Hint: Use the Frisch–Waugh–Lovell theorem.

23.3 Control function in the linear instrumental variable model

Definition 23.5 below parallels Definition 23.3 above.

Definition 23.5 (control function) Define the control function estimator $\hat{\beta}_{\text{CF}}$ as follows.

1. Run OLS of D on Z , and obtain the residual \check{D}_i ($i = 1, \dots, n$). If D_i is a vector, then we need to run component-wise OLS to obtain \check{D}_i . Put the residual vectors in a matrix \check{D} with rows \check{D}_i^T ;
2. Run OLS of Y on D and \check{D} , and obtain the coefficient of D , $\hat{\beta}_{\text{CF}}$.

Show that $\hat{\beta}_{\text{CF}} = \hat{\beta}_{\text{TSLS}}$.

Remark: In Definition 23.5, \check{D} from Step 1 is called the control function for Step 2. Hausman (1978) pointed out this result. Wooldridge (2015) provided more general discussion of the control function methods in more complex models.

Hint: Use the results in Problems A2.3 and A2.4.

23.4 Data analysis: Efron and Feldman (1991)

Efron and Feldman (1991) was one of the early studies dealing with noncompliance under the potential outcomes framework. The original randomized experiment, the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT), was designed to evaluate the effect of the drug cholestyramine on cholesterol levels. In the dataset `EF.csv`, the first column contains the binary indicators for treatment and control, the second column contains the proportions of the nominal cholestyramine dose actually taken, the last three columns are cholesterol levels. Note that the individuals did not know whether they were assigned to cholestyramine or to the placebo, but differences in adverse side effects could induce differences in compliance behavior by treatment status. All individuals were assigned the same nominal dose of the drug or placebo, for the same time period. Column 3, C_3 , was taken prior to a communication about the benefits of a low-cholesterol diet, Column 4, C_4 , was taken after this suggestion, but prior to the random assignment to cholestyramine or placebo, and Column 5, C_5 , an average of post-randomization cholesterol readings, averaged over two-month readings for a period of time averaging 7.3 years for all the individuals in the study. Efron and Feldman (1991) used the change in cholesterol level as the final outcome of interest, defined as $C_5 - 0.25C_3 - 0.75C_4$. The original paper contains more detailed descriptions.

This dataset is more complicated than the noncompliance problem discussed in class. You can analyze it based on your understanding of the problem, but you need to justify your choice of method. There is no gold-standard solution for this problem.

23.5 Recommended reading

Imbens (2014) gave an econometrician's perspective of IV.



24

Application of the Instrumental Variable Method: Fuzzy Regression Discontinuity

The regression discontinuity introduced in Chapter 20 and the instrumental variable introduced in Chapters 21–23 are two important examples of *natural experiments*. The study designs are not as ideal as the randomized experiments in Part II, but they have features similar to the experiments. That's why they are called natural experiments.

Compounding regression discontinuity with instrumental variable yields the *fuzzy regression discontinuity*, another important natural experiment. I will start with examples and then provide a mathematical formulation.

24.1 Motivating examples

Chapter 20 introduces the regression discontinuity. The following two examples are slightly different because the treatments received are not deterministic functions of the running variables. Rather, the running variables discontinuously change the probabilities of the treatments received at the cutoff point.

Example 24.1 *In 2000, the Government of India launched the Prime Minister's Village Road Program, and by 2015, this program had funded the construction of all-weather roads to nearly 200,000 villages. Based on village level data, Asher and Novosad (2020) use a regression discontinuity to estimate the effect of new feeder roads on various economic variables. The national program guidelines prioritized larger villages according to arbitrary thresholds based on the 2001 Population Census. The treatment variable equals one if the village received a new road before the year in which the outcomes were measured. The difference between the population size of a village and the threshold did not determine the treatment variable but affected its probability discontinuously at the cutoff point zero.*

Example 24.2 *Li et al. (2015) used the data on the first-year students enrolled in 2004 to 2006 from two Italian universities to evaluate the causal effect of a university grant on the drop out rate. The students were eligible for this grant if their standardized family income was below 15,000 euros. For simplicity, we use the running variable defined as 15,000 minus the standardized*

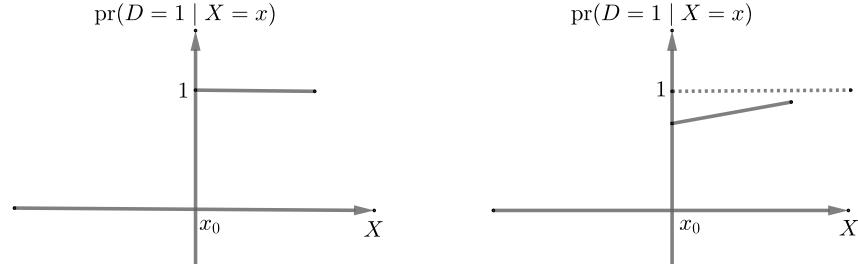


FIGURE 24.1: The treatment assignments of sharp regression discontinuity (left) and fuzzy regression discontinuity (right)

family income. To receive this grant, the students must apply first. Therefore, the eligibility and the application status jointly determined the final treatment status. The running variable alone did not determine the treatment status although it changed the treatment probability at the cutoff point zero.

Example 24.3 Amarante et al. (2016) estimated the impact of *in utero* exposure to a social assistance program on children's birth outcomes. They used a regression discontinuity induced by the Uruguayan Plan de Atención Nacional a la Emergencia Social. It was a temporary social assistance program targeted to the poorest 10 percent of households, implemented between April 2005 and December 2007. Households with a predicted low income score below a predetermined threshold were assigned to the program. The predicted income score did not determine whether the mother received at least one program transfer during the pregnancy but it changed the probability of the final treatment received. The birth outcomes included birth weight, weeks of gestation, etc.

The above examples are called fuzzy regression discontinuity in contrast to the (sharp) regression discontinuity in Chapter 20. I will analyze the data in Examples 24.1 and 24.2 in Section 24.3 below.

24.2 Mathematical formulation

Let X_i denote the running variable which determines $Z_i = 1(X_i \geq x_0)$ with the cutoff point x_0 . The treatment received D_i may not equal Z_i , but $\text{pr}(D_i = 1 | X_i = x)$ has a jump at x_0 . Figure 24.1 compares the treatment received probabilities of the sharp regression discontinuity and fuzzy regression discontinuity. It shows a special case of fuzzy regression discontinuity with $\text{pr}(D = 1 | X < x_0) = 0$, which is coherent to Example 24.2.

Let Y_i denote the outcome of interest. Viewing Z_i as the treatment as-

signed, we can define potential outcomes $\{D_i(1), D_i(0), Y_i(1), Y_i(0)\}$. The sharp regression discontinuity of Z allows for identification of

$$\begin{aligned}\tau_D(x_0) &= E\{D(1) - D(0) \mid X = x_0\} \\ &= \lim_{\varepsilon \rightarrow 0+} E(D \mid Z = 1, X = x_0 + \varepsilon) - \lim_{\varepsilon \rightarrow 0+} E(D \mid Z = 0, X = x_0 - \varepsilon)\end{aligned}$$

and

$$\begin{aligned}\tau_Y(x_0) &= E\{Y(1) - Y(0) \mid X = x_0\} \\ &= \lim_{\varepsilon \rightarrow 0+} E(Y \mid Z = 1, X = x_0 + \varepsilon) - \lim_{\varepsilon \rightarrow 0+} E(Y \mid Z = 0, X = x_0 - \varepsilon)\end{aligned}$$

based on Theorem 20.2. Using Z as an IV for D and imposing the IV assumptions at $X = x_0$, we can identify the local complier average causal effect by applying Theorem 21.1.

Theorem 24.1 *Assume*

$$D_i(1) \geq D_i(0)$$

and

$$D_i(1) = D_i(0) \implies Y_i(1) = Y_i(0)$$

in the infinitesimal neighborhood of x_0 . The local complier average causal effect equals

$$\begin{aligned}\tau_c(x_0) &\equiv E\{Y(1) - Y(0) \mid D(1) > D(0), X = x_0\} \\ &= \frac{E\{Y(1) - Y(0) \mid X = x_0\}}{E\{D(1) - D(0) \mid X = x_0\}}.\end{aligned}$$

Further assume that $E\{D(1) \mid X = x\}$ and $E\{Y(1) \mid X = x\}$ are continuous from the right at $X = x_0$, and $E\{D(0) \mid X = x\}$ and $E\{Y(0) \mid X = x\}$ are continuous from the left at $X = x_0$. The local complier average causal effect can be identified by

$$\tau_c(x_0) = \frac{\lim_{\varepsilon \rightarrow 0+} E(Y \mid Z = 1, X = x_0 + \varepsilon) - \lim_{\varepsilon \rightarrow 0+} E(Y \mid Z = 0, X = x_0 - \varepsilon)}{\lim_{\varepsilon \rightarrow 0+} E(D \mid Z = 1, X = x_0 + \varepsilon) - \lim_{\varepsilon \rightarrow 0+} E(D \mid Z = 0, X = x_0 - \varepsilon)}$$

if the $E(D \mid Z = 1, X = x)$ has a non-zero jump at $X = x_0$.

Theorem 24.1 is a superposition of Theorems 20.2 and 21.1. I leave its proof as Problem 24.1.

In both sharp and fuzzy regression discontinuity, the key is to specify the neighborhood around the cutoff point. Practically, a smaller neighborhood leads to smaller bias but larger variance, while a larger neighborhood leads to larger bias but smaller variance. That is, we face a bias-variance trade-off. Some automatic procedures exist based on some statistical criteria, which relies on some strong conditions. It seems wiser to conduct sensitivity analysis over a range of the choice of h .

Assume that we have specified the neighborhood of x_0 determined by a bandwidth h . For data with $X_i \in [x_0 - h, x_0 + h]$, we can estimate $\tau_D(x_0)$ by

$\hat{\tau}_D(x_0)$ = the coefficient of Z_i in the OLS fit of D_i on $\{1, Z_i, R_i, L_i\}$,

and estimate $\tau_Y(x_0)$

$\hat{\tau}_Y(x_0)$ = the coefficient of Z_i in the OLS fit of Y_i on $\{1, Z_i, R_i, L_i\}$,

recalling the definitions $R_i = \max(X_i - x_0, 0)$ and $L_i = \min(X_i - x_0, 0)$. Then we can estimate the local complier average causal effect by

$$\hat{\tau}_c(x_0) = \hat{\tau}_Y(x_0)/\hat{\tau}_D(x_0).$$

This is an indirect least squares estimator. By Theorem 23.1, it is numerically identical to

the coefficient of D_i in the TSLS fit of Y_i on $\{1, D_i, R_i, L_i\}$

with D_i instrumented by Z_i . In sum, after specifying h , the estimation of $\tau_c(x_0)$ reduces to a TSLS procedure with the local data around the cutoff point.

24.3 Application

24.3.1 Re-analyzing Asher and Novosad (2020)'s data

Figure 24.2 shows the result using `occupation_index_andrsn` as the outcome.

The package `rdrobust` selects the bandwidth automatically. The results suggest that receiving a new road did not affect the outcome significantly.

```
> road_dat = read.csv("indianroad.csv")
> road_dat$runv = road_dat$left + road_dat$right
> library("rdrobust")
> frd_road = with(road_dat,
+ {
+   rdrobust(y = occupation_index_andrsn,
+             x = runv,
+             c = 0,
+             fuzzy = r2012)
+ })
> res = cbind(frd_road$coef, frd_road$se)
> round(res, 3)
      Coeff Std. Err.
Conventional -0.253    0.301
Bias-Corrected -0.283    0.301
Robust        -0.283    0.359
```

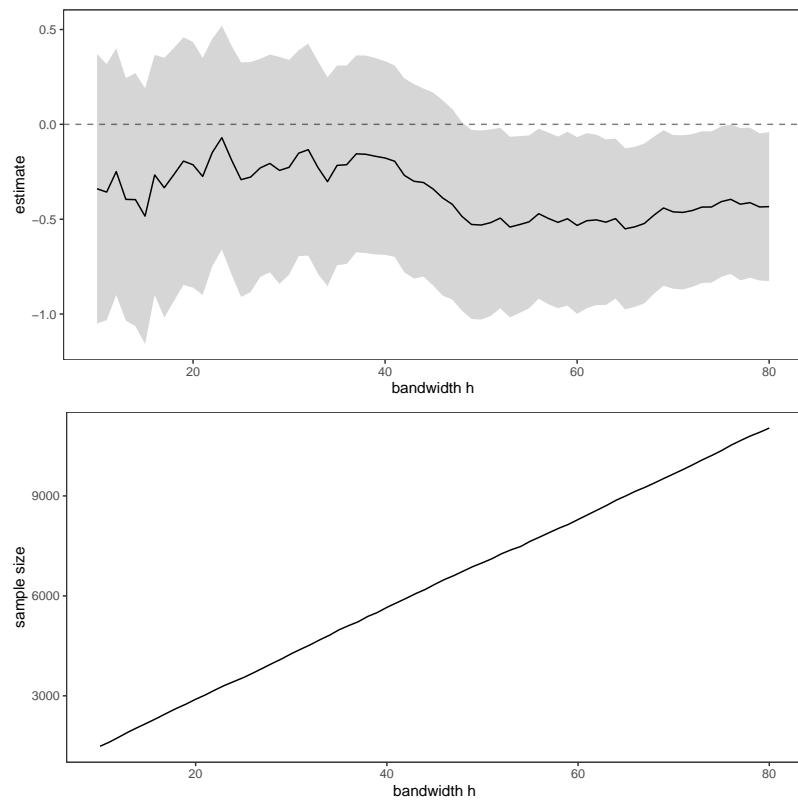


FIGURE 24.2: Re-analyzing Asher and Novosad (2020)'s data, with point estimates and standard errors from TSLS.

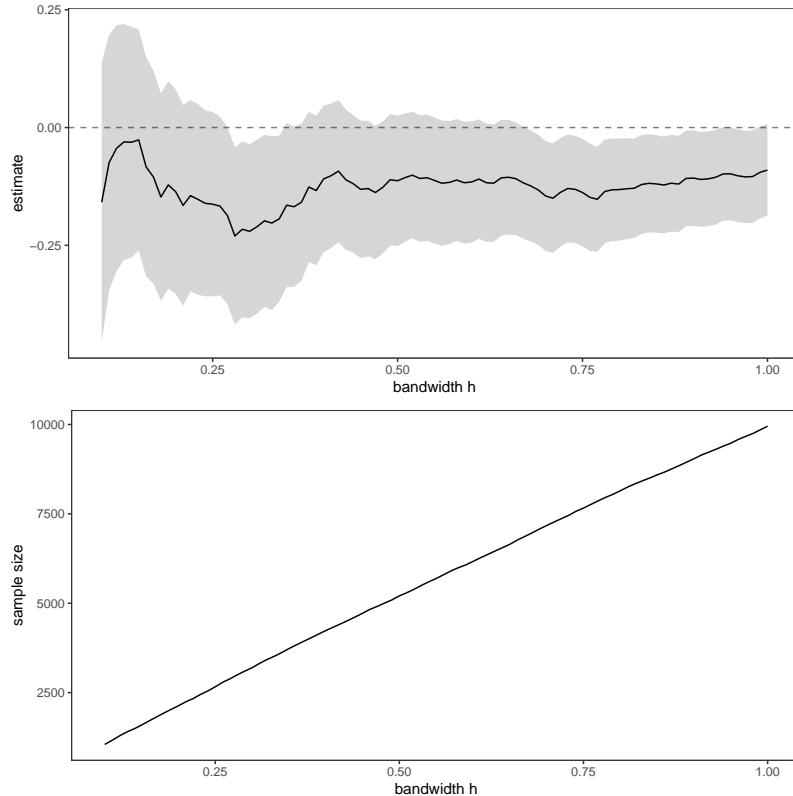


FIGURE 24.3: Re-analyzing Li et al. (2015)'s data, with point estimates and standard errors from TSLS.

24.3.2 Re-analyzing Li et al. (2015)'s data

Recall that the running variable is 15,000 minus the standardized income in Example 24.2. In the analysis, I restrict the data to a subset with this running between $[-5,000, 5,000]$, and then divide the running variable by 5,000 so that the running variable is bounded between $[-1, 1]$ at cutoff point zero.

The results based on the package `rdrobust` suggest that the university grant did not affect the dropout rate significantly.

```
> italy = read.csv("italy.csv")
> library("rdrobust")
> frd_italy = with(italy,
+                   {
+                     rdrobust(y = outcome,
+                               x = rv0,
+                               c = 0,
+                               fuzzy = D)
```

```

+           })
> res = cbind(frd_italy$coef, frd_italy$se)
> round(res, 3)
      Coeff Std. Err.
Conventional -0.149   0.101
Bias-Corrected -0.155   0.101
Robust       -0.155   0.121

```

24.4 Discussion

Both Chapter 20 and this chapter formulate regression discontinuity based on the continuity of the conditional expectations of the potential outcomes given the running variables. This perspective is mathematically simpler but it only identifies the local effects precisely at the cutoff point of the running variable. Hahn et al. (2001) started this line of literature.

An alternative, not so dominant perspective is based on *local randomization* (Cattaneo et al., 2015; Li et al., 2015). If we view the running variable as a noisy measure of some underlying truth and the cutoff point is somewhat arbitrarily chosen, the units near the cutoff point do not differ systematically. This suggests that in a small neighborhood of the cutoff point, the units receive the treatment and the control in a random fashion just as in a randomized experiment. Similar to the issue of choosing h in the first perspective, it is crucial to decide how local should the randomized experiment be under the regression discontinuity. It is not easy to quantify the intuition mathematically, and again conducting sensitivity analysis with a range of h seems a reasonable approach in the second perspective as well.

See Sekhon and Titunik (2017) for more conceptual discussion of regression discontinuity.

24.5 Homework Problems

24.1 Proof of Theorem 24.1

Prove Theorem 24.1.

24.2 Data analysis

Section 24.3.1 estimated the effect on `occupation_index_andrsn`. Four other outcome variables are `transport_index_andrsn`, `firms_index_andrsn`,

30024 Application of the Instrumental Variable Method: Fuzzy Regression Discontinuity

`consumption_index_andrsn`, and `agriculture_index_andrsn`, with meanings defined in the original paper. Estimate the effects on these outcomes.

24.3 Reflection on the analysis of Li et al. (2015)'s data

In Li et al. (2015), a key variable determining the treatment status is the binary application status A , which has potential outcomes $A(1)$ and $A(0)$ corresponding to the treatment $Z = 1$ and control $Z = 0$. By definition,

$$D(1) = A(1), \quad D(0) = 0,$$

so the compliers $\{D(1), D(0)\} = (1, 0)$ is equivalent to $A(1) = 1$. So

$$\tau_c(x_0) = E\{Y(1) - Y(0) \mid A(1) = 1, X = x_0\}.$$

Section 24.3.2 used the whole data set to estimate $\tau_c(x_0)$.

An alternative analysis is based on units with $A = 1$ only. Then the treatment status is determined by X . However, this analysis can be problematic because

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0+} E\{Y \mid A = 1, X = x_0 + \varepsilon\} - \lim_{\varepsilon \rightarrow 0+} E\{Y \mid A = 1, X = x_0 - \varepsilon\} \\ &= E\{Y(1) \mid A(1) = 1, X = x_0\} - E\{Y(0) \mid A(0) = 1, X = x_0\}. \end{aligned} \quad (24.1)$$

Prove (24.1) and explain why this analysis can be problematic.

Remark: The left-hand side of (24.1) is the identification formula of the local average treatment effect at $X = x_0$, conditioning on $A = 1$. The right-hand side of (24.1) is the difference in means of the potential outcomes for subgroup of units with $(A(1) = 1, X = x_0)$ and $(A(0) = 1, X = x_0)$, respectively.

24.4 Recommended reading

Imbens and Lemieux (2008) gave a practical guidance to regression discontinuity based on the potential outcomes framework. Lee and Lemieux (2010) reviewed regression discontinuity and its applications in economics.

25

Application of the Instrumental Variable Method: Mendelian Randomization

Katan (1986) was concerned with the observational studies suggesting that low serum cholesterol levels were associated with the risk of cancer. As we have discussed, however, observational studies suffer from unmeasured confounding. Consequently, it is difficult to interpret the apparent association as causality. In the particular problem studied by Katan (1986), it is even possible that early stages of cancer reversely cause low serum cholesterol levels. Disentangling the causal effect of the serum cholesterol level on cancer seems a hard problem using standard epidemiologic studies. Katan (1986) argued that Apolipoprotein E genes are associated with the serum cholesterol levels but do not directly affect the cancer status. So if low serum cholesterol levels causes cancer, we should observe differences in cancer risks among people with and without the genotype that leads to different serum cholesterol levels. Using our language for causal inference, Katan (1986) proposed to use Apolipoprotein E genes as IVs.

Katan (1986) did not conduct any data analysis but just proposed a conceptual design that could address not only *unmeasured confounding* but also *reverse causality*. Since then, more complicated and sophisticated studies have been conducted thanks to the modern genome-wide association studies. These studies used genetic information as IVs for exposures in epidemiologic studies to estimate causal effects of exposures on outcomes. They were all motivated by *Mendel's second law*, the law of *random assortment*, which suggests the inheritance of one trait is independent of the inheritance of other traits. Therefore, the method of using genetic information as IV is called *Mendelian Randomization* (MR).

25.1 Background and motivation

Graphically, Figure 25.1 shows the causal diagram on the treatment D , outcome Y , unmeasured confounder U , as well as the genetic IVs G_1, \dots, G_p . In many Mendelian Randomization studies, the genetic IVs are single nucleotide polymorphisms (SNPs). Because of pleiotropy, it is possible that the genetic

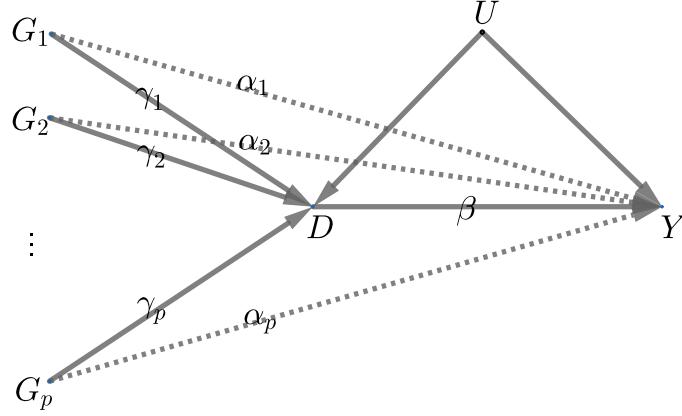


FIGURE 25.1: Causal graph for Mendelian randomization

IVs have direct effect on the outcome of interest, so Figure 25.1 also allows for the violation of the exclusion restriction assumption.

The standard linear IV model assumes away the direct effect of the IVs on the outcome. Definition 25.1 below gives both the structural and reduced forms.

Definition 25.1 (linear IV model) *The standard linear IV model*

$$Y = \beta_0 + \beta D + \beta_u U + \varepsilon_Y, \quad (25.1)$$

$$D = \gamma_0 + \gamma_1 G_1 + \cdots + \gamma_p G_p + \gamma_u U + \varepsilon_D, \quad (25.2)$$

has reduced form

$$Y = \beta_0 + \beta \gamma_0 + \beta \gamma_1 G_1 + \cdots + \beta \gamma_p G_p + (\beta_u + \beta_0 \gamma_u) U + \varepsilon_Y, \quad (25.3)$$

$$D = \gamma_0 + \gamma_1 G_1 + \cdots + \gamma_p G_p + \gamma_u U + \varepsilon_D, \quad (25.4)$$

Definition 25.2 below allows for the violation of exclusion restriction. Then, \$G_1, \dots, G_p\$ are not valid IVs.

Definition 25.2 (linear model with possibly invalid IVs) *The linear model*

$$Y = \beta_0 + \beta D + \alpha_1 G_1 + \cdots + \alpha_p G_p + \beta_u U + \varepsilon_Y, \quad (25.5)$$

$$D = \gamma_0 + \gamma_1 G_1 + \cdots + \gamma_p G_p + \gamma_u U + \varepsilon_D, \quad (25.6)$$

has reduced form

$$Y = (\beta_0 + \beta \gamma_0) + (\alpha_1 + \beta \gamma_1) G_1 + \cdots + (\alpha_p + \beta \gamma_p) G_p + (\beta_u + \beta \gamma_u) U + \varepsilon_Y, \quad (25.7)$$

$$D = \gamma_0 + \gamma_1 G_1 + \cdots + \gamma_p G_p + \gamma_u U + \varepsilon_D. \quad (25.8)$$

Therefore, in Definition 25.1 with exclusion restriction, we have

$$\Gamma_j = \beta\gamma_j, \quad (j = 1, \dots, p);$$

in Definition 25.2 without exclusion restriction, we have

$$\Gamma_j = \alpha_j + \beta\gamma_j, \quad (j = 1, \dots, p).$$

If we have individual data, we can apply the classic TSLS estimator to estimate β under the linear IV model in Definition 25.1. However, most Mendelian Randomization studies do not have individual data but rather summary statistics from multiple genome-wide association studies. A canonical setting consists of the regression coefficients of the treatment on the genetic IVs:

$$\hat{\gamma}_1 \rightarrow \gamma_1, \dots, \hat{\gamma}_p \rightarrow \gamma_p \tag{25.9}$$

in probability with standard errors

$$\text{se}_{D1}, \dots, \text{se}_{Dp}, \tag{25.10}$$

and the regression coefficients of the outcome on the genetic IVs:

$$\hat{\Gamma}_1 \rightarrow \Gamma_1, \dots, \hat{\Gamma}_p \rightarrow \Gamma_p \tag{25.11}$$

in probability with standard errors

$$\text{se}_{Y1}, \dots, \text{se}_{Yp}. \tag{25.12}$$

I will focus on the statistical inference of β based on the above summary statistics. For simplicity, we assume that the estimates in (25.9) and (25.11) are jointly independent, they are all asymptotically normal, and the standard errors in (25.10) and (25.12) are all fixed and known. The asymptotic normality can often be justified by central limit theorems of the regression coefficients. The standard errors are accurate estimates of the true standard errors. Therefore, the only subtle assumption is the joint independence of the $\hat{\gamma}_j$'s and the $\hat{\Gamma}_j$'s are reasonable because they are often calculated based on different samples. The independence among the $\hat{\gamma}_j$'s can be reasonable if the G_j 's are independent and the true linear model for D holds with homoskedastic error terms¹. The independence among the $\hat{\Gamma}_j$'s follows from a similar argument.

25.2 MR based on summary statistics

25.2.1 Fixed-effect estimator

Under Definition 25.1, $\alpha_j = 0$ which implies that $\beta = \Gamma_j/\gamma_j$ for all j . A simple approach is based on the so-called *meta-analysis* (Bowden et al., 2018), that is,

¹This can be tricky if the error term of the linear model is heteroskedastic. Without the independence of the G_j 's, it is hard to justify the independence.

30425 Application of the Instrumental Variable Method: Mendelian Randomization

combining multiple estimates $\hat{\beta}_j = \hat{\Gamma}_j/\hat{\gamma}_j$ for the common parameter β . Using delta method (see Example A1.3), $\hat{\beta}_j$ has approximate squared standard error

$$\text{se}_j^2 = (\text{se}_{Yj}^2 + \hat{\beta}_j^2 \text{se}_{Dj}^2)/\hat{\gamma}_j^2.$$

Therefore, the best linear combination to estimate β is the Fisher weighting based on inverse of the variances:

$$\hat{\beta}_{\text{fisher0}} = \frac{\sum_{j=1}^p \hat{\beta}_j / \text{se}_j^2}{\sum_{j=1}^p 1 / \text{se}_j^2}$$

which has variance $(\sum_{j=1}^p 1 / \text{se}_j^2)^{-1}$. Ignoring the uncertainty due to $\hat{\gamma}_j$ quantified by se_{Dj} , the estimator reduces to

$$\hat{\beta}_{\text{fisher1}} = \frac{\sum_{j=1}^p \hat{\beta}_j \hat{\gamma}_j^2 / \text{se}_{Yj}^2}{\sum_{j=1}^p \hat{\gamma}_j^2 / \text{se}_{Yj}^2} = \frac{\sum_{j=1}^p \hat{\Gamma}_j \hat{\gamma}_j / \text{se}_{Yj}^2}{\sum_{j=1}^p \hat{\gamma}_j^2 / \text{se}_{Yj}^2},$$

which has variance $(\sum_{j=1}^p 1 \hat{\gamma}_j^2 / \text{se}_{Yj}^2)^{-1}$. Inference based on $\hat{\beta}_{\text{fisher1}}$ is suboptimal although it is more widely used in practice (Bowden et al., 2018).

Focus on the suboptimal yet simpler estimator $\hat{\beta}_{\text{fisher1}}$. Under Definition 25.2, we can show that

$$\hat{\beta}_{\text{fisher1}} \rightarrow \frac{\sum_{j=1}^p \Gamma_j \gamma_j / \text{se}_{Yj}^2}{\sum_{j=1}^p \gamma_j^2 / \text{se}_{Yj}^2} = \beta + \frac{\sum_{j=1}^p \alpha_j \gamma_j / \text{se}_{Yj}^2}{\sum_{j=1}^p \gamma_j^2 / \text{se}_{Yj}^2}$$

in probability. If $\alpha_j = 0$ for all j , $\hat{\beta}_{\text{fisher1}}$ is consistent. Even this does not hold, it is still possible that $\hat{\beta}_{\text{fisher1}}$ is consistent as long as the inner product between α_j and γ_j weighted by $1/\text{se}_{Yj}^2$ is zero. This holds if we have many genetic instruments and violation of the exclusion restriction, captured by α_j , is an independent random draw from a distribution with mean zero.

25.2.2 Egger regression

Start with Definition 25.1. With the true parameters, we have

$$\Gamma_j = \beta \gamma_j \quad (j = 1, \dots, p);$$

with the estimates, the above identify holds only approximately

$$\hat{\Gamma}_j \approx \beta \hat{\gamma}_j \quad (j = 1, \dots, p).$$

This seems a classic OLS problem of $\{\hat{\Gamma}_j\}_{j=1}^p$ on $\{\hat{\gamma}_j\}_{j=1}^p$. We can fit an OLS of $\hat{\Gamma}_j$ on $\hat{\gamma}_j$, with or without an intercept, possibly weighted by w_j , to estimate β . The following results hold thanks to the algebraic properties of the WLS reviewed in Section A2.5.

Without an intercept, the coefficient of $\hat{\gamma}_j$ is

$$\hat{\beta}_{\text{egger1}} = \frac{\sum_{j=1}^p \hat{\gamma}_j \hat{\Gamma}_j w_j}{\sum_{j=1}^p \hat{\gamma}_j^2 w_j},$$

which reduces to $\hat{\beta}_{\text{fisher1}}$ if $w_j = 1/\text{se}_{Yj}^2$. So the Egger regression is more general than the fixed-effect estimator in Section 25.2.1.

With an intercept, the coefficient of $\hat{\gamma}_j$ is

$$\hat{\beta}_{\text{egger0}} = \frac{\sum_{j=1}^p (\hat{\gamma}_j - \hat{\gamma}_w)(\hat{\Gamma}_j - \hat{\Gamma}_w)w_j}{\sum_{j=1}^p (\hat{\gamma}_j - \hat{\gamma}_w)^2 w_j}$$

where $\hat{\gamma}_w = \sum_{j=1}^p \hat{\gamma}_j w_j / \sum_{j=1}^p w_j$ and $\hat{\Gamma}_w = \sum_{j=1}^p \hat{\Gamma}_j w_j / \sum_{j=1}^p w_j$ are the weighted averages of the $\hat{\gamma}_j$'s and $\hat{\Gamma}_j$'s, respectively. Even without assuming that all γ_j 's are zero under Definition 25.2, we have

$$\hat{\beta}_{\text{egger0}} \rightarrow \frac{\sum_{j=1}^p (\gamma_j - \gamma_w)(\Gamma_j - \Gamma_w)w_j}{\sum_{j=1}^p (\gamma_j - \gamma_w)^2 w_j} = \beta + \frac{\sum_{j=1}^p (\gamma_j - \gamma_w)(\alpha_j - \alpha_w)w_j}{\sum_{j=1}^p (\gamma_j - \gamma_w)^2 w_j}$$

in probability, where γ_w, Γ_w and α_w are the corresponding weighted averages of the true parameters. So $\hat{\beta}_{\text{egger0}}$ is consistent for β as long as the weighted least squares coefficient of α_j on γ_j is zero. This is weaker than $\alpha_j = 0$ for all j . This weaker assumption holds if γ_j and α_j are realizations of independent random variables, which is called the Instrument Strength Independent of Direct Effect assumption (Bowden et al., 2015). More interestingly, the intercept from the Egger regression is

$$\hat{\alpha}_{\text{egger0}} = \hat{\Gamma}_w - \hat{\beta}_{\text{egger0}} \hat{\gamma}_w,$$

which, under the InSIDE assumption converges to

$$\Gamma_w - \beta \gamma_w = \alpha_w$$

in probability. So the intercept estimates the weighted average of the direct effects.

25.3 An example

I use the `bmi.sbp` data in the `mr.raps` package to illustrate the Egger regressions.

```
> library("mr.raps")
> bmisbp = subset(bmi.sbp,
```

30625 Application of the Instrumental Variable Method: Mendelian Randomization

```
+           select = c("beta.exposure",
+                      "beta.outcome",
+                      "se.exposure",
+                      "se.outcome"))
```

The Egger regressions with or without the intercept give very similar results.

```
> mr.egger = lm(beta.outcome ~ 0 + beta.exposure,
+                 data = bmisbp,
+                 weights = 1/se.outcome^2)
> summary(mr.egger)

Call:
lm(formula = beta.outcome ~ 0 + beta.exposure, data = bmisbp,
    weights = 1/se.outcome^2)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-5.6999 -1.1691 -0.0199  1.0073 11.3449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
beta.exposure  0.3173     0.1106   2.869  0.00468 **

Residual standard error: 2.052 on 159 degrees of freedom
Multiple R-squared:  0.04921,   Adjusted R-squared:  0.04323
F-statistic: 8.229 on 1 and 159 DF,  p-value: 0.004682

>
> mr.egger.w = lm(beta.outcome ~ beta.exposure,
+                   data = bmisbp,
+                   weights = 1/se.outcome^2)
> summary(mr.egger.w)

Call:
lm(formula = beta.outcome ~ beta.exposure, data = bmisbp, weights = 1/se.outcome^2)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-5.7099 -1.1774 -0.0296  0.9969 11.3393

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0001133  0.0020794   0.055  0.95660
beta.exposure 0.3172989  0.1109485   2.860  0.00481 **

Residual standard error: 2.059 on 158 degrees of freedom
Multiple R-squared:  0.04922,   Adjusted R-squared:  0.0432
F-statistic: 8.179 on 1 and 158 DF,  p-value: 0.004811
```

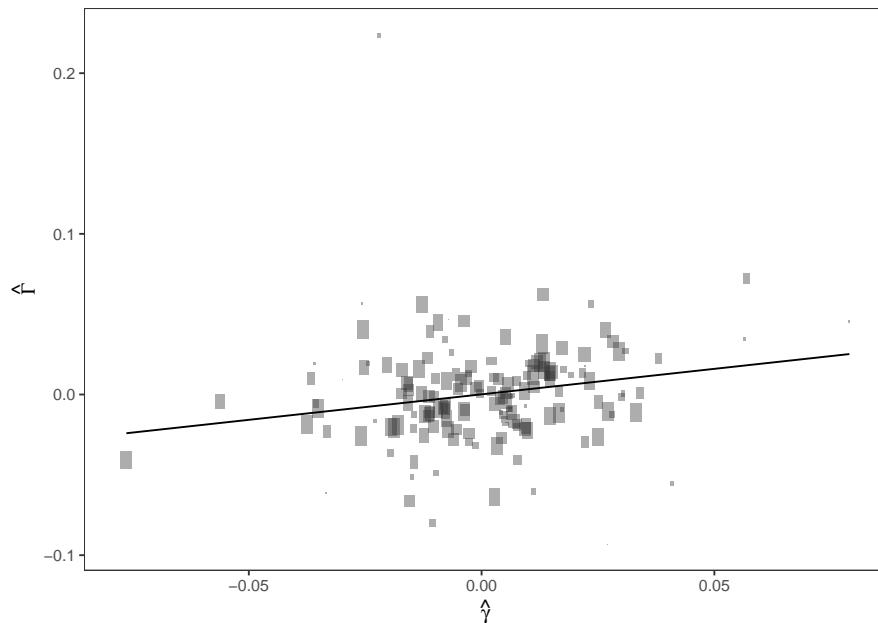


FIGURE 25.2: Scatter plot proportional to the inverse of the variance, with the Egger regression line

Figure 25.2 shows the raw data as well as the fitted Egger regression line.

25.4 Critiques of the analysis based on Mendelian randomization

MR is an application of the idea of IV. It relies on strong assumptions. I provide three sets of critiques from the conceptual, biological and technical perspectives.

Conceptually, most studies based on MR have ill-defined treatments from the potential outcomes perspective. For instance, the treatments are often defined as the cholesterol level or body mass index. They are composite variables and can correspond to complex, non-unique definitions of the hypothetical experiments. The SUTVA often does not hold for these treatments.

Biologically, the fundamental assumptions for the IV analysis may not hold. Mendel's second law ensures that the inheritances of different traits are independent. However, it does not ensure that the candidate IVs are independent of the hidden confounders between the treatment and the outcome.

30825 *Application of the Instrumental Variable Method: Mendelian Randomization*

It is possible that these IVs have direct effects on the confounders. It is also possible that some unmeasured genes affect both the IVs and the confounders. Mendel's second law does not ensure the exclusion restriction assumption either. It is possible that the IVs have other causal pathways to the outcome, beyond the pathway through the treatment of interest.

Technically, the statistical assumptions for MR are quite strong. Clearly, the linear IV model is a strong modeling assumption. The independence of the $\hat{\gamma}_j$'s and the $\hat{\Gamma}_j$'s is also quite strong. Other issues in the data collecting process can further complicate the interpretation of the IV assumptions. For instance, the treatments and outcomes are often measured with errors, and the genome wide associate studies are often based on the case-control design.

VanderWeele et al. (2014) is an excellent review paper that discusses the methodological challenges in MR.

25.5 Homework Problems

25.1 Data analysis

Analyze the `bmi.bmi` data in the R package `mr.raps`. See the package and Zhao et al. (2020, Section 7.2) for more details.

25.2 Recommended reading

Davey Smith and Ebrahim (2003) reviewed the potentials and limitations of Mendelian randomization.

Part VI

Causal Mechanisms with Post-Treatment Variables



26

Principal Stratification

Parts II–V focus on causal effects of a treatment on an outcome, possibly adjusting for observed pretreatment covariates. Many applications also have some post-treatment variable M which happens after the treatment but before the outcome. An important question is how to use the post-treatment variable M appropriately. I will start with several motivating examples and then introduce Frangakis and Rubin (2002)'s formulation of this problem based on potential outcomes.

26.1 Motivating Examples

Example 26.1 (noncompliance) *In randomized experiments with noncompliance, we can use M to represent the treatment received, which is affected by the treatment assignment Z and affects the outcome Y . In this example, M has the same meaning as D in Chapter 21.*

Example 26.2 (truncation by death) *In randomized experiments to patients with severe diseases, some patients may die before the measurement of the outcome Y , e.g., the quality of life. The post-treatment variable M in this example is the binary indicator of the survival status.*

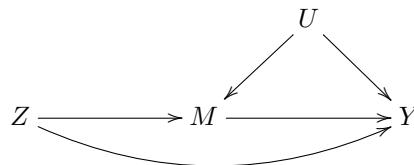
Example 26.3 (unemployment) *In job training programs, units are randomly assigned to treatment and control groups, and report their employment status M and wage Y . Then the post-treatment variable is the binary indicator of the employment status M .*

Example 26.4 (surrogate endpoint) *In clinical trials, the outcomes of interest (e.g., 30 years of survival) require a long and costly follow-up. Practitioners instead collect data on some other variables early in the follow-up that are easy to measure. These variables are called the “surrogate endpoints.” A concrete example is from clinical trials on HIV patients, where the candidate surrogate endpoint M is the CD4 cell count.*

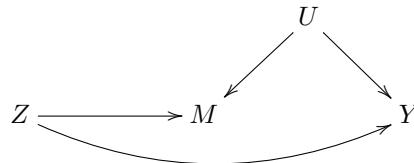
Examples 26.1–26.4 have the similarity that an intermediate variable M occurs between the treatment and the outcome. Here “between” can mean that

1. M is on the causal pathway from Z to Y as Figure 26.1(a);
2. M is not on the causal pathway from Z to Y as Figure 26.1(b).

Example 26.1 corresponds to Figure 26.1(a). Examples 26.2 and 26.3 correspond to Figure 26.1(b). Example 26.4 can correspond to Figure 26.1(a) or (b), depending on the choice of the surrogate end point.



(a) M is on the causal pathway from Z to Y , with Z randomized and U representing unmeasured confounding



(b) M is not on the causal pathway from Z to Y , with Z randomized and U representing unmeasured confounding

FIGURE 26.1: Causal diagrams with a post-treatment variable M

26.2 The Problem of Conditioning on the Post-Treatment Variable

A naive method to deal with the post-treatment variable M is to condition on its observed value as if it were a pretreatment covariate. However, M is fundamentally different from X , because the former is affected by the treatment in general but the latter is not. It is also a “rule of thumb” that data analyzers should not condition on any post-treatment variables in evaluating the average effect of the treatment on the outcome (Cochran, 1957; Rosenbaum, 1984). Based on potential outcomes, Frangakis and Rubin (2002) gave the following insightful explanation.

For simplicity, we focus on completely randomized experiment in this chapter.

Assumption 26.1 (complete randomization with an intermediate variable)
 $Z \perp\!\!\!\perp \{M(1), M(0), Y(1), Y(0)\}$.

Conditioning on $M = m$, we compare

$$\text{pr}(Y | Z = 1, M = m)$$

and

$$\text{pr}(Y | Z = 0, M = m).$$

This comparison seems intuitive which measures the difference in the outcome distributions in treated and control groups given the same value of the post-treatment variable. When M is a pre-treatment covariate, this comparison yields a reasonable subgroup effect. However, when M is a post-treatment variable, the interpretation of this comparison is problematic. Under Assumption 26.1, we can re-write

$$\begin{aligned} \text{pr}(Y | Z = 1, M = m) &= \text{pr}\{Y(1) | Z = 1, M(1) = m\} \\ &= \text{pr}\{Y(1) | M(1) = m\} \end{aligned}$$

and

$$\begin{aligned} \text{pr}(Y | Z = 0, M = m) &= \text{pr}\{Y(0) | Z = 0, M(0) = m\} \\ &= \text{pr}\{Y(0) | M(0) = m\}. \end{aligned}$$

Therefore, we are comparing the distributions of $Y(1)$ and $Y(0)$ for different subset of units because the units with $M(1) = m$ are different from the units with $M(0) = m$ if the Z affects M . Consequently, the comparison conditioning on $M = m$ does not have a causal interpretation in general unless $M(1) = M(0)$.¹

Revisit Example 26.1. Comparing $\text{pr}(Y | Z = 1, M = 1)$ and $\text{pr}(Y | Z = 0, M = 1)$ is equivalent to comparing the treated potential outcomes for compliers and always-takers and control potential outcomes for always-takers, under the monotonicity assumption that $M(1) \geq M(0)$. Part 3 of Problem 22.7 has pointed out the drawbacks of this analysis.

Revisit Example 26.2. If the treatment improves the survival status, the treatment can save more weak patients than the control. In this case, units with $M(1) = 1$ are weaker than units with $M(0) = 1$, so the naive comparison gives biased results that is in favor of the control.

¹Based on the causal diagrams, we can reach the same conclusion. In Figure 26.1, even though $Z \perp\!\!\!\perp U$ by randomization of Z , conditioning on M introduces the “collider bias” that causes $Z \not\perp\!\!\!\perp U$.

26.3 Conditioning on the Potential Values of the Post-Treatment Variable

Frangakis and Rubin (2002) proposed to condition on the joint potential value of the post-treatment variable $U = \{M(1), M(0)\}$ and compare

$$\text{pr}\{Y(1) | M(1) = m_1, M(0) = m_0\}$$

and

$$\text{pr}\{Y(0) | M(1) = m_1, M(0) = m_0\}$$

for some (m_1, m_0) . This is a comparison of the potential outcomes under treatment and control for the same subset of units with $M(1) = m_1$ and $M(0) = m_0$. Frangakis and Rubin (2002) called this strategy *principal stratification*, viewing $\{M(1), M(0)\}$ as a pre-treatment covariate. Based on this idea, we can define

$$\tau(m_1, m_0) = E\{Y(1) - Y(0) | M(1) = m_1, M(0) = m_0\}$$

as the principal stratification average causal effect for the subgroup with $M(1) = m_1, M(0) = m_0$. For a binary M , we have four subgroups

$$\begin{cases} \tau(1, 1) &= E\{Y(1) - Y(0) | M(1) = 1, M(0) = 1\}, \\ \tau(1, 0) &= E\{Y(1) - Y(0) | M(1) = 1, M(0) = 0\}, \\ \tau(0, 1) &= E\{Y(1) - Y(0) | M(1) = 0, M(0) = 1\}, \\ \tau(0, 0) &= E\{Y(1) - Y(0) | M(1) = 0, M(0) = 0\}. \end{cases} \quad (26.1)$$

Since $\{M(1), M(0)\}$ is unaffected by the treatment, it is a covariate so $\tau(m_1, m_0)$ is a subgroup causal effect. For subgroups with $M(1) = M(0)$, the treatment does not change the intermediate variable, so $\tau(1, 1)$ and $\tau(0, 0)$ measure the *dissociative effects*. For other subgroups with $m_1 \neq m_0$, the principal stratification average causal effects $\tau(m_1, m_0)$ measure the *associative effects*. These terminologies are from Frangakis and Rubin (2002), which do not assume that M is on the causal pathway from Z to Y . When we have Figure 26.1(a), we can interpret the dissociative effects as direct effects of Z on Y that act independent of M , although we cannot simply interpret the associative effects as direct or indirect effects of Z on Y .

Example 26.1 (noncompliance) With noncompliance, (26.1) consists of the average causal effects for the always takers, compliers, defiers, and never takers (Imbens and Angrist, 1994; Angrist et al., 1996).

Example 26.2 (truncation by death) Because the outcome is well defined only if the patient survives, three subgroup causal effects in (26.1) are not meaningful, and the only well-defined subgroup effect is

$$\tau(1, 1) = E\{Y(1) - Y(0) | M(1) = 1, M(0) = 1\}. \quad (26.2)$$

It is called the survivor average causal effect (Rubin, 2006a). It is the average causal effect of the treatment on the outcome for those units who survive regardless of the treatment status.

Example 26.3 (unemployment) *The unemployment problem is isomorphic to the truncation by death problem because the wage is well-defined only if the unit is employed in the first place. Therefore, the only well defined subgroup effect is (26.2), the employed average causal effect. Previously, Heckman (1979) proposed a model, now called the Heckman Selection Model, to deal with unemployment in modeling the wage, viewing the wages of those unemployed as missing values². However, Zhang and Rubin (2003) and Zhang et al. (2009) argued that $\tau(1, 1)$ is a more meaningful quantity under the potential outcomes framework.*

Example 26.4 (surrogate endpoint) *Intuitively, we want to assess the effect of the treatment on the outcome via the effect of the treatment on the surrogate endpoint. Therefore, a good surrogate endpoint should satisfy two conditions: first, if the treatment does not affect the surrogate, then it does not affect the outcome either; second, if the treatment affects the surrogate, then it affects the outcome too. The first condition is called the “causal necessity” by Frangakis and Rubin (2002), and the second condition is called the “causal sufficiency” by Gilbert and Hudgens (2008). Based on (26.1) for a binary surrogate endpoint, causal necessity requires that $\tau(1, 1)$ and $\tau(0, 0)$ are zero, and causal sufficiency requires that $\tau(1, 0)$ and $\tau(0, 1)$ are not zero.*

26.4 Statistical Inference and Its Difficulty

In Example 26.1, if we have randomization, monotonicity and exclusion restriction, then we can identify the complier average causal effect. This is the key result derived in Chapter 21.

However, in other examples, we cannot impose the exclusion restriction assumption. For instance, $\tau(1, 1)$ is the main parameter of interest in Examples 26.2 and 26.3, and $\tau(1, 1)$ and $\tau(0, 0)$ are both of interest in Example 26.4.

²Heckman won nobel prize of economics in 2000 “for his development of theory and methods for analyzing selective samples.” His model contains two stages. First, the employment status is determined by a latent linear model

$$M_i = 1(X_i^\top \beta + u_i \geq 0).$$

Second, the latent log wage is determined by a linear model

$$Y_i^* = W_i^\top \gamma + v_i$$

and Y_i^* is observed as Y_i only if $M_i = 1$. In his two-stage model, the covariates X_i and W_i may differ, and the errors (u_i, v_i) are correlated bivariate Normal.

Without the exclusion restriction assumption, it is very challenging to identify the principal stratification average causal effect. Sometimes, we cannot even impose the monotonicity assumption, and thus cannot identify the proportions of the latent strata in the first place.

26.4.1 Special case: truncation by death with binary outcome

I use the simple setting with a binary treatment, binary survival status and binary outcome to illustrate the idea and especially the difficulty of statistical inference based on principal stratification.

In addition to Assumption 26.1, we impose the monotonicity.

Assumption 26.2 (monotonicity) $M(1) \geq M(0)$.

Theorem 22.1 demonstrates that under Assumptions 26.1 and 26.2, we can identify the proportions of the three latent strata by

$$\begin{aligned}\pi_{(1,1)} &= \text{pr}(M = 1 | Z = 0), \\ \pi_{(0,0)} &= \text{pr}(M = 0 | Z = 1), \\ \pi_{(1,0)} &= \text{pr}(M = 1 | Z = 1) - \text{pr}(M = 1 | Z = 0).\end{aligned}$$

Our goal is to identify the survivor average causal effect $\tau(1, 1)$. First, we can easily identify $E\{Y(0) | M(1) = 1, M(0) = 1\}$ because the observed group ($Z = 0, M = 1$) consists of only survivors:

$$E\{Y(0) | M(1) = 1, M(0) = 1\} = E(Y | Z = 0, M = 1).$$

The key is then to identify $E\{Y(1) | M(1) = 1, M(0) = 1\}$. The observed group ($Z = 1, M = 1$) is a mixture of two strata $(1, 1)$ and $(1, 0)$, therefore we have

$$\begin{aligned}E(Y | Z = 1, M = 1) &= \frac{\pi_{(1,1)}}{\pi_{(1,1)} + \pi_{(1,0)}} E\{Y(1) | M(1) = 1, M(0) = 1\} \\ &\quad + \frac{\pi_{(1,0)}}{\pi_{(1,1)} + \pi_{(1,0)}} E\{Y(1) | M(1) = 1, M(0) = 0\}.\end{aligned}$$

We have two unknown parameters but only one equation. So we cannot uniquely determine $E\{Y(1) | M(1) = 1, M(0) = 1\}$ from the above equation. Nevertheless, this equation contains some information about the quantity of interest. That is, $E\{Y(1) | M(1) = 1, M(0) = 1\}$ is partially identified by Definition 18.1.

For a binary outcome Y , we know that $E\{Y(1) | M(1) = 1, M(0) = 0\}$ is bounded between 0 and 1, and consequently, $E\{Y(1) | M(1) = 1, M(0) = 1\}$ is bounded between the solutions to the following two equations:

$$\begin{aligned}E(Y | Z = 1, M = 1) &= \frac{\pi_{(1,1)}}{\pi_{(1,1)} + \pi_{(1,0)}} E\{Y(1) | M(1) = 1, M(0) = 1\} \\ &\quad + \frac{\pi_{(1,0)}}{\pi_{(1,1)} + \pi_{(1,0)}}\end{aligned}$$

and

$$E(Y | Z = 1, M = 1) = \frac{\pi_{(1,1)}}{\pi_{(1,1)} + \pi_{(1,0)}} E\{Y(1) | M(1) = 1, M(0) = 1\}.$$

Therefore, $E\{Y(1) | M(1) = 1, M(0) = 1\}$ has lower bound

$$\frac{\{\pi_{(1,1)} + \pi_{(1,0)}\}E(Y | Z = 1, M = 1) - \pi_{(1,0)}}{\pi_{(1,1)}},$$

and upper bound

$$\frac{\{\pi_{(1,1)} + \pi_{(1,0)}\}E(Y | Z = 1, M = 1)}{\pi_{(1,1)}}.$$

We can then derive the bounds on $\tau(1, 1)$, summarized below.

Theorem 26.1 *Under Assumptions 26.1 and 26.2 with a binary Y , we have*

$$\begin{aligned} & \frac{\{\pi_{(1,1)} + \pi_{(1,0)}\}E(Y | Z = 1, M = 1) - \pi_{(1,0)}}{\pi_{(1,1)}} - E(Y | Z = 0, M = 1) \\ & \leq \tau(1, 1) \\ & \leq \frac{\{\pi_{(1,1)} + \pi_{(1,0)}\}E(Y | Z = 1, M = 1)}{\pi_{(1,1)}} - E(Y | Z = 0, M = 1). \end{aligned}$$

In most truncation by death problems, the lower and upper bounds are quite different, and they are bounded away from the extreme values -1 and 1 . So we can use Imbens and Manski (2004)'s confidence interval for $\tau(1, 1)$ which involves two steps: first, we obtain the estimated lower and upper bounds $[\hat{l}, \hat{u}]$ with estimated standard errors (se_l, se_u); second, we construct the confidence interval as $[\hat{l} - z_\alpha se_l, \hat{u} + z_\alpha se_u]$, where z_α is the $1 - \alpha$ quantile of the standard normal distribution.

To summarize, this is a challenging problem since we cannot identify the parameter based on the observed data even with infinite sample size. We can derive large-sample bounds for $\tau(1, 1)$ but the statistical inference based on the bounds are not standard. If we do not have monotonicity, the large-sample bounds have even more complex forms (Zhang and Rubin, 2003; Jiang et al., 2016).

26.4.2 An application

I use the data in Yang and Small (2016) from the Acute Respiratory Distress Syndrome Network study involving 861 patients with lung injury and acute respiratory distress syndrome. Patients were randomized to receive mechanical ventilation with either lower tidal volumes or traditional tidal volumes. The outcome is the binary indicator for whether patients could breathe without assistance by day 28. Table 26.1 summarizes the observed data.

TABLE 26.1: Data truncated by death with * indicating the outcomes for dead patients

Treatment $Z = 1$			Control $Z = 0$				
	$Y = 1$	$Y = 0$		$Y = 1$	$Y = 0$		
$M = 1$	54	268	322	$M = 1$	59	218	277
$M = 0$	*	*	109	$M = 0$	*	*	152

We first obtain the point estimators of the latent strata:

$$\hat{\pi}_{(1,1)} = \frac{277}{277 + 152} = 0.646, \quad \hat{\pi}_{(0,0)} = \frac{109}{109 + 322} = 0.253, \quad \hat{\pi}_{(1,0)} = 0.101.$$

The sample means of the outcome for survived patients are

$$\hat{E}(Y | Z = 1, M = 1) = \frac{54}{302} = 0.168, \quad \hat{E}(Y | Z = 0, M = 1) = \frac{59}{277} = 0.213.$$

The estimates for the bounds on $E\{Y(1) | M(1) = 1, M(0) = 1\}$ are

$$\left[\frac{(0.646 + 0.101) \times 0.168 - 0.101}{0.101}, \frac{(0.646 + 0.101) \times 0.168}{0.101} \right] = [0.037, 0.194],$$

so the bounds on $\tau(1, 1)$ are

$$[0.037 - 0.213, 0.194 - 0.213] = [-0.176, -0.019].$$

Incorporating the sampling uncertainty based on the bootstrap, the upper bound becomes positive.

26.4.3 Extensions

Zhang and Rubin (2003) started the literature of large-sample bounds. Imai (2008a) and Lee (2009) were two follow-up papers. Cheng and Small (2006) derived the bounds with multiple treatment arms. Yang and Small (2016) used a secondary outcome to sharpen the bounds on the survivor average causal effect.

26.5 Principal score method

Without additional assumptions, we can only derive bounds on the causal effects within principal strata, but cannot identify them in general. We must impose additional assumptions to achieve nonparametric identification of the $\tau(m_1, m_0)$'s. There is no consensus on the choice of the assumptions. Those

additional assumptions are not testable, and their plausibility depends on the application. A line of research parallels causal inference with unconfounded observational studies. For simplicity, I focus on the case with strong monotonicity.

26.5.1 Principal score method under strong monotonicity

Assumption 26.3 (strong monotonicity) $M(0) = 0$.

Similar to the ignorability assumption, we now assume the *principal ignorability* assumption.

Assumption 26.4 (principal ignorability) $E\{Y(0) \mid M(1) = 1, X\} = E\{Y(0) \mid M(1) = 0, X\}$.

These assumptions ensures nonparametric identification of the causal effects within principal strata.

Theorem 26.2 *Under Assumptions 26.1, 26.3 and 26.4, the principal stratification average causal effects can be identified by*

$$\tau(1, 0) = E(Y \mid Z = 1, M = 1) - E\{\pi(X)Y \mid Z = 0\}/\pi$$

and

$$\tau(0, 0) = E(Y \mid Z = 1, M = 0) - E\{(1 - \pi(X))Y \mid Z = 0\}/(1 - \pi)$$

where $\pi(X) = \text{pr}\{M(1) = 1 \mid X\}$ and $\pi = \text{pr}\{M(1) = 1\}$ can be identified by

$$\pi(X) = \text{pr}(M = 1 \mid Z = 1, X)$$

and

$$\pi = \text{pr}(M = 1 \mid Z = 1).$$

The conditional probability $\pi(X) = \text{pr}\{M(1) = 1 \mid X\}$ is called the *principal score*. Theorem 26.2 states that $\tau(1, 0)$ and $\tau(0, 0)$ can be identified by difference in means with appropriate weights depending on the principal score.

Proof of Theorem 26.2: I will only prove that

$$E\{Y(0) \mid M(1) = 1\} = E\{\pi(X)Y \mid Z = 0\}/\pi.$$

The left-hand side equals

$$\begin{aligned} E\{M(1)Y(0)\}/\pi &= E[E\{M(1) \mid X\}E\{Y(0) \mid X\}]/\pi \\ &= E[\pi(X)E\{Y(0) \mid X\}]/\pi \\ &= E[E\{\pi(X)Y(0) \mid X\}]/\pi \\ &= E\{\pi(X)Y(0)\}/\pi \\ &= E\{\pi(X)Y \mid Z = 0\}/\pi. \end{aligned}$$

□

Theorem 26.2 motivates the following simple estimators for $\tau(1, 0)$ and $\tau(0, 0)$, respectively:

1. fit a logistic regression of M on X using only data from the treated group to obtain $\hat{\pi}(X_i)$;
2. estimate π by $\hat{\pi} = \sum_{i=1}^n Z_i M_i / \sum_{i=1}^n Z_i$;
3. obtain moment estimators:

$$\hat{\tau}(1, 0) = \frac{\sum_{i=1}^n Z_i M_i Y_i}{\sum_{i=1}^n Z_i M_i} - \frac{\sum_{i=1}^n (1 - Z_i) \hat{\pi}(X_i) Y_i}{\hat{\pi} \sum_{i=1}^n (1 - Z_i)}$$

and

$$\hat{\tau}(0, 0) = \frac{\sum_{i=1}^n Z_i (1 - M_i) Y_i}{\sum_{i=1}^n Z_i (1 - M_i)} - \frac{\sum_{i=1}^n (1 - Z_i) (1 - \hat{\pi}(X_i)) Y_i}{(1 - \hat{\pi}) \sum_{i=1}^n (1 - Z_i)};$$

4. use the bootstrap to approximate the variances of $\hat{\tau}(1, 0)$ and $\hat{\tau}(0, 0)$.

26.5.2 Extensions

Follmann (2000), Hill et al. (2002), Jo and Stuart (2009), Jo et al. (2011) and Stuart and Jo (2015) started the literature of using the principal score to identify causal effects within principal strata. Ding and Lu (2017) provided theoretical foundation for this strategy. They prove Theorem 26.2 as well as a more general version under monotonicity; see Problem 26.1. Jiang et al. (2022) give a unified discussion of this strategy for observational studies and propose multiply robust estimators for causal effects within principal strata.

26.6 Other methods

To estimate principal stratification average causal effects without the exclusion restriction assumption, Zhang et al. (2009) proposed to use the normal mixture models. However, the inference based on the normal mixture models can be quite fragile. A strategy is to use additional information to improve the inference under some restrictions (Ding et al., 2011; Mealli and Pacini, 2013; Mattei et al., 2013; Jiang et al., 2016).

Conceptually, the principal stratification framework works for general M . A multi-valued M generates many latent principal strata, and a continuous M generates infinitely many latent principal strata. In those cases, identifying the probability of the principal strata is non-trivial in the first place let alone identifying the principal stratification average causal effects. Jiang and Ding (2021) reviewed some useful strategies.

26.7 Homework problems

26.1 Principal score method under monotonicity

This problem extends Theorem 26.2, with Assumption 26.3 replaced by Assumption 26.2 and Assumption 26.4 replaced by the assumption below.

Assumption 26.5 (principal ignorability) *We have*

$$E\{Y(1) \mid M(1) = 1, M(0) = 0, X\} = E\{Y(1) \mid M(1) = 1, M(0) = 1, X\}$$

and

$$E\{Y(0) \mid M(1) = 1, M(0) = 0, X\} = E\{Y(0) \mid M(1) = 0, M(0) = 0, X\}.$$

Theorem 26.3 *Under Assumptions 26.1, 26.2 and 26.5, the principal stratification average causal effects can be identified by*

$$\tau(1,0) = E\{w_{1,(1,0)}(X)Y \mid Z = 1, M = 1\} - E\{w_{0,(1,0)}(X)Y \mid Z = 0, M = 0\},$$

$$\tau(0,0) = E(Y \mid Z = 1, M = 0) - E\{w_{0,(0,0)}(X)Y \mid Z = 0, M = 0\},$$

$$\tau(1,1) = E\{w_{1,(1,1)}(X)Y \mid Z = 1, M = 1\} - E(Y \mid Z = 0, M = 1)$$

with

$$\begin{aligned} w_{1,(1,0)}(X) &= \frac{\pi_{(1,0)}(X)}{\pi_{(1,0)}(X) + \pi_{(1,1)}(X)} / \frac{\pi_{(1,0)}}{\pi_{(1,0)} + \pi_{(1,1)}}, \\ w_{0,(1,0)}(X) &= \frac{\pi_{(1,0)}(X)}{\pi_{(1,0)}(X) + \pi_{(0,0)}(X)} / \frac{\pi_{(1,0)}}{\pi_{(1,0)} + \pi_{(0,0)}}, \\ w_{0,(0,0)}(X) &= \frac{\pi_{(0,0)}(X)}{\pi_{(1,0)}(X) + \pi_{(0,0)}(X)} / \frac{\pi_{(0,0)}}{\pi_{(1,0)} + \pi_{(0,0)}}, \\ w_{1,(1,1)}(X) &= \frac{\pi_{(1,1)}(X)}{\pi_{(1,0)}(X) + \pi_{(1,1)}(X)} / \frac{\pi_{(1,1)}}{\pi_{(1,0)} + \pi_{(1,1)}}. \end{aligned}$$

Moreover, the conditional and marginal principal scores are all identifiable by

$$\pi_{(0,0)}(X) = \text{pr}(M = 0 \mid Z = 1, X),$$

$$\pi_{(1,1)}(X) = \text{pr}(M = 1 \mid Z = 0, X),$$

$$\pi_{(1,0)}(X) = \text{pr}(M = 1 \mid Z = 1, X) - \text{pr}(M = 1 \mid Z = 0, X).$$

and

$$\pi_{(0,0)} = \text{pr}(M = 0 \mid Z = 1),$$

$$\pi_{(1,1)} = \text{pr}(M = 1 \mid Z = 0),$$

$$\pi_{(1,0)} = \text{pr}(M = 1 \mid Z = 1) - \text{pr}(M = 1 \mid Z = 0).$$

Remark: Based on Theorem 26.3, we can construct weighting estimators. Theorem 26.3 is Proposition 2 in Ding and Lu (2017), which also provided more details for the estimation.

26.2 Recommended reading

Frangakis and Rubin (2002) proposed the principal stratification framework. Zhang and Rubin (2003) derived large-sample bounds on the survivor average causal effect. Jiang and Ding (2021) reviewed various strategies to identify the causal effects within principal strata.

27

Mediation Analysis: Natural Direct and Indirect Effects

With an intermediate variable M between the treatment Z and outcome Y , the causal effects within principal strata defined by $U = \{M(1), M(0)\}$ can assess the treatment effect heterogeneity across latent groups U . When M is indeed on the causal pathway from Z to Y , causal effects within some principal strata, $\tau(1, 1)$ and $\tau(0, 0)$, can give information about the direct effect of Z on Y . However, these direct effects are only for two latent groups. The causal effects within the other two principal strata, $\tau(1, 0)$ and $\tau(0, 1)$, contain both the direct and indirect effects. Fundamentally, principal stratification does not provide any information about the indirect effect of Z on Y through M because it does not even assume that M can be intervened.

In the above discussion, I use the notions of “direct effect” and “indirect effect” in a casual way. When M lies on the pathway from Z to Y , researchers often want to assess the extent to which the effect of Z on Y is through M and the extent to which the effect is through other pathways. This is called *mediation analysis*. It is the topic of this chapter.

27.1 Motivating Examples

In mediation analysis, we have a treatment Z , an outcome Y , a mediator M , and some background covariates X . Figure 27.3 illustrates their relationship. Below we give some concrete examples.

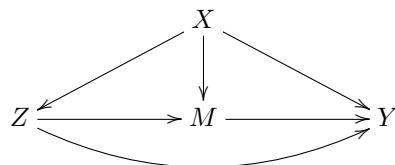


FIGURE 27.1: Directed acyclic graph for mediation

Example 27.1 VanderWeele et al. (2012) conducted mediation analysis to assess the extent to which the effect of variants on chromosome 15q25.1 on lung cancer is mediated through smoking and to which it operates through other causal pathways. The exposure levels correspond to changes from 0 to 2 C alleles, smoking intensity is measured by the square root of cigarettes per day, and the outcome is the lung cancer indicator. VanderWeele et al. (2012)'s study contained many sociodemographic covariates.

Example 27.2 Rudolph et al. (2018) studies the causal mechanism from neighborhood poverty to adolescent substance use, mediated by the school and peer environment. They used data from the National Comorbidity Survey Replication Adolescent Supplement, a nationally representative survey of U.S. adolescents conducted during 2001–2004. The treatment is the binary indicator of neighborhood disadvantage, defined as living in the lowest tertile of neighborhood socioeconomic status based on data from the 2000 U.S. Census. Four binary mediators are measures of school and peer environments, and six binary outcomes are measures of substance use. Baseline covariates included the adolescent's sex, age, race, immigration generation, family income, etc.

Example 27.3 The *mediation* package in R contains a dataset called *jobs*, which is from JOBS II, a randomized field experiment that investigates the efficacy of a job training intervention on unemployed workers. We used this dataset in Chapter 21.5. The program is designed to not only increase reemployment among the unemployed but also enhance the mental health of the job seekers. It is therefore of interest to assess the indirect effect of the intervention on the mental health through job search efficacy and its direct effect acting through other pathways. We will revisit this example later.

27.2 Nested Potential Outcomes

27.2.1 Natural Direct and Indirect Effects

Below we drop the index i for unit i and assume all random variables are iid draws from a super population. For simplicity, we focus on a binary treatment Z .

We first consider the hypothetical intervention on z and define potential mediators and outcomes corresponding to the intervention on z :

$$\{M(z), Y(z) : z = 0, 1\}.$$

We then consider hypothetical intervention on both z and m and define potential outcomes corresponding to the interventions on z and m :

$$\{Y(z, m) : z = 0, 1; m \in \mathcal{M}\},$$

where \mathcal{M} contains all possible values of m . Robins and Greenland (1992) and Pearl (2001) further consider the nested potential outcomes corresponding to intervention on z and $m = M(z') \equiv M_{z'}$:

$$\{Y(z, M_{z'}) : z = 0, 1; z' = 0, 1\}$$

where we write $M(z')$ as $M_{z'}$ to avoid excessive parentheses. The notation $Y(z, M_{z'})$ is the hypothetical outcome if the treatment were set at level z and the mediator were set at its potential level $M(z')$ under treatment z' . Importantly, z and z' can be different. With a binary treatment, we have four nested potential outcomes in total:

$$\{Y(1, M_1), Y(1, M_0), Y(0, M_1), Y(0, M_0)\}.$$

The nested potential outcome $Y(1, M_1)$ is the hypothetical outcome if the treatment were set at $z = 1$ and the mediator were set at what would happen under $z = 1$. Similarly, $Y(0, M_0)$ is the outcome if the treatment were set at $z = 0$ and the mediator were set at what would happen under $z = 0$. It would be surprising if $Y(1, M_1) \neq Y(1)$ or $Y(0, M_0) \neq Y(0)$. Therefore, we make the following assumption throughout this chapter.

Assumption 27.1 (composition) $Y(z, M_z) = Y(z)$ for $z = 0, 1$.

The composition assumption cannot be proved. It is indeed an assumption. Without causing philosophical debates, we can even define $Y(1)$ as $Y(1, M_1)$, and define $Y(0)$ as $Y(0, M_0)$.

The nested potential outcome $Y(1, M_0)$ is the hypothetical outcome if the unit received treatment 1 but its mediator were set at its natural value M_0 without the treatment. Similarly, $Y(0, M_1)$ is the hypothetical outcome if the unit received control 0 but its mediator were set at its natural value M_1 under the treatment. They are two cross-world counterfactual terms and useful for defining the direct and indirect effects.

Definition 27.1 (total, direct and indirect effects) Define the total effect of the treatment on the outcome as

$$\tau = E\{Y(1) - Y(0)\}.$$

Define the natural direct effect as

$$\text{NDE} = E\{Y(1, M_0) - Y(0, M_0)\}.$$

Define the natural indirect effect as

$$\text{NIE} = E\{Y(1, M_1) - Y(1, M_0)\}.$$

The total effect is the standard average causal effect of Z on Y . The natural direct effect measures the effect of the treatment on the outcome if the

mediator were set at the natural value M_0 without the intervention. The natural indirect effect measures the effect of the treatment through changing the mediator if the treatment itself were set at $z = 1$. Under the composition assumption, the natural direct and indirect effects reduce to

$$\text{NDE} = E\{Y(1, M_0) - Y(0)\}, \quad \text{NIE} = E\{Y(1) - Y(1, M_0)\},$$

and therefore, we can decompose the total effect as the sum of the natural direct and indirect effects.

Proposition 27.1 *By Definition 27.1 and Assumption 27.1, $\tau = \text{NDE} + \text{NIE}$.*

Mathematically, we can also define the natural indirect effect as $E\{Y(0, M_1) - Y(0, M_0)\}$ where the treatment is fixed at 0. However, this definition does not lead to the decomposition in Proposition 27.1.

Unfortunately, the nest potential outcome $Y(1, M_0)$ is not an easy quantity to understand due to the cross-world nature of the interventions: the treatment is set at $z = 1$ but the mediator is set at its natural value M_0 under treatment $z = 0$. Clearly, these two interventions on the treatment cannot simultaneously happen in any realized experiment. To understand the cross-world potential outcome $Y(1, M_0)$, we need to imagine the existence of parallel worlds as shown in Figure 27.2. Let's focus on $Y(1, M_0)$. When the treatment is set at $z = 1$, the mediator must take value M_1 . If at the same time we want to set the mediator at $m = M_0$, we must know the value of M_0 for the same unit from another experiment in the parallel world. This can be an unrealistic physical experiment because it requires that the same unit is intervened at two different levels of the treatment. Under some strong assumptions about the homogeneity of units, we may use another unit's mediator value under control as a proxy for M_0 .

27.2.2 Metaphysics or Science

Causal inference is hard, and there is no agreement even on its mathematical notation. Robins and Greenland (1992) and Pearl (2001) used the nested potential outcomes to define the natural direct and indirect effects. However, Frangakis and Rubin (2002) called $Y(1, M_0)$ and $Y(0, M_1)$ *a priori counterfactuals* because we cannot observe them in any physical experiments. In this sense, they do not exist *a priori*. According to Popper (1963), a way to distinguish science and metaphysics is the falsifiability of the statements. That is, if a statement is not falsifiable based on any physical experiments or observations, then it is not a scientific but rather a metaphysical statement. Because we cannot observe $Y(1, M_0)$ and $Y(0, M_1)$ in any experiments, we cannot falsify any statements involving them except for the trivial ones (e.g., some outcomes are binary, or continuous, or bounded). Therefore, a strict Popperian statistician would view mediation analysis as metaphysics.

More strikingly, Dawid (2000) criticized the potential outcomes framework

to be metaphysical, and he called Rubin's Science Table a "metaphysical array." This is a critique on not only the a priori counterfactuals $Y(1, M_0)$ and $Y(0, M_1)$ but also the simple potential outcomes $Y(1)$ and $Y(0)$. Dawid (2000) argued that because we can never observe $Y(1)$ and $Y(0)$ jointly, then introducing the notation $\{Y(1), Y(0)\}$ is a metaphysical activity. He is correct about the metaphysical nature of the joint distribution of $\text{pr}\{Y(1), Y(0)\}$, but he is incorrect about the marginal distributions. Based on the observed data, we indeed can falsify some statement about the marginal distributions, although we cannot falsify any statements about the joint distribution.¹ Therefore, even according to Popper (1963), Rubin's Science Table is not metaphysical because it has some nontrivial falsifiable implications although not all implications are falsifiable. This is the fundamental difference between $\{Y(1), Y(0)\}$ and $\{Y(1, M_0), Y(0, M_1)\}$.

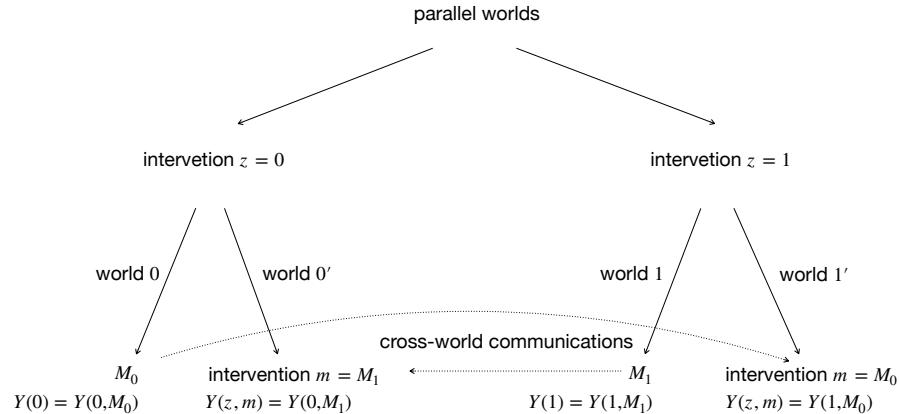


FIGURE 27.2: Crossworld potential outcomes $Y(1, M_0)$ and $Y(0, M_1)$

¹By the probability theory, given the marginal distributions of $\text{pr}(Y(1) \leq y_1)$ and $\text{pr}(Y(0) \leq y_0)$, we can bound the joint distribution of $\text{pr}(Y(1) \leq y_1, Y(0) \leq y_0)$ by the Frechet–Hoeffding inequality:

$$\begin{aligned} & \max\{0, \text{pr}(Y(1) \leq y_1) + \text{pr}(Y(0) \leq y_0) - 1\} \\ & \leq \text{pr}(Y(1) \leq y_1, Y(0) \leq y_0) \\ & \leq \min\{\text{pr}(Y(1) \leq y_1), \text{pr}(Y(0) \leq y_0)\}. \end{aligned}$$

This is often a loose inequality. Unfortunately, we do not have any information beyond this inequality without imposing additional assumptions.

27.3 The Mediation Formula

Pearl (2001)'s mediation formula relies on the following four assumptions. The first three essentially assumes that the treatment and the mediator are both randomized conditional on observed covariates.

Assumption 27.2 *There is no treatment-outcome confounding:*

$$Z \perp\!\!\!\perp Y(z, m) \mid X$$

for all z and m .

Assumption 27.3 *There is no mediator-outcome confounding:*

$$M \perp\!\!\!\perp Y(z, m) \mid (X, Z)$$

for all z and m .

Assumptions 27.2 and 27.3 together are often called *sequential ignorability*. They are equivalent to the assumption that (Z, M) are jointly randomized conditioning on X :

$$(Z, M) \perp\!\!\!\perp Y(z, m) \mid X \tag{27.1}$$

for all z and m . I leave the proof as Problem 27.1.

Assumption 27.4 *There is no treatment-mediator confounding:*

$$Z \perp\!\!\!\perp M(z) \mid X$$

for all z .

The last assumption is the cross-world independence.

Assumption 27.5 *There is no cross-world independence between the potential outcomes and potential mediators:*

$$Y(z, m) \perp\!\!\!\perp M(z') \mid X$$

for all z, z' and m .

Assumptions 27.2–27.4 are very strong, but at least they hold under experiments with randomized treatment and mediator. Assumption 27.5 is stronger because no physical experiment can ensure it. Because we can never observe $Y(z, m)$ and $M(z')$ in any experiment if $z \neq z'$, Assumption 27.5 can never be validated so it is fundamentally meta-physical.

I give an example below in which Assumptions 27.2–27.5 all hold.

Example 27.4 Given X , we generate

$$\begin{aligned} Z &= 1\{f_Z(X, \varepsilon_Z)\}, \\ M(z) &= 1\{f_M(X, z, \varepsilon_M)\}, \\ Y(z, m) &= f_Y(X, z, m, \varepsilon_Y), \end{aligned}$$

for $z, m = 0, 1$, where $\varepsilon_Z, \varepsilon_M, \varepsilon_Y$ are all independent. Consequently, we generate the observed values of M and Y from

$$\begin{aligned} M &= M(Z) = 1\{f_M(X, Z, \varepsilon_M)\}, \\ Y &= Y(Z, M) = f_Y(X, Z, M, \varepsilon_Y). \end{aligned}$$

We can verify that Assumptions 27.2–27.5 hold under this data generating process; see Problem 27.2.

Pearl (2001) proved the following key result for mediation analysis.

Theorem 27.1 Under Assumptions 27.2–27.5, we have

$$E\{Y(z, M_{z'}) \mid X = x\} = \sum_m E(Y \mid Z = z, M = m, X = x) \text{pr}(M = m \mid Z = z', X = x)$$

and therefore,

$$E\{Y(z, M_{z'})\} = \sum_x E\{Y(z, M_{z'}) \mid X = x\} \text{pr}(X = x).$$

Theorem 27.1 assumes that both M and X are discrete. With general M and X , the mediation formulas become

$$E\{Y(z, M_{z'}) \mid X = x\} = \int E(Y \mid Z = z, M = m, X = x) f_M(m \mid Z = z', X = x) dm$$

and

$$E\{Y(z, M_{z'})\} = \int E\{Y(z, M_{z'}) \mid X = x\} f_X(x) dx.$$

From Theorem 27.1, the identification formulas for the means of the nested potential outcomes depend on the conditional mean of the outcome given the treatment, mediator, and covariates, as well as the conditional mean of the mediator given the treatment and covariates. We need to evaluate these two conditional means at different treatment levels if the nested potential outcome involves cross-world interventions.

If we drop the cross-world independence assumption, we can modify the definition of the natural direct and indirect effects and the same formulas hold. See Problem 27.8 for more details.

I give the proof below.

Proof of Theorem 27.1: By the tower property, $E\{Y(z, M_{z'})\} =$

$E[E\{Y(z, M_{z'}) \mid X\}]$, so we need only to prove the formula for $E\{Y(z, M_{z'}) \mid X = x\}$. Starting with the law of total probability, we have

$$\begin{aligned} & E\{Y(z, M_{z'}) \mid X = x\} \\ &= \sum_m E\{Y(z, M_{z'}) \mid M_{z'} = m, X = x\} \text{pr}(M_{z'} = m \mid X = x) \\ &= \sum_m E\{Y(z, m) \mid M_{z'} = m, X = x\} \text{pr}(M_{z'} = m \mid X = x) \\ &= \sum_m \underbrace{E\{Y(z, m) \mid X = x\}}_{\text{Assumption 27.5}} \underbrace{\text{pr}(M = m \mid Z = z', X = x)}_{\text{Assumption 27.4}} \\ &= \sum_m \underbrace{E(Y \mid Z = z, M = m, X = x)}_{\text{Assumptions 27.2 and 27.3}} \text{pr}(M = m \mid Z = z', X = x). \end{aligned}$$

□

The above proof is actually trivial from a mathematical perspective. It illustrates the necessity of Assumptions 27.2–27.5.

Conditioning on $X = x$, the mediation formulas for $Y(1, M_1)$ and $Y(0, M_0)$ simplifies to

$$\begin{aligned} & E\{Y(1, M_1) \mid X = x\} \\ &= \sum_m E(Y \mid Z = 1, M = m, X = x) \text{pr}(M = m \mid Z = 1, X = x) \\ &= E(Y \mid Z = 1, X = x) \end{aligned}$$

and

$$\begin{aligned} & E\{Y(0, M_0) \mid X = x\} \\ &= \sum_m E(Y \mid Z = 0, M = m, X = x) \text{pr}(M = m \mid Z = 0, X = x) \\ &= E(Y \mid Z = 0, X = x) \end{aligned}$$

based on the law of total probability; the mediation formula for $Y(1, M_0)$ simplifies to

$$E\{Y(1, M_0) \mid X = x\} = \sum_m E(Y \mid Z = 1, M = m, X = x) \text{pr}(M = m \mid Z = 0, X = x),$$

where the conditional expectation of the outcome is given $Z = 1$ but the conditional distribution of the mediator is given $Z = 0$. This leads to the identification formulas of the natural direct and indirect effects.

Corollary 27.1 *Under Assumptions 27.2–27.5, the conditional natural direct and indirect effects are identified by*

$$\begin{aligned} \text{NDE}(x) &= E\{Y(1, M_0) - Y(0, M_0) \mid X = x\} \\ &= \sum_m \{E(Y \mid Z = 1, M = m, X = x) - E(Y \mid Z = 0, M = m, X = x)\} \\ &\quad \times \text{pr}(M = m \mid Z = 0, X = x) \end{aligned}$$

and

$$\begin{aligned}\text{NIE}(x) &= E\{Y(1, M_1) - Y(1, M_0) \mid X = x\} \\ &= \sum_m E(Y \mid Z = 1, M = m, X = x) \\ &\quad \times \{\text{pr}(M = m \mid Z = 1, X = x) - \text{pr}(M = m \mid Z = 0, X = x)\};\end{aligned}$$

the unconditional ones can be identified by $\text{NDE} = \sum_x \text{NDE}(x)\text{pr}(X = x)$ and $\text{NIE} = \sum_x \text{NIE}(x)\text{pr}(X = x)$.

As a special case, with a binary M , the formula of the NIE reduces to a product form below.

Corollary 27.2 *Under Assumptions 27.2–27.5, for a binary mediator M , we have*

$$\text{NIE}(x) = \tau_{Z \rightarrow M}(x)\tau_{M \rightarrow Y}(1, x)$$

and $\text{NIE} = E\{\text{NIE}(X)\}$, where

$$\tau_{Z \rightarrow M}(x) = \text{pr}(M = 1 \mid Z = 1, X = x) - \text{pr}(M = 1 \mid Z = 0, X = x).$$

and

$$\tau_{M \rightarrow Y}(z, x) = E(Y \mid Z = z, M = 1, X = x) - E(Y \mid Z = z, M = 0, X = x)$$

I leave the proof of Corollary 27.2 as Problem 27.4. Corollary 27.2 gives a simple formula in the case of a binary M . With randomized Z conditional on X , we can view $\tau_{Z \rightarrow M}(x)$ as the conditional average causal effect of Z on M . With randomized M conditional on (X, Z) , we can view $\tau_{M \rightarrow Y}(z, x)$ as the conditional average causal effect of M on Y . The conditional natural indirect effect equals their product. This is coherent with our intuition that the indirect effect acts from Z to M and then from M to Y .

27.4 The Mediation Formula Under Linear Models

Theorem 27.1 gives the nonparametric identification formula for mediation analysis. It allows us to derive various formulas for mediation analysis under different models. I will introduce the famous Baron–Kenny method under linear models below. VanderWeele (2015) gives explicit formulas for the natural direct and indirect effects for many commonly-used models. I relegate the details of other models to Section 27.6.

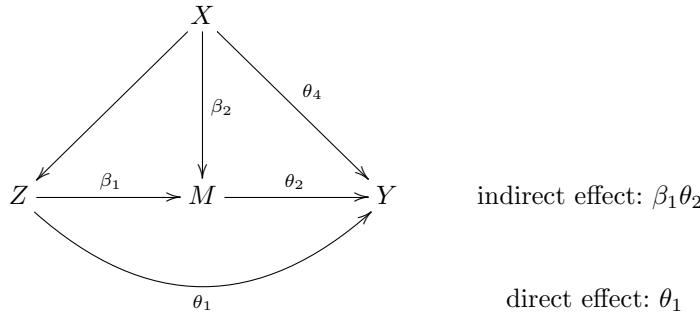


FIGURE 27.3: The Baron–Kenny Method for mediation under linear models

27.4.1 The Baron–Kenny Method

The Baron–Kenny method assumes the following linear models for the mediator and outcome given the treatment and covariates.

Assumption 27.6 (linear models for the Baron–Kenny method) *Both the mediator and outcome follow linear models:*

$$\begin{cases} E(M | Z, X) = \beta_0 + \beta_1 Z + \beta_2^T X, \\ E(Y | Z, M, X) = \theta_0 + \theta_1 Z + \theta_2 M + \theta_4^T X. \end{cases}$$

Under these linear models, the formulas for the natural direct and indirect effects simplify to functions of the coefficients.

Corollary 27.3 (Baron–Kenny formulas for mediation) *Under Assumptions 27.2–27.5 and 27.6,*

$$\text{NDE} = \theta_1, \quad \text{NIE} = \theta_2 \beta_1.$$

Proof of Corollary 27.3: We have

$$\text{NDE}(x) = \sum_m \theta_1 \text{pr}(M = m | Z = 0, X = x) = \theta_1$$

and

$$\begin{aligned} \text{NIE}(x) &= \sum_m (\theta_0 + \theta_1 + \theta_2 m + \theta_4^T x) \\ &\quad \times \{\text{pr}(M = m | Z = 1, X = x) - \text{pr}(M = m | Z = 0, X = x)\} \\ &= \theta_2 \{E(M = m | Z = 1, X = x) - E(M = m | Z = 0, X = x)\} \\ &= \theta_2 \beta_1, \end{aligned}$$

which do not depend on x . Therefore, they are also the formulas for the unconditional natural direct and indirect effects. \square

If we obtain OLS estimators of these coefficients, we can estimate the direct and indirect effects by

$$\hat{NDE} = \hat{\theta}_1, \quad \hat{NIE} = \hat{\theta}_2 \hat{\beta}_1,$$

which is called the Baron–Kenny method (Judd and Kenny, 1981; Baron and Kenny, 1986) although it had several antecedents (e.g., Hyman, 1955; Alwin and Hauser, 1975; Judd and Kenny, 1981; Sobel, 1982).

Standard software packages report the standard error of \hat{NDE} from OLS. Sobel (1982, 1986) used the delta method to obtain the standard error of \hat{NIE} . Based on the formula in Example A1.2, the asymptotic variance of $\hat{\theta}_2 \hat{\beta}_1$ equals $\text{var}(\hat{\theta}_2)\hat{\beta}_1^2 + \hat{\theta}_2^2\text{var}(\hat{\beta}_1)$. So the estimated variance is

$$\hat{\text{var}}(\hat{\theta}_2)\hat{\beta}_1^2 + \hat{\theta}_2^2\hat{\text{var}}(\hat{\beta}_1).$$

Testing the null hypothesis of NIE based on $\hat{\theta}_2 \hat{\beta}_1$ and the estimated variance above is called *Sobel's test* in the literature of mediation analysis.

27.4.2 An Example

We can easily implement the Baron–Kenny method via the following code.

```
library("car")
BKmediation = function(Z, M, Y, X)
{
  ## two regressions and coefficients
  mediator.reg = lm(M ~ Z + X)
  mediator.Zcoef = mediator.reg$coef[2]
  mediator.Zse = sqrt(hccm(mediator.reg)[2, 2])

  outcome.reg = lm(Y ~ Z + M + X)
  outcome.Zcoef = outcome.reg$coef[2]
  outcome.Zse = sqrt(hccm(outcome.reg)[2, 2])
  outcome.Mcoef = outcome.reg$coef[3]
  outcome.Mse = sqrt(hccm(outcome.reg)[3, 3])

  ## Baron-Kenny point estimates
  NDE = outcome.Zcoef
  NIE = outcome.Mcoef*mediator.Zcoef

  ## Sobel's variance estimate based the delta method
  NDE.se = outcome.Zse
  NIE.se = sqrt(outcome.Mse^2*mediator.Zcoef^2 +
                outcome.Mcoef^2*mediator.Zse^2)

  res = matrix(c(NDE, NIE,
```

```

NDE.se, NIE.se,
NDE/NDE.se, NIE/NIE.se),
2, 3)
rownames(res) = c("NDE", "NIE")
colnames(res) = c("est", "se", "t")

res
}

```

Revisiting Example 27.3, we obtain the following estimates for the direct and indirect effects:

```

> library(mediation)
> Z = jobs$treat
> M = jobs$job_seek
> Y = jobs$depress2
> getX      = lm(treat ~ econ_hard + depress1 +
+                  sex + age + occp + marital +
+                  nonwhite + educ + income,
+                  data = jobs)
> X = model.matrix(getX)[, -1]
> res = BKmediation(Z, M, Y, X)
> round(res, 3)
    est      se      t
NDE -0.037  0.042 -0.885
NIE -0.014  0.009 -1.528

```

Both the estimates for the direct and indirect effects are negative although they are insignificant.

27.5 Sensitivity analysis

Mediation analysis relies on strong and untestable assumptions. One crucial assumption is that there is no unmeasured confounding among the treatment, mediator and outcome. Various sensitivity analysis methods appeared in the literature. In particular, Ding and Vanderweele (2016) proposed Cornfield-type sensitivity bounds and Zhang and Ding (2022) proposed a sensitivity analysis method tailored to the Baron–Kenny method based on linear structural equation models.

27.6 Homework problems

27.1 Sequential randomization and joint randomization

Show (27.1) is equivalent to Assumptions 27.2 and 27.3.

27.2 Verifying the assumptions for mediation analysis

Show that Assumptions 27.2–27.5 hold under the data generating process in Example 27.4.

27.3 Another set of assumptions for the mediation formula

Imai et al. (2010) invoked the following set of assumptions to derive the mediation formula.

Assumption 27.7 Assume

$$\{Y(z, m), M(z')\} \perp\!\!\!\perp Z \mid X$$

and

$$Y(z, m) \perp\!\!\!\perp M(z') \mid (Z = z', X)$$

for all z, z', m .

Theorem 27.2 Under Assumption 27.7, the mediation formula holds.

Prove Theorem 27.2.

27.4 Natural indirect effect with a binary mediator

Prove Corollary 27.2.

27.5 With Treatment-Outcome Interaction on the Outcome

VanderWeele (2015) suggested using the following linear models:

$$\begin{cases} E(M \mid Z, X) &= \beta_0 + \beta_1 Z + \beta_2^T X, \\ E(Y \mid Z, M, X) &= \theta_0 + \theta_1 Z + \theta_2 M + \theta_3 ZM + \theta_4^T X, \end{cases}$$

where the outcome model has the interaction term between the treatment and the mediator.

Under the above linear models, show that

$$\text{NDE} = \theta_1 + \theta_3 \{\beta_0 + \beta_2^T E(X)\}, \quad \text{NIE} = (\theta_2 + \theta_3) \beta_1.$$

How do we estimate NDE and NIE with IID data?

Remark: Consider the simple case with a binary Z and binary M . Under

the linear models, the average causal effect of Z of M equals β_1 , and the average causal effect of M on Y equals $\theta_2 + \theta_3 E(Z)$. Therefore, it is possible that both of these effects are positive, but the natural indirect effect is negative. For instance:

$$\beta_1 = 1, \quad \theta_2 = 1, \quad \theta_3 = -1.5, \quad E(Z) = 0.5.$$

This is somewhat paradoxical, and can be called the *mediator paradox*. Chen et al. (2007) reported a related *surrogate endpoint paradox* or *intermediate variable paradox*.

27.6 Logistic Model for Binary Mediator

Consider the following Logistic model for the binary mediator and linear model for the outcome:

$$\begin{cases} \text{logit}\{\text{pr}(M = 1 | Z, X)\} &= \beta_0 + \beta_1 Z + \beta_2^T X, \\ E(Y | Z, M, X) &= \theta_0 + \theta_1 Z + \theta_2 M + \theta_4^T X, \end{cases}$$

where $\text{logit}(w) = \log\{w/(1-w)\}$ with inverse $\text{expit}(w) = (1 + e^{-w})^{-1}$.

Under these models, show that

$$\text{NDE} = \theta_1, \quad \text{NIE} = \theta_2 E \left\{ \text{expit}(\beta_0 + \beta_1 + \beta_2^T X) - \text{expit}(\beta_0 + \beta_2^T X) \right\}.$$

How do we estimate NDE and NIE with IID data?

27.7 Mediation analysis with binary mediator and outcome

Consider the following Logistic models for the binary mediator and outcome:

$$\begin{cases} \text{logit}\{\text{pr}(M = 1 | Z, X)\} &= \beta_0 + \beta_1 Z + \beta_2^T X, \\ \text{logit}\{\text{pr}(Y = 1 | Z, M, X)\} &= \theta_0 + \theta_1 Z + \theta_2 M + \theta_4^T X. \end{cases}$$

Express NDE and NIE in terms of the model parameters and the distribution of X . How do we estimate NDE and NIE with IID data?

27.8 Modify the definitions to drop the cross-world independence

Define

$$Y(z, F_{M_{z'}|X}) = \int Y(z, m) f_{M_{z'}}(m | X) dm$$

as the potential outcome under treatment z and a random draw from the distribution of $M_{z'} | X$. The key difference between $Y(z, M_{z'})$ and $Y(z, F_{M_{z'}|X})$ is that $M_{z'}$ is the potential mediator for the same unit whereas $F_{M_{z'}|X}$ is a random draw from the conditional distribution of the potential mediator in the whole population. Define the natural direct and indirect effects as

$$\text{NDE} = E\{Y(1, F_{M_0|X}) - Y(0, F_{M_0|X})\}, \quad \text{NIE} = E\{Y(1, F_{M_1|X}) - Y(1, F_{M_0|X})\}.$$

Show that under Assumptions 27.2–27.4, the identification formulas for NDE and NIE remain the same as in the main text.

Remark: Modifying the definitions of the nested potential outcomes allows us to relax the strong cross-world independence assumption but weakens the interpretation of the natural direct and indirect effects. See VanderWeele (2015) for more discussion and VanderWeele and Tchetgen Tchetgen (2017) for an application to a more complex setting with time varying treatment and mediator.

27.9 Connections between principal stratification and mediation analysis

VanderWeele (2008) and Forastiere et al. (2018) reviewed and compared principal stratification and mediation analysis.



28

Controlled Direct Effect

The formulation of mediation analysis in Chapter 27 relies on the nested potential outcomes, and fundamentally, some nested potential outcomes are not observable in any physical experiments. If we stick to the Popperian philosophy of science, we should only define causal parameters in terms of quantities that are observable under some experiments. This chapter discusses an alternative view of causal inference with an intermediate variable. In this view, we only define the direct effect but can not define the indirect effect.

28.1 Identification and estimation of the controlled direct effect

We view Z and M as two factors, and define potential outcomes $Y(z, m)$ for $z = 0, 1$ and $m \in \mathcal{M}$. Based on these potential outcomes, we can define the *controlled direct effect* (CDE) below.

Definition 28.1 (CDE) Define

$$\text{CDE}(m) = E\{Y(1, m) - Y(0, m)\}.$$

By definition, $\text{CDE}(m)$ is the average causal effect of the treatment if the intermediate variable is fixed at m . The parameter $\text{CDE}(m)$ can capture the direct effect of the treatment holding the mediator at m . However, this formulation cannot capture the indirect effect. In particular, the parameter $E\{Y(z, 1) - Y(z, 0)\}$ only measures the effect of the mediator on the outcome holding the treatment at z . This is not a meaningful definition of the indirect effect.

To identify $\text{CDE}(m)$, we need the following assumption, which basically requires that Z and M are jointly randomized given X .

Assumption 28.1 *Sequential ignorability requires*

$$Z \perp\!\!\!\perp Y(z, m) \mid X, \quad M \perp\!\!\!\perp Y(z, m) \mid (Z, X)$$

or, equivalently,

$$(Z, M) \perp\!\!\!\perp Y(z, m) \mid X.$$

I will focus on the case with a binary Z and M . Mathematically, we can just view this problem as an observational study with four treatment levels

$$(z, m) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}.$$

The following theorem extends the results for observational studies with a binary treatment, identifying

$$\mu_{zm} = E\{Y(z, m)\}$$

based on outcome regression, inverse probability weighting, and doubly robust estimation.

Define

$$\mu_{zm}(x) = E(Y | Z = z, M = m, X = x)$$

as the outcome mean conditional on the treatment, mediator and covariates. Define

$$e_{zm}(x) = \text{pr}(Z = z, M = m | X = x) = \text{pr}(Z = z | X = x)\text{pr}(M = m | Z = z, X = x)$$

as the probability of the joint value of Z and M conditional on the covariates.

Theorem 28.1 *Under Assumption 28.1, we have*

$$\mu_{zm} = E\{\mu_{zm}(X)\}$$

or

$$\mu_{zm} = E\left\{\frac{I(Z = z, M = m)Y}{e_{zm}(X)}\right\}.$$

Moreover, based on the working models $e_{zm}(X, \alpha)$ and $\mu_{zm}(X, \beta)$, we have the doubly robust formula

$$\mu_{zm}^{\text{dr}} = E\{\mu_{zm}(X, \beta)\} + E\left[\frac{I(Z = z, M = m)\{Y - \mu_{zm}(X, \beta)\}}{e_{zm}(X, \alpha)}\right],$$

which equals μ_{zm} if either $e_{zm}(X, \alpha) = e_{zm}(X)$ or $\mu_{zm}(X, \beta) = \mu_{zm}(X)$.

The proof of Theorem 28.1 is similar to those for the standard unconfounded observational studies. Problem 28.2 gives a general result. Based on the outcome mean model, we can obtain $\hat{\mu}_{zm}(x)$ for $\mu_{zm}(x)$. Based on the treatment model, we can obtain $\hat{e}_z(x)$ for $\text{pr}(Z = z | X = x)$; based on the intermediate variable model, we can obtain $\hat{e}_m(z, x)$ for $\text{pr}(M = m | Z = z, X = x)$. We can then estimate μ_{zm} by outcome regression

$$\hat{\mu}_{zm}^{\text{reg}} = n^{-1} \sum_{i=1}^n \hat{\mu}_{zm}(X_i),$$

by inverse probability weighting

$$\begin{aligned}\hat{\mu}_{zm}^{\text{ht}} &= n^{-1} \sum_{i=1}^n \frac{I(Z_i = z, M_i = m) Y_i}{\hat{e}_z(X_i) \hat{e}_m(z, X_i)}, \\ \hat{\mu}_{zm}^{\text{haj}} &= \sum_{i=1}^n \frac{I(Z_i = z, M_i = m) Y_i}{\hat{e}_z(X_i) \hat{e}_m(z, X_i)} / \sum_{i=1}^n \frac{I(Z_i = z, M_i = m)}{\hat{e}_z(X_i) \hat{e}_m(z, X_i)},\end{aligned}$$

or by augmented inverse probability weighting

$$\hat{\mu}_{zm}^{\text{dr}} = \hat{\mu}_{zm}^{\text{reg}} + n^{-1} \sum_{i=1}^n \frac{I(Z_i = z, M_i = m) \{Y_i - \hat{\mu}_{zm}(X_i)\}}{\hat{e}_z(X_i) \hat{e}_m(z, X_i)}.$$

We can then estimate $\text{CDE}(m)$ by $\hat{\mu}_{1m} - \hat{\mu}_{0m}$ and use the bootstrap to approximate the standard error.

If we are willing to assume a linear outcome model, the controlled direct effect simplifies to the coefficient of the treatment. Example 28.1 below gives the details.

Example 28.1 Under Assumption 28.1 and a linear outcome model,

$$E(Y | Z, M, X) = \theta_0 + \theta_1 Z + \theta_2 M + \theta_4^T X,$$

we can show that $\text{CDE}(m)$ equals the coefficient θ_1 , which coincides with the natural direct effect in the Baron–Kenny method. I relegate the proof to Problem 28.3.

28.2 Discussion

The formulation of the controlled direct effect does not involve nested or a priori counterfactual potential outcomes, and its identification does not require the cross-world counterfactual independence assumption. The parameter $\text{CDE}(m)$ can capture the direct effect of the treatment holding the mediator at m . However, this formulation cannot capture the indirect effect. I summarize the causal frameworks for intermediate variables below.

chapter	framework	direct effect	indirect effect
26	principal stratification	$\tau(1, 1), \tau(0, 0)$?
27	mediation analysis	NDE	NIE
29	controlled direct effect	$\text{CDE}(m)$?

The mediation analysis framework can decompose the total effect into natural direct and indirect effects, but it requires nested potential outcomes and cross-world independence. The principal stratification and controlled direct effect frameworks cannot define indirect effects but they do not involve

nested potential outcomes and cross-world independence. Moreover, the principal stratification framework does not necessarily require that M lies on the causal pathway from the treatment to the outcome. But its identification and estimation involves disentangling mixture distributions, which is a nontrivial task in statistics.

28.3 Homework problems

28.1 CDE and NDE

Show that under cross-world independence $Y(z, m) \perp\!\!\!\perp M(z') \mid X$ for all z, z' and m , the conditional CDE($m \mid x$) = $E\{Y(1, m) - Y(0, m) \mid X = x\}$ and NDE(x) = $E\{Y(1, M_0) - Y(0, M_0) \mid X = x\}$ have the following relationship:

$$\text{NDE}(x) = E\{\text{CDE}(M_0 \mid x)\},$$

which reduces to

$$\text{NDE}(x) = \sum_m \text{CDE}(m \mid x) \text{pr}(M_0 = m \mid X = x)$$

for a discrete M . Without the cross-world independence, does this relationship still hold in general?

28.2 Observational studies with a multi-valued treatment

Theorem 28.1 is a special case of the following theorem for unconfounded observational studies with multiple treatment levels (Imai and Van Dyk, 2004; Cattaneo, 2010). Below, I state the general problem and theorem.

Consider an observational study with a multi-valued treatment $Z \in \{1, \dots, K\}$, covariates X , and outcome Y . Unit i has K potential outcomes $Y_i(1), \dots, Y_i(K)$ corresponding to the K treatment levels. Causal effects can be defined as comparisons of the potential outcomes. In general, we can define causal effect in terms of contrasts of the potential outcomes:

$$\tau_c = \sum_{k=1}^K c_k E\{Y(k)\}$$

where $\sum_{k=1}^K c_k = 0$. The canonical choice of the pairwise comparison

$$\tau_{k,k'} = E\{Y(k) - Y(k')\}.$$

Therefore, the key is to identify and estimate the means of the potential outcomes $\mu_k = E\{Y(k)\}$ under the ignorability assumption below based on IID data of $(Z_i, X_i, Y_i)_{i=1}^n$.

Assumption 28.2 $Z \perp\!\!\!\perp \{Y(1), \dots, Y(K)\} \mid X$.

Define the generalized propensity score as

$$e_k(X) = \text{pr}(Z = k \mid X),$$

and define the conditional outcome mean as

$$\mu_k(X) = E(Y \mid Z = k, X)$$

for $k = 1, \dots, K$. We have the following theorem.

Theorem 28.2 *Under Assumption 28.2, we have*

$$\mu_k = E\{\mu_k(X)\}$$

or

$$\mu_k = E\left\{\frac{I(Z = k)Y}{e_k(X)}\right\}.$$

Moreover, based on the working models $e_k(X, \alpha)$ and $\mu_k(X, \beta)$, we have the doubly robust formula

$$\mu_k^{\text{dr}} = E\{\mu_k(X, \beta)\} + E\left[\frac{I(Z = k)\{Y - \mu_k(X, \beta)\}}{e_k(X, \alpha)}\right],$$

which equals μ_k if either $e_k(X, \alpha) = e_k(X)$ or $\mu_k(X, \beta) = \mu_k(X)$.

Prove Theorem 28.2.

Remark: Theorem 28.1 is a special case of Theorem 28.2 if we view the (Z, M) in Theorem 28.1 as a treatment with four levels. The CDE(m) is a special case of τ_c .

28.3 CDE in the linear outcome model

Show that under Assumption 28.1, if $E(Y \mid Z, M, X) = \theta_0 + \theta_1 Z + \theta_2 M + \theta_4^T X$, then

$$\text{CDE}(m) = \theta_1$$

for all m ; if $E(Y \mid Z, M, X) = \theta_0 + \theta_1 Z + \theta_2 M + \theta_3 ZM + \theta_4^T X$, then

$$\text{CDE}(m) = \theta_1 + \theta_3 m.$$

28.4 CDE in the logit outcome model

Show that for a binary outcome, under Assumption 28.1, if

$$\text{logit}\{\text{pr}(Y = 1 \mid Z, M, X)\} = \theta_0 + \theta_1 Z + \theta_2 M + \theta_4^T X,$$

then

$$\text{CDE}(m) = E\{\text{expit}(\theta_0 + \theta_1 + \theta_2 m + \theta_4^T X) - \text{expit}(\theta_0 + \theta_2 m + \theta_4^T X)\};$$

if

$$\text{logit}\{\text{pr}(Y = 1 | Z, M, X)\} = \theta_0 + \theta_1 Z + \theta_2 M + \theta_3 ZM + \theta_4^T X,$$

then

$$\text{CDE}(m) = E\{\text{expit}(\theta_0 + \theta_1 + \theta_2 m + \theta_3 m + \theta_4^T X) - \text{expit}(\theta_0 + \theta_2 m + \theta_4^T X)\}.$$

28.5 Recommended reading

Nguyen et al. (2021) provided a friendly review of the topics in Chapters 27 and 29.

29

Time-Varying Treatment and Confounding

Studies with time-varying treatments are common in biomedical and social sciences. James Robins championed the research in biostatistics. A classic example is that HIV patients may take the azidothymidine, an antiretroviral medication, on and off over time (Robins et al., 2000; Hernán et al., 2000). Similar problems also exist in other fields. In education, a classic example is that students may receive different types of instructions over time (Hong and Raudenbush, 2008). In political science, a classic example is that candidates continuously recalibrate their campaign strategy based on time-varying polls and opponent actions (Blackwell, 2013).

Causal inference with a time-varying treatment is not a simple extension of causal inference with a treatment at a single time point. The main challenge is time-varying confounding. Even if we assume all time-varying confounders are observed, we still face statistical challenges in adjusting for those confounders. On the one hand, we should stratify on these confounders to adjust for confounding; on the other hand, stratifying on post-treatment variables will cause bias. Due to these two conflicting goals, causal inference with time-varying treatments and confounding requires more sophisticated statistical methods. It is the main topic of this chapter.

To minimize the notational burden, I will use the setting with a treatment at two time points to convey the most important ideas. Extensions to treatments at multiple time points can be conceptually straightforward although technical complexities will arise in finite samples. I will discuss the complexities and relegate general results to Problems 29.6–29.9.

29.1 Basic setup and sequential ignorability

Start with a treatment at two time points. The temporal order of the variables with two time points is below:

$$X_0 \rightarrow Z_1 \rightarrow X_1 \rightarrow Z_2 \rightarrow Y$$

where

- X_0 denotes the baseline pre-treatment covariates;

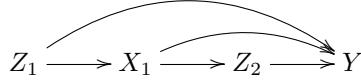


FIGURE 29.1: Without unmeasured confounding U between X_1 and Y . The causal diagram conditions on the pre-treatment covariates X_0 .

- Z_1 denotes the treatment at time point 1;
- X_1 denotes the time-varying covariates between the treatments at time points 1 and 2;
- Z_2 denotes the treatment at time point 2;
- Y denotes the outcome.

With binary treatment (Z_1, Z_2) , each unit has four potential outcomes

$$Y(z_1, z_2) \text{ for } z_1, z_2 = 0, 1.$$

The observed outcome equals

$$Y = Y(Z_1, Z_2) = \sum_{z_1=0,1} \sum_{z_2=0,1} 1(Z_1 = z_1) 1(Z_2 = z_2) Y(z_1, z_2).$$

I will focus on the canonical setting with sequential ignorability, that is, the treatments are sequentially randomized given the observed history.

Assumption 29.1 (sequential ignorability) (1) Z_1 is randomized given X_0 :

$$Z_1 \perp\!\!\!\perp Y(z_1, z_2) \mid X_0 \text{ for } z_1, z_2 = 0, 1.$$

(2) Z_2 is randomized given (Z_1, X_1, X_0) :

$$Z_2 \perp\!\!\!\perp Y(z_1, z_2) \mid (Z_1, X_1, X_0) \text{ for } z_1, z_2 = 0, 1.$$

Figure 29.1 is a simple causal diagram corresponding to Assumption 29.1, which does not contain any unmeasured confounding.

Figure 29.2 is a more complex causal diagram corresponding to Assumption 29.1. Sequential ignorability rules out only the confounding between the treatment (Z_1, Z_2) and the outcome Y , but allows for unmeasured confounding between the time-varying covariate X_1 and the outcome Y . The possible existence of U causes many subtle issues even under sequential ignorability.

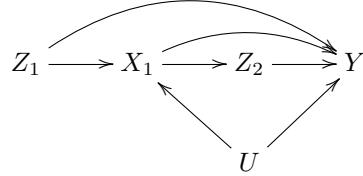


FIGURE 29.2: With unmeasured confounding between X_1 and Y . The causal diagram conditions on the pre-treatment covariates X_0 .

29.2 g-formula and outcome modeling

Recall the outcome-based identification formula with a treatment at a single time point:

$$E\{Y(z)\} = E\{E(Y | Z = z, X)\}.$$

With discrete X , it reduces to

$$E\{Y(z)\} = \sum_x E(Y | Z = z, X = x) \text{pr}(X = x);$$

with continuous X , it reduces to

$$E\{Y(z)\} = \int E(Y | Z = z, X = x) f_X(x) dx.$$

The following result extends it to the setting with a treatment at two time points.

Theorem 29.1 *Under Assumption 29.1,*

$$E\{Y(z_1, z_2)\} = E\left[E\{E(Y | z_2, z_1, X_1, X_0) | z_1, X_0\}\right]. \quad (29.1)$$

In Theorem 29.1, I simplify the notation ‘ $Z_2 = z_2$ ’ to ‘ z_2 ’ for simplicity. To void complex formulas in this Chapter, I will use the lower case letter to represent the event that the random variable takes the corresponding value. With discrete X_0 and X_1 , the identification formula (29.1) reduces to

$$E\{Y(z_1, z_2)\} = \sum_{x_0} \sum_{x_1} E(Y | z_2, z_1, x_1, x_0) \text{pr}(x_1 | z_1, x_0) \text{pr}(x_0); \quad (29.2)$$

with continuous X_0 and X_1 , the identification formula (29.1) reduces to

$$E\{Y(z_1, z_2)\} = \int \int E(Y | z_2, z_1, x_1, x_0) f(x_1 | z_1, x_0) f(x_0) dx_1 dx_0. \quad (29.3)$$

Compare (29.2) with the formula based on the law of total probability to gain more insights:

$$\begin{aligned} E(Y) &= \sum_{x_0} \sum_{z_1} \sum_{x_1} \sum_{z_2} E(Y | z_2, z_1, x_1, x_0) \\ &\quad \text{pr}(z_1 | z_2, x_1, x_0) \text{pr}(x_1 | z_2, x_0) \text{pr}(z_2 | x_0) \text{pr}(x_0). \end{aligned} \quad (29.4)$$

Erasing the probabilities of z_2 and z_1 in (29.4), we can obtain the formula (29.3). This is intuitive because the potential outcome $Y(z_1, z_2)$ has the meaning of fixing Z_1 and Z_2 at z_1 and z_2 , respectively.

Robins called (29.2) and (29.3) the g-formulas. Now I will prove Theorem 29.1.

Proof of Theorem 29.1: By the tower property,

$$E\{Y(z_1, z_2)\} = E\left[E\{Y(z_1, z_2) | X_0\}\right],$$

so we focus on $E\{Y(z_1, z_2) | X_0\}$. By Assumption 29.1(1) and the tower property,

$$\begin{aligned} E\{Y(z_1, z_2) | X_0\} &= E\{Y(z_1, z_2) | z_1, X_0\} \\ &= E\left[E\{Y(z_1, z_2) | z_1, X_1, X_0\} | z_1, X_0\right]. \end{aligned}$$

By Assumption 29.1(2),

$$\begin{aligned} E\{Y(z_1, z_2) | X_0\} &= E\left[E\{Y(z_1, z_2) | z_2, z_1, X_1, X_0\} | z_1, X_0\right] \\ &= E\left[E\{Y | z_2, z_1, X_1, X_0\} | z_1, X_0\right]. \end{aligned}$$

The formula (29.1) follows. \square

29.2.1 Plug-in estimation based on outcome modeling

The g-formulas (29.2) and (29.3) suggest that to estimate the means of the potential outcomes, we need to model $E(Y | z_2, z_1, x_1, x_0)$, $\text{pr}(x_1 | z_1, x_0)$ and $\text{pr}(x_0)$. With these fitted models, we can plug them in the g-formulas.

With some special functional forms, this task can be simplified. Example 29.1 below gives the results under a linear model for the outcome.

Example 29.1 Assume a linear outcome model

$$E(Y | z_2, z_1, x_1, x_0) = \beta_0 + \beta_1 z_2 + \beta_2 z_1 + \beta_3 x_1 + \beta_4 x_0.$$

We can verify that

$$\begin{aligned} E\{Y(z_1, z_2)\} &= \sum_{x_0} \sum_{x_1} (\beta_0 + \beta_1 z_2 + \beta_2 z_1 + \beta_3 x_1 + \beta_4 x_0) \text{pr}(x_1 | z_1, x_0) \text{pr}(x_0) \\ &= \beta_0 + \beta_1 z_2 + \beta_2 z_1 + \beta_3 \sum_{x_0} E(X_1 | z_1, x_0) \text{pr}(x_0) + \beta_4 E(X_0). \end{aligned}$$

Define

$$E\{X_1(z_1)\} = \sum_{x_0} E(X_1 | z_1, x_0) \text{pr}(x_0) \quad (29.5)$$

to simplify the formula as

$$E\{Y(z_1, z_2)\} = \beta_0 + \beta_1 z_2 + \beta_2 z_1 + \beta_3 E\{X_1(z_1)\} + \beta_4 E(X_0).$$

In (29.5), I introduce the potential outcome of X_1 under the treatment $Z_1 = z_1$ at time point 1. It is reasonable because the right-hand side of (29.5) is the identification formula of $E\{X_1(z_1)\}$ under ignorability $X_1(z_1) \perp\!\!\!\perp Z_1 | X_0$ for $z_1 = 0, 1$. We do not really need the potential outcome $X_1(z_1)$ and the ignorability, but it is a convenient notation and matches with our discussion before.

Define $\tau_{Z_1 \rightarrow X_1} = E\{X_1(1) - X_1(0)\}$. We can verify that

$$\begin{aligned} E\{Y(1, 0) - Y(0, 0)\} &= \beta_2 + \beta_3 \tau_{Z_1 \rightarrow X_1}, \\ E\{Y(0, 1) - Y(0, 0)\} &= \beta_1, \\ E\{Y(1, 1) - Y(0, 0)\} &= \beta_1 + \beta_2 + \beta_3 \tau_{Z_1 \rightarrow X_1}. \end{aligned}$$

Therefore, we can estimate the effect of (Z_1, Z_2) on Y based on the above formulas by first estimating the regression coefficients β s and the average causal effect of Z_1 on X_1 using standard methods.

However, Robins and Wasserman (1997) pointed out a surprising drawback of the plug-in estimation based on outcome modeling. They showed that with model misspecification in this strategy, data analyzers may falsely reject the null hypothesis of zero causal effect of (Z_1, Z_2) on Y even when the true effect is zero in the data-generating process. They called it the *g-null paradox*. Perhaps surprisingly, they show that the g-null paradox may even arise in the simple linear outcome model in Example 29.1. McGrath et al. (2021) revisited this paradox. See Problem 29.1 for more details.

29.2.2 Recursive estimation based on outcome modeling

The plug-in estimation in Section 29.2.1 involves modeling the time-varying confounder X_1 and causes the unpleasant g-null paradox. It is not a desirable method.

Recall the outcome regression estimator with a treatment at a single time based on $E\{Y(z)\} = E\{E(Y | Z = z, X)\}$. We first fit a model of Y on X using the subset of the data with $Z = z$, and obtain the fitted values $\hat{Y}_i(z)$ for all units. We then obtain the estimator

$$\hat{E}\{Y(z)\} = n^{-1} \sum_{i=1}^n \hat{Y}_i(z).$$

Similarly, the recursive expectation formula in (29.1) motivates a simpler

method for estimation. Start from the inner conditional expectation, denoted by

$$\tilde{Y}_2(z_1, z_2) = E(Y | Z_2 = z_2, Z_1 = z_1, X_1, X_0).$$

We can fit a model of Y on (X_1, X_0) using the subset of the data with $(Z_2 = z_2, Z_1 = z_1)$, and obtain the fitted values $\hat{Y}_{2i}(z_1, z_2)$ for all units. Move on to outer conditional expectation, denoted by

$$\tilde{Y}_1(z_1, z_2) = E\{\tilde{Y}_2(z_1, z_2) | Z_1 = z_1, X_0\}.$$

We can fit a model of $\hat{Y}_2(z_1, z_2)$ on X_0 using the subset of data with $Z_1 = z_1$, and obtain the fitted values $\hat{Y}_{1i}(z_1, z_2)$ for all units. The final estimator for $E\{Y(z_1, z_2)\}$ is then

$$\hat{E}\{Y(z_1, z_2)\} = n^{-1} \sum_{i=1}^n \hat{Y}_{1i}(z_1, z_2).$$

The above recursive estimation does not involve fitting a model for X_1 and avoids the g-null paradox. See Problem 29.2 for a special case.

29.3 Inverse propensity score weighting

Recall the propensity-score-based identification formula with a treatment at a single time point:

$$E\{Y(z)\} = E\left\{ \frac{1(Z=z)Y}{\text{pr}(Z=z | X)} \right\}.$$

The following result extends it to the setting with a treatment at two time points. Define

$$e(z_1, X_0) = \text{pr}(Z_1 = z_1 | X_0)$$

and

$$e(z_2, Z_1, X_1, X_0) = \text{pr}(Z_2 = z_2 | Z_1, X_1, X_0)$$

as the propensity scores at time points 1 and 2, respectively.

Theorem 29.2 *Under Assumption 29.1,*

$$E\{Y(z_1, z_2)\} = E\left\{ \frac{1(Z_1 = z_1)1(Z_2 = z_2)Y}{e(z_1, X_0)e(z_2, Z_1, X_1, X_0)} \right\}. \quad (29.6)$$

Theorem 29.2 reveals the omitted overlap assumption:

$$0 < e(z_1, X_0) < 1, \quad 0 < e(z_2, Z_1, X_1, X_0) < 1$$

for all z_1 and z_2 . If some propensity scores are 0 or 1, then the identification formula (29.6) blows up to infinity.

Proof of Theorem 29.2: Conditioning on (Z_1, X_1, X_0) and using Assumption 29.1(2), we can simplify the right-hand side of (29.6) as

$$\begin{aligned} & E \left\{ \frac{1(Z_1 = z_1)1(Z_2 = z_2)Y(z_1, z_2)}{\text{pr}(Z_1 = z_1 | X_0)\text{pr}(Z_2 = z_2 | Z_1, X_1, X_0)} \right\} \\ &= E \left\{ \frac{1(Z_1 = z_1)\text{pr}(Z_2 = z_2 | Z_1, X_1, X_0)E(Y(z_1, z_2) | Z_1, X_1, X_0)}{\text{pr}(Z_1 = z_1 | X_0)\text{pr}(Z_2 = z_2 | Z_1, X_1, X_0)} \right\} \\ &= E \left\{ \frac{1(Z_1 = z_1)}{\text{pr}(Z_1 = z_1 | X_0)}E(Y(z_1, z_2) | Z_1, X_1, X_0) \right\} \\ &= E \left\{ \frac{1(Z_1 = z_1)}{\text{pr}(Z_1 = z_1 | X_0)}Y(z_1, z_2) \right\}, \end{aligned} \quad (29.7)$$

where (29.7) follows from the tower property.

Conditioning on X_0 and using Assumption 29.1(1), we can simplify the right-hand side of (29.7) as

$$\begin{aligned} & E \left\{ \frac{\text{pr}(Z_1 = z_1 | X_0)}{\text{pr}(Z_1 = z_1 | X_0)}E(Y(z_1, z_2) | X_0) \right\} \\ &= E\{E(Y(z_1, z_2) | X_0)\} \\ &= E\{Y(z_1, z_2)\}, \end{aligned}$$

where, again, the last line follows from the tower property. \square

The estimator based on IPW is much simpler which only involves modeling two binary indicators. First, we can fit a model of Z_1 on X_0 to obtain the fitted values $\hat{e}_1(z_1, X_{0i})$ and fit a model of Z_2 on (Z_1, X_1, X_0) to obtain the fitted values $\hat{e}_2(z_2, Z_{1i}, X_{1i}, X_{0i})$ for all units. Then, we obtain the following IPW estimator:

$$\hat{E}^{\text{ht}}\{Y(z_1, z_2)\} = n^{-1} \sum_{i=1}^n \frac{1(Z_{1i} = z_1)1(Z_{2i} = z_2)Y_i}{\hat{e}_1(z_1, X_{0i})\hat{e}_2(z_2, Z_{1i}, X_{1i}, X_{0i})}.$$

Similar to the discussion in Chapter 11, the Horvitz–Thompson-type estimator is not invariant to location shift of the outcome and suffers from instability in finite samples. A modified Hajek-type estimator is $\hat{E}^{\text{haj}}\{Y(z_1, z_2)\} = \hat{E}^{\text{ht}}\{Y(z_1, z_2)\}/\hat{1}^{\text{ht}}(z_1, z_2)$, where

$$\hat{1}^{\text{ht}}(z_1, z_2) = n^{-1} \sum_{i=1}^n \frac{1(Z_{1i} = z_1)1(Z_{2i} = z_2)}{\hat{e}_1(z_1, X_{0i})\hat{e}_2(z_2, Z_{1i}, X_{1i}, X_{0i})}.$$

29.4 Multiple time points

Extending the estimation strategies in Sections 29.2 and 29.3 is not immediate with multiple time points. Even with a binary treatment and K time points, the number of treatment combination grows exponentially with K (for example, $2^5 = 32$ and $2^{10} = 1024$). Consequently, the outcome regression and IPW estimators in Sections 29.2 and 29.3 are not feasible in finite samples.

29.4.1 Marginal structural model

A powerful approach is based on the marginal structural model (MSM) (Robins et al., 2000; Hernán et al., 2000). For simplicity of notation, I will only present the MSM with $K = 2$ although its main use is in the general case.

Definition 29.1 (MSM) *The marginal mean of $Y(z_1, z_2)$ equals*

$$E\{Y(z_1, z_2)\} = f(z_1, z_2; \beta).$$

A leading example of Definition 29.1 is $E\{Y(z_1, z_2)\} = \beta_0 + \beta_1 z_1 + \beta_2 z_2$. It is also straightforward to include the baseline covariates in the model. Definition 29.2 below extends Definition 29.1.

Definition 29.2 (MSM with baseline covariates) *The mean of $Y(z_1, z_2)$ conditional on X_0 equals*

$$E\{Y(z_1, z_2) | X_0\} = f(z_1, z_2, X_0; \beta).$$

A leading example of Definition 29.2 is

$$E\{Y(z_1, z_2) | X_0\} = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3^\top X_0. \quad (29.8)$$

If we observe all potential outcomes, we can solve β from the following minimization problem:

$$\beta = \arg \min_b \sum_{z_2} \sum_{z_1} E\{Y(z_1, z_2) - f(z_1, z_2, X_0; b)\}^2.$$

For simplicity, I focus on the least squares formulation. We can also extend the discussion to a general loss function.

Under sequential ignorability, we can solve β from the following minimization problem that only involves observables.

Theorem 29.3 (IPW under MSM) *Under Assumption 29.1 and Definition 29.2,*

$$\beta = \arg \min_b \sum_{z_2} \sum_{z_1} E \left[\frac{1(Z_1 = z_1)1(Z_2 = z_2)}{e(z_1, X_0)e(z_2, Z_1, X_1, X_0)} \{Y - f(z_1, z_2, X_0; b)\}^2 \right].$$

The proof of Theorem 29.3 is similar to that of Theorem 29.2. I relegate it to Problem 29.3.

Theorem 29.3 implies a simple estimation strategy based on weighted regressions. For instance, under (29.8), we can fit WLS of Y_i on $(1, Z_{1i}, Z_{2i}, X_{0i})$ with weights $\hat{e}_1^{-1}(Z_{1i}, X_{0i})\hat{e}_{2i}^{-1}(Z_{2i}, Z_{i1}, X_{1i}, X_{0i})$.

29.4.2 Structural nested model

A key problem of IPW is that it is not applicable if the overlap assumption is violated. To address this challenge, Robins proposed the structural nested model. Again, to simplify the presentation, I only review the version with two time points.

Definition 29.3 (structural nested model) *The conditional effect at time point 1 is*

$$E\{Y(z_1, 0) - Y(0, 0) \mid Z_1 = z_1, X_0\} = g_1(z_1, X_0; \beta) \text{ for all } z_1$$

and the conditional effect at time point 2 is

$$E\{Y(z_1, z_2) - Y(z_1, 0) \mid Z_2 = z_2, Z_1 = z_1, X_1, X_0\} = g_2(z_2, z_1, X_1, X_0; \beta) \text{ for all } z_1, z_2.$$

In Definition 29.3, two logical restrictions are

$$g_1(0, X_0; \beta) = 0$$

and

$$g_2(0, z_1, X_1, X_0; \beta) = 0 \text{ for all } z_1.$$

Two leading choices of Definition 29.3 are below.

Example 29.2 Assume

$$\begin{cases} g_1(z_1, X_0; \beta) = \beta_1 z_1, \\ g_2(z_2, z_1, X_1, X_0; \beta) = (\beta_2 + \beta_3 z_1) z_2. \end{cases}$$

Example 29.3 Assume

$$\begin{cases} g_1(z_1, X_0; \beta) = (\beta_1 + \beta_4^T X_0) z_1, \\ g_2(z_2, z_1, X_1, X_0; \beta) = (\beta_2 + \beta_3 z_1 + \beta_4^T X_1) z_2. \end{cases}$$

Compare Definitions 29.2 and 29.3. The structural nested model allows for adjusting for the time-varying covariates whereas the marginal structural model only allows for adjusting for baseline covariates. The estimation under Definition 29.3 is more involved. A strategy is to estimate the parameter based on estimating equations.

I first introduce two important building blocks for the discussing the estimation. Define

$$U_2(\beta) = Y - g_2(Z_2, Z_1, X_1, X_0; \beta)$$

and

$$U_1(\beta) = Y - g_2(Z_2, Z_1, X_1, X_0; \beta) - g_1(Z_1, X_0; \beta).$$

They are not directly computable from the data because they depend on the true value of the parameter β . At the true value, they have the following properties.

Lemma 29.1 *Under Assumption 29.1 and Definition 29.3, we have*

$$\begin{aligned} E\{U_2(\beta) \mid Z_2, Z_1, X_1, X_0\} &= E\{U_2(\beta) \mid Z_1, X_1, X_0\} \\ &= E\{Y(Z_1, 0) \mid Z_1, X_1, X_0\} \end{aligned}$$

and

$$\begin{aligned} E\{U_1(\beta) \mid Z_1, X_0\} &= E\{U_1(\beta) \mid X_0\} \\ &= E\{Y(0, 0) \mid X_0\}. \end{aligned}$$

Lemma 29.1 involves a subtle notation $Y(Z_1, 0)$ because Z_1 is random. It should be read as $Y(Z_1, 0) = Z_1 Y(1, 0) + (1 - Z_1) Y(0, 0)$. Based on the definitions and Lemma 29.1, $U_1(\beta)$ acts as the control potential outcome before receiving any treatment and $U_2(\beta)$ acts as the control potential outcome after receiving the treatment at time point 1.

Proof of Lemma 29.1: First, we have

$$\begin{aligned} E\{U_2(\beta) \mid Z_2 = 1, Z_1, X_1, X_0\} &= E\{Y(Z_1, 1) - g_2(1, Z_1, X_1, X_0; \beta) \mid Z_2 = 1, Z_1, X_1, X_0\} \\ &= E\{Y(Z_1, 0) \mid Z_2 = 1, Z_1, X_1, X_0\} \\ E\{U_2(\beta) \mid Z_2 = 0, Z_1, X_1, X_0\} &= E\{Y(Z_1, 0) - g_2(0, Z_1, X_1, X_0; \beta) \mid Z_2 = 0, Z_1, X_1, X_0\} \\ &= E\{Y(Z_1, 0) \mid Z_2 = 0, Z_1, X_1, X_0\} \end{aligned}$$

so

$$E\{U_2(\beta) \mid Z_2, Z_1, X_1, X_0\} = E\{Y(Z_1, 0) \mid Z_2, Z_1, X_1, X_0\} = E\{Y(Z_1, 0) \mid Z_1, X_1, X_0\}$$

where the last identity follows from sequential ignorability. Since the last term does not depend on Z_2 , we also have

$$E\{U_2(\beta) \mid Z_2, Z_1, X_1, X_0\} = E\{U_2(\beta) \mid Z_1, X_1, X_0\}.$$

Using the above results, we have

$$\begin{aligned} E\{U_1(\beta) \mid Z_1, X_0\} &= E\{U_2(\beta) - g_1(Z_1, X_0; \beta) \mid Z_1, X_0\} \\ &= E[E\{U_2(\beta) - g_1(Z_1, X_0; \beta) \mid X_1, Z_1, X_0\} \mid Z_1, X_0] \\ &= E[E\{Y(Z_1, 0) - g_1(Z_1, X_0; \beta) \mid X_1, Z_1, X_0\} \mid Z_1, X_0] \\ &= E\{Y(Z_1, 0) - g_1(Z_1, X_0; \beta) \mid Z_1, X_0\} \\ &= E\{Y(0, 0) \mid Z_1, X_0\} \\ &= E\{Y(0, 0) \mid X_0\} \end{aligned}$$

where the last identity follows from sequential ignorability. Since the last term does not depend on Z_1 , we also have

$$E\{U_1(\beta) | Z_1, X_0\} = E\{U_1(\beta) | X_0\}.$$

□

With Lemma 29.1, we can prove Theorem 29.4 below.

Theorem 29.4 *Under Assumption 29.1 and Definition 29.3,*

$$E\left[h_2(Z_1, X_1, X_0)\{Z_2 - e(1, Z_1, X_1, X_0)\}U_2(\beta)\right] = 0$$

and

$$E\left[h_1(X_0)\{Z_1 - e(1, X_0)\}U_1(\beta)\right] = 0.$$

for any functions h_1 and h_2 , provided that the moments exist.

Proof of Theorem 29.2: Use the tower property by conditioning on (Z_2, Z_1, X_1, X_0) and Lemma 29.1 to obtain

$$\begin{aligned} & E[h_2(Z_1, X_1, X_0)\{Z_2 - e(1, Z_1, X_1, X_0)\}E\{U_2(\beta) | Z_2, Z_1, X_1, X_0\}] \\ &= E[h_2(Z_1, X_1, X_0)\{Z_2 - e(1, Z_1, X_1, X_0)\}E\{U_2(\beta) | Z_1, X_1, X_0\}]. \end{aligned}$$

Use the tower property by conditioning on (Z_1, X_1, X_0) to show that the last identity equals 0.

Similarly, use the tower property by conditioning on (Z_1, X_0) and Lemma 29.1 to obtain

$$\begin{aligned} & E[h_1(X_0)\{Z_1 - e(1, X_0)\}E\{U_1(\beta) | Z_1, X_0\}] \\ &= E[h_1(X_0)\{Z_1 - e(1, X_0)\}E\{U_1(\beta) | X_0\}]. \end{aligned}$$

Use the tower property by conditioning on X_0 to show that the last identity equals 0. □

To use Theorem 29.4, we must specify h_1 and h_2 to ensure that there are enough equations for solving β . Example 29.4 below revisits Example 29.2.

Example 29.4 *Under Example 29.2, we can choose $h_1 = 1$ and $h_2 = (1, Z_1)$ to obtain*

$$\begin{aligned} E[\{Z_2 - e(1, Z_1, X_1, X_0)\}\{Y - (\beta_2 + \beta_3 Z_1)Z_2\}] &= 0, \\ E[Z_1\{Z_2 - e(1, Z_1, X_1, X_0)\}\{Y - (\beta_2 + \beta_3 Z_1)Z_2\}] &= 0, \\ E[\{Z_1 - e(1, X_0)\}\{Y - (\beta_2 + \beta_3 Z_1)Z_2 - \beta_1 Z_1\}] &= 0. \end{aligned}$$

We can then solve for the β 's from the above linear equations; see Problem 29.5. A natural question is that whether alternative choices of (h_1, h_2) can lead to more efficient estimators. The answer is yes. For example, we can choose many (h_1, h_2) and use the generalized method of moment (Hansen, 1982). The technical details are beyond this book.

Naimi et al. (2017) and Vansteelandt and Joffe (2014) provided tutorials on the structural nested models.

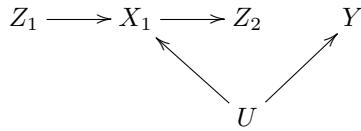


FIGURE 29.3: With unmeasured confounding between X_1 and Y . The causal diagram ignores the pre-treatment covariates X_0 .

29.5 Homework problems

29.1 *g-null paradox*

Consider the simple causal diagram in Figure 29.3 without pre-treatment covariates X_0 and without the arrows from (Z_1, Z_2) to Y . So the effect of (Z_1, Z_2) on Y is zero.

Revisit Example 29.1. Show that the expectation $E\{Y(z_1, z_2)\}$ does not depend on (z_1, z_2) if

$$\beta_1 = \beta_3 = 0 \text{ and } \beta_2 = 0$$

or

$$\beta_1 = \beta_3 = 0 \text{ and } E\{X_1(z_1)\} \text{ does not depend on } z_1.$$

holds.

Remark: However, $\beta_2 = 0$ in the first condition rules out the dependence of Y on X_1 , contradicting with the existence of unmeasured confounder U between X_1 and Y ; the independence of $E\{X_1(z_1)\}$ on z_1 rules out the dependence of X_1 on Z_1 , contradicting with the existence of the arrow from Z_1 on X_1 . That is, if there is an unmeasured confounder U between X_1 and Y and there is an arrow from Z_1 on X_1 , then the formula of $E\{Y(z_1, z_2)\}$ in Example 29.1 must depend on (z_1, z_2) , which leads to a contradiction with the absence of arrows from (Z_1, Z_2) to Y .

29.2 *Recursive estimation under the null model*

Consider the recursive estimation method in 29.2.2 under the causal diagram in Problem 29.1. Show that based linear models, the estimator converges to 0.

29.3 *IPW under MSM*

Prove Theorem 29.3.

29.4 Structural nested model with a single time point

Recall the standard setting of observational studies with IID data draw from $\{X, Z, Y(1), Y(0)\}$. Define the propensity score as $e(X) = \text{pr}(Z = 1 | X)$. Assume

$$Z \perp\!\!\!\perp Y(0) | X$$

and the following structural nested model.

Definition 29.4 (structural nested model with a single time point) *The conditional mean of the individual effect is*

$$E\{Y(z) - Y(0) | Z = z, X\} = g(z, X; \beta).$$

In Definition 29.4, a logical restriction is $g(0, X; \beta) = 0$. Prove the following results.

1. We have

$$E\{Y - g(Z, X; \beta) | X, Z\} = E\{Y - g(Z, X; \beta) | X\} = E\{Y(0) | X\}.$$

2. We have

$$E\left[h(X)\{Z - e(X)\}\{Y - g(Z, X; \beta)\}\right] = 0 \quad (29.9)$$

for any function h , provided that the moment exists.

Remark: (29.9) is the basis for parameter estimation. Consider a special case of Definition 29.4 with $g(z, X; \beta) = \beta z$. Choose $h(X) = 1$ to obtain

$$E\{(Z - e(X))(Y - \beta Z)\} = 0.$$

Solve for β to obtain

$$\beta = \frac{E\{(Z - e(X))Y\}}{E\{(Z - e(X))Z\}}.$$

That is, β equals the coefficient of Z in the two-stage least squares of Y on Z with $Z - e(X)$ being the instrument variable for Z .

Consider a special case of Definition 29.4 with $g(z, X; \beta) = (\beta_0 + \beta_1^T X)z$. Choose $h(X) = (1, X)$ to obtain

$$E\left\{\left(\frac{Z - e(X)}{(Z - e(X))X}\right)(Y - \beta_0 Z - \beta_1^T XZ)\right\} = 0.$$

That is, (β_0, β_1) equal the coefficients in the two-stage least squares of Y on (Z, XZ) with $(Z - e(X), (Z - e(X))X)$ being the instrument variable for (Z, XZ) .

29.5 Estimation under Example 29.4

We can estimate the β 's by solving the empirical version of the estimating equations in Example 29.4. We first estimate the two propensity scores and obtain the centered treatment

$$\check{Z}_{1i} = Z_{1i} - \hat{e}(1, X_{0i})$$

at time point 1 and

$$\check{Z}_{2i} = Z_{2i} - \hat{e}(1, Z_{1i}, X_{1i}, X_{0i})$$

at time point 2.

Show that we can estimate β_2 and β_3 by running two-stage least squares of Y_i on $(Z_{2i}, Z_{1i}Z_{2i})$ with $(\check{Z}_{2i}, Z_{1i}\check{Z}_{2i})$ as the instrumental variable for $(Z_{2i}, Z_{1i}Z_{2i})$, and then we can estimate β_1 by running two-stage least squares of $Y_i - (\hat{\beta}_2 + \hat{\beta}_3 Z_{1i})Z_{2i}$ on Z_{1i} with \check{Z}_{1i} as the instrumental variable for Z_{1i} .

29.6 g-formula with a treatment at multiple time points

Extend the discussion to the setting with K time points. The temporal ordering of the variables is

$$X_0 \rightarrow Z_1 \rightarrow X_1 \rightarrow Z_2 \rightarrow \cdots \rightarrow X_{K-1} \rightarrow Z_K.$$

Introduce the notation $\bar{Z}_k = (Z_1, \dots, Z_k)$ and $\bar{X}_k = (X_0, X_1, \dots, X_k)$ with lower case \bar{z}_k and \bar{x}_k denoting the corresponding realized values. With $k = 0$, we have $\bar{X}_0 = X_0$ and \bar{Z}_0 is empty. Each unit has 2^K potential outcomes:

$$Y(\bar{z}_K) \text{ for all } z_1, \dots, z_K = 0, 1.$$

Assume sequential ignorability below.

Assumption 29.2 (sequential ignorability at multiple time points) *We have*

$$Z_k \perp\!\!\!\perp Y(\bar{z}_K) \mid (\bar{Z}_{k-1}, \bar{X}_{k-1})$$

for all $k = 1, \dots, K$ and all $z_1, \dots, z_K = 0, 1$.

Prove Theorem 29.5 below.

Theorem 29.5 (g-formula with multiple time points) *Under Assumption 29.2,*

$$E\{Y(\bar{z}_K)\} = E\left[\cdots E\{E(Y \mid \bar{z}_K, \bar{X}_{K-1}) \mid \bar{z}_{K-1}, \bar{X}_{K-2}\} \cdots \mid z_1, X_0\right].$$

Remark: In Theorem 29.5, I use the simplified notation “ \bar{z}_k ” for “ $\bar{Z}_k = \bar{z}_k$.” With discrete X , Theorem 29.5 reduces to

$$\begin{aligned} E\{Y(\bar{z}_K)\} &= \sum_{x_0} \sum_{x_1} \cdots \sum_{x_{K-1}} E(Y \mid \bar{z}_K, \bar{x}_{K-1}) \\ &\quad \cdot \text{pr}(x_{K-1} \mid \bar{z}_{K-1}, \bar{x}_{K-2}) \cdots \text{pr}(x_1 \mid z_1, x_0) \text{pr}(x_0); \end{aligned}$$

with continuous X , Theorem 29.5 reduces to

$$\begin{aligned} E\{Y(\bar{z}_K)\} &= \int E(Y | \bar{z}_K, \bar{x}_{K-1}) \\ &\quad \cdot f(x_{K-1} | \bar{z}_{K-1}, \bar{x}_{K-2}) \cdots f(x_1 | z_1, x_0) f(x_0) d\bar{x}_{K-1}. \end{aligned}$$

29.7 IPW with a treatment at multiple time points

Inherit the setting of Problem 29.6. Define the propensity score at K time points as

$$\begin{aligned} e(z_1, X_0) &= \text{pr}(Z_1 = z_1 | X_0), \\ &\vdots \\ e(z_k, \bar{Z}_{k-1}, \bar{X}_{k-1}) &= \text{pr}(Z_k = z_k | \bar{Z}_{k-1}, \bar{X}_{k-1}), \\ &\vdots \\ e(z_K, \bar{Z}_{K-1}, \bar{X}_{K-1}) &= \text{pr}(Z_K = z_K | \bar{Z}_{K-1}, \bar{X}_{K-1}). \end{aligned}$$

Prove Theorem 29.7 below assuming overlap implicitly.

Theorem 29.6 (IPW with multiple time points) *Under Assumption 29.2,*

$$E\{Y(\bar{z}_K)\} = E\left\{\frac{1(Z_1 = z_1) \cdots 1(Z_K = z_K) Y}{e(z_1, X_0) \cdots e(z_K, \bar{Z}_{K-1}, \bar{X}_{K-1})}\right\}.$$

Based on Theorem 29.7, construct the Horvitz–Thompson and Hajek estimators.

29.8 MSM with a treatment at multiple time points

The number of potential outcomes grows exponentially with K . The formulas in Problems 29.6 and 29.7 are not directly applicable in finite samples. We can impose the following structural assumptions on the potential outcomes.

Definition 29.5 (MSM with multiple time points) *Assume*

$$E\{Y(\bar{z}_K) | X_0\} = f(\bar{z}_K, X_0; \beta).$$

Two leading examples of Definition 29.5 are $E\{Y(\bar{z}_K) | X_0\} = \beta_0 + \beta_1 \sum_{k=1}^K z_k + \beta_2^\top X_0$ and $E\{Y(\bar{z}_K) | X_0\} = \beta_0 + \sum_{k=1}^K \beta_k z_k + \beta_{K+1}^\top X_0$.

If we know all the potential outcomes, we can solve β from the following minimization problem:

$$\beta = \arg \min_b \sum_{\bar{z}_K} E\{Y(\bar{z}_K) - f(\bar{z}_K, X_0; \beta)\}^2.$$

Theorem 29.7 below shows that under Assumption 29.2, we can solve β from a minimization problem that only involves observables.

Theorem 29.7 (IPW for MSM with multiple time points) *Under Assumption 29.2,*

$$\beta = \arg \min_b \sum_{\bar{z}_K} E \left[\frac{1(Z_1 = z_1) \cdots 1(Z_K = z_K)}{e(z_1, X_0) \cdots e(z_K, \bar{Z}_{K-1}, \bar{X}_{K-1})} \{Y - f(\bar{z}_K, X_0; \beta)\}^2 \right].$$

29.9 Structural nested model with a treatment at multiple time points

Inherit the setting from Problem 29.6 and the notation from Problem 29.7. This problem presents a general structural nested model.

Definition 29.6 (structural nested model with multiple time points) *The conditional effect at time k is*

$$E\{Y(\bar{z}_k, 0) - Y(\bar{z}_{k-1}, 0) \mid \bar{z}_k, \bar{X}_{k-1}\} = g_k(\bar{z}_k, \bar{X}_{k-1}; \beta)$$

for all \bar{z}_k and all $k = 1, \dots, K$.

In Definition 29.6, a logical restriction is

$$g_k(0, \bar{z}_{k-1}, \bar{X}_{k-1}; \beta) = 0$$

for all \bar{z}_{k-1} and all $k = 1, \dots, K$.

Define

$$U_k(\beta) = Y - \sum_{s=1}^k g_s(\bar{Z}_s, \bar{X}_{s-1}; \beta)$$

for all $k = 1, \dots, K$. Theorem 29.8 below extends Theorem 29.4.

Theorem 29.8 *Under Assumption 29.2 and Definition 29.6,*

$$E [h_k(\bar{Z}_{k-1}, \bar{X}_{k-1}) \{Z_k - e(1, \bar{Z}_{k-1}, \bar{X}_{k-1})\} U_k(\beta)] = 0$$

for all $k = 1, \dots, K$.

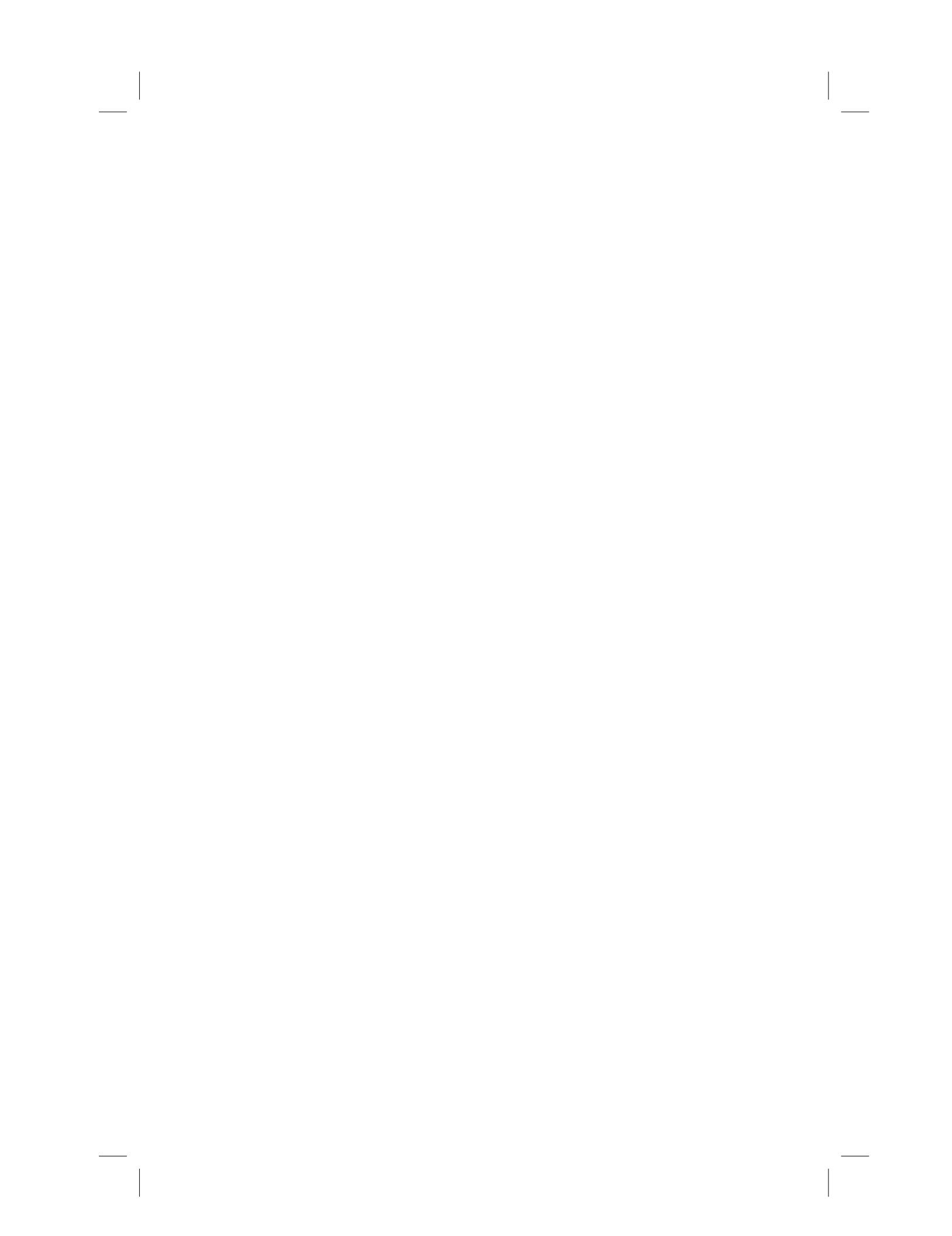
Remark: Choosing appropriate h_k 's, we can estimate β by solving the empirical version of Theorem 29.8.

29.10 Recommended reading

Robins et al. (2000) reviewed the MSM. Naimi et al. (2017) reviewed the g-methods.

Part VII

Appendices



A1

Probability and Statistics

A1.1 Probability

A1.1.1 Tower property and variance decomposition

Given random variables or vectors A, B, C , we have

$$E(A) = E\{E(A | B)\}$$

and

$$E(A | C) = E\{E(A | B, C) | C\}.$$

Given a random variable A and random variables or vectors B, C , we have

$$\text{var}(A) = E\{\text{var}(A | B)\} + \text{var}\{E(A | B)\}$$

and

$$\text{var}(A | C) = E\{\text{var}(A | B, C) | C\} + \text{var}\{E(A | B, C) | C\}.$$

Similarly, we can decompose the covariance as

$$\text{cov}(A_1, A_2) = E\{\text{cov}(A_1, A_2 | B)\} + \text{cov}\{E(A_1 | B), E(A_2 | B)\}$$

and

$$\text{cov}(A_1, A_2 | C) = E\{\text{cov}(A_1, A_2 | B, C) | C\} + \text{cov}\{E(A_1 | B, C), E(A_2 | B, C) | C\}.$$

A1.1.2 Limiting theorems

Definition A1.1 (convergence in probability) A sequence of random variables $(X_n)_{n \geq 1}$ converges to X in probability, if for every $\varepsilon > 0$, we have

$$\text{pr}(|X_n - X| > \varepsilon) \rightarrow 0$$

as $n \rightarrow \infty$.

Definition A1.2 (convergence in distribution) A sequence of random variables $(X_n)_{n \geq 1}$ converges to X in distribution, if

$$\text{pr}(X_n \leq x) \rightarrow \text{pr}(X \leq x)$$

for all continuity point x of $\text{pr}(X \leq x)$, as $n \rightarrow \infty$.

Convergence in probability is stronger than convergence in distribution. Definitions A1.1 and A1.2 are useful for stating the following two fundamental theorems in probability theory.

Theorem A1.1 (law of large numbers) *If $X_1, \dots, X_n \stackrel{IID}{\sim} X$ with $E|X| < \infty$, then $\bar{X} = n^{-1} \sum_{i=1}^n X_i \rightarrow E(X)$ in probability.*

The law of large numbers in Theorem A1.1 states that the sample average is close to the population mean in the limit.

Theorem A1.2 (central limit theorem) *If $X_1, \dots, X_n \stackrel{IID}{\sim} X$ with $\text{var}(X) < \infty$, then*

$$\frac{\bar{X} - E(X)}{\sqrt{\text{var}(X)/n}} \rightarrow N(0, 1)$$

in distribution.

The central limit theorem in Theorem A1.2 states that the standardized sample average is close to a standard Normal random variable in the limit.

Theorems A1.1 and A1.2 assume IID random variables for convenience. There are also many law of large numbers and central limit theorems for the sample mean of independent random variable (e.g., Durrett, 2019).

A1.1.3 Delta method

Delta method is a power tool to derive asymptotic Normality of nonlinear functions of an asymptotically Normal random vector. I review a special case of delta method below.

Theorem A1.3 (delta method) *Assume $\sqrt{n}(X_n - \mu) \rightarrow N(0, \Sigma)$ in distribution and the function $g(x)$ has non-zero derivative $\nabla g(\mu)$ at μ . Then*

$$\sqrt{n}\{g(X_n) - g(\mu)\} \rightarrow N(0, (\nabla g(\mu))^T \Sigma \nabla g(\mu))$$

in distribution.

I will omit the proof of Theorem A1.3. It is intuitive based on the first-order Taylor expansion:

$$g(X_n) - g(\mu) \cong (\nabla g(\mu))^T (X_n - \mu).$$

A leading example of delta method is to obtain the asymptotic Normality of ratio.

Example A1.1 (asymptotic normality for ratio) *Assume*

$$\sqrt{n} \begin{pmatrix} Y_n - \mu_Y \\ X_n - \mu_X \end{pmatrix} \rightarrow N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{YX} \\ \sigma_{YX} & \sigma_X^2 \end{pmatrix} \right) \quad (\text{A1.1})$$

in distribution with $\mu_X \neq 0$. Apply Theorem A1.3 to obtain that

$$\sqrt{n} \left(\frac{Y_n}{X_n} - \frac{\mu_Y}{\mu_X} \right) \rightarrow N \left(0, \frac{\sigma_Y^2}{\mu_X^2} + \frac{\mu_Y^2 \sigma_X^2}{\mu_X^4} - \frac{2\mu_Y \sigma_{YX}}{\mu_X^3} \right) \quad (\text{A1.2})$$

in distribution. In the special case that X_n and Y_n are asymptotically independent, the asymptotic variance of Y_n/X_n simplifies to $\sigma_Y^2/\mu_X^2 + \mu_Y^2 \sigma_X^2/\mu_X^4$. I leave the details to Problem A1.2.

The asymptotic variance in Example A1.1 is a little cumbersome. An easier way to memorize it is based on the following approximation:

$$\frac{Y_n}{X_n} - \frac{\mu_Y}{\mu_X} = \frac{Y_n - \mu_Y/\mu_X \cdot X_n}{X_n} \cong \frac{Y_n - \mu_Y/\mu_X \cdot X_n}{\mu_X}, \quad (\text{A1.3})$$

so the asymptotic variance of the ratio equals the asymptotic variance of

$$\frac{Y_n - \mu_Y/\mu_X \cdot X_n}{\mu_X},$$

which is a linear combination of Y_n and X_n . Slutsky's theorem can make the approximation in (A1.3) rigorous; it is beyond this book.

Example A1.2 (asymptotic normality for product) Assume (A1.1). Apply Theorem A1.3 to obtain that

$$\sqrt{n} (X_n Y_n - \mu_X \mu_Y) \rightarrow N(0, \mu_Y^2 \sigma_X^2 + \mu_X^2 \sigma_Y^2 + 2\mu_X \mu_Y \sigma_{XY}) \quad (\text{A1.4})$$

in distribution. In the special case that X_n and Y_n are asymptotically independent, the asymptotic variance of $X_n Y_n$ simplifies to $\mu_Y^2 \sigma_X^2 + \mu_X^2 \sigma_Y^2$. I leave the details to Problem A1.3.

A1.2 Statistical inference

A1.2.1 Point estimation

Assume that θ is the parameter of interest. Oftentimes, the problem also contain other parameters not of interest, denoted by η . Statisticians call η the nuisance parameter. Based on data, we can compute an estimator $\hat{\theta}$. Throughout this book, we take the frequentist's perspective by assuming that θ is a fixed number and $\hat{\theta}$ is random due to the randomness of data. Two basic requirements for an estimator are below.

Definition A1.3 (unbiasedness) The estimator $\hat{\theta}$ is unbiased for θ if

$$E(\hat{\theta}) = \theta$$

for all possible values of θ and η .

Definition A1.4 (consistency) *The estimator $\hat{\theta}$ is consistent for θ if*

$$\hat{\theta} \rightarrow \theta$$

in probability as the sample size approaches to infinity, for all possible values of θ and η .

Unbiasedness requires that the mean of the estimator is identical to the parameter of interest. Consistency requires that the estimator is close to the true parameter in the limit. Unbiased does not imply consistency, and consistency does not imply unbiasedness either. Unbiasedness can be restrictive because it is impossible even in some simple statistics problems. Consistency is often the basic requirement in most statistics problems.

A1.2.2 Confidence interval

A point estimator $\hat{\theta}$ is a random variable which differs from the true parameter. Statisticians are often interested in finding an interval that covers the true parameter with certain given probability. This interval is computed based on the data, and it is random.

Definition A1.5 (confidence interval) *A data-dependent interval $[\hat{\theta}_L, \hat{\theta}_U]$ is a confidence interval for θ with coverage probability $1 - \alpha$ if*

$$\text{pr}(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) \geq 1 - \alpha.$$

Definition A1.6 (asymptotic confidence interval) *A data-dependent interval $[\hat{\theta}_L, \hat{\theta}_U]$ is an asymptotic confidence interval for θ with coverage probability $1 - \alpha$ if*

$$\text{pr}(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) \rightarrow 1 - \alpha'$$

with $\alpha' \geq \alpha$, as $n \rightarrow \infty$.

A standard choice is $\alpha = 0.05$. In Definitions A1.5 and A1.6, the coverage probabilities can be larger than the nominal level $1 - \alpha$. That is, the definitions allow for over coverage but do not allow for under coverage. With over coverage, we say that the confidence interval is conservative. Of course, we hope the confidence interval to be as narrow as possible. Otherwise, the definition of confidence interval can be arbitrary.

A1.2.3 Hypothesis testing

Many applied problems can be formulated as testing a hypothesis:

$$H_0 : \theta = 0.$$

The decision rule ϕ is a binary function of the data: $\phi = 1$ if we reject H_0 ; $\phi = 0$ if we fail to reject H_0 . The type one error rate of the test is the probability of rejection if the null hypothesis holds. I review the definition below.

Definition A1.7 When H_0 holds, define the type one error rate of the test ϕ as the maximum possible value of the probability

$$\text{pr}(\phi = 1).$$

A standard choice is to make sure that the type one error rate is below $\alpha = 0.05$. The type two error rate of the test is the probability of no rejection if the null hypothesis does not hold. I review the definition below.

Definition A1.8 When H_0 does not hold, define the type two error rate of the test ϕ as the probability

$$\text{pr}(\phi = 0).$$

Given the control of the type one error rate under H_0 , we hope the type two error rate is as low as possible when H_0 does not hold.

A1.2.4 Wald-type confidence interval and test

Many statistics problems have the following structure. The parameter of interest is θ . We first find a consistent estimator $\hat{\theta}$ that converges in probability to θ , and show that it is asymptotically Normal with mean θ and variance v which may depends on θ as well as other parameters. We then find a consistent estimator \hat{v} for v , based on analytic formulas or the bootstrap reviewed in Chapter A1.5. We finally construct the Wald-type confidence interval for θ as

$$\hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{v}}$$

which covers θ with probability approximately $1 - \alpha$. When this interval excludes a particular c , for example, $c = 0$, we reject the null hypothesis $H_0(c) : \theta = c$, which is called the Wald test.

A1.2.5 Duality between constructing confidence sets and testing null hypotheses

Consider the statistical inference problem for a scalar parameter θ . A fundamental result in statistics is that constructing confidence sets for θ is equivalent to testing null hypotheses about θ . This is often called the duality between constructing confidence sets and testing null hypotheses.

Section A1.2.4 has reviewed the duality based on the Wald-type confidence interval and test. The duality also holds in general. Assume that $\hat{\Theta}$ is a $(1 - \alpha)$ -level confidence set for θ :

$$\text{pr}(\theta \in \hat{\Theta}) = 1 - \alpha.$$

Then we can reject the null hypothesis $H_0(c) : \theta = c$ if c is not in the set $\hat{\Theta}$. This is a valid test because when θ indeed equals c , we have correct type one error rate $\text{pr}(\theta \notin \hat{\Theta}) = \alpha$. Conversely, if we test a sequence of null hypotheses

$H_0(c) : \theta = c$, we can obtain the corresponding p -values, $p(c)$, as a function of c . Then the values of c that we fail to reject at level α form a confidence set for θ :

$$\hat{\Theta} = \{c : p(c) \geq \alpha\} = \{c : \text{fail to reject } H_0(c) \text{ at level } \alpha\}.$$

It is a valid confidence set because

$$\text{pr}(\theta \in \hat{\Theta}) = \text{pr}\{\text{fail to reject } H_0(\theta) \text{ at level } \alpha\} = 1 - \alpha.$$

Here I use “confidence set” instead of “confidence interval” because $\hat{\Theta}$ based on inverting tests may not be an interval. See the use of the duality in Sections A1.4.2 and 3.6.1.

A1.3 Inference with 2×2 tables

A1.3.1 Fisher’s exact test

Fisher proposed an exact test for $H_0 : p_1 = p_0$ under the statistical model:

$$n_{11} \sim \text{Binomial}(n_1, p_1), \quad n_{01} \sim \text{Binomial}(n_0, p_0), \quad n_{11} \perp\!\!\!\perp n_{01}.$$

The table below summarizes the data.

	1	0	row sum
sample 1	n_{11}	n_{10}	n_1
sample 0	n_{01}	n_{00}	n_0
column sum	$n_{\cdot 1}$	$n_{\cdot 0}$	n

He argued that the sum $n_{11} + n_{01} \equiv n_{\cdot 1}$ contains little information for the difference between p_1 and p_0 , and n_{11} conditioning on the sum has Hypergeometric distribution that does not depend on the unknown parameter $p_1 = p_0$ under H_0 :

$$\text{pr}(n_{11} = k) = \frac{\binom{n_{\cdot 1}}{k} \binom{n-n_{\cdot 1}}{n_1-k}}{\binom{n}{n_1}}.$$

In R, the function `fisher.test` implement this test.

A1.3.2 Estimation with 2×2 tables

Based on the model in Section A1.3.1, we can estimate the parameters p_1 and p_0 by sample frequencies:

$$\hat{p}_1 = \frac{n_{11}}{n_1}, \quad \hat{p}_0 = \frac{n_{01}}{n_0}.$$

Therefore, we can estimate the risk difference, log risk ratio, and log odds ratio

$$\begin{aligned} \text{RD} &= p_1 - p_0, \\ \log \text{RR} &= \log \frac{p_1}{p_0}, \\ \log \text{OR} &= \log \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \end{aligned}$$

by the sample analogues

$$\begin{aligned} \hat{\text{RD}} &= \hat{p}_1 - \hat{p}_0, \\ \log \hat{\text{RR}} &= \log \frac{\hat{p}_1}{\hat{p}_0}, \\ \log \hat{\text{OR}} &= \log \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)} = \log \frac{n_{11}n_{00}}{n_{10}n_{01}}. \end{aligned}$$

Based on the asymptotic approximation (see Problem A1.4), the estimated variance for the above parameters are

$$\begin{aligned} &\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}, \\ &\frac{1-\hat{p}_1}{n_1\hat{p}_1} + \frac{1-\hat{p}_0}{n_0\hat{p}_0}, \\ &\frac{1}{n_1\hat{p}_1(1-\hat{p}_1)} + \frac{1}{n_0\hat{p}_0(1-\hat{p}_0)}, \end{aligned}$$

respectively. The log transformation above yields better Normal approximations because the risk ratio and odds ratio are always positive.

A1.4 Two famous problems in statistics

A1.4.1 Behrens–Fisher problem

Consider the two-sample problem with n_1 units under the treatment and n_0 units under the control, respectively. Assume the outcomes under the treatment $\{Y_i : Z_i = 1\}$ are IID from $N(\mu_1, \sigma_1^2)$ and the outcomes under the control $\{Y_i : Z_i = 0\}$ are IID from $N(\mu_0, \sigma_0^2)$, respectively. The goal is to test $H_0 : \mu_1 = \mu_0$.

Start with the easier case with $\sigma_1^2 = \sigma_0^2$. Coherent with Chapter 3, let $\hat{Y}(1)$ and $\hat{Y}(0)$ denote the sample means of the outcomes under the treatment and

control, respectively. A standard result is that

$$t_{\text{equal}} = \frac{\hat{Y}(1) - \hat{Y}(0)}{\sqrt{\frac{n}{n_1 n_0 (n-2)} \left[\sum_{Z_i=1} \{Y_i - \hat{Y}(1)\}^2 + \sum_{Z_i=0} \{Y_i - \hat{Y}(0)\}^2 \right]}} \sim t_{n-2}.$$

Based on t_{equal} , we can construct a test for H_0 .

Now consider the more difficult case with possibly different σ_1^2 and σ_0^2 . The distribution of t_{equal} is no longer t_{n-2} . Estimating the variances separately, we can also define

$$t_{\text{unequal}} = \frac{\hat{Y}(1) - \hat{Y}(0)}{\sqrt{\frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}}},$$

where

$$\hat{S}^2(1) = (n_1 - 1)^{-1} \sum_{Z_i=1} \{Y_i - \hat{Y}(1)\}^2, \quad \hat{S}^2(0) = (n_0 - 1)^{-1} \sum_{Z_i=0} \{Y_i - \hat{Y}(0)\}^2$$

are the sample variances of the outcomes under the treatment and control, respectively. Unfortunately, the exact distribution of t_{unequal} depends on the known variances. Testing H_0 without assuming equal variances is the famous Behrens–Fisher problem. With large sample sizes n_1 and n_0 , the central limit theorem ensures that t_{unequal} is approximately $N(0, 1)$. So we can construct approximate test for H_0 .

A1.4.2 Fieller–Creasy problem

Consider the two-sample problem with n_1 units under the treatment and n_0 units under the control, respectively. Assume the outcomes under the treatment $\{Y_i : Z_i = 1\}$ are IID from $N(\mu_1, 1)$ and the outcomes under the control $\{Y_i : Z_i = 0\}$ are IID from $N(\mu_0, 1)$, respectively. The goal is to estimate $\gamma = \mu_1 / \mu_0$. We can use $\hat{\gamma} = \hat{Y}(1) / \hat{Y}(0)$ to estimate γ . But the point estimator has a complicated distribution, which does not yield a simple procedure to construct the confidence interval for γ .

Fieller’s confidence interval can be formulated as inverting tests for a sequence of null hypotheses: $H_0(c) : \gamma = c$. Under $H_0(c)$, we have

$$\frac{\hat{Y}(1) - c\hat{Y}(0)}{\sqrt{1/n_1 + c^2/n_0}} \sim N(0, 1)$$

which motivates the confidence interval

$$\left\{ c : \left| \frac{\hat{Y}(1) - c\hat{Y}(0)}{\sqrt{1/n_1 + c^2/n_0}} \right| \leq z_\alpha \right\}$$

where z_α is the upper $1 - \alpha/2$ quantile of a standard Normal random variable.

A1.5 Bootstrap

It is often very tedious to derive the variance formulas for complex estimators. Efron (1979) proposed the bootstrap as a general tool for variance estimation. There are many versions of the bootstrap (Davison and Hinkley, 1997). In this book, we only need the most basic one: the nonparametric bootstrap, which will be simply called the bootstrap.

Consider the generic setting with

$$Y_1, \dots, Y_n \stackrel{\text{IID}}{\sim} Y,$$

where Y_i can be a general random element denoting the observed data for unit i . An estimator $\hat{\theta}$ is a function of the observed data: $\hat{\theta} = T(Y_1, \dots, Y_n)$. When T is a complex function, it may not be easy to obtain the variance or asymptotic variance of $\hat{\theta}$.

The uncertainty of $\hat{\theta}$ is driven by the IID sampling of Y_1, \dots, Y_n from the true distribution. Although the true distribution is unknown, it can be well approximated by its empirical version

$$\hat{F}_n(y) = n^{-1} \sum_{i=1}^n I(Y_i \leq y),$$

when the sample size n is large. If we believe this approximation, we can simulate $\hat{\theta}$ by sampling

$$(Y_1^*, \dots, Y_n^*) \stackrel{\text{IID}}{\sim} \hat{F}_n(y).$$

Because $\hat{F}_n(y)$ is a discrete distribution with mass $1/n$ on each observed data point, the simulation of $\hat{\theta}$ reduces to the following procedure:

1. sample (Y_1^*, \dots, Y_n^*) from $\{Y_1, \dots, Y_n\}$ with replacement;
2. compute $\hat{\theta}^* = T(Y_1^*, \dots, Y_n^*)$;
3. repeat the above two steps B times to obtain the bootstrap replicates $\{\hat{\theta}_1^*, \dots, \hat{\theta}_B^*\}$.

We can then approximate the (asymptotic) variance of $\hat{\theta}$ by the sample variance of the bootstrap replicates:

$$\hat{V}_{\text{boot}} = (B - 1)^{-1} \sum_{b=1}^B (\hat{\theta}_b^* - \bar{\theta}^*)^2,$$

where $\bar{\theta}^* = B^{-1} \sum_{b=1}^B \hat{\theta}_b^*$. The bootstrap confidence interval based on the Normal approximation is then

$$\hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{\text{boot}}},$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ upper quantile of $N(0, 1)$.

A1.6 Homework problems

A1.1 Independent but not IID data

Assume that the X_i 's are independent with mean μ_i and variances σ_i^2 for $i = 1, \dots, n$. The parameter of interest is $\mu = n^{-1} \sum_{i=1}^n \mu_i$. Show that $\hat{\mu} = n^{-1} \sum_{i=1}^n X_i$ is unbiased for μ and find its variance. Show that the usual variance estimator for IID data

$$\hat{v} = \{n(n-1)\}^{-1} \sum_{i=1}^n (X_i - \hat{\mu})^2$$

is a conservative estimator for the variance of $\hat{\mu}$ in the sense that

$$E(\hat{v}) - \text{var}(\hat{\mu}) = \{n(n-1)\}^{-1} \sum_{i=1}^n (\mu_i - \mu)^2 \geq 0.$$

Remark: Consider a simpler case with $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$ for all $i = 1, \dots, n$. The sample mean is unbiased for μ with variance σ^2/n . Moreover, an unbiased estimator for the variance σ^2/n is $\hat{\sigma}^2/n = \hat{v}$, where $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \hat{\mu})^2$.

A1.2 Asymptotic Normality of ratio

Prove (A1.2).

A1.3 Asymptotic Normality of product

Prove (A1.4).

A1.4 Variance estimators in two-by-two tables

Use delta method to show the variance estimators in Section A1.3.2.

A2

Linear and Logistic Regressions

A2.1 Population ordinary least squares

Assume that $(x_i, y_i)_{i=1}^n \stackrel{\text{IID}}{\sim} (x, y)$, where x is a p -dimensional random scalar or vector and y is a random scalar. Below I will use (x, y) to denote a general observation, dropping the subscript i for simplicity. Define the population ordinary least squares (OLS) coefficient as

$$\beta = \arg \min_b E \{(y - x^\top b)^2\}.$$

The objective function is quadratic in b , so we can show that the minimizer is

$$\beta = \{E(x x^\top)\}^{-1} E(xy)$$

if the moments exist and $E(x x^\top)$ is invertible.

With β , we can define

$$\varepsilon = y - x^\top \beta \tag{A2.1}$$

as the *population residual*. By the definition of β , we can verify that

$$E(x\varepsilon) = E\{x(y - x^\top \beta)\} = E(xy) - E(x x^\top) \beta = 0.$$

Example A2.1 (population OLS with an intercept) If we include 1 as a component of x , then

$$E(\varepsilon) = E(y - x^\top \beta) = 0$$

which further implies that $\text{cov}(x, \varepsilon) = 0$. So with an intercept in β , the mean of the population residual must be zero, and it is uncorrelated with other covariates by construction.

Example A2.2 (univariate population OLS with an intercept) An important special case is that for scalars x and y , we can define

$$(\alpha, \beta) = \arg \min_{a, b} E\{(y - a - bx)^2\}$$

which have explicit formulas

$$\beta = \frac{\text{cov}(x, y)}{\text{var}(x)}, \quad \alpha = E(y) - \beta E(x).$$

Example A2.3 (univariate population OLS without an intercept) *Without intercept, we can define*

$$\gamma = \arg \min_c E\{(y - cx)^2\}$$

which equals

$$\gamma = \frac{E(xy)}{E(x^2)}.$$

When x has mean zero, $\beta = \gamma$ in the above two population OLS.

We can also rewrite (A2.1) as

$$y = x^\top \beta + \varepsilon, \quad (\text{A2.2})$$

which holds by the definition of the population OLS coefficient and residual without any modeling assumption. We call (A2.2) the population OLS decomposition.

A2.2 Sample OLS

Based on data $(x_i, y_i)_{i=1}^n \stackrel{\text{IID}}{\sim} (x, y)$, we can easily obtain the moment estimator for the population OLS coefficient

$$\hat{\beta} = \left(n^{-1} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(n^{-1} \sum_{i=1}^n x_i y_i \right),$$

and the residuals $\hat{\varepsilon}_i = y_i - x_i^\top \hat{\beta}$. This is called the sample OLS or simply the OLS. The OLS coefficient $\hat{\beta}$ minimizes the residual sum of squares

$$\hat{\beta} = \arg \min_b n^{-1} \sum_{i=1}^n (y_i - x_i^\top b)^2,$$

which satisfies the following Normal equation:

$$\sum_{i=1}^n x_i (y_i - x_i^\top \hat{\beta}) = 0.$$

The fitted values equal

$$\hat{y}_i = x_i^\top \hat{\beta} \quad (i = 1, \dots, n).$$

Using the matrix notation

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

we can write the OLS coefficient as

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

and the fitted vector as

$$\hat{Y} = X \hat{\beta} = X(X^T X)^{-1} X^T Y.$$

Define the hat matrix as

$$H = X(X^T X)^{-1} X^T.$$

Then we also have $\hat{Y} = HY$, justifying the name ‘‘hat matrix.’’

Assuming finite fourth moments of (x, y) , we can use the law of large numbers and the central limit theorem to show that

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, V = B^{-1} M B^{-1})$$

in distribution, where $B = E(xx^T)$ and $M = E(\varepsilon^2 xx^T)$. So a moment estimator for the asymptotic variance of $\hat{\beta}$ is

$$\hat{V}_{\text{EHW}} = n^{-1} \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i^T \right) \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1},$$

which is called the Eicker–Huber–White (EHW) robust covariance estimator (Eicker, 1967; Huber, 1967; White, 1980). We can show that $n\hat{V}_{\text{EHW}} \rightarrow V$ in probability. Based on $\hat{\beta}$ and \hat{V}_{EHW} , we can make inference about the population OLS coefficient β . In R, the `lm` function can compute $\hat{\beta}$, and the `hccm` function in the package `car` can compute \hat{V}_{EHW} .

There are many variants of the EHW robust covariance estimator (Long and Ervin, 2000). In particular, the HC1 variant modifies $\hat{\varepsilon}_i^2$ to $\hat{\varepsilon}_i^2/(n-p)$, the HC2 variant modifies $\hat{\varepsilon}_i^2$ to $\hat{\varepsilon}_i^2/(1-h_{ii})$, the HC3 variant modifies $\hat{\varepsilon}_i^2$ to $\hat{\varepsilon}_i^2/(1-h_{ii})^2$, in the definition of \hat{V}_{EHW} , where h_{ii} is the (i, i) th diagonal element of H , also called the leverage scores.

A2.3 Frisch–Waugh–Lovell Theorem

The Frisch–Waugh–Lovell (FWL) theorem has two versions: one at the population level and the other at the sample level. It reduces multivariate OLS to univariate OLS and therefore facilitate the understanding and calculation of the OLS coefficients. Below I will present special cases of the FWL Theorem which are enough for this book.

Theorem A2.1 (population FWL) *The coefficient of x_1 in the OLS fit of y on (x_1, x_2, \dots, x_p) equals the coefficient of \tilde{x}_1 in the OLS fit of y or \tilde{y} on \tilde{x}_1 , where \tilde{y} is the residual from the OLS fit of y on (x_2, \dots, x_p) and \tilde{x}_1 is the residual from the OLS fit of x_1 on (x_2, \dots, x_p) .*

In Theorem A2.1, residualizing x_1 is crucial but residualizing y is not.

Theorem A2.2 (sample FWL) *With data $(Y, X_1, X_2, \dots, X_p)$ containing column vectors, the coefficient of X_1 equals the coefficient of \tilde{X}_1 in the OLS fit of Y or \tilde{Y} on \tilde{X}_1 , where \tilde{Y} is the residual vector from the OLS fit of Y on (X_2, \dots, X_p) and \tilde{X}_1 is the residual from the OLS fit of X_1 on (X_2, \dots, X_p) .*

Again, in Theorem A2.2, residualizing X_1 is crucial but residualizing Y is not.

A2.4 Linear model

Sometimes, we impose a stronger model assumption which requires the conditional mean of y given x is linear:

$$E(y | x) = x^\top \beta$$

or, equivalently,

$$y = x^\top \beta + \varepsilon, \quad E(\varepsilon | x) = 0,$$

which is called the restricted mean model. Under this model, the population OLS coefficient is the true parameter of interest:

$$\begin{aligned} \{E(xx^\top)\}^{-1} E(xy) &= \{E(xx^\top)\}^{-1} E\{xE(y | x)\} \\ &= \{E(xx^\top)\}^{-1} E(xx^\top \beta) \\ &= \beta. \end{aligned}$$

Moreover, the population OLS coefficient does not depend on the distribution of x . The asymptotic inference in Section A2.1 applies to this model too.

In the special case with $\text{var}(\varepsilon | x) = \sigma^2$, the asymptotic variance of the OLS coefficient reduces to

$$V = \sigma^2 \{E(xx^\top)\}^{-1}$$

so a simpler moment estimator for the asymptotic variance of $\hat{\beta}$ is

$$\hat{V}_{\text{OLS}} = \hat{\sigma}^2 \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1}$$

where $\hat{\sigma}^2 = (n-p)^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2$. This is the standard covariance estimator from the `lm` function.

A2.5 Weighted least squares

Assuming that $(w_i, x_i, y_i) \stackrel{\text{IID}}{\sim} (w, x, y)$ with $w \neq 0$. At the population level, we can define weighted least squares (WLS) coefficient as

$$\beta_w = \arg \min_b E\{w(y - x^T b)^2\},$$

which satisfies

$$E\{wx(y - x^T \beta_w)\} = 0$$

and thus equals

$$\beta_w = \{E(wxx^T)\}^{-1}E(wxy)$$

if $E(wxx^T)$ is invertible.

At the sample level, we can define the WLS coefficient as

$$\hat{\beta}_w = \arg \min_b \sum_{i=1}^n w_i(y_i - x_i^T b)^2,$$

which satisfies

$$\sum_{i=1}^n w_i x_i (y_i - x_i^T \hat{\beta}_w) = 0$$

and thus equals

$$\hat{\beta}_w = \left(n^{-1} \sum_{i=1}^n w_i x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n w_i x_i y_i \right)$$

if $\sum_{i=1}^n w_i x_i x_i^T$ is invertible.

A2.6 Logistic regression

A2.6.1 Model

Technically, we can use apply the OLS procedure even if the outcome y is binary. However, it is a little awkward to have predicted probabilities outside the range of $[0, 1]$. This motivates us to consider the following model:

$$\text{pr}(y_i = 1 | x_i) = g(x_i^T \beta),$$

where $g(\cdot) : \mathbb{R} \rightarrow [0, 1]$ is a monotone function, and its inverse is often called the *link function*. The $g(\cdot)$ function can be any distribution function of a random variable, but we will focus on the logistic form:

$$g(z) = \frac{e^z}{1 + e^z} = (1 + e^{-z})^{-1}.$$

We can also write the logistic model as

$$\text{pr}(y_i = 1 | x_i) \equiv \pi(x_i, \beta) = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}},$$

or, equivalently,

$$\text{logit}\{\text{pr}(y_i = 1 | x_i)\} \equiv \log \frac{\text{pr}(y_i = 1 | x_i)}{1 - \text{pr}(y_i = 1 | x_i)} = x_i^\top \beta.$$

Assume that x_{i1} is binary. Under the logistic model, we have

$$\begin{aligned} \beta_1 &= \text{logit}\{\text{pr}(y_i = 1 | x_{i1} = 1, \dots)\} - \text{logit}\{\text{pr}(y_i = 1 | x_{i1} = 0, \dots)\} \\ &= \log \frac{\text{pr}(y_i = 1 | x_{i1} = 1, \dots)/\text{pr}(y_i = 0 | x_{i1} = 1, \dots)}{\text{pr}(y_i = 1 | x_{i1} = 0, \dots)/\text{pr}(y_i = 0 | x_{i1} = 0, \dots)}, \end{aligned}$$

where \dots contains all other regressor x_{i2}, \dots, x_{ip} . Therefore, the coefficient β_1 equals the log odds ratio of x_{i1} on y_i conditional on other regressors.

A2.6.2 Maximum likelihood estimate

To estimate the parameter β , we can maximize the following likelihood function:

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n \{\pi(x_i, \beta)\}^{y_i} \{1 - \pi(x_i, \beta)\}^{1-y_i} \\ &= \prod_{i=1}^n \left\{ \frac{\pi(x_i, \beta)}{1 - \pi(x_i, \beta)} \right\}^{y_i} \{1 - \pi(x_i, \beta)\} \\ &= \prod_{i=1}^n \left(\frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}} \right)^{y_i} \frac{1}{1 + e^{x_i^\top \beta}} \\ &= \prod_{i=1}^n \frac{e^{y_i x_i^\top \beta}}{1 + e^{x_i^\top \beta}}. \end{aligned}$$

Let $\hat{\beta}$ denote the maximizer, which is called the maximum likelihood estimate (MLE). Taking the log of $L(\beta)$ and differentiating it with respect to β , we can show that the MLE must satisfy the first order condition:

$$\sum_{i=1}^n x_i \{y_i - \pi(x_i, \hat{\beta})\} = 0.$$

So if x_i contains an intercept, the MLE must satisfy

$$\sum_{i=1}^n \{y_i - \pi(x_i, \hat{\beta})\} = 0,$$

that is, the average of the observed y_i 's must be identical to the average of the fitted probabilities $\pi(x_i, \hat{\beta})$'s.

Using the general theory for the MLE, we can show that it is consistent for the true parameter β and is asymptotically normal:

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, V)$$

in distribution, where $V = E[\pi(x_i, \beta)\{1 - \pi(x_i, \beta)\}xx^\top]$. So we can approximate the covariance matrix of $\hat{\beta}$ by

$$n^{-1} \sum_{i=1}^n \pi(x_i, \hat{\beta})\{1 - \pi(x_i, \hat{\beta})\}x_i x_i^\top.$$

In R, the `glm` function can find the MLE and report the estimated covariance matrix.

A2.6.3 Extension to the case-control study

In case-control studies, sampling is conditional on the binary outcome, that is, units with outcomes $y_i = 1$ and $y_i = 0$ are sampled with different probabilities. Let s_i be the sampling indicator. In case control studies, we have

$$\text{pr}(s_i = 1 | x_i, y_i) = \text{pr}(s_i = 1 | y_i)$$

as a function of y_i , and we only observe units with $s_i = 1$.

Prentice and Pyke (1979) showed that logistic regression is applicable in case-control studies although the above discussion assume IID sampling.

A2.6.4 Logistic regression with weights

Sometimes, unit i has weight w_i , then we can fit a weighted logistic regression by solving

$$\sum_{i=1}^n w_i x_i \{y_i - \pi(x_i, \hat{\beta})\} = 0.$$

A2.7 Homework problems

A2.1 Sample OLS with intercept

Assume the regressor x_i contains an intercept. Show that

$$\bar{y} = \bar{x}^\top \hat{\beta}. \quad (\text{A2.3})$$

A2.2 Univariate weighed least squares

As a special case of WLS, define

$$(\hat{\alpha}_w, \hat{\beta}_w) = \arg \min_{(a,b)} \sum_{i=1}^n w_i (y_i - a - bx_i)^2$$

where $w_i \geq 0$. Show that

$$\hat{\beta}_w = \frac{\sum_{i=1}^n w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2} \quad (\text{A2.4})$$

and

$$\hat{\alpha}_w = \bar{y}_w - \hat{\beta}_w \bar{x}_w, \quad (\text{A2.5})$$

where $\bar{x}_w = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i$ and $\bar{y}_w = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i$ are the weighted averages of the x_i 's and y_i 's.

Further assume that the x_i 's are binary. Show that

$$\hat{\beta}_w = \frac{\sum_{i=1}^n w_i x_i y_i}{\sum_{i=1}^n w_i x_i} - \frac{\sum_{i=1}^n w_i (1 - x_i) y_i}{\sum_{i=1}^n w_i (1 - x_i)}.$$

That is, if the regressor is binary in the univariate WLS, the coefficient of the regressor equals the difference in the weighted means.

Hint: Think about an appropriate reparametrization of the WLS problem. Otherwise, the derivation can be tedious.

A2.3 OLS with orthogonal regressors

Consider sample OLS fit of an n -vector Y on an $n \times p$ matrix X , with coefficient $\hat{\beta}$. Partition X into $X = (X_1, X_2)$, where X_1 is an $n \times k$ matrix and X_2 is an $n \times l$ matrix, with $p = k + l$. Correspondingly, partition $\hat{\beta}$ into

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}.$$

Assume X_1 and X_2 are orthogonal, that is, $X_1^\top X_2 = 0$. Show that $\hat{\beta}_1$ equals the coefficient from OLS of Y on X_1 and $\hat{\beta}_2$ equals the coefficient from OLS of Y on X_2 , respectively.

A2.4 OLS with a non-degenerate transformation of the regressors

Define $\hat{\beta}$ as the coefficient from the sample OLS fit of an n -vector Y on an $n \times p$ matrix X . Let Γ be a $p \times p$ non-degenerate matrix, and define $X' = X\Gamma$. Define $\hat{\beta}'$ as the coefficient from the sample OLS fit of Y on X' .

Show that

$$\hat{\beta} = \Gamma \hat{\beta}'.$$

A3

Some Useful Lemmas for Simple Random Sampling

A3.1 Lemmas

Simple random sampling is a basic topic in standard survey sampling textbooks (e.g., Cochran, 1953). Below I review some results for simple random sampling that are useful for design-based inference in the CRE in Chapters 3 and 4.

A simple random sample of size n_1 consists of a subset from a finite population of n units indexed by $i = 1, \dots, n$. Let $\mathbf{Z} = (Z_1, \dots, Z_n)$ be the inclusion indicators of the n units with $Z_i = 1$ if unit i is sampled and $Z_i = 0$ otherwise. The vector \mathbf{Z} can take $\binom{n}{n_1}$ possible permutations of a vector of n_1 1's and n_0 0's, and each has equal probability. The following lemma summarizes the first two moments of the inclusion indicators.

Lemma A3.1 *Under simple random sampling, we have*

$$E(Z_i) = \frac{n_1}{n}, \quad \text{var}(Z_i) = \frac{n_1 n_0}{n^2}, \quad \text{cov}(Z_i, Z_j) = -\frac{n_1 n_0}{n^2(n-1)}.$$

In more compact forms, we have

$$E(\mathbf{Z}) = \frac{n_1}{n} \mathbf{1}_n, \quad \text{cov}(\mathbf{Z}) = \frac{n_1 n_0}{n(n-1)} \mathbf{P}_n,$$

where $\mathbf{1}_n$ is a n -dimensional vector of 1's, and $\mathbf{P}_n = \mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^\top$ is the $n \times n$ projection matrix orthogonal to $\mathbf{1}_n$.

Let $\{c_1, \dots, c_n\}$ be a finite population with mean $\bar{c} = \sum_{i=1}^n c_i/n$ and variance

$$S_c^2 = (n-1)^{-1} \sum_{i=1}^n (c_i - \bar{c})^2;$$

let $\{d_1, \dots, d_n\}$ be another finite population with mean $\bar{d} = \sum_{i=1}^n d_i/n$ and variance

$$S_d^2 = (n-1)^{-1} \sum_{i=1}^n (d_i - \bar{d})^2;$$

their covariance is

$$S_{cd} = (n - 1)^{-1} \sum_{i=1}^n (c_i - \bar{c})(d_i - \bar{d}).$$

Based on the simple random sample, the sample means are

$$\hat{c} = n_1^{-1} \sum_{i=1}^n Z_i c_i, \quad \hat{d} = n_1^{-1} \sum_{i=1}^n Z_i d_i;$$

sample variances are

$$\hat{S}_c^2 = (n_1 - 1)^{-1} \sum_{i=1}^n Z_i (c_i - \hat{c})^2, \quad \hat{S}_d^2 = (n_1 - 1)^{-1} \sum_{i=1}^n Z_i (d_i - \hat{d})^2;$$

the sample covariance is

$$\hat{S}_{cd} = (n_1 - 1)^{-1} \sum_{i=1}^n Z_i (c_i - \hat{c})(d_i - \hat{d}).$$

Lemma A3.2 below gives the moments of the sample means \hat{c} and \hat{d} .

Lemma A3.2 *The sample means are unbiased for the population means:*

$$E(\hat{c}) = \bar{c}, \quad E(\hat{d}) = \bar{d}.$$

Their variances and covariance are

$$\text{var}(\hat{c}) = \frac{n_0}{nn_1} S_c^2, \quad \text{var}(\hat{d}) = \frac{n_0}{nn_1} S_d^2, \quad \text{cov}(\hat{c}, \hat{d}) = \frac{n_0}{nn_1} S_{cd}.$$

In the variance formula in Lemma A3.2, the coefficient $n_0/(nn_1) = 1/n_1 \times (1 - n_1/n)$ in Lemma A3.2 is different from $1/n_1$ under IID sampling. The additional factor $1 - n_1/n = n_0/n$ is called the *finite population correction*.

Lemma A3.3 below gives the unbiasedness of the sample variances and covariance for estimating the population analogs.

Lemma A3.3 *The sample variances and covariance are unbiased for their population versions:*

$$E(\hat{S}_c^2) = S_c^2, \quad E(\hat{S}_d^2) = S_d^2, \quad E(\hat{S}_{cd}) = S_{cd}.$$

An important practical question is to make inference about \bar{c} based on the simple random sample. This requires a more precise characterization of the distribution of its unbiased estimator \hat{c} . The finite-sample exact distribution of \hat{c} depends on the whole finite population $\{c_1, \dots, c_n\}$, which is intractable in general. The following finite population central limit theorem characterizes the asymptotic distribution of \hat{c} based on its first two moments.

Lemma A3.4 (finite population central limit theorem) *As $n \rightarrow \infty$, if*

$$\frac{\max_{1 \leq i \leq n} (c_i - \bar{c})^2}{\min(n_1, n_0) S_c^2} \rightarrow 0,$$

then

$$\frac{\hat{c} - \bar{c}}{\sqrt{\frac{n_0}{nn_1} S_c^2}} \rightarrow N(0, 1)$$

in distribution, and $\hat{S}_c^2 / S_c^2 \rightarrow 1$ in probability.

Lemma A3.4 justifies the Wald-type $1 - \alpha$ confidence interval for \bar{c} :

$$\hat{c} \pm z_{1-\alpha/2} \sqrt{\frac{n_0}{nn_1} \hat{S}_c^2}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ upper quantile of the standard Normal random variable.

A3.2 Proofs

Proof of Lemma A3.1: By symmetry, the Z_i 's have the same mean, so

$$n_1 = \sum_{i=1}^n Z_i = E \left(\sum_{i=1}^n Z_i \right) = nE(Z_i) \implies E(Z_i) = n_1/n.$$

Because Z_i is a Bernoulli random variable, its variance is

$$\text{var}(Z_i) = \frac{n_1}{n} \left(1 - \frac{n_1}{n}\right) = \frac{n_1 n_0}{n^2}.$$

By symmetry again, the Z_i 's have the same variance and the pairs (Z_i, Z_j) 's have the same covariance, so

$$0 = \text{var} \left(\sum_{i=1}^n Z_i \right) = n\text{var}(Z_i) + n(n-1)\text{cov}(Z_i, Z_j)$$

which implies that

$$\text{cov}(Z_i, Z_j) = -\frac{n_1 n_0}{n^2(n-1)} \quad (i \neq j).$$

□

Proof of Lemma A3.1: The unbiasedness of the sample mean follows from linearity. For example,

$$E(\hat{c}) = E\left(\frac{1}{n_1} \sum_{i=1}^n Z_i c_i\right) = \frac{1}{n_1} \sum_{i=1}^n E(Z_i) c_i = \bar{c}.$$

The covariance of the sample means is

$$\begin{aligned} & \text{cov}(\hat{c}, \hat{d}) \\ &= \text{cov}\left\{\frac{1}{n_1} \sum_{i=1}^n Z_i(c_i - \bar{c}), \frac{1}{n_1} \sum_{i=1}^n Z_i(d_i - \bar{d})\right\} \\ &= \frac{1}{n_1^2} \left[\sum_{i=1}^n \text{var}(Z_i)(c_i - \bar{c})(d_i - \bar{d}) + \sum_{i \neq j} \text{cov}(Z_i, Z_j)(c_i - \bar{c})(d_j - \bar{d}) \right] \\ &= \frac{1}{n_1^2} \left[\frac{n_1 n_0}{n^2} \sum_{i=1}^n (c_i - \bar{c})(d_i - \bar{d}) - \frac{n_1 n_0}{n^2(n-1)} \sum_{i \neq j} (c_i - \bar{c})(d_j - \bar{d}) \right]. \end{aligned}$$

Because

$$0 = \sum_{i=1}^n (c_i - \bar{c}) \sum_{i=1}^n (d_i - \bar{d}) = \sum_{i=1}^n (c_i - \bar{c})(d_i - \bar{d}) + \sum_{i \neq j} (c_i - \bar{c})(d_j - \bar{d}),$$

the covariance of the sample means reduces to

$$\begin{aligned} & \text{cov}(\hat{c}, \hat{d}) \\ &= \frac{1}{n_1^2} \left[\frac{n_1 n_0}{n^2} \sum_{i=1}^n (c_i - \bar{c})(d_i - \bar{d}) + \frac{n_1 n_0}{n^2(n-1)} \sum_{i=1}^n (c_i - \bar{c})(d_i - \bar{c}) \right] \\ &= \frac{n_0}{n n_1} S_{cd}. \end{aligned}$$

The variance formulas are just special cases with $\hat{c} = \hat{d}$. \square

Proof of Lemma A3.3: We prove only the sample covariance term, because the formulas for sample variances are special cases. We have the following decomposition:

$$\begin{aligned} (n_1 - 1)\hat{S}_{cd} &= \sum_{i=1}^n Z_i(c_i - \hat{c})(d_i - \hat{d}) \\ &= \sum_{i=1}^n Z_i\{(c_i - \bar{c}) - (\hat{c} - \bar{c})\}\{(d_i - \bar{d}) - (\hat{d} - \bar{d})\} \\ &= \sum_{i=1}^n Z_i(c_i - \bar{c})(d_i - \bar{d}) - n_1(\hat{c} - \bar{c})(\hat{d} - \bar{d}). \end{aligned}$$

Taking expectation on both sides, we have

$$\begin{aligned}
 E\{(n_1 - 1)\hat{S}_{cd}\} &= \sum_{i=1}^n E(Z_i)(c_i - \bar{c})(d_i - \bar{d}) - n_1 E\{(\hat{c} - \bar{c})(\hat{d} - \bar{d})\} \\
 &= \frac{n_1}{n} \sum_{i=1}^n (c_i - \bar{c})(d_i - \bar{d}) - n_1 \frac{n_0}{nn_1} S_{cd} \\
 &= S_{cd} \left\{ \frac{n_1(n-1)}{n} - \frac{n_0}{n} \right\} \\
 &= (n_1 - 1)S_{cd},
 \end{aligned}$$

and the conclusion follows by dividing both sides by $n_1 - 1$. \square

Proof of Lemma A3.4: Hájek (1960) gave a proof of the central limit theorem for simple random sampling, and Lehmann (1975) gave a more accessible version of the proof. Li and Ding (2017) modified the central limit theorem as presented in Lemma A3.4, and proved the consistency of the sample variance. Due to the technical complexities, I omit the proof. \square

A3.3 Comments on the literature

Survey sampling and experimental design are deeply connected ever since Neyman (1934, 1935)'s seminal work. Li and Ding (2017) and Mukerjee et al. (2018) made many theoretical ties between these two areas.

A3.4 Homework Problems

A3.1 Vector form of the results

Assume the c_i 's are vectors and modify

$$S_c^2 = (n-1)^{-1} \sum_{i=1}^n (c_i - \bar{c})(c_i - \bar{c})^\top, \quad \hat{S}_c^2 = (n_1 - 1)^{-1} \sum_{i=1}^n Z_i(c_i - \hat{c})(c_i - \hat{c})^\top.$$

Show that

$$E(\hat{c}) = \bar{c}, \quad \text{cov}(\hat{c}) = \frac{n_0}{nn_1} S_c^2, \quad E(\hat{S}_c^2) = S_c^2.$$



Bibliography

- Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267.
- Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76:1537–1557.
- Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, 29:1–11.
- Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84:781–807.
- Alwin, D. F. and Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, 40:37–47.
- Amarante, V., Manacorda, M., Miguel, E., and Vigorito, A. (2016). Do cash transfers improve birth outcomes? evidence from matched vital statistics, program, and social security data. *American Economic Journal: Economic Policy*, 8:1–43.
- Anderson, T. W. and Rubin, H. (1950). The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, 21:570–582.
- Angrist, J., Lang, D., and Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1:136–163.
- Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review*, 99:1384–1414.
- Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *American Economic Review*, 80:313–336.
- Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66:249–288.

- Angrist, J. D. and Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88:450–477.
- Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90:431–442.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, 91:444–455.
- Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106:979–1014.
- Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.
- Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'Metrics: The Path from Cause to Effect*. Princeton: Princeton University Press.
- Aronow, P. M., Green, D. P., and Lee, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *Annals of Statistics*, 42:850–871.
- Asher, S. and Novosad, P. (2020). Rural roads and local economic development. *American Economic Review*, 110:797–823.
- Baker, S. G. and Lindeman, K. S. (1994). The paired availability design: a proposal for evaluating epidural analgesia during labor. *Statistics in Medicine*, 13:2269–2278.
- Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92:1171–1176.
- Ball, S., Bogatz, G., Rubin, D., and Beaton, A. (1973). Reading with television: An evaluation of the electric company. a report to the children's television workshop. volumes 1 and 2.
- Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973.
- Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrika*, 34:123–138.
- Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182.

- Basman, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica*, 25:77–83.
- Bazzano, L. A., He, J., Muntner, P., Vupputuri, S., and Whelton, P. K. (2003). Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the United States. *Annals of Internal Medicine*, 138:891–897.
- Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., and Zhao, L. (2013). Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*, 37:170–196.
- Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94:991–1013.
- Bickel, P. J., Hammel, E. A., and O’Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–404.
- Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.
- Bind, M.-A. C. and Rubin, D. B. (2020). When possible, report a fisher-exact p value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences of the United States of America*, 117:19151–19158.
- Blackwell, M. (2013). A framework for dynamic causal inference in political science. *American Journal of Political Science*, 57:504–520.
- Bloniarz, A., Liu, H., Zhang, C. H., Sekhon, J., and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 113:7383–7390.
- Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, 8:225–246.
- Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L., and Bärnighausen, T. (2014). Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology*, 25:729.
- Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44:512–525.
- Bowden, J., Spiller, W., Del Greco M, F., Sheehan, N., Thompson, J., Minelli, C., and Davey Smith, G. (2018). Improving the visualization, interpretation

- and analysis of two-sample summary data mendelian randomization via the radial plot and radial regression. *International Journal of Epidemiology*, 47:1264–1278.
- Bradford Hill, A. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300.
- Bradford Hill, A. (2020). The environment and disease: association or causation? (with discussion). *Observational Studies*, 6:1–65.
- Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1:200–232.
- Butler, C. C. (1969). A test for symmetry using the sample distribution function. *Annals of Mathematical Statistics*, 40:2209–2210.
- Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–734.
- Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research.
- Carpenter, C. and Dobkin, C. (2009). The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, 1:164–182.
- Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155:138–154.
- Cattaneo, M. D., Frandsen, B. R., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, 3:1–24.
- Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:673–700.
- Charig, C. R., Webb, D. R., Payne, S. R., and Wickham, J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal*, 292:879–882.
- Chen, H., Geng, Z., and Jia, J. (2007). Criteria for surrogate end points. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:919–932.

- Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:815–836.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68.
- Chong, A., Cohen, I., Field, E., Nakasone, E., and Torero, M. (2016). Iron deficiency and schooling attainment in peru. *American Economic Journal: Applied Economics*, 8:222–55.
- Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Supplement to the Journal of the Royal Statistical Society*, 5:171–176.
- Cochran, W. G. (1953). *Sampling Techniques*. New York: Wiley.
- Cochran, W. G. (1957). Analysis of covariance: its nature and uses. *Biometrics*, 13:261–281.
- Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society: Series A (General)*, 128:234–266.
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24:295–313.
- Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā*, 35:417–446.
- Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203.
- Cox, D. R. (1982). Randomization and concomitant variables in the design of experiments. In G. Kallianpur, P. R. K. and Ghosh, J. K., editors, *Statistics and Probability: Essays in Honor of C. R. Rao*, pages 197–202. North-Holland, Amsterdam.
- Cox, D. R. (2007). On a generalization of a result of W. G. Cochran. *Biometrika*, 94:755–759.
- Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96:187–199.
- Cuzick, J., Edwards, R., and Segnan, N. (1997). Adjusting for non-compliance and contamination in randomized clinical trials. *Statistics in Medicine*, 16:1017–1029.

- D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221:644–654.
- Davey Smith, G. and Ebrahim, S. (2003). “Mendelian randomization”: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32:1–22.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
- Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41:1–15.
- Dawid, A. P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association*, 95:407–424.
- Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94:1053–1062.
- Ding, P. (2016). A paradox from randomization-based causal inference (with discussion). *Statistical Science*, 32:331–345.
- Ding, P. (2021). The Frisch–Waugh–Lovell theorem for standard errors. *Statistics and Probability Letters*, 168:108945.
- Ding, P. and Dasgupta, T. (2016). A potential tale of two by two tables from completely randomized experiments. *Journal of American Statistical Association*, 111:157–168.
- Ding, P. and Dasgupta, T. (2017). A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. *Biometrika*, 105:45–56.
- Ding, P., Feller, A., and Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114:304–317.
- Ding, P., Geng, Z., Yan, W., and Zhou, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association*, 106:1578–1591.
- Ding, P. and Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*, 33:214–237.
- Ding, P., Li, X., and Miratrix, L. W. (2017a). Bridging finite and super population causal inference. *Journal of Causal Inference*, 5:20160027.
- Ding, P. and Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79:757–777.

- Ding, P. and Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference*, 3:41–57.
- Ding, P. and VanderWeele, T. J. (2014). Generalized Cornfield conditions for the risk difference. *Biometrika*, 101:971–977.
- Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27:368–377.
- Ding, P. and Vanderweele, T. J. (2016). Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Biometrika*, 103:483–490.
- Ding, P., VanderWeele, T. J., and Robins, J. M. (2017b). Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, 104:291–302.
- Doll, R. and Hill, A. B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, 2:739.
- Dorn, H. F. (1953). Philosophy of inferences from retrospective studies. *American Journal of Public Health and the Nations Health*, 43:677–683.
- Durrett, R. (2019). *Probability: Theory and Examples*. Cambridge: Cambridge University Press.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26.
- Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials (with discussion). *Journal of the American Statistical Association*, 86:9–17.
- Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 59–82. Berkeley, CA: University of California Press.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. New York: Chapman and Hall/CRC.
- Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16:175–185.
- Firth, D. and Bennett, K. E. (1998). Robust models in probability sampling (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60:3–21.
- Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh by Oliver and Boyd, 1st edition.

- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh, London: Oliver and Boyd, 1st edition.
- Fisher, R. A. (1957). Dangers of cigarette smoking [letter]. *British Medical Journal*, 2:297–298.
- Fogarty, C. B. (2018a). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 80:1035–1056.
- Fogarty, C. B. (2018b). Regression assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105:994–1000.
- Follmann, D. A. (2000). On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association*, 95:1101–1109.
- Forastiere, L., Mattei, A., and Ding, P. (2018). Principal ignorability in mediation analysis: through and beyond sequential ignorability. *Biometrika*, 105:979–986.
- Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58:21–29.
- Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *Annals of Applied Statistics*, 2:176–196.
- Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40:180–193.
- Freedman, D. A. (2008c). Randomization does not justify logistic regression. *Statistical Science*, 23:237–249.
- Freedman, D. A. and Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, 32:392–409.
- Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173:761–767.
- Gastwirth, J. L., KRIEGER, A. M., and ROSENBAUM, P. R. (1998). Cornfield’s inequality. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. New York: Wiley.
- Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.
- Gerber, A. S., Green, D. P., and Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102:33–48.

- Gilbert, P. B. and Hudgens, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics*, 64:1146–1154.
- Gould, A. L. (1998). Multi-centre trial analysis revisited. *Statistics in Medicine*, 17:1779–1797.
- Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics*, 5:263–275.
- Guo, K. and Basse, G. (2023). The generalized Oaxaca–Blinder estimator. *Journal of American Statistical Association*, 118:524–536.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331.
- Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69:201–209.
- Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15:965–1056.
- Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20:25–46.
- Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5:361–74.
- Hájek, J. (1971). Comment on “an essay on the logical foundations of survey sampling, part one”. *The foundations of survey sampling*, 236.
- Hammond, E. C. and Horn, D. (1958). Smoking and death rates: report on forty four months of follow-up of 187,783 men. *Journal of the American Medical Association*, 166:1159–1172, 1294–1308.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054.
- Hartley, H. O., Rao, J. N. K., and Kiefer, G. (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association*, 64:841–851.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46:1251–1271.
- Hearst, N., Newman, T. B., and Hulley, S. B. (1986). Delayed effects of the military draft on mortality. *New England Journal of Medicine*, 314:620–624.

- Heckman, J. and Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics*, 86:30–57.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.
- Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C., and Sarkar, P. (2016). A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, 4:61–80.
- Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 11:561–570.
- Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
- Hill, J., Waldfogel, J., and Brooks-Gunn, J. (2002). Differential effects of high-quality child care. *Journal of Policy Analysis and Management*, 21:601–627.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240.
- Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2:259–278.
- Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1:69–88.
- Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236.
- Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2011). Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42:1–28.
- Hodges, J. L. and Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics*, 33:482–497.
- Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American statistical Association*, 81:945–960.
- Hong, G. and Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, 33:333–362.

- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47:663–685.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Cam, L. M. L. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. Berkeley, California: University of California Press.
- Hyman, H. H. (1955). *Survey Design and Analysis: Principles, Cases, and Procedures*. Glencoe, IL: Free Press.
- Imai, K. (2008a). Sharp bounds on the causal effects in randomized experiments with “truncation-by-death”. *Statistics and Probability Letters*, 78:144–149.
- Imai, K. (2008b). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, 27:4857–4873.
- Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25:51–71.
- Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99:854–866.
- Imbens, G. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58:1129–1179.
- Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93:126–132.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86:4–29.
- Imbens, G. W. (2014). Instrumental variables: An econometrician’s perspective. *Statistical Science*, 29:323–358.
- Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, 50:373–419.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62:467–475.
- Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142:615–635.

- Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72:1845–1857.
- Imbens, G. W. and Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies*, 64:555–574.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.
- Investigators, I. T. et al. (2014). Endovascular or open repair strategy for ruptured abdominal aortic aneurysm: 30 day outcomes from improve randomised trial. *British Medical Journal*, 348:f7661.
- Ioannidis, J. P. A., Tan, Y. J., and Blum, M. R. (2019). Limitations and misinterpretations of E-values for sensitivity analyses of observational studies. *Annals of Internal Medicine*, 170:108–111.
- Jackson, L. A., Jackson, M. L., Nelson, J. C., Neuzil, K. M., and Weiss, N. S. (2006). Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *International Journal of Epidemiology*, 35:337–344.
- Jiang, Z. and Ding, P. (2020). Measurement errors in the binary instrumental variable model. *Biometrika*, 107:238–245.
- Jiang, Z. and Ding, P. (2021). Identification of causal effects within principal strata using auxiliary variables. *Statistical Science*, 36:493–508.
- Jiang, Z., Ding, P., and Geng, Z. (2016). Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:829–848.
- Jiang, Z., Yang, S., and Ding, P. (2022). Multiply robust estimation of causal effects under principal ignorability. *Journal of the Royal Statistical Society - Series B (Statistical Methodology)*, 84:1423–1445.
- Jo, B. and Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, 28:2857–2875.
- Jo, B., Stuart, E. A., MacKinnon, D. P., and Vinokur, A. D. (2011). The use of propensity scores in mediation analysis. *Multivariate Behavioral Research*, 46:425–452.
- Judd, C. M. and Kenny, D. A. (1981). Process analysis estimating mediation in treatment evaluations. *Evaluation Review*, 5:602–619.
- Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22:523–539.

- Katan, M. B. (1986). Apoprotein E isoforms, serum cholesterol, and cancer. *Lancet*, 327:507–508.
- King, G. and Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, 14:131–159.
- Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, 83:2043–2063.
- Koenker, R. and Xiao, Z. (2002). Inference on the quantile regression process. *Econometrica*, 70:1583–1612.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116:4156–4165.
- Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., and Robins, J. M. (2005). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, 163:262–270.
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76:604–620.
- Lee, D. S. (2008). Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*, 142:675–697.
- Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76:1071–1102.
- Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48:281–355.
- Lee, M.-J. (2018). Simple least squares estimator for treatment effects using propensity score residuals. *Biometrika*, 105:149–164.
- Lee, W.-C. (2011). Bounding the bias of unmeasured factors with confounding and effect-modifying potentials. *Statistics in Medicine*, 30:1007–1017.
- Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. California: Holden-Day, Inc.
- Lei, L. and Ding, P. (2021). Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, 108:815–828.

- Li, F., Mattei, A., and Mealli, F. (2015). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *Annals of Applied Statistics*, 9:1906–1931.
- Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018a). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113:390–400.
- Li, F., Thomas, L. E., and Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188:250–257.
- Li, X. and Ding, P. (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine*, 35:957–960.
- Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112:1759–1769.
- Li, X. and Ding, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 82:241–268.
- Li, X., Ding, P., and Rubin, D. B. (2018b). Asymptotic theory of rerandomization in treatment-control experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 115:9157–9162.
- Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique. *Annals of Applied Statistics*, 7:295–318.
- Lin, Z., Ding, P., and Han, F. (2023). Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*.
- Lind, J. (1753). A treatise of the scurvy. *Three Parts. Containing an Inquiry into the Nature, Causes and Cure, of that Disease. Together with a Critical and Chronological View of what has been Published on the Subject*.
- Lipsitch, M., Tchetgen Tchetgen, E., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21:383–388.
- Little, R. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14:949–968.
- Liu, H. and Yang, Y. (2020). Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 107:935–948.
- Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, 54:217–224.

- Lu, S. and Ding, P. (2023). Flexible sensitivity analysis for causal inference in observational studies subject to unmeasured confounding. <https://arxiv.org/abs/2305.17643>.
- Lumley, T., Shaw, P. A., and Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79:200–220.
- Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23:2937–2960.
- Luo, X., Dasgupta, T., Xie, M., and Liu, R. Y. (2021). Leveraging the fisher randomization test using confidence distributions: Inference, combination and fusion learning. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83:777–797.
- Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review*, 2:319–323.
- Manski, C. F. (2003). *Partial Identification of Probability Distributions*. New York: Springer.
- Mattei, A., Li, F., and Mealli, F. (2013). Exploiting multiple outcomes in bayesian principal stratification analysis with application to the evaluation of a job training program. *Annals of Applied Statistics*, 7:2336–2360.
- McCrory, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142:698–714.
- McDonald, C. J., Hui, S. L., and Tierney, W. M. (1992). Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD Computing: Computers in Medical Practice*, 9:304–312.
- McGrath, S., Young, J. G., and Hernán, M. A. (2021). Revisiting the g-null paradox. *Epidemiology*, 33:114–120.
- Mealli, F. and Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108:1120–1131.
- Meinert, C. L., Knatterud, G. L., Prout, T. E., and Klimt, C. R. (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. ii. mortality results. *Diabetes*, 19:Suppl–789.
- Mercatanti, A. and Li, F. (2014). Do debit cards increase household spending? evidence from a semiparametric causal analysis of a survey. *Annals of Applied Statistics*, 8:2485–2508.

- Ming, K. and Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, 56:118–124.
- Ming, K. and Rosenbaum, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, 10:455–463.
- Miratrix, L. W., Sekhon, J. S., and Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:369–396.
- Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40:1263–1282.
- Mukerjee, R., Dasgupta, T., and Rubin, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113:868–881.
- Naimi, A. I., Cole, S. R., and Kennedy, E. H. (2017). An introduction to g methods. *International Journal of Epidemiology*, 46:756–762.
- Negi, A. and Wooldridge, J. M. (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 40:504–534.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles (with discussion). section 9 (translated). reprinted ed. *Statistical Science*, 5:465–472.
- Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society*, 97:558–625.
- Neyman, J. (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement to the Journal of the Royal Statistical Society*, 2:107–180.
- Nguyen, T. Q., Schmid, I., Ogburn, E. L., and Stuart, E. A. (2021). Clarifying causal mediation analysis for the applied researcher: Effect identification via three assumptions and five potential outcomes. *Psychological Methods*, 26:255–271.
- Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112:1720–1732.
- Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika*, 82:669–688.

- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.
- Pearl, J. (2001). Direct and indirect effects. In Breese, J. S. and Koller, D., editors, *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 411–420. pp. 411–420. San Francisco: Morgan Kaufmann Publishers Inc.
- Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In Grunwald, P. and Spirtes, P., editors, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, Corvallis, OR: 425–432. Association for Uncertainty in Artificial Intelligence.
- Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, 174:1223–1227.
- Pearl, J. (2018). Does obesity shorten life? Or is it the soda? On non-manipulable causes. *Journal of Causal Inference*, 6:20182001.
- Pearl, J. and Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29:579–595.
- Permutt, T. and Hebel, J. R. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 45:619–622.
- Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9:Article39.
- Pimentel, S. D., Yoon, F., and Keele, L. (2015). Variable-ratio matching with fine balance in a study of the Peer Health Exchange. *Statistics in Medicine*, 34:4070–4082.
- Poole, C. (2010). On the origin of risk relativism. *Epidemiology*, 21:3–9.
- Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge.
- Powers, D. E. and Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology*, 76:266–278.
- Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411.
- Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 65:161–172.
- Reichenbach, H. (1957). *The Direction of Time*. University of California Press.

- Rigdon, J. and Hudgens, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, 34:924–935.
- Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science*, 22:544–559.
- Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese*, 121:151–179.
- Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155.
- Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.
- Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495.
- Robins, J. M. and Wasserman, L. A. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, volume 409–420.
- Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A*, 147:656–666.
- Rosenbaum, P. R. (1987a). Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387–394.
- Rosenbaum, P. R. (1987b). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74:13–26.
- Rosenbaum, P. R. (1989). The role of known effects in observational studies. *Biometrics*, 45:557–569.
- Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies (with discussion). *Statistical Science*, 17:286–327.
- Rosenbaum, P. R. (2002b). *Observational Studies*. Springer, 2nd edition.
- Rosenbaum, P. R. (2015). Two R packages for sensitivity analysis in observational studies. *Observational Studies*, 1:1–17.
- Rosenbaum, P. R. (2018). Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. *Annals of Applied Statistics*, 12:2312–2334.

- Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, 7:143–176.
- Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society - Series B (Statistical Methodology)*, 45:212–218.
- Rosenbaum, P. R. and Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.
- Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524.
- Rosenbaum, P. R. and Rubin, D. B. (2023). Propensity scores in the design of observational studies for causal effects. *Biometrika*, 110:1–13.
- Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.
- Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *The Proceedings of the social statistics section of the American Statistical Association*, volume 233, page 239. American Statistical Association Alexandria, VA.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.
- Rubin, D. B. (1980). Comment on “Randomization analysis of experimental data: the Fisher randomization test” by D. Basu. *Journal of American Statistical Association*, 75:591–593.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Degism, modeling, decisions. *Journal of American Statistical Association*, 100:322–331.
- Rubin, D. B. (2006a). Causal inference through potential outcomes and principal stratification: application to studies with “censoring” due to death (with discussion). *Statistical Science*, 21:299–309.
- Rubin, D. B. (2006b). *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26:20–36.

- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2:808–840.
- Rudolph, K. E., Goin, D. E., Paksarian, D., Crowder, R., Merikangas, K. R., and Stuart, E. A. (2018). Causal mediation analysis with observational data: considerations and illustration examining mechanisms linking neighborhood poverty to adolescent substance use. *American Journal of Epidemiology*, 188:598–608.
- Sabbaghi, A. and Rubin, D. B. (2014). Comments on the Neyman–Fisher controversy and its consequences. *Statistical Science*, 29:267–284.
- Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Henry Holt and Company.
- Sanders, E., Gustafson, P. and Karim, M. E. (2021). Incorporating partial adherence into the principal stratification analysis framework. *Statistics in Medicine*, 40:3625–3644.
- Sanderson, E., Macdonald-Wallis, C., and Davey Smith, G. (2017). Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. *International Journal of Epidemiology*, 47:587–596.
- Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1120.
- Schlesselman, J. J. (1978). Assessing effects of confounding variables. *American Journal of Epidemiology*, 108:3–8.
- Schochet, P. Z., Burghardt, J., and McConnell, S. (2008). Does job corps work? impact findings from the national job corps study. *American Economic Review*, 98:1864–1886.
- Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, 12:487–508.
- Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, 47:1–52.
- Sekhon, J. S. and Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. In *Regression Discontinuity Designs*, volume 38. Emerald Publishing Limited.
- Shinozaki, T. and Matsuyama, Y. (2015). Doubly robust estimation of standardized risk difference and ratio in the exposed population. *Epidemiology*, 26:873–877.

- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13:290–312.
- Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. *Sociological Methodology*, 16:159–186.
- Sommer, A. and Zeger, S. L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine*, 10:45–52.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25:1–21.
- Stuart, E. A. and Jo, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Statistical Methods in Medical Research*, 24:657–674.
- Tao, Y. and Fu, H. (2019). Doubly robust estimation of the weighted average treatment effect for a target population. *Statistics in Medicine*, 38:315–325.
- Theil, H. (1953). Estimation and simultaneous correlation in complete equation systems. central planning bureau. Technical report, Mimeo, The Hague.
- Thistlewaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51:309.
- Thistlewaite, D. L. and Campbell, D. T. (2016). Regression-discontinuity analysis: An alternative to the ex-post facto experiment (with discussion). *Observational Studies*, 2:119–209.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288.
- Titterington, D. (2013). *Biometrika* highlights from volume 28 onwards. *Biometrika*, 100:17–73.
- Valeri, L. and Vanderweele, T. J. (2014). The estimation of direct and indirect causal effects in the presence of misclassified binary mediator. *Biostatistics*, 15:498–512.
- Van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.
- Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
- Van Elteren, P. (1960). On the combination of independent two-sample tests of wilcoxon. *Bulletin of the Institute of International Statistics*, 37:351–361.

- VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*, 78:2957–2962.
- VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.
- VanderWeele, T. J., Asomaning, K., and Tchetgen Tchetgen, E. J. (2012). Genetic variants on 15q25.1, smoking, and lung cancer: An assessment of mediation and interaction. *American Journal of Epidemiology*, 175:1013–1020.
- VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, 167:268–274.
- VanderWeele, T. J. and Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67:1406–1413.
- VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79:917–938.
- VanderWeele, T. J., Tchetgen Tchetgen, E. J., Cornelis, M., and Kraft, P. (2014). Methodological challenges in Mendelian randomization. *Epidemiology*, 25:427.
- Vansteelandt, S. and Daniel, R. M. (2014). On regression adjustment for the propensity score. *Statistics in Medicine*, 33:4053–4072.
- Vansteelandt, S. and Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 84:657–685.
- Vansteelandt, S. and Joffe, M. (2014). Structural nested models and G-estimation: the partially realized promise. *Statistical Science*, 29:707–731.
- Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110:1024–1036.
- Voight, B. F., Peloso, G. M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M. K., Hindy, G., Hólm, H., Ding, E. L., and Johnson, T. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *The Lancet*, 380:572–580.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113:1228–1242.

- Wager, S., Du, W., Taylor, J., and Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 113:12673–12678.
- Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11:284–300.
- Wang, L., Zhang, Y., Richardson, T. S., and Zhou, X.-H. (2020). Robust estimation of propensity score weights via subclassification. *arXiv preprint arXiv:1602.06366*.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838.
- Wooldridge, J. (2016). Should instrumental variables be used as matching variables? *Research in Economics*, 70:232–237.
- Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50:420–445.
- Wu, J. and Ding, P. (2021). Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, 116:1898–1913.
- Yang, F. and Small, D. S. (2016). Using post-outcome measurement information in censoring-by-death problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:299–318.
- Yang, S. and Ding, P. (2018). Asymptotic causal inference with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105:487–493.
- Zelen, M. (1979). A new design for randomized clinical trials. *New England Journal of Medicine*, 300:1242–1245.
- Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *Journal of Educational and Behavioral Statistics*, 28:353–368.
- Zhang, J. L., Rubin, D. B., and Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104:166–176.
- Zhang, M. and Ding, P. (2022). Interpretable sensitivity analysis for the baron-kenny approach to mediation with unmeasured confounding. *arXiv preprint arXiv:2205.08030*.
- Zhao, A. and Ding, P. (2021a). Covariate-adjusted Fisher randomization tests for the average treatment effect. *Journal of Econometrics*, 225:278–294.

Zhao, A. and Ding, P. (2021b). No star is good news: A unified look at rerandomization based on p -values from covariate balance tests. *arXiv preprint arXiv:2112.10545*.

Zhao, Q., Wang, J., Heman, G., Bowden, J., and Small, D. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of Statistics*, 48:1742–1769.