

Linear Model and Extensions

Peng Ding
Department of Statistics
University of California, Berkeley



*To students and readers
who are interested in linear models
and their applications to real-world problems*



Contents

Acronyms	xiii
Symbols	xv
Useful R packages	xvii
Preface	xix
I Introduction	1
1 Motivations for Statistical Models	3
1.1 Data and statistical models	3
1.2 Why linear models?	6
2 Ordinary Least Squares with a Univariate Covariate	7
2.1 Ordinary least squares with a univariate covariate	7
2.2 Final comments	9
2.3 Homework problems	10
II Ordinary Least Squares and Statistical Inference	13
3 Ordinary Least Squares with Multiple Covariates	15
3.1 The OLS formula	15
3.2 The geometry of OLS	17
3.3 The projection matrix from OLS	19
3.4 Homework problems	20
4 Gauss–Markov Model and Gauss–Markov Theorem	25
4.1 Gauss–Markov model	25
4.2 Properties of the OLS estimator	26
4.3 Variance estimation	28
4.4 Gauss–Markov Theorem	28
4.5 Homework problems	31
5 Normal Linear Model: Inference and Prediction	35
5.1 Joint distribution of the OLS coefficient and variance estimator	36
5.2 Pivotal quantities and statistical inference	37
5.2.1 Scalar parameters	37
5.2.2 Vector parameters	38
5.3 Prediction based on pivotal quantities	40
5.4 Examples	41

5.4.1	Univariate regression	41
5.4.2	Anscombe's Quartet: the importance of graphical diagnostics	42
5.4.3	Multivariate regression	45
5.5	Homework problems	47
6	Asymptotic Inference in OLS: Eicker–Huber–White (EHW) robust standard error	51
6.1	Motivation	51
6.1.1	Numerical examples	51
6.1.2	Goal of this chapter	52
6.2	Consistency of OLS	54
6.3	Asymptotic Normality of the OLS estimator	55
6.4	Eicker–Huber–White standard error	56
6.4.1	Sandwich variance estimator	56
6.4.2	Other heteroskedasticity-consistent (HC) standard errors	58
6.4.3	Special case with homoskedasticity	59
6.5	Examples	60
6.5.1	LaLonde experimental data	60
6.5.2	Data from King and Roberts (2015)	60
6.5.3	Boston housing data	62
6.6	Final remarks	64
6.7	Homework problems	64
III	Interpretation of Ordinary Least Squares Based on Partial Regressions	67
7	Frisch–Waugh–Lovell Theorem	69
7.1	Long and short regressions	69
7.2	FWL Theorem for the regression coefficients	70
7.3	FWL Theorem for standard errors	72
7.4	Gram–Schmidt orthogonalization, QR decomposition, and computation of OLS	73
7.5	Homework problems	76
8	Applications of the Frisch–Waugh–Lovell Theorem	79
8.1	Centering regressors	79
8.1.1	Intercept and centering regressors	79
8.1.2	Dummy variables and centering regressors within groups	81
8.2	Partial correlation coefficient and Simpson's paradox	81
8.3	Hypothesis testing and analysis of variance	84
8.4	Homework problems	88
9	Cochran's Formula and Omitted-Variable Bias	93
9.1	Cochran's formula	93
9.2	Omitted-variable bias	95
9.3	Homework problems	96

IV	Model Fitting, Checking, and Misspecification	99
10	Multiple Correlation Coefficient	101
10.1	Equivalent definitions of the multiple correlation coefficient	101
10.2	The multiple correlation coefficient and F statistic	102
10.3	Numerical examples	103
10.4	Homework problems	104
11	Leverage Scores and Leave-One-Out Formulas	107
11.1	Leverage scores	107
11.1.1	The average leverage score equals p/n	107
11.1.2	The leverage scores are all bounded between 0 and 1	108
11.1.3	The i th leverage score measures the impact of the i th observation in prediction	108
11.1.4	The i th leverage score measures whether x_i is an outlier compared with other covariates	109
11.1.5	Other properties of the leverage scores	109
11.2	Leave-one-out formulas	110
11.3	Applications of the leave-one-out formulas	112
11.3.1	Gauss updating formula	112
11.3.2	Outlier detection based on residuals	113
11.3.3	Jackknife	114
11.4	Homework problems	117
12	Population Ordinary Least Squares and Misspecified Linear Model	119
12.1	Population OLS	119
12.2	Population FWL Theorem and Cochran's formula	121
12.3	Population R^2 and partial correlation coefficient	123
12.4	Inference for the population OLS	125
12.4.1	Inference with the Eicker–Huber–White standard errors	125
12.5	To model or not to model?	125
12.5.1	Population OLS and the restricted mean model	125
12.5.2	More on residual plots	127
12.6	Conformal prediction based on exchangeability	130
12.7	Homework problems	133
V	Overfitting, Regularization, and Model Selection	139
13	Perils of Overfitting	141
13.1	David Freedman's simulation	141
13.2	Variance inflation factor	143
13.3	Bias-variance trade-off	144
13.4	Model selection criteria	146
13.4.1	RSS, R-squared and adjusted R-squared	146
13.4.2	Information criteria	148
13.4.3	Cross-validation (CV)	148
13.5	Best subset and forward/backward selection	149
13.6	Homework problems	151
14	Ridge Regression	153
14.1	Introduction to ridge regression	153

14.2	Ridge regression via the SVD of the covariate matrix	155
14.3	Statistical properties	156
14.4	Selection of the tuning parameter	158
14.4.1	Based on parameter estimation	158
14.4.2	Based on prediction	158
14.5	Computation of ridge regression	159
14.6	Numerical examples	160
14.6.1	Uncorrelated covariates	160
14.6.2	Correlated covariates	161
14.7	Further comments on OLS, ridge, and PCA	162
14.8	Homework problems	164
15	Lasso	169
15.1	Introduction to the lasso	169
15.2	Comparing the lasso and ridge: a geometric perspective	170
15.3	Computing the lasso coefficients via coordinate descent	172
15.3.1	The soft-thresholding lemma	172
15.3.2	Coordinate descent for the lasso	172
15.4	Example: comparing OLS, ridge and lasso	174
15.5	Other shrinkage estimators	176
15.5.1	Bridge estimator	176
15.5.2	Elastic net	176
15.6	Homework problems	177
VI	Transformation and Weighting	181
16	Transformations in OLS	183
16.1	Transformation of the outcome	183
16.1.1	Log transformation	183
16.1.2	Box–Cox transformation	184
16.2	Transformation of the covariates	186
16.2.1	Polynomial, basis expansion, and generalized additive model	186
16.2.2	Regression discontinuity and regression kink	188
16.3	Homework problems	191
17	Interactions in OLS	193
17.1	Two binary covariates interact	193
17.2	A binary covariate interacts with a general covariate	194
17.2.1	Treatment effect heterogeneity	194
17.2.2	Johnson–Neyman technique	194
17.2.3	Blinder–Oaxaca decomposition	194
17.2.4	Chow test	196
17.3	Difficulties of interaction	196
17.3.1	Removable interaction	196
17.3.2	Main effect in the presence of interaction	197
17.3.3	Power	198
17.4	Homework problems	199
18	Restricted OLS	201
18.1	Examples	201

18.2 Algebraic properties	202
18.3 Statistical inference	203
18.4 Final remarks	204
18.5 Homework problems	204
19 Weighted Least Squares	209
19.1 Generalized least squares	209
19.2 Weighted least squares	211
19.3 WLS motivated by heteroskedasticity	212
19.3.1 Feasible generalized least squares	212
19.3.2 Aggregate data and ecological regression	214
19.4 WLS Motivated by Survey Weights	216
19.5 WLS as a Building Block for Local linear regression	218
19.6 Homework problems	220
VII Generalized Linear Models	225
20 Logistic Regression for Binary Outcomes	227
20.1 Regression with binary outcomes	227
20.1.1 Linear probability model	227
20.1.2 General link functions	228
20.2 Maximum likelihood estimator of the logistic model	230
20.3 Statistics with the logit model	232
20.3.1 Inference	232
20.3.2 Prediction	234
20.4 More on the interpretations of the coefficients	235
20.5 Does the link function matter?	236
20.6 Extensions of the logistic regression	239
20.6.1 Penalized logistic regression	239
20.6.2 Case-control study	239
20.7 Other model formulations	241
20.7.1 Latent linear model	241
20.7.2 Inverse model	241
20.8 Homework problems	243
21 Logistic Regressions for Categorical Outcomes	247
21.1 Multinomial distribution	247
21.2 Multinomial logistic model for nominal outcomes	248
21.2.1 Modeling	248
21.2.2 MLE	249
21.3 A latent variable representation for the multinomial logistic regression	251
21.4 Proportional odds model for ordinal outcomes	252
21.5 A case study	254
21.5.1 Binary logistic for the treatment	255
21.5.2 Binary logistic for the outcome	256
21.5.3 Multinomial logistic for the outcome	256
21.5.4 Proportional odds logistic for the outcome	257
21.6 Discrete choice models	259
21.6.1 Model	259
21.6.2 MLE	260

21.6.3	Example	260
21.6.4	More comments	263
21.7	Homework problems	263
22	Regression Models for Count Outcomes	265
22.1	Some random variables for counts	265
22.1.1	Poisson	265
22.1.2	Negative-Binomial	266
22.1.3	Zero-inflated count distributions	267
22.2	Regression models for counts	268
22.2.1	Poisson regression	268
22.2.2	Negative-Binomial regression	270
22.2.3	Zero-inflated regressions	271
22.3	A case study	272
22.3.1	Linear, Poisson, and Negative-Binomial regressions	272
22.3.2	Zero-inflated regressions	273
22.4	Homework problems	276
23	Generalized Linear Models: A Unification	279
23.1	Generalized Linear Models	279
23.1.1	Exponential family	279
23.1.2	Generalized linear model	281
23.2	MLE for GLM	282
23.3	Other GLMs	284
23.4	Homework problems	286
24	Misspecified Generalized Linear Models: Restricted Mean Models and Sandwich Covariance Matrix	289
24.1	Restricted mean model	289
24.2	Sandwich covariance matrix	290
24.3	Applications of the sandwich standard errors	292
24.3.1	Linear regression	292
24.3.2	Logistic regression	293
24.3.2.1	An application	293
24.3.2.2	A misspecified logistic regression	294
24.3.3	Poisson regression	294
24.3.3.1	A correctly specified Poisson regression	294
24.3.3.2	A Negative-Binomial regression model	295
24.3.3.3	Misspecification of the conditional mean	295
24.3.4	How robust are the robust standard errors?	296
24.4	Homework problems	296
25	Generalized Estimating Equation for Correlated Multivariate Data	299
25.1	Examples of correlated data	299
25.1.1	Longitudinal data	299
25.1.2	Clustered data: a neuroscience experiment	299
25.1.3	Clustered data: a public health intervention	300
25.2	Marginal model and the generalized estimating equation	301
25.3	Statistical inference with GEE	303
25.3.1	Computation using the Gauss–Newton method	303
25.3.2	Asymptotic inference	303

25.3.3	Implementation: choice of the working covariance matrix	304
25.4	A special case: cluster-robust standard error	305
25.4.1	OLS	305
25.4.2	Logistic regression	306
25.5	Application	307
25.5.1	Clustered data: a neuroscience experiment	307
25.5.2	Clustered data: a public health intervention	308
25.5.3	Longitudinal data	309
25.6	Critiques on the key assumptions	310
25.6.1	Assumption (25.4)	310
25.6.2	Assumption (25.5)	312
25.7	Final comments	313
25.7.1	Explanation versus prediction	313
25.7.2	Small number of clusters	313
25.8	Homework problems	313

VIII Beyond Modeling the Conditional Mean 315

26 Quantile Regression 317

26.1	From the mean to the quantile	317
26.2	From the conditional mean to conditional quantile	320
26.3	Sample regression quantiles	322
26.3.1	Computation	322
26.3.2	Asymptotic inference	323
26.4	Numerical examples	324
26.4.1	Sample quantiles	324
26.4.2	OLS versus LAD	325
26.5	Application	327
26.5.1	Parents' and children's heights	327
26.5.2	U.S. wage structure	327
26.6	Extensions	328
26.6.1	Cluster-robust standard error for quantile regression	328
26.6.2	High-dimensional quantile regression	329
26.7	Homework problems	329

27 Modeling Time-to-Event Outcomes 331

27.1	Examples	331
27.1.1	Survival analysis	331
27.1.2	Duration analysis	332
27.2	Time-to-event data	333
27.3	Some examples of random variables for time to event	334
27.4	Kaplan–Meier survival curve	336
27.5	Cox model for time-to-event outcome	339
27.5.1	Cox model and its interpretation	341
27.5.2	Partial likelihood	342
27.5.3	Examples	344
27.5.4	Log-rank test as a score test from Cox model	347
27.6	Extensions	350
27.6.1	Stratified Cox proportional hazards model	350
27.6.2	Clustered Cox model	351

27.6.3 Penalized Cox model	352
27.7 Critiques on survival analysis	352
27.8 Homework problems	353

IX Appendices 355

A Linear Algebra 357

A.1 Basics of vectors and matrices	357
A.2 Vector calculus	366
A.3 Homework problems	368

B Random Variables 371

B.1 Some important univariate random variables	371
B.1.1 Normal, chi-squared, t and F	371
B.1.2 Beta–Gamma duality	372
B.1.3 Exponential, Laplace, and Gumbel distributions	373
B.2 Multivariate distributions	374
B.3 Multivariate Normal and its properties	376
B.4 Quadratic forms of random vectors	377
B.5 Homework problems	379

C Limiting Theorems and Basic Asymptotics 383

C.1 Convergence in probability	383
C.2 Convergence in distribution	384
C.3 Tools for proving convergence in probability and distribution	387

D M-Estimation and MLE 389

D.1 M-estimation	389
D.2 Maximum likelihood estimator	392
D.3 Homework problems	395

Bibliography 397

Acronyms

I try hard to avoid using acronyms to reduce the unnecessary burden for reading. However, the following acronyms are standard and will be used repeatedly.

ANOVA	(Fisher's) analysis of variance
BLUE	best linear unbiased estimator (in Gauss–Markov Theorem)
CATE	conditional average treatment effect
CDF	cumulative distribution function
CLT	central limit theorem
CV	cross-validation
EHW	Eicker–Huber–White (robust covariance matrix or standard error)
FDA	U.S. Food and Drug Administration
FWL	Frisch–Waugh–Lovell (theorem)
GEE	generalized estimating equation
GLM	generalized linear model
HC	heteroskedasticity-consistent (covariance matrix or standard error)
IID	independent and identically distributed
LAD	least absolute deviations
lasso	least absolute shrinkage and selection operator
MLE	maximum likelihood estimate
OLS	ordinary least squares
PCA	principal component analysis
RCT	randomized controlled trial
RSS	residual sum of squares
SVD	singular value decomposition
WLS	weighted least squares



Symbols

All vectors are column vectors as in \mathbb{R} unless stated otherwise. Let the superscript “ T ” denote the transpose of a vector or matrix.

\mathbb{R}	the set of all real numbers
\mathbb{R}^p	the set of p -dimensional vectors with real components
$\perp\!\!\!\perp$	independence and conditional independence
$\overset{\text{IID}}{\sim}$	independent and identically distributed (IID)
$\overset{\text{a}}{\sim}$	approximation in distribution
I_n	identity matrix of dimension $n \times n$
x_i	covariate vector for unit i
y_i	outcome for unit i
X	covariate matrix
Y	outcome vector
H	hat matrix $H = X(X^{\text{T}}X)^{-1}X^{\text{T}}$
h_{ii}	leverage score: the (i, i) th element of the hat matrix H
β	regression coefficient
ε	error term



Useful R packages

This book uses the following R packages and functions.

package	function or data	use
car	hccm	Eicker–Huber–White robust standard error
	linearHypothesis	testing linear hypotheses in linear models
foreign	read.dta	read stata data
gee	gee	Generalized estimating equation
HistData	GaltonFamilies	Galton’s data on parents’ and children’s heights
MASS	lm.ridge	ridge regression
	glm.nb	Negative-Binomial regression
glmnet	cv.glmnet	Lasso with cross-validation
mlbench	BostonHousing	Boston housing data
	polr	proportional odds logistic regression
Matching	lalonge	LaLonde data
nnet	multinom	Multinomial logistic regression
quantreg	rq	quantile regression
survival	coxph	Cox proportional hazards regression
	survdif	log rank test
	survfit	Kaplan–Meier curve
ElemStatLearn	prostate	data for Hastie et al. (2009)
wooldridge	mroz	data for Wooldridge (2016)



Preface

The importance of studying the linear model

A central task in statistics is to use data to build models to make inferences about the underlying data-generating processes or make predictions of future observations. Although real problems are very complex, the linear model can often serve as a good approximation to the true data-generating process. Sometimes, although the true data-generating process is nonlinear, the linear model can be a useful approximation if we properly transform the data based on domain knowledge. Even in highly nonlinear problems, the linear model can still be a useful first attempt in the data analysis process.

Moreover, the linear model has many elegant algebraic and geometric properties. Under the linear model, we can derive many explicit formulas to gain insights about various aspects of statistical modeling. In more complicated models, deriving explicit formulas may be impossible. Nevertheless, we can use the linear model to build intuition and make conjectures about more complicated models.

Pedagogically, the linear model serves as a building block in the whole statistical training. This book builds on my lecture notes for a master's level "Linear Model" course at UC Berkeley, taught over the past ten years. Most students are master's students in statistics. Some are undergraduate students with strong technical preparations. Some are Ph.D. students in statistics. Some are master's or Ph.D. students in other departments. This book requires the readers to have basic training in linear algebra, probability theory, and statistical inference.

Recommendations for instructors

This book has twenty-seven chapters in the main text and four appendices. As I mentioned before, this book grows out of my teaching of "Linear Model" at UC Berkeley. In different years, I taught the course in different ways, and this book is a union of my lecture notes over the past ten years. Below I make some recommendations for instructors based on my own teaching experience. Since UC Berkeley is on the semester system, instructors on the quarter system should make some adjustments to my recommendations below.

Version 1: a basic linear model course assuming minimal technical preparations

If you want to teach a basic linear model course without assuming strong technical preparations from the students, you can start from the appendices by reviewing basic linear algebra, probability theory, and statistical inference. Then you can cover Chapters 2–17. If time permits, you can consider covering Chapter 20 due to the importance of the logistic model for binary data.

Version 2: an advanced linear model course assuming strong technical preparations

If you want to teach an advanced linear model course assuming strong technical preparations from the students, you can start with the main text directly. When I did this, I asked my teaching assistants to review the appendices in the first two lab sessions and assigned

homework problems from the appendices to remind the students to review the background materials. Then you can cover Chapters 2–24. You can omit Chapter 18 and some sections in other chapters due to their technical complications. If time permits, you can consider covering Chapter 25 due to the importance of the generalized estimating equation as well as its byproduct called the “cluster-robust standard error,” which is important for many social science applications. Furthermore, you can consider covering Chapter 27 due to the importance of the Cox proportional hazards model.

Version 3: an advanced generalized linear models course

If you want to teach a course on generalized linear models, you can use Chapters 20–27.

Additional recommendations for readers and students

Readers and students can first read my recommendations for instructors above. In addition, I have three other recommendations.

More simulation studies

This book contains some basic simulation studies. I encourage the readers to conduct more intensive simulation studies to deepen their understanding of the theory and methods.

Practical data analysis

Box wrote wisely that “all models are wrong but some are useful.” The usefulness of models strongly depends on the applications. When teaching “Linear Model,” I sometimes replaced the final exam with the final project to encourage students to practice data analysis and make connections between the theory and applications.

Homework problems

This book contains many homework problems. It is important to try some homework problems. Moreover, some homework problems contain useful theoretical results, and some are even stated as theorems. Even if you do not have time to figure out the details for those problems, it is helpful to at least read the statements of the problems.

Omitted topics

Although “Linear Model” is a standard course offered by most statistics departments, it is not entirely clear what we should teach as the field of statistics is evolving. When I was teaching “Linear Model,” I asked myself the following question many times:

What are the most widely used statistics methods?

Arguably, in R, the following five functions are the top choices by empirical researchers:

- `lm()`,
- `glm()`,
- `coxph()` in the package `survival`,
- `gee()` in the package `gee`,
- `rq()` in the package `quantreg`,

which are for

- linear model,
- generalized linear model including logistic regression and Poisson regression,
- Cox proportional hazards model,
- generalized estimating equations,
- quantile regression,

respectively. Readers can view this book as a tutorial for these five functions. However, my selection of the topics was biased by my own experience with applied statistics. Some readers may feel that this book has omitted some important topics related to the linear model. I make some brief comments below.

Bootstrap

Efron (1979) proposed the bootstrap as a powerful method to estimate the variance of general estimators. Of course, it is also useful for estimating the variances of all estimators discussed in this book. For instance, the bootstrap works well for generalized linear models. If we sample with replacement from the data $\{x_i, y_i\}_{i=1}^n$, we can recalculate the maximum likelihood estimators to obtain the bootstrap variance estimators. It turns out that the bootstrap variance estimators approximate the sandwich variance estimators discussed in Chapter 24. Therefore, bootstrap variance estimators are robust to misspecified models (Buja et al., 2019a,b). The discussion extends to clustered data. In particular, we can the cluster bootstrap by resampling the clusters to approximate the cluster-robust standard error in Chapter 25.4. Overall, the bootstrap seems like a magic!

I largely ignore the discussion of the bootstrap for at least the following four reasons. First, the statistical models in this book are all “nice” models that enjoy explicit analytic approximations of the variances. The bootstrap is not crucial for them. Second, this book aims to provide insights into specific models, including deriving explicit formulas of the variance estimators. Third, this book can be viewed as a tutorial for the `R` functions mentioned above. Those `R` functions use the explicit formulas. Fourth, although the bootstrap is intuitive, its rigorous theory requires advanced proving techniques. These technical issues are beyond the scope of this book; see Wu (1986); Mammen (1992); Shao and Tu (1995) for further discussion.

Nevertheless, it is impossible to completely ignore the bootstrap in this book. In Chapter 26.4, the small simulation study on quantile regression demonstrates the advantage of the bootstrap: it performs better than other existing variance estimators based on analytic formulas because those estimators are sensitive to the choices of tuning parameters. If your computational resource permits, the bootstrap is attractive for variance estimation.

Bayesian methods

Bayesian methods are powerful in applied data analysis. However, Bayesian methods are more demanding for computation and require specifications of prior distributions by the users. Therefore, I feel that they are more advanced statistical methods and thus are beyond the scope of this book. Hoff (2009) and Gelman et al. (2013) are two excellent books on Bayesian statistics.

Advanced econometric models

After the linear model, many econometric textbooks cover the instrumental variable models and panel data models. For these more specialized topics, Wooldridge (2010) is a canonical textbook.

Advanced biostatistics models

This book covers the generalized estimating equation in Chapter 25. For analyzing longitudinal data, linear and generalized linear mixed effects models are powerful tools. Fitzmaurice et al. (2012) is a canonical textbook on applied longitudinal data analysis. Song (2007) is a textbook on general correlated data analysis.

This book also covers the Cox proportional hazards model in Chapter 27. For more advanced methods for survival analysis, Kalbfleisch and Prentice (2011) is a canonical textbook.

Causal inference

I do not cover causal inference in this book intentionally. To minimize the overlap of the materials, I wrote another textbook on causal inference (Ding, 2024). However, at UC Berkeley, I did teach a version of “Linear Model” with a causal inference unit after introducing the basics of linear model and logistic model. Students seemed to like it because of the connections between statistical models and causal inference.

Features of the book

The linear model is an old topic in statistics. There are already many excellent textbooks on the linear model. This book has the following features.

- This book provides an intermediate-level introduction to the linear model. It balances rigorous proofs and intuitive explanations.
- This book introduces the theory of misspecified models and emphasizes its implications for practical data analysis. The mathematical theory of misspecified models dated back at least to Huber (1967) and attracted research interest from academic statisticians ever since. However, it is not common to see the theory introduced in standard textbooks on statistical modeling and data analysis.
- This book provides not only theory but also simulation studies and case studies.
- This book provides the R code and data to replicate all simulation studies and case studies at Harvard Dataverse:

<https://doi.org/10.7910/DVN/DBDYVJ>

- This book covers the theory of the linear model related to social sciences and biomedical studies.
- This book provides homework problems with different technical difficulties.
- For instructors who teach related courses using this book, the solutions to the problems are available upon request.

Other textbooks may also have one or two of the above features. This book has the above features simultaneously. I hope that instructors and readers find these features attractive.

Acknowledgments

Many students at UC Berkeley made critical and constructive comments on early versions of my lecture notes. As teaching assistants for the “Linear Model” course, Sizhu Lu, Yaxuan Huang, Andy Shen, Chaoran Yu, and Jason Wu read early versions of my book carefully and helped me to improve the book a lot.

Professors Hongyuan Cao, Zhichao Jiang, and Richard Guo taught related courses based on an early version of the book. They made very valuable suggestions.

My research collaborations with Anqi Zhao, Lihua Lei, Zhichao Jiang, Dennis Shen, Wen Zhou, Zifeng Zhang, and Mingrui Zhang used a lot of results related to this book. They helped me sharpen many statements in this book.

I am also very grateful for the suggestions from Professors Nianqiao Ju, Alan Agresti, and Peter XK Song. My colleague, Professor Bin Yu, and more broadly, the intellectual environment at UC Berkeley encouraged me to teach and write more about the theory of misspecified statistical models and its implications for practical data analysis.

When I was a student, I took a linear model course based on Weisberg (2005). In my early years of teaching, I used Christensen (2002) and Agresti (2015) as reference books. When I was a junior faculty at UC Berkeley, I sat in Professor Jim Powell’s econometrics courses and got access to his wonderful lecture notes. They all heavily impacted my understanding and formulation of the linear model.

The U.S. National Science Foundation partially supported my research over the years (grant numbers # 1712714, # 1745640, and # 1945136).

Contacting me

Please feel free to email me at

pengdingpku@berkeley.edu

if you identify any errors in the book, or if you use the book for teaching and want the solutions to the homework problems.



Part I

Introduction



1

Motivations for Statistical Models

This book is about the linear model and its extensions. Before delving into the mathematical details of specific models, I will briefly provide some motivations for studying statistical models.

1.1 Data and statistical models

A wide range of problems in statistics and machine learning have the following data structure:

Unit	outcome/response	covariates/features/predictors			
i	Y	X_1	X_2	\cdots	X_p
1	y_1	x_{11}	x_{12}	\cdots	x_{1p}
2	y_2	x_{21}	x_{22}	\cdots	x_{2p}
\vdots	\vdots	\vdots	\vdots		\vdots
n	y_n	x_{n1}	x_{n2}	\cdots	x_{np}

For each unit i , we observe the outcome of interest (also called the response), y_i , as well as p covariates (also called features or predictors), x_{i1}, \dots, x_{ip} . We often use

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

to denote the n -dimensional outcome vector, and

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

to denote the $n \times p$ covariate matrix, also called the *design matrix*. In most cases, the first column of X contains constants 1s.

Based on the data (X, Y) , we can ask the following questions:

- (Q1) Describe the relationship between X and Y , i.e., their association or correlation. For example, how is the patients' average height related to the children's average height? How is one's height related to one's weight? How are one's education and working experience related to one's income?

- (Q2) Predict Y^* with new data X^* , based on the old data (X, Y) . In particular, we want to use the current data (X, Y) to train a predictor, and then use it to predict future Y^* based on future X^* . This is called *supervised learning* in the field of machine learning. For example, how do we predict whether an email is spam or not based on the frequencies of the most commonly occurring words and punctuation marks in the email? How do we predict cancer patients' survival time based on their clinical measures?
- (Q3) Estimate the causal effect of some components in X on Y . What if we change some components of X ? How do we measure the impact of the hypothetical intervention of some components of X on Y ? This is a much harder question because most statistical tools are designed to infer association, not causation. For example, the U.S. Food and Drug Administration (FDA) approves drugs based on randomized controlled trials (RCT) because RCTs are most credible to infer causal effects of drugs on health outcomes. Economists are interested in evaluating the effect of a job training program on employment and wages. However, this is a notoriously difficult problem because participation in the job training program is not randomized in observational data.

The above descriptions are about generic X and Y , which can have many different types. We often use different statistical models to capture the features of different types of data. Below I give a brief overview of models that will appear in later parts of this book.

- (T1) X and Y are univariate and continuous. In Francis Galton's¹ classic example, X is the parents' average height and Y is the children's average height (Galton, 1886). Let \hat{y}_i denote the “fitted value” of the outcome for unit i with covariate value x_i . Galton derived the following formula:

$$\hat{y}_i = \bar{y} + \hat{\rho} \frac{\hat{\sigma}_y}{\hat{\sigma}_x} (x_i - \bar{x})$$

which is equivalent to

$$\frac{\hat{y}_i - \bar{y}}{\hat{\sigma}_y} = \hat{\rho} \frac{x_i - \bar{x}}{\hat{\sigma}_x}, \quad (1.1)$$

where

$$\bar{x} = n^{-1} \sum_{i=1}^n x_i, \quad \bar{y} = n^{-1} \sum_{i=1}^n y_i$$

are the sample means,

$$\hat{\sigma}_x^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \hat{\sigma}_y^2 = (n-1)^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

are the sample variances, and $\hat{\rho} = \hat{\sigma}_{xy} / (\hat{\sigma}_x \hat{\sigma}_y)$ is the sample Pearson correlation coefficient with the sample covariance

$$\hat{\sigma}_{xy} = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

The identity (1.1) is the famous formula of “regression towards mediocrity” or “regression towards the mean”. Galton first introduced the terminology “regression.”² Galton

¹Who was Francis Galton? He was Charles Darwin's nephew and was famous for his pioneer work in statistics and for devising a method for classifying fingerprints that proved useful in forensic science. He also invented the term *eugenics*, a field that causes a lot of controversies nowadays.

²The name “regression” is widely used in statistics now. For instance, we sometimes use “linear regression” interchangeably with “linear model.” We also extend the name to “logistic regression” or “Cox regression,” which will be discussed in later chapters of this book.

called regression because the relative deviation of the children's average height is smaller than that of the parents' average height if $|\hat{\rho}| < 1$. We will derive (1.1) in Chapter 2.

- (T2) Y is univariate and continuous, and X is multivariate of mixed types. In the `R` package `ElemStatLearn`, the dataset `prostate` has an outcome of interest as the log of the prostate-specific antigen `lpsa` and some potential predictors including the log cancer volume `lcavol`, the log prostate weight `lweight`, age `age`, etc. The main chapters of this book, Chapters 3–19, will discuss linear regression with multidimensional covariates.
- (T3) Y is binary or indicator of two classes, and X is multivariate of mixed types. For example, in the `R` package `wooldridge`, the dataset `mroz` contains an outcome of interest being the binary indicator for whether a woman was in the labor force in the year 1975, and some useful covariates are in the table below:

covariate name	covariate meaning
kidslt6	number of kids younger than six years old
kidsge6	number of kids between six and eighteen years old
age	age
educ	years of education
husage	husband's age
huseduc	husband's years of education

Chapter 20 will discuss *logistic regression* for binary outcomes.

- (T4) Y is categorical without ordering. For example, the choice of housing type, single-family house, townhouse, or condominium, is a categorical variable. Chapter 21 will discuss *multinomial logistic regression* for categorical outcomes without ordering.
- (T5) Y is categorical and ordered. For example, the final course evaluation at UC Berkeley can take value in $\{1, 2, 3, 4, 5, 6, 7\}$. These numbers have clear ordering but they are not the usual real numbers. Chapter 21 will discuss *proportional odds regression* for ordered categorical outcomes.
- (T6) Y represents counts. For example, the number of times one went to the gym last week is a non-negative integer representing counts. Chapter 22 will discuss regression models for count outcomes.
- (T7) Y is multivariate and correlated. In medical trials, the data are often longitudinal, meaning that the patient's outcomes are measured repeatedly over time. So each patient has a multivariate outcome. In field experiments of public health and development economics, the randomized interventions are often at the village level but the outcome data are collected at the household level. So within villages, the outcomes are correlated. Chapter 25 will discuss the *generalized estimating equation* for correlated data.
- (T8) Y represent time-to-event outcomes. For example, in medical trials, a major outcome of interest is the survival time; in labor economics, a major outcome of interest is the time to find the next job. The former is called *survival analysis* in biostatistics and the latter is called *duration analysis* in econometrics. Chapter 27 will discuss *Cox proportional hazards regression*.

1.2 Why linear models?

Why do we study linear models if many real problems may have nonlinear structures? There are important reasons.

- (R1) Linear models are simple but non-trivial starting points for learning.
- (R2) Linear models can provide insights because we can derive explicit formulas based on elegant algebra and geometry.
- (R3) Linear models can handle nonlinearity by incorporating nonlinear terms of covariates, for example, X can contain the polynomials or nonlinear transformations of the original covariates. In statistics, “linear” means *linear in parameters*, not *linear in covariates*.
- (R4) Linear models can be good approximations of nonlinear data-generating processes.
- (R5) Linear models are simpler than nonlinear models, but they do not necessarily perform worse than more complicated nonlinear models. We have finite data so we cannot fit arbitrarily complicated models.

If you are interested in nonlinear models, you can read another book on machine learning.

2

Ordinary Least Squares with a Univariate Covariate

This chapter discusses ordinary least squares (OLS) with a single covariate. It can provide insights into later chapters because it is the building block of OLS with multiple covariates.

2.1 Ordinary least squares with a univariate covariate

Figure 2.1 shows the scatterplot of Galton's dataset which can be found in the `R` package `HistData` as `GaltonFamilies`. In this dataset, `father` denotes the height of the father and `mother` denotes the height of the mother. The x-axis denotes the mid-parent height, calculated as $(\text{father} + 1.08 \cdot \text{mother})/2$, and the y-axis denotes the height of a child.

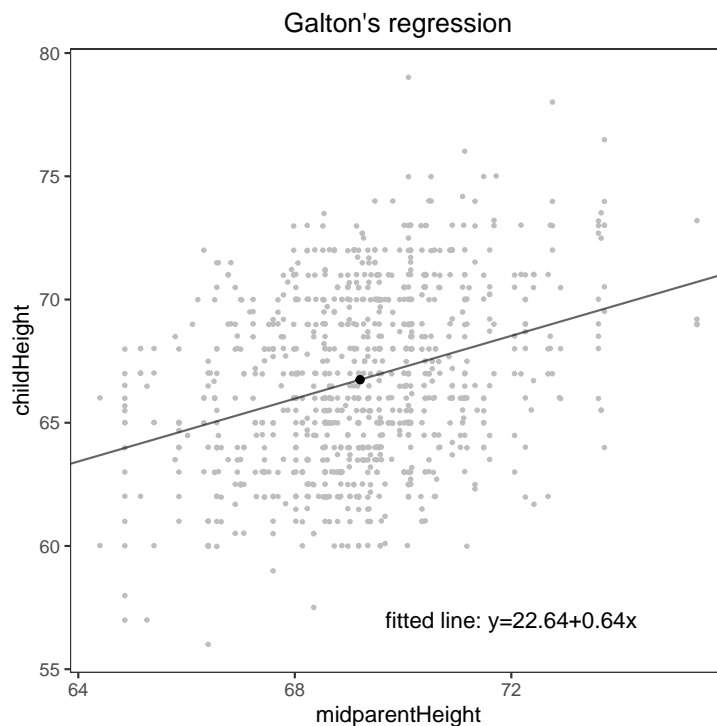


FIGURE 2.1: Galton's dataset

With n data points $(x_i, y_i)_{i=1}^n$, our goal is to find the best linear fit of the data

$$(x_i, \hat{y}_i = \hat{\alpha} + \hat{\beta}x_i)_{i=1}^n,$$

where the coefficients $\hat{\alpha}$ and $\hat{\beta}$ are determined from the data. What do we mean by the “best” fit? Gauss proposed to use the following OLS criterion:¹

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{a, b} n^{-1} \sum_{i=1}^n (y_i - a - bx_i)^2. \quad (2.1)$$

The OLS criterion is based on the squared “misfits” $y_i - a - bx_i$. Another intuitive criterion is based on the absolute values of those misfits, which is called the least absolute deviation (LAD). However, OLS is simpler because the objective function is smooth in (a, b) , and we can obtain close-form solutions. I will discuss LAD in Chapter 26 as a special case of *quantile regression*.

How do we solve the OLS minimization problem in (2.1)? The objective function is quadratic, and as a and b diverge, it diverges to infinity. So it must have a minimizer $(\hat{\alpha}, \hat{\beta})$, which satisfies the first-order condition:

$$-\frac{2}{n} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0, \quad (2.2)$$

$$-\frac{2}{n} \sum_{i=1}^n x_i (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0. \quad (2.3)$$

The equations (2.2) and (2.3) are called the Normal Equations of OLS. The first equation (2.2) implies

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}, \quad (2.4)$$

that is, the OLS line must go through the sample mean of the data (\bar{x}, \bar{y}) . The second equation (2.3) implies

$$\overline{xy} = \hat{\alpha}\bar{x} + \hat{\beta}\overline{x^2}, \quad (2.5)$$

where \overline{xy} is the sample mean of the $x_i y_i$'s, and $\overline{x^2}$ is the sample mean of the x_i^2 's. Subtracting (2.4) $\times \bar{x}$ from (2.5), we have

$$\overline{xy} - \bar{x}\bar{y} = \hat{\beta}(\overline{x^2} - \bar{x}^2),$$

which is

$$\hat{\sigma}_{xy} = \hat{\beta}\hat{\sigma}_x^2,$$

and implies

$$\hat{\beta} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}. \quad (2.6)$$

So the OLS coefficient of x equals the sample covariance between x and y divided by the sample variance of x . From (2.4), we obtain that

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \quad (2.7)$$

¹The idea of OLS is often attributed to Gauss and Legendre. Gauss used it in the process of discovering Ceres, and his work was published in 1809. Legendre's work appeared in 1805 but Gauss claimed that he had been using it since 1794 or 1795. Stigler (1981) reviews the history of OLS.

By (2.7), the fitted line $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ simplifies to $\hat{y}_i = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i$, and more symmetrically, $\hat{y}_i - \bar{y} = \hat{\beta}(x_i - \bar{x})$. With (2.6), we can further simplify the fitted line as

$$\begin{aligned}\hat{y}_i - \bar{y} &= \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}(x_i - \bar{x}) \\ &= \frac{\hat{\rho}_{xy}\hat{\sigma}_x\hat{\sigma}_y}{\hat{\sigma}_x^2}(x_i - \bar{x}),\end{aligned}$$

which implies

$$\frac{\hat{y}_i - \bar{y}}{\hat{\sigma}_y} = \hat{\rho}_{xy} \frac{x_i - \bar{x}}{\hat{\sigma}_x},$$

the Galtonian formula mentioned in Chapter 1.

We can obtain the fitted line based on Galton's data using the R code below.

```
> library("HistData")
> xx = GaltonFamilies$midparentHeight
> yy = GaltonFamilies$childHeight
>
> center_x = mean(xx)
> center_y = mean(yy)
> sd_x = sd(xx)
> sd_y = sd(yy)
> rho_xy = cor(xx, yy)
>
> beta_fit = rho_xy*sd_y/sd_x
> alpha_fit = center_y - beta_fit*center_x
> alpha_fit
[1] 22.63624
> beta_fit
[1] 0.6373609
```

I then generate Figure 2.1 based on the original data and the OLS coefficients.

2.2 Final comments

I make two final comments on OLS.

(C1) We can write the sample mean as the solution to the OLS with only the intercept:

$$\bar{y} = \arg \min_{\mu} n^{-1} \sum_{i=1}^n (y_i - \mu)^2. \quad (2.8)$$

(C2) We can fit OLS of y_i on x_i without the intercept:

$$\hat{\beta} = \arg \min_b n^{-1} \sum_{i=1}^n (y_i - bx_i)^2$$

which equals

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{\langle x, y \rangle}{\langle x, x \rangle}, \quad (2.9)$$

where $x = (x_1, \dots, x_n)^T$ and $y = (y_1, \dots, y_n)^T$ are the n -dimensional vectors containing all observations, and $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$ denotes the inner product between x and y .

Although it is rare to fit the above OLS in practical data analysis, the formulas in (2.8) and (2.9) will be the building blocks for many discussions in later parts of the book. I leave the proof of (2.8) and (2.9) as Problem 2.1.

2.3 Homework problems

2.1 Univariate OLS

Prove (2.8) and (2.9).

2.2 Pairwise slopes

Prove Theorem 2.1 below.

Theorem 2.1 Given $(x_i, y_i)_{i=1}^n$ with univariate x_i and y_i , show that $\hat{\beta}$ in (2.6) equals

$$\hat{\beta} = \sum_{(i,j)} w_{ij} b_{ij},$$

where the summation is over all pairs of observations (i, j) ,

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j}$$

is the slope determined by two points (x_i, y_i) and (x_j, y_j) , and

$$w_{ij} = \frac{(x_i - x_j)^2}{\sum_{(i', j')} (x_{i'} - x_{j'})^2}$$

is the weight proportional to the squared distance between x_i and x_j . In the above formulas, if $x_i = x_j$, then we define $b_{ij} = 0$, and the corresponding weight w_{ij} equals 0.

Remark: Wu (1986) and Gelman and Park (2009) used Theorem 2.1. Problem 3.10 gives a more general result.

2.3 No regression

Woolley (1941) proposed a method to minimize the sum of the areas formed between the data points and fitted line. The right panel of Figure 2.2 illustrates the area formed between data point (x_i, y_i) and the line $y = a + bx$.

Prove that if $\hat{\rho}_{xy} > 0$, then the minimizer $(\hat{\alpha}', \hat{\beta}')$ satisfies

$$\hat{\beta}' = \frac{\hat{\sigma}_y}{\hat{\sigma}_x}$$

and

$$\hat{\alpha}' = \bar{y} - \hat{\beta}\bar{x}.$$

Remark: The fitted line is

$$\begin{aligned} \hat{y}_i &= \hat{\alpha}' + \hat{\beta}'x_i \\ &= \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}'x_i, \end{aligned}$$

which is equivalent to

$$\frac{\hat{y}_i - \bar{y}}{\hat{\sigma}_y} = \frac{x_i - \bar{x}}{\hat{\sigma}_x}.$$

It does not have the regression factor $\hat{\rho}_{xy}$, compared with the Galtonian formula, which is derived by minimizing the residual sum of squares as illustrated by the left panel of Figure 2.2.

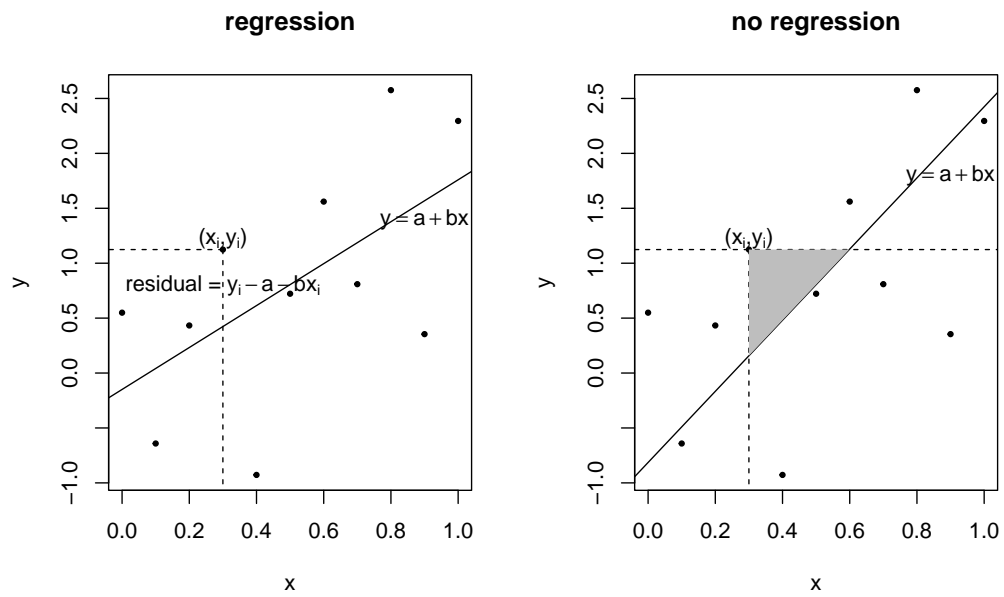


FIGURE 2.2: Regression (left) and no regression (right)



Part II

Ordinary Least Squares and Statistical Inference



3

Ordinary Least Squares with Multiple Covariates

This chapter provides algebraic results about ordinary least squares (OLS). The results in this chapter do not rely on any stochastic assumptions.

3.1 The OLS formula

Recall that we have the outcome vector and covariate matrix:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Depending on the purpose, it is convenient to view X as a collection of row or column vectors:

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = (X_1, \dots, X_p)$$

where $x_i^T = (x_{i1}, \dots, x_{ip})$ is the row vector consisting of the covariates of unit i , and $X_j = (x_{1j}, \dots, x_{nj})^T$ is the column vector of the j -th covariate for all units.

We want to find the “best” linear fit of the data $(x_i, y_i)_{i=1}^n$ with

$$\hat{y}_i = x_i^T \hat{\beta} = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

in the sense that

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^p} n^{-1} \sum_{i=1}^n (y_i - x_i^T b)^2 \quad (3.1)$$

$$= \arg \min_{b \in \mathbb{R}^p} n^{-1} \|Y - Xb\|^2, \quad (3.2)$$

where $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$ is called the OLS coefficient, the \hat{y}_i ’s are called the fitted values, and the $\hat{\varepsilon}_i = y_i - \hat{y}_i$ ’s are called the residuals.

The objective function in (3.1) is quadratic in b , which diverges to infinity when b diverges to infinity. So it must have a minimizer $\hat{\beta}$ satisfying the first-order condition

$$-\frac{2}{n} \sum_{i=1}^n x_i (y_i - x_i^T \hat{\beta}) = 0,$$

which simplifies to

$$\sum_{i=1}^n x_i (y_i - x_i^T \hat{\beta}) = 0, \quad (3.3)$$

or, equivalently, in matrix form:

$$X^T(Y - X\hat{\beta}) = 0. \quad (3.4)$$

The above equations (3.3) and (3.4) are called the *Normal equation* of the OLS, which implies the main theorem:

Theorem 3.1 *The OLS coefficient in (3.1) and (3.2) equals*

$$\begin{aligned} \hat{\beta} &= \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \left(\sum_{i=1}^n x_i y_i \right) \\ &= (X^T X)^{-1} X^T Y \end{aligned}$$

if $X^T X = \sum_{i=1}^n x_i x_i^T$ is non-degenerate.

Comment on the two equivalent forms in Theorem 3.1. The equivalence of the two forms of the OLS coefficient follows from

$$X^T X = (x_1, \dots, x_n) \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \sum_{i=1}^n x_i x_i^T,$$

and

$$X^T Y = (x_1, \dots, x_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^n x_i y_i.$$

Depending on the purpose, both forms can be useful in later discussions.

Comment on the condition in Theorem 3.1. The non-degeneracy of $X^T X$ in Theorem 3.1 requires that for any non-zero vector $\alpha = (\alpha_1, \dots, \alpha_p)^T \in \mathbb{R}^p$, we must have

$$\alpha^T X^T X \alpha = \|X\alpha\|^2 \neq 0$$

which is equivalent to

$$X\alpha = \alpha_1 X_1 + \dots + \alpha_p X_p \neq 0,$$

i.e., the columns of X are *linearly independent*.¹ This effectively rules out redundant columns in the design matrix X . If X_1 can be represented by other columns $X_1 = c_2 X_2 + \dots + c_p X_p$ for some (c_2, \dots, c_p) , then $X^T X$ is degenerate.

Throughout the book, we invoke the following condition unless stated otherwise.

Condition 3.1 *The column vectors of X are linearly independent.*

¹This book uses different notions of “independence,” which can be confusing sometimes. In linear algebra, a set of vectors is linearly *independent* if any nonzero linear combination of them is not zero; see Appendix A. In probability theory, two random variables are *independent* if their joint density factorizes into the product of the marginal distributions; see Appendix B.

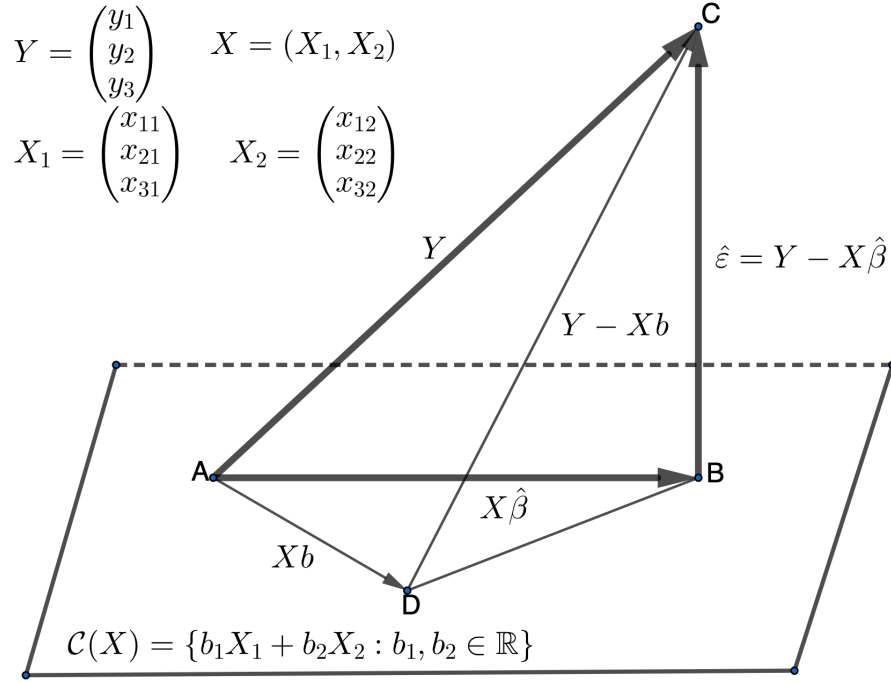


FIGURE 3.1: The geometry of OLS

3.2 The geometry of OLS

The OLS has clear geometric interpretations. Figure 3.1 illustrates its geometry with $n = 3$ and $p = 2$. For any $b = (b_1, \dots, b_p)^T \in \mathbb{R}^p$ and $X = (X_1, \dots, X_p) \in \mathbb{R}^{n \times p}$,

$$Xb = (X_1, \dots, X_p) \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} = b_1X_1 + \dots + b_pX_p$$

represents a linear combination of the column vectors of the design matrix X . So the OLS problem is to find the best linear combination of the column vectors of X to approximate the response vector Y . Recall that all linear combinations of the column vectors of X constitute the column space of X , denoted by²

$$\mathcal{C}(X) = \{b_1X_1 + \dots + b_pX_p : b_1, \dots, b_p \in \mathbb{R}\}.$$

So the OLS problem is to find the vector in $\mathcal{C}(X)$ that is the closest to Y . Geometrically, the vector must be the projection of Y onto $\mathcal{C}(X)$. By projection, the residual vector $\hat{\varepsilon} = Y - X\hat{\beta}$ must be orthogonal to $\mathcal{C}(X)$, or, equivalently, the residual vector is orthogonal to X_1, \dots, X_p . This geometric intuition implies that

$$X_1^T \hat{\varepsilon} = 0, \dots, X_p^T \hat{\varepsilon} = 0.$$

²Please review Appendix A for some basic linear algebra background.

In matrix form, we have

$$X^T \hat{\varepsilon} = \begin{pmatrix} X_1^T \hat{\varepsilon} \\ \vdots \\ X_p^T \hat{\varepsilon} \end{pmatrix} = 0,$$

which is equivalent to

$$X^T(Y - X\hat{\beta}) = 0,$$

the Normal equation in (3.4). The above argument gives a geometric derivation of the OLS formula in Theorem 3.1.

In Figure 3.1, since the triangle ABC is rectangular, the fitted vector $\hat{Y} = X\hat{\beta}$ is orthogonal to the residual vector $\hat{\varepsilon}$, and moreover, the Pythagorean Theorem implies that

$$\|Y\|^2 = \|X\hat{\beta}\|^2 + \|\hat{\varepsilon}\|^2.$$

The following theorem states an algebraic fact that gives an alternative proof of the OLS formula. It is essentially the Pythagorean Theorem for the rectangular triangle BCD in Figure 3.1.

Theorem 3.2 *For any $b \in \mathbb{R}^p$, we have the following decomposition*

$$\|Y - Xb\|^2 = \|Y - X\hat{\beta}\|^2 + \|X(\hat{\beta} - b)\|^2,$$

where implies that $\|Y - Xb\|^2 \geq \|Y - X\hat{\beta}\|^2$ with equality holding if and only if $b = \hat{\beta}$.

Proof of Theorem 3.2: We have the following decomposition:

$$\begin{aligned} \|Y - Xb\|^2 &= (Y - Xb)^T(Y - Xb) \\ &= (Y - X\hat{\beta} + X\hat{\beta} - Xb)^T(Y - X\hat{\beta} + X\hat{\beta} - Xb) \\ &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}) + (X\hat{\beta} - Xb)^T(X\hat{\beta} - Xb) \\ &\quad + (Y - X\hat{\beta})^T(X\hat{\beta} - Xb) + (X\hat{\beta} - Xb)^T(Y - X\hat{\beta}). \end{aligned}$$

The first term equals $\|Y - X\hat{\beta}\|^2$ and the second term equals $\|X(\hat{\beta} - b)\|^2$. We need to show the last two terms are zero. By symmetry of these two terms, we only need to show that the last term is zero. This is true by the Normal equation (3.4) of the OLS:

$$(X\hat{\beta} - Xb)^T(Y - X\hat{\beta}) = (\hat{\beta} - b)^T X^T(Y - X\hat{\beta}) = 0.$$

□

I end this section by commenting on the role of the intercept in OLS.

The role of the intercept in OLS. In most applications, X contains a column of $1_n = (1, \dots, 1)^T$. In those cases, we have

$$1_n^T \hat{\varepsilon} = 0,$$

and therefore,

$$n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i = 0.$$

That is, with the intercept in OLS, the residuals are automatically centered to have mean 0.

3.3 The projection matrix from OLS

The geometry in Section 3.2 also shows that $\hat{Y} = X\hat{\beta}$ is the solution to the following problem

$$\hat{Y} = \arg \min_{v \in \mathcal{C}(X)} \|Y - v\|^2.$$

Using Theorem 3.1, we have $\hat{Y} = X\hat{\beta} = HY$, where

$$H = X(X^T X)^{-1} X^T$$

is an $n \times n$ matrix. It is called the *hat matrix* because it puts a hat on Y when multiplying Y . Algebraically, we can show that H is a projection matrix³ because

$$\begin{aligned} H^2 &= X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T \\ &= X(X^T X)^{-1} X^T \\ &= H, \end{aligned}$$

and

$$\begin{aligned} H^T &= \{X(X^T X)^{-1} X^T\}^T \\ &= X(X^T X)^{-1} X^T \\ &= H. \end{aligned}$$

Its rank equals its trace, so

$$\begin{aligned} \text{rank}(H) = \text{trace}(H) &= \text{trace}\{X(X^T X)^{-1} X^T\} \\ &= \text{trace}\{(X^T X)^{-1} X^T X\} \\ &= \text{trace}(I_p) \\ &= p. \end{aligned}$$

The projection matrix H has the following geometric interpretations.

Proposition 3.1 *The projection matrix $H = X(X^T X)^{-1} X^T$ satisfies*

- (G1) $Hv = v$ if and only if $v \in \mathcal{C}(X)$;
- (G2) $Hw = 0$ if and only if $w \perp \mathcal{C}(X)$.

Recall that $\mathcal{C}(X)$ is the column space of X . (G1) states that projecting any vector in $\mathcal{C}(X)$ onto $\mathcal{C}(X)$ does not change the vector. (G2) states that projecting any vector orthogonal to $\mathcal{C}(X)$ onto $\mathcal{C}(X)$ results in a zero vector.

Proof of Proposition 3.1: I first prove (G1). If $v \in \mathcal{C}(X)$, then $v = Xb$ for some b , which implies

$$Hv = X(X^T X)^{-1} X^T Xb = Xb = v.$$

Conversely, if $v = Hv$, then $v = X(X^T X)^{-1} X^T v = Xb$ with $b = (X^T X)^{-1} X^T v$, which ensures $v \in \mathcal{C}(X)$.

I then prove (G2). If $w \perp \mathcal{C}(X)$, then w is orthogonal to all column vectors of X , that is, $X_j^T w = 0$ ($j = 1, \dots, p$). In matrix form, we have $X^T w = 0$, which implies

$$Hw = X(X^T X)^{-1} X^T w = 0.$$

³Review the definition and properties of projection matrices in Appendix A.

Conversely, if $Hw = X(X^T X)^{-1} X^T w = 0$, then $w^T X(X^T X)^{-1} X^T w = 0$. Because $(X^T X)^{-1}$ is positive definite under Condition 3.1, we have $X^T w = 0$, which implies $w \perp \mathcal{C}(X)$. \square

Writing $H = (h_{ij})_{1 \leq i, j \leq n}$ and $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T$, we have another basic identity

$$\begin{aligned}\hat{y}_i &= \sum_{j=1}^n h_{ij} y_j \\ &= h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j.\end{aligned}$$

It shows that the predicted value \hat{y}_i is a linear combination of the outcomes of all units and the coefficients depend on H . Moreover, if X contains a column of intercepts $1_n = (1, \dots, 1)^T$, then

$$H 1_n = 1_n, \quad (3.5)$$

which implies

$$\sum_{j=1}^n h_{ij} = 1 \quad (i = 1, \dots, n) \quad (3.6)$$

and therefore, \hat{y}_i is a weighted average of the outcomes of all units. Although the sum of the weights is 1, some of them can be negative. Readers, make sure the claims of (3.5) and (3.6) make sense to you. See Problem 3.6.

In general, the hat matrix has complex forms, but when the covariates are dummy variables for group indicators, it has more explicit forms. I give two examples below.

Example 3.1 *In a treatment-control experiment with n_1 treated and n_0 control units, the matrix X contains 1 and a dummy variable for the treatment:*

$$X = \begin{pmatrix} 1_{n_1} & 1_{n_1} \\ 1_{n_0} & 0_{n_0} \end{pmatrix}.$$

We can show that

$$H = \text{diag}\{n_1^{-1} 1_{n_1} 1_{n_1}^T, n_0^{-1} 1_{n_0} 1_{n_0}^T\}.$$

Example 3.2 *In an experiment with n_j units receiving treatment level j ($j = 1, \dots, J$), the covariate matrix X contains J dummy variables for the treatment levels:*

$$X = \text{diag}\{1_{n_1}, \dots, 1_{n_J}\}.$$

We can show that

$$H = \text{diag}\{n_1^{-1} 1_{n_1} 1_{n_1}^T, \dots, n_J^{-1} 1_{n_J} 1_{n_J}^T\}.$$

I leave the proofs of Examples 3.1 and 3.2 as Problem 3.7.

3.4 Homework problems

3.1 Univariate and multivariate OLS

Derive the univariate OLS based on the multivariate OLS formula with

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

where the x_i 's are scalars.

3.2 OLS via vector and matrix calculus

Use vector and matrix calculus to prove that the OLS coefficient $\hat{\beta}$ minimizes $(Y - Xb)^T(Y - Xb)$.

3.3 OLS based on pseudo inverse

Prove that $\hat{\beta} = X^+Y$.

Remark: Recall the definition of the pseudo inverse in Appendix A. Under Condition 3.1, we have $X^+ = (X^T X)^{-1} X^T$.

3.4 Invariance of OLS

Theorem 3.3 below states the invariance properties of OLS. Prove Theorem 3.3.

Theorem 3.3 Assume that $X^T X$ is non-degenerate and Γ is a $p \times p$ non-degenerate matrix. Define $\tilde{X} = X\Gamma$. From the OLS fit of Y on X , we obtain the coefficient $\hat{\beta}$, the fitted value \hat{Y} , and the residual $\hat{\varepsilon}$; from the OLS fit of Y on \tilde{X} , we obtain the coefficient $\tilde{\beta}$, the fitted value \tilde{Y} , and the residual $\tilde{\varepsilon}$.

We have

$$\hat{\beta} = \Gamma \tilde{\beta}, \quad \hat{Y} = \tilde{Y}, \quad \hat{\varepsilon} = \tilde{\varepsilon}.$$

Remark: From a linear algebra perspective, X and $X\Gamma$ have the same column space if Γ is a non-degenerate matrix:

$$\{Xb : b \in \mathbb{R}^p\} = \{X\Gamma c : c \in \mathbb{R}^p\}.$$

Consequently, there must be a unique projection of Y onto the common column space.

3.5 Invariance of the hat matrix

This problem extends Theorem 3.3 in Problem 3.4.

Prove that H does not change if we change X to $X\Gamma$ where $\Gamma \in \mathbb{R}^{p \times p}$ is a non-degenerate matrix.

3.6 Hat matrix with the intercept

Prove (3.5) and (3.6).

3.7 Special hat matrices

Verify the formulas of the hat matrices in Examples 3.1 and 3.2.

3.8 OLS with multiple responses

For each unit $i = 1, \dots, n$, we have multiple responses $y_i = (y_{i1}, \dots, y_{iq})^T \in \mathbb{R}^q$ and multiple covariates $x_i = (x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^p$. Define

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix} = \begin{pmatrix} y_1^T \\ \vdots \\ y_n^T \end{pmatrix} = (Y_1, \dots, Y_q) \in \mathbb{R}^{n \times q}$$

and

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = (X_1, \dots, X_p) \in \mathbb{R}^{n \times p}$$

as the response and covariate matrices, respectively. Define the multiple OLS coefficient matrix as

$$\hat{B} = \arg \min_{B \in \mathbb{R}^{p \times q}} \sum_{i=1}^n \|y_i - B^T x_i\|^2$$

Show that $\hat{B} = (\hat{B}_1, \dots, \hat{B}_q)$ has column vectors

$$\begin{aligned} \hat{B}_1 &= (X^T X)^{-1} X^T Y_1, \\ &\vdots \\ \hat{B}_q &= (X^T X)^{-1} X^T Y_q. \end{aligned}$$

Remark: This result tells us that the OLS fit with a vector outcome reduces to multiple separate OLS fits, or, the OLS fit of a matrix Y on a matrix X reduces to the column-wise OLS fits of Y on X .

3.9 Full sample and subsample OLS coefficients

Partition the full sample into K subsamples:

$$X = \begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(K)} \end{pmatrix}, \quad Y = \begin{pmatrix} Y_{(1)} \\ \vdots \\ Y_{(K)} \end{pmatrix},$$

where the k th sample consists of $(X_{(k)}, Y_{(k)})$ with $X_{(k)} \in \mathbb{R}^{n_k \times p}$ and $Y_{(k)} \in \mathbb{R}^{n_k}$ being the covariate matrix and outcome vector. The sample sizes satisfy $n = \sum_{k=1}^K n_k$. Let $\hat{\beta}$ be the OLS coefficient based on the full sample, and $\hat{\beta}_{(k)}$ be the OLS coefficient based on the k th sample.

Prove that

$$\hat{\beta} = \sum_{k=1}^K W_{(k)} \hat{\beta}_{(k)},$$

where the weight matrix equals

$$W_{(k)} = (X^T X)^{-1} X_{(k)}^T X_{(k)}.$$

Remark: In the special case of a univariate y_i and x_i , the OLS of y_i on x_i without the intercept gives the coefficient

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

Partition the units into K disjoint parts: $\{1, \dots, n\} = I_1 \cup \dots \cup I_K$. Run OLS of y_i on x_i without the intercept using units in I_k to obtain the coefficient $\hat{\beta}_{(k)}$. The above formula implies that

$$\hat{\beta} = \sum_{k=1}^K W_{(k)} \hat{\beta}_{(k)}$$

where

$$W_{(k)} = \frac{\sum_{i \in I_k} x_i^2}{\sum_{i=1}^n x_i^2}$$

is proportional to the sum of squares of the regressor x_i 's in I_k .

3.10 Jacobi's theorem

Prove Theorem 3.4 below.

Theorem 3.4 (Jacobi's Theorem) *The set $\{1, \dots, n\}$ has $\binom{n}{p}$ size- p subsets. Each subset S defines a linear equation for $b \in \mathbb{R}^p$:*

$$Y_S = X_S b$$

where $Y_S \in \mathbb{R}^p$ is the subvector of Y and $X_S \in \mathbb{R}^{p \times p}$ is the submatrix of X , corresponding to the units in S . Define the subset coefficient

$$\hat{\beta}_S = X_S^{-1} Y_S$$

if X_S is invertible and $\hat{\beta}_S = 0$ otherwise.

The OLS coefficient equals a weighted average of these subset coefficients:

$$\hat{\beta} = \sum_S w_S \hat{\beta}_S$$

where the summation is over all subsets, and the weights are

$$w_S = \frac{|\det(X_S)|^2}{\sum_{S'} |\det(X_{S'})|^2}.$$

Remark: Theorem 3.4 extends Problem 2.2. Subrahmanyam (1972) reported Theorem 3.4 although Berman (1988) attributed it to Jacobi. Wu (1986) used it in analyzing the statistical properties of OLS.

To prove Theorem 3.4, you can use Cramer's rule to express the OLS coefficient and use the Cauchy–Binet formula to expand the determinant of $X^T X$.



4

Gauss–Markov Model and Gauss–Markov Theorem

Without any stochastic assumptions, the OLS in Chapter 3 is purely algebraic. If we want to discuss the statistical properties of OLS, we must invoke some statistical modeling assumptions. This chapter focuses on the Gauss–Markov model.

4.1 Gauss–Markov model

A simple starting point is the following Gauss–Markov model with a fixed design matrix X and unknown parameters (β, σ^2) .

Assumption 4.1 (Gauss–Markov model) *We have*

$$Y = X\beta + \varepsilon$$

where the design matrix X is fixed with linearly independent column vectors, and the random error term ε has the first two moments:

$$\begin{aligned} E(\varepsilon) &= 0, \\ \text{cov}(\varepsilon) &= \sigma^2 I_n. \end{aligned}$$

The unknown parameters are (β, σ^2) .

The Gauss–Markov model assumes that Y has mean $X\beta$ and covariance matrix $\sigma^2 I_n$. At the individual level, we can also write it as

$$y_i = x_i^T \beta + \varepsilon_i, \quad (i = 1, \dots, n)$$

where the error terms are uncorrelated with mean 0 and variance σ^2 .

The assumption that X is fixed is not essential, because we can condition on X even if we think X is random. The mean of each y_i is linear in x_i with the same β coefficient, which can be a strong assumption. So is the *homoskedasticity*¹ assumption that the error terms have the same variance σ^2 . The critiques on the assumptions aside, I will derive the properties of $\hat{\beta}$ under the Gauss–Markov model.

¹In this book, I do not spell it as *homoscedasticity* since “k” better indicates the meaning of variance. McCulloch (1985) gave a convincing argument. See also Paloyo (2014).

4.2 Properties of the OLS estimator

I first derive the mean and covariance of $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Theorem 4.1 *Under Assumption 4.1, we have*

$$\begin{aligned} E(\hat{\beta}) &= \beta, \\ \text{cov}(\hat{\beta}) &= \sigma^2 (X^T X)^{-1}. \end{aligned}$$

Proof of Theorem 4.1: Because $E(Y) = X\beta$, we have

$$\begin{aligned} E(\hat{\beta}) &= E\{(X^T X)^{-1} X^T Y\} \\ &= (X^T X)^{-1} X^T E(Y) \\ &= (X^T X)^{-1} X^T X\beta \\ &= \beta. \end{aligned}$$

Because $\text{cov}(Y) = \sigma^2 I_n$, we have

$$\begin{aligned} \text{cov}(\hat{\beta}) &= \text{cov}\{(X^T X)^{-1} X^T Y\} \\ &= (X^T X)^{-1} X^T \text{cov}(Y) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}. \end{aligned}$$

□

We can decompose the response vector as

$$Y = \hat{Y} + \hat{\varepsilon},$$

where the fitted vector is

$$\hat{Y} = X\hat{\beta} = HY$$

and the residual vector is

$$\hat{\varepsilon} = Y - \hat{Y} = (I_n - H)Y.$$

The two matrices H and $I_n - H$ are the keys, which have the following properties.

Lemma 4.1 *Both H and $I_n - H$ are projection matrices. They satisfy*

$$\begin{aligned} HX &= X, \\ (I_n - H)X &= 0. \end{aligned}$$

They are orthogonal:

$$H(I_n - H) = (I_n - H)H = 0.$$

Lemma 4.1 follows from simple linear algebra, and I leave its proof as Problem 4.1. It states that H and $I_n - H$ are projection matrices onto the column space of X and its complement. Algebraically, \hat{Y} and $\hat{\varepsilon}$ are orthogonal by the OLS projection because Lemma 4.1 implies

$$\begin{aligned} \hat{Y}^T \hat{\varepsilon} &= Y^T H^T (I_n - H) Y \\ &= Y^T H (I_n - H) Y \\ &= 0. \end{aligned}$$

This is also coherent with the geometry in Figure 3.1.

Moreover, we can derive the mean and covariance matrix of \hat{Y} and $\hat{\varepsilon}$.

Theorem 4.2 Under Assumption 4.1, we have

$$E \begin{pmatrix} \hat{Y} \\ \hat{\varepsilon} \end{pmatrix} = \begin{pmatrix} X\beta \\ 0 \end{pmatrix}$$

and

$$\text{cov} \begin{pmatrix} \hat{Y} \\ \hat{\varepsilon} \end{pmatrix} = \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix}.$$

So \hat{Y} and $\hat{\varepsilon}$ are uncorrelated.

Please do not be confused with the two statements about \hat{Y} and $\hat{\varepsilon}$:

(S1) \hat{Y} and $\hat{\varepsilon}$ are orthogonal.

(S2) \hat{Y} and $\hat{\varepsilon}$ are uncorrelated.

They have different meanings. The first statement (S1) is an algebraic fact of the OLS procedure. It is about a relationship between two vectors \hat{Y} and $\hat{\varepsilon}$ which holds without assuming the Gauss–Markov model. The second statement (S2) is stochastic. It is about a relationship between two random vectors \hat{Y} and $\hat{\varepsilon}$, which requires the Gauss–Markov model assumption.

Proof of Theorem 4.2: The conclusion follows from the simple fact that

$$\begin{aligned} \begin{pmatrix} \hat{Y} \\ \hat{\varepsilon} \end{pmatrix} &= \begin{pmatrix} HY \\ (I_n - H)Y \end{pmatrix} \\ &= \begin{pmatrix} H \\ I_n - H \end{pmatrix} Y \end{aligned}$$

is a linear transformation of Y .

It has mean

$$\begin{aligned} E \begin{pmatrix} \hat{Y} \\ \hat{\varepsilon} \end{pmatrix} &= \begin{pmatrix} H \\ I_n - H \end{pmatrix} E(Y) \\ &= \begin{pmatrix} H \\ I_n - H \end{pmatrix} X\beta \\ &= \begin{pmatrix} HX\beta \\ (I_n - H)X\beta \end{pmatrix} \\ &= \begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \end{aligned}$$

because $HX = X$ and $(I_n - H)X = 0$ by Lemma 4.1.

It has covariance matrix

$$\begin{aligned} \text{cov} \begin{pmatrix} \hat{Y} \\ \hat{\varepsilon} \end{pmatrix} &= \begin{pmatrix} H \\ I_n - H \end{pmatrix} \text{cov}(Y) \begin{pmatrix} H^\top & (I_n - H)^\top \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} H \\ I_n - H \end{pmatrix} \begin{pmatrix} H & I_n - H \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} H^2 & H(I_n - H) \\ (I_n - H)H & (I_n - H)^2 \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix}, \end{aligned}$$

because $H^2 = 2$, $(I_n - H)^2 = I_n - H$, and $H(I_n - H) = 0$ by Lemma 4.1. \square

Assume the Gauss–Markov model. Although the original responses and error terms are uncorrelated between units with

$$\text{cov}(y_i, y_j) = 0, \quad \text{cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j,$$

the fitted values and the residuals are correlated with

$$\text{cov}(\hat{y}_i, \hat{y}_j) = \sigma^2 h_{ij}, \quad \text{cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\sigma^2 h_{ij} \text{ for } i \neq j,$$

based on Theorem 4.2.

4.3 Variance estimation

Theorem 4.1 quantifies the uncertainty of $\hat{\beta}$ by its covariance matrix. However, it is not directly useful because σ^2 is still unknown. Our next task is to estimate σ^2 based on the observed data. It is the variance of each ε_i , but the ε_i 's are not observable either. Their empirical analogues are the residuals $\hat{\varepsilon}_i = y_i - x_i^\top \hat{\beta}$. It seems intuitive to estimate σ^2 by

$$\tilde{\sigma}^2 = \text{RSS}/n,$$

where

$$\text{RSS} = \sum_{i=1}^n \hat{\varepsilon}_i^2$$

is the residual sum of squares. However, Theorem 4.2 shows that $\hat{\varepsilon}_i$ has mean zero and variance $\sigma^2(1 - h_{ii})$, which is not the same as the variance of the original ε_i . Consequently, RSS has mean

$$\begin{aligned} E(\text{RSS}) &= \sum_{i=1}^n \sigma^2(1 - h_{ii}) \\ &= \sigma^2\{n - \text{trace}(H)\} \\ &= \sigma^2(n - p), \end{aligned}$$

which implies Theorem 4.3 below.

Theorem 4.3 *Define*

$$\hat{\sigma}^2 = \text{RSS}/(n - p) = \sum_{i=1}^n \hat{\varepsilon}_i^2/(n - p).$$

Then $E(\hat{\sigma}^2) = \sigma^2$ under Assumption 4.1.

Theorem 4.3 implies that $\tilde{\sigma}^2$ is a biased estimator for σ^2 because $E(\tilde{\sigma}^2) = \sigma^2(n - p)/n$. It underestimates σ^2 but with a large sample size n , the bias is small.

4.4 Gauss–Markov Theorem

So far, we have focused on the OLS estimator. It is intuitive, but we have not answered the fundamental question yet. Why should we focus on it? Are there any other better

estimators? Under the Gauss–Markov model, the answer is definite: we focus on the OLS estimator because it is optimal in the sense of having the smallest covariance matrix among all linear unbiased estimators. The following famous Gauss–Markov theorem quantifies this claim, which was named after Carl Friedrich Gauss and Andrey Markov.² It is for this reason that I call the corresponding model the Gauss–Markov model. The textbook by Monahan (2008) also uses this name.

Theorem 4.4 (Gauss–Markov Theorem) *Under Assumption 4.1, the OLS estimator $\hat{\beta}$ for β is the best linear unbiased estimator (BLUE) in the sense that³*

$$\text{cov}(\tilde{\beta}) \succeq \text{cov}(\hat{\beta})$$

for any estimator $\tilde{\beta}$ satisfying

(C1) $\tilde{\beta} = AY$ for some $A \in \mathbb{R}^{p \times n}$ not depending on Y ;

(C2) $E(\tilde{\beta}) = \beta$ for any β .

Before proving Theorem 4.4, we need to understand its meaning and immediate implications. We do not compare the OLS estimator with any arbitrary estimators. In fact, we restrict to the estimators that are linear and unbiased. Condition (C1) requires that $\tilde{\beta}$ is a linear estimator. More precisely, it is a linear transformation of the response vector Y , where A can be any complex and possibly nonlinear function of X . Condition (C2) requires that $\tilde{\beta}$ is an unbiased estimator for β , no matter what true value β takes.

Why do we restrict the estimator to be linear? The class of linear estimator is actually quite large because A can be any nonlinear function of X , and the only requirement is that the estimator is linear in Y . The unbiasedness is a natural requirement for many problems. However, in modern applications with many covariates, some biased estimators can perform better than unbiased estimators if they have smaller variances. We will discuss these estimators in Part V of this book.

We compare the estimators based on their covariances, which are natural extensions of variances for scalar random variables. The conclusion $\text{cov}(\tilde{\beta}) \succeq \text{cov}(\hat{\beta})$ implies that for any vector $c \in \mathbb{R}^p$, we have

$$c^T \text{cov}(\tilde{\beta}) c \geq c^T \text{cov}(\hat{\beta}) c$$

which is equivalent to

$$\text{var}(c^T \tilde{\beta}) \geq \text{var}(c^T \hat{\beta}). \quad (4.1)$$

So any linear transformation of the OLS estimator has a variance smaller than or equal to the same linear transformation of any other estimator. In particular, if $c = (0, \dots, 1, \dots, 0)^T$ with only the j th coordinate being 1, then the inequality (4.1) implies that

$$\text{var}(\tilde{\beta}_j) \geq \text{var}(\hat{\beta}_j), \quad (j = 1, \dots, p).$$

So the OLS estimator has a smaller variance than other estimators for all coordinates.

Now we prove Theorem 4.4.

Proof of Theorem 4.4: We must verify that the OLS estimator itself satisfies (C1) and (C2). We have $\hat{\beta} = \hat{A}Y$ with $\hat{A} = (X^T X)^{-1} X^T$, and it is unbiased by Theorem 4.1.

First, the unbiasedness requires that $E(\tilde{\beta}) = \beta$ for any value of β . Under the Gauss–Markov Model, it requires that

$$E(AY) = AE(Y) = AX\beta = \beta$$

²David and Neyman (1938) used the name “Markoff Theorem”. Lehmann (1951) appeared to first use the name “Gauss–Markov Theorem.”

³We write $M_1 \succeq M_2$ if $M_1 - M_2$ is positive semi-definite. See Appendix A for a review.

for any value of β . The requirement $AX\beta = \beta$ for any value of β implies that

$$AX = I_p \quad (4.2)$$

must hold. In particular, the OLS estimator satisfies $\hat{A}X = (X^T X)^{-1} X^T X = I_p$.

Second, we can decompose the covariance of $\tilde{\beta}$ as

$$\begin{aligned} \text{cov}(\tilde{\beta}) &= \text{cov}(\hat{\beta} + \tilde{\beta} - \hat{\beta}) \\ &= \text{cov}(\hat{\beta}) + \text{cov}(\tilde{\beta} - \hat{\beta}) + \text{cov}(\hat{\beta}, \tilde{\beta} - \hat{\beta}) + \text{cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}). \end{aligned}$$

The last two terms are in fact 0. By symmetry, we only need to show that the third term is 0:

$$\begin{aligned} \text{cov}(\hat{\beta}, \tilde{\beta} - \hat{\beta}) &= \text{cov}\{\hat{A}Y, (A - \hat{A})Y\} \\ &= \hat{A}\text{cov}(Y)(A - \hat{A})^T \\ &= \sigma^2 \hat{A}(A - \hat{A})^T \\ &= \sigma^2 (\hat{A}A^T - \hat{A}\hat{A}^T) \\ &= \sigma^2 \{(X^T X)^{-1} X^T A^T - (X^T X)^{-1} X^T X (X^T X)^{-1}\} \\ &= \sigma^2 \{(X^T X)^{-1} I_p - (X^T X)^{-1}\} \quad (\text{by (4.2)}) \\ &= 0. \end{aligned}$$

The above covariance decomposition simplifies to

$$\text{cov}(\tilde{\beta}) = \text{cov}(\hat{\beta}) + \text{cov}(\tilde{\beta} - \hat{\beta}),$$

which implies

$$\text{cov}(\tilde{\beta}) - \text{cov}(\hat{\beta}) = \text{cov}(\tilde{\beta} - \hat{\beta}) \succeq 0.$$

□

In the process of the proof, we have shown two stronger results

$$\text{cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}) = 0$$

and

$$\text{cov}(\tilde{\beta} - \hat{\beta}) = \text{cov}(\tilde{\beta}) - \text{cov}(\hat{\beta}).$$

They hold only when $\hat{\beta}$ is BLUE. They do not hold when comparing two general estimators.

Theorem 4.4 is elegant but abstract. It says that in some sense, we can focus on the OLS estimator because it is the best one in terms of the covariance among all linear unbiased estimators. Then we do not need to consider other estimators. However, we have not mentioned any other estimators for β yet, which makes Theorem 4.4 not concrete enough. From the proof above, a linear unbiased estimator $\tilde{\beta} = AY$ only needs to satisfy $AX = I_p$, which imposes p^2 constraints on the $p \times n$ matrix A . Therefore, we have $p(n - p)$ free parameters to choose from and have infinitely many linear unbiased estimators in general. A class of linear unbiased estimators that will be discussed more thoroughly in Chapter 19 are the weighted least squares estimators

$$\tilde{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y,$$

where Σ is a positive definite matrix not depending on Y such that Σ and $X^T \Sigma^{-1} X$ are invertible. It is linear in Y , and we can show that it is unbiased for β :

$$\begin{aligned} E(\tilde{\beta}) &= E\{(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y\} \\ &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} X \beta \\ &= \beta. \end{aligned}$$

Different choices of Σ give different $\tilde{\beta}$, but Theorem 4.4 states that the OLS estimator with $\Sigma = I_n$ has the smallest covariance matrix under the Gauss–Markov model.

I will give an extension and some applications of the Gauss–Markov Theorem in Problems 4.3–4.7.

4.5 Homework problems

4.1 Projection matrices

Prove Lemma 4.1.

4.2 Univariate OLS and the optimal design

Assume the Gauss–Markov model $y_i = \alpha + \beta x_i + \varepsilon_i$ ($i = 1, \dots, n$) with a scalar x_i . Show that the variance of the OLS coefficient for x_i equals

$$\text{var}(\hat{\beta}) = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2.$$

Assume x_i is in the interval $[0, 1]$. We want to choose their values to minimize $\text{var}(\hat{\beta})$. Assume that n is an even number. Find the x_i 's that minimize $\text{var}(\hat{\beta})$.

Remark: You may find the following probability result useful. For a random variable ξ in the interval $[0, 1]$, we have the following inequality

$$\begin{aligned} \text{var}(\xi) &= E(\xi^2) - \{E(\xi)\}^2 \\ &\leq E(\xi) - \{E(\xi)\}^2 \\ &= E(\xi)\{1 - E(\xi)\} \\ &\leq 1/4. \end{aligned}$$

The first inequality becomes an equality if and only if $\xi = 0$ or 1 ; the second inequality becomes an equality if and only if $E(\xi) = 1/2$.

4.3 BLUE estimator for the mean

Assume that y_i has mean μ and variance σ^2 , and the y_i 's are uncorrelated ($i = 1, \dots, n$). A linear estimator of the mean μ has the form $\hat{\mu} = \sum_{i=1}^n a_i y_i$, which is unbiased as long as $\sum_{i=1}^n a_i = 1$. So there are infinitely many linear unbiased estimators for μ .

Find the BLUE for μ and prove why it is BLUE.

4.4 More variance estimators under the Gauss–Markov model

Assume the Gauss–Markov model in Assumption 4.1.

With the OLS estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$, we can obtain the residual vector $\hat{\varepsilon} = Y - X\hat{\beta}$. For any symmetric matrix A , prove that

$$\hat{\sigma}_A^2 = \frac{\hat{\varepsilon}^T A \hat{\varepsilon}}{\text{trace}(A(I_n - H))}$$

is unbiased for σ^2 as long as $\text{trace}(A(I_n - H)) \neq 0$.

Remark: This chapter focuses on $\hat{\sigma}_A^2$ with $A = I_n$. In general, $\hat{\sigma}_A^2$ gives infinitely many unbiased estimators for σ^2 . A natural question is: what is the optimal choice of A with minimum variance? This question is more complicated because the variance of $\hat{\sigma}_A^2$ depends on not only the mean and covariance of ε as specified in Assumption 4.1 but also higher-order moments of ε . We will revisit this problem in Problem 5.2.

4.5 Consequence of useless regressors

Partition the covariate matrix and parameter into

$$X = (X_1, X_2), \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

where $X_1 \in \mathbb{R}^{n \times k}$, $X_2 \in \mathbb{R}^{n \times l}$, $\beta_1 \in \mathbb{R}^k$ and $\beta_2 \in \mathbb{R}^l$ with $k + l = p$. Assume the Gauss–Markov model with $\beta_2 = 0$. Let $\hat{\beta}_1$ be the first k coordinates of $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\tilde{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y$ be the coefficient based on the OLS fit of Y on X_1 only.

Prove that

$$\text{cov}(\hat{\beta}_1) \succeq \text{cov}(\tilde{\beta}_1).$$

4.6 Simple average of subsample OLS coefficients

Inherit the setting of Problem 3.9. Define the simple average of the subsample OLS coefficients as

$$\bar{\beta} = K^{-1} \sum_{k=1}^K \hat{\beta}_{(k)}.$$

Assume the Gauss–Markov model. Prove that

$$\text{cov}(\bar{\beta}) \succeq \text{cov}(\hat{\beta}).$$

4.7 Gauss–Markov Theorem for prediction

Theorem 4.5 below extends Theorem 4.4. Prove Theorem 4.5.

Theorem 4.5 (Gauss–Markov Theorem for Prediction) *Under Assumption 4.1, the OLS predictor $\hat{Y} = X\hat{\beta}$ for the mean $X\beta$ is the best linear unbiased predictor in the sense that $\text{cov}(\tilde{Y}) \succeq \text{cov}(\hat{Y})$ for any predictor \tilde{Y} satisfying*

(C1) $\tilde{Y} = \tilde{H}Y$ for some $\tilde{H} \in \mathbb{R}^{n \times n}$ not depending on Y ;

(C2) $E(\tilde{Y}) = X\beta$ for any β .

4.8 Nonlinear unbiased estimator under the Gauss–Markov model

Under Assumption 4.1, prove that if the matrices Q_j 's satisfy

$$X^T Q_j X = 0, \quad \text{trace}(Q_j) = 0 \text{ for all } j = 1, \dots, p \quad (4.3)$$

then

$$\tilde{\beta} = \hat{\beta} + \begin{pmatrix} Y^T Q_1 Y \\ \vdots \\ Y^T Q_p Y \end{pmatrix}$$

is unbiased for β .

Remark: The above estimator $\tilde{\beta}$ is a quadratic function of Y . It is a nonlinear unbiased estimator for β . It is not difficult to show the unbiasedness. More remarkably, Koopmann

(1982, Theorem 4.3) showed that under Assumption 4.1, any unbiased estimator for β must have the form of $\tilde{\beta}$.

The condition in (4.3) is not a trivial condition. Can you find Q_j 's to satisfy (4.3)?



5

Normal Linear Model: Inference and Prediction

Under the Gauss–Markov model in Assumption 4.1 in Chapter 4, we have calculated the first two moments of the OLS estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$:

$$\begin{aligned} E(\hat{\beta}) &= \beta, \\ \text{cov}(\hat{\beta}) &= \sigma^2 (X^T X)^{-1}, \end{aligned}$$

and have shown that $\hat{\sigma}^2 = \hat{\varepsilon}^T \hat{\varepsilon} / (n - p)$ is unbiased for σ^2 , where $\hat{\varepsilon} = Y - X\hat{\beta}$ is the residual vector. The Gauss–Markov theorem further ensures that the OLS estimator is the best linear unbiased estimator (BLUE). Although these results characterize the nice properties of the OLS estimator, they do not fully determine its distribution and are thus inadequate for statistical inference.

This chapter will derive the joint distribution of $(\hat{\beta}, \hat{\sigma}^2)$ under the Normal linear model with stronger distribution assumptions.

Assumption 5.1 (Normal linear model) *We have*

$$Y \sim N(X\beta, \sigma^2 I_n),$$

or, equivalently,

$$y_i \stackrel{\text{i.i.d.}}{\sim} N(x_i^T \beta, \sigma^2), \quad (i = 1, \dots, n),$$

where the design matrix X is fixed with linearly independent column vectors. The unknown parameters are (β, σ^2) .

We can also write the Normal linear model as a linear function of covariates with error terms:

$$Y = X\beta + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2 I_n)$$

or, equivalently,

$$y_i = x_i^T \beta + \varepsilon_i \text{ with } \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2) \quad (i = 1, \dots, n).$$

Assumption 5.1 implies Assumption 4.1. Beyond the Gauss–Markov model, it further requires independent and identically distributed (IID) Normal error terms. Assumption 5.1 is strong, but it is canonical in statistics. It allows us to derive elegant formulas and also justifies the outputs of the linear regression functions in many statistical packages.

More interestingly, even though Assumption 5.1 is strong, the statistical procedures derived in this chapter are robust to various violations.¹ Chapter 6 will discuss the robustness of the t test in this chapter even when the error terms are not Normally distributed. Moreover, Chapter 6 will relax Assumption 5.1 and propose a modified procedure that is particularly robust to heteroskedasticity.

¹As another example, Zhang et al. (2025) discussed the robustness of the t test in this chapter even when the error terms are correlated in an unknown way.

5.1 Joint distribution of the OLS coefficient and variance estimator

We first state the main theorem on the joint distribution of $(\hat{\beta}, \hat{\sigma}^2)$ via the joint distribution of $(\hat{\beta}, \hat{\varepsilon})$.

Theorem 5.1 *Under Assumption 5.1, we have*

$$\begin{pmatrix} \hat{\beta} \\ \hat{\varepsilon} \end{pmatrix} \sim N \left\{ \begin{pmatrix} \beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} (X^T X)^{-1} & 0 \\ 0 & I_n - H \end{pmatrix} \right\},$$

and

$$\hat{\sigma}^2 / \sigma^2 \sim \chi_{n-p}^2 / (n-p).$$

So

$$\hat{\beta} \perp\!\!\!\perp \hat{\varepsilon}, \quad \hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2.$$

Proof of Theorem 5.1: First,

$$\begin{aligned} \begin{pmatrix} \hat{\beta} \\ \hat{\varepsilon} \end{pmatrix} &= \begin{pmatrix} (X^T X)^{-1} X^T Y \\ (I_n - H) Y \end{pmatrix} \\ &= \begin{pmatrix} (X^T X)^{-1} X^T \\ I_n - H \end{pmatrix} Y \end{aligned}$$

is a linear transformation of Y , so they are jointly Normal. We have verified their means and variances in Chapter 4, so we only need to show that their covariance is zero:

$$\begin{aligned} \text{cov}(\hat{\beta}, \hat{\varepsilon}) &= (X^T X)^{-1} X^T \text{cov}(Y) (I_n - H)^T \\ &= \sigma^2 (X^T X)^{-1} X^T (I_n - H^T) \\ &= 0 \end{aligned}$$

which holds because $(I_n - H)X = 0$ by Lemma 4.1. The joint Normality with 0 covariance implies $\hat{\beta} \perp\!\!\!\perp \hat{\varepsilon}$.

Second, because $\hat{\sigma}^2 = \text{RSS}/(n-p) = \hat{\varepsilon}^T \hat{\varepsilon} / (n-p)$ is a quadratic function of $\hat{\varepsilon}$, it is independent of $\hat{\beta}$. We only need to show that it is a scaled chi-squared distribution. This follows from Theorem B.10 in Appendix B due to the Normality of $\hat{\varepsilon}/\sigma$ with the projection matrix $I_n - H$ as its covariance matrix. \square

The second theorem is on the joint distribution of $(\hat{Y}, \hat{\varepsilon})$. We have shown their means and covariance matrix in Chapter 4. Because they are linear transformations of Y , they are jointly Normal and independent.

Theorem 5.2 *Under Assumption 5.1, we have*

$$\begin{pmatrix} \hat{Y} \\ \hat{\varepsilon} \end{pmatrix} \sim N \left\{ \begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix} \right\},$$

so

$$\hat{Y} \perp\!\!\!\perp \hat{\varepsilon}.$$

Now we have seen several properties of \hat{Y} and $\hat{\varepsilon}$:

- (P1) In Chapter 3, we have shown that $Y = \hat{Y} + \hat{\varepsilon}$ with $\hat{Y} \perp \hat{\varepsilon}$ by the OLS properties, which is a pure linear algebra fact without assumptions.

- (P2) In Chapter 4, Theorem 4.2 ensures that \hat{Y} and $\hat{\varepsilon}$ are uncorrelated under Assumption 4.1.
- (P3) Now Theorem 5.2 further ensures that $\hat{Y} \perp\!\!\!\perp \hat{\varepsilon}$ under Assumption 5.1.

The first result (P1) states that \hat{Y} and $\hat{\varepsilon}$ are orthogonal. The second result (P2) states that \hat{Y} and $\hat{\varepsilon}$ are uncorrelated. The third result (P3) states \hat{Y} and $\hat{\varepsilon}$ are independent. They hold under different assumptions.

5.2 Pivotal quantities and statistical inference

5.2.1 Scalar parameters

We first consider statistical inference for $c^T\beta$, a one-dimensional linear function of β where $c \in \mathbb{R}^p$. For example, if $c = e_j \equiv (0, \dots, 1, \dots, 0)^T$ with only the j th element being one, then $c^T\beta = \beta_j$ is the j th element of β , which measures the impact of x_{ij} on y_i on average. Standard software packages report statistical inference for each element of β . Sometimes we may also be interested in $\beta_j - \beta_{j'}$, the difference between the coefficients of two covariates, which corresponds to $c = (0, \dots, 0, 1, 0, \dots, 0, -1, 0, \dots, 0)^T = e_j - e_{j'}$.

Theorem 5.1 implies that

$$c^T\hat{\beta} \sim N\{c^T\beta, \sigma^2 c^T(X^T X)^{-1}c\}.$$

However, this is not directly useful because σ^2 is unknown. With σ^2 replaced by $\hat{\sigma}^2$, the standardized distribution

$$T_c = \frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\hat{\sigma}^2 c^T(X^T X)^{-1}c}}$$

does not follow $N(0, 1)$ anymore. In fact, it is a t distribution as shown in Theorem 5.3 below.

Theorem 5.3 *Under Assumption 5.1, for a fixed vector $c \in \mathbb{R}^p$, we have*

$$T_c \sim t_{n-p}.$$

Proof of Theorem 5.3: From Theorem 5.1, the standardized distribution with the true σ^2 follows

$$\frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\sigma^2 c^T(X^T X)^{-1}c}} \sim N(0, 1),$$

$\hat{\sigma}^2/\sigma^2 \sim \chi_{n-p}^2/(n-p)$, and they are independent. These facts imply that

$$\begin{aligned} T_c &= \frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\hat{\sigma}^2 c^T(X^T X)^{-1}c}} \\ &= \frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\sigma^2 c^T(X^T X)^{-1}c}} \bigg/ \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \\ &\sim \frac{N(0, 1)}{\sqrt{\chi_{n-p}^2/(n-p)}}, \end{aligned}$$

where $N(0, 1)$ and χ_{n-p}^2 denote independent standard Normal and χ_{n-p}^2 random variables, respectively. Therefore, $T_c \sim t_{n-p}$ by the definition of the t distribution in Appendix B. \square

In Theorem 5.3, the T_c on the left-hand side depends on the observed data and the unknown true parameters, but the t_{n-p} on the right-hand side is a random variable depending on only the dimension (n, p) of X , but neither the data nor the true parameters. Because of this, we call the quantity on the left-hand side a *pivotal quantity*. Based on the quantiles of the t_{n-p} random variable, we can tie the data and the true parameter via the following probability statement

$$\text{pr} \left\{ \left| \frac{c^T \hat{\beta} - c^T \beta}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} \right| \leq t_{1-\alpha/2, n-p} \right\} = 1 - \alpha$$

for any $0 < \alpha < 1$, where $t_{1-\alpha/2, n-p}$ is the $1 - \alpha/2$ quantile of t_{n-p} . When $n - p$ is large (e.g. larger than 30), the $1 - \alpha/2$ quantile of t_{n-p} is close to that of $N(0, 1)$. In particular, when $n - p$ is large, $t_{97.5\%, n-p} \approx 1.96$, the 97.5% quantile of $N(0, 1)$, which is the critical value for the 95% confidence interval.

Define

$$\hat{\text{se}}_c = \sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}$$

which is often called the (estimated) standard error of $c^T \hat{\beta}$. Using the definition of $\hat{\text{se}}_c$, we can equivalently write the above probability statement as

$$\text{pr} \left\{ c^T \hat{\beta} - t_{1-\alpha/2, n-p} \hat{\text{se}}_c \leq c^T \beta \leq c^T \hat{\beta} + t_{1-\alpha/2, n-p} \hat{\text{se}}_c \right\} = 1 - \alpha.$$

We use

$$c^T \hat{\beta} \pm t_{1-\alpha/2, n-p} \hat{\text{se}}_c$$

as a $1 - \alpha$ level confidence interval for $c^T \beta$. By duality of confidence interval and hypothesis testing, we can also construct a level α test for $c^T \beta$. More precisely, we reject the null hypothesis $c^T \beta = d$ if the above confidence interval does not cover d , for a fixed number d .

As an important case, $c = e_j$ so $c^T \beta = \beta_j$. Standard software packages, for example, **R**, report the point estimator $\hat{\beta}_j$, the standard error $\hat{\text{se}}_j = \sqrt{\hat{\sigma}^2 [(X^T X)^{-1}]_{jj}}$, the t statistic $T_j = \hat{\beta}_j / \hat{\text{se}}_j$, and the two-sided p -value $\text{pr}(|t_{n-p}| \geq |T_j|)$ for testing whether β_j equals zero or not. Section 5.4 below gives some examples.

5.2.2 Vector parameters

We then consider statistical inference for $C\beta$, a multi-dimensional linear function of β where $C \in \mathbb{R}^{l \times p}$. If $l = 1$, then it reduces to the one-dimensional case. If

$$C = \begin{pmatrix} c_1^T \\ \vdots \\ c_l^T \end{pmatrix}$$

with $l > 1$, then

$$C\beta = \begin{pmatrix} c_1^T \beta \\ \vdots \\ c_l^T \beta \end{pmatrix}$$

corresponds to the joint values of l parameters $c_1^T \beta, \dots, c_l^T \beta$.

Example 5.1 If

$$C = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

then

$$C\beta = \begin{pmatrix} \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

contains all the coefficients except for the first one (the intercept in most cases). Most software packages report the test of the joint significance of $(\beta_2, \dots, \beta_p)$. Section 5.4 below gives some examples.

Example 5.2 Another leading application is to test whether $\beta_2 = 0$ in the following regression partitioned by $X = (X_1, X_2)$, where X_1 and X_2 are $n \times k$ and $n \times l$ matrices:

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

with

$$C = \begin{pmatrix} 0_{l \times k} & I_l \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

and therefore, $C\beta = \beta_2$. We will discuss this partitioned regression in more detail in Chapters 7 and 8.

Now we will focus on the generic problem of inferring $C\beta$. To avoid degeneracy, we assume that C does not have redundant rows, quantified below.

Assumption 5.2 C has linearly independent rows.

Theorem 5.1 implies that

$$C\hat{\beta} - C\beta \sim N\{0, \sigma^2 C(X^T X)^{-1} C^T\},$$

and therefore, by Theorem B.10 in Appendix B, the standardized quadratic form has a chi-squared distribution²

$$(C\hat{\beta} - C\beta)^T \{\sigma^2 C(X^T X)^{-1} C^T\}^{-1} (C\hat{\beta} - C\beta) \sim \chi_l^2.$$

Again this is not directly useful with unknown σ^2 . Replacing σ^2 with the unbiased estimator $\hat{\sigma}^2$ and using a scaling factor l , we can obtain a pivotal quantity that has an F distribution as summarized in Theorem 5.4 below.

Theorem 5.4 Under Assumptions 5.1 and 5.2, the F statistic

$$F_C = \frac{(C\hat{\beta} - C\beta)^T \{C(X^T X)^{-1} C^T\}^{-1} (C\hat{\beta} - C\beta)}{l\hat{\sigma}^2}$$

follows the F distribution with degrees of freedom l and $n - p$:

$$F_C \sim F_{l, n-p}.$$

²Technically, we need to prove that $\sigma^2 C(X^T X)^{-1} C^T$ is a positive definite matrix.

Because X has linearly independent columns, $X^T X$ is a non-degenerate and thus positive definite matrix. Since $u^T C(X^T X)^{-1} C^T u \geq 0$, to show that $C(X^T X)^{-1} C^T$ is non-degenerate, we only need to show that $u^T C(X^T X)^{-1} C^T u = 0$ must imply $u = 0$. From $u^T C(X^T X)^{-1} C^T u = 0$, we know $C^T u = u_1 c_1 + \cdots + u_l c_l = 0$. Since the rows of C are linearly independent, we must have $u = 0$. The proof is complete.

Proof of Theorem 5.4: Similar to the proof of Theorem 5.3, we apply Theorem 5.1 to derive that

$$F_C = \frac{(C\hat{\beta} - C\beta)^T \{\sigma^2 C(X^T X)^{-1} C^T\}^{-1} (C\hat{\beta} - C\beta)/l}{\hat{\sigma}^2/\sigma^2} \\ \sim \frac{\chi_l^2/l}{\chi_{n-p}^2/(n-p)},$$

where χ_l^2 and χ_{n-p}^2 denote independent χ_l^2 and χ_{n-p}^2 random variables, respectively. Therefore, $F_C \sim F_{l,n-p}$ by the definition of the F distribution in Appendix B. \square

Theorem 5.4 motivates the following confidence region for $C\beta$:

$$\left\{ r : (C\hat{\beta} - r)^T \{C(X^T X)^{-1} C^T\}^{-1} (C\hat{\beta} - r) \leq l\hat{\sigma}^2 f_{1-\alpha, l, n-p} \right\},$$

where $f_{1-\alpha, l, n-p}$ is the upper α quantile of the $F_{l, n-p}$ distribution. By duality of the confidence region and hypothesis testing, we can also construct a level α test for $C\beta$. Most statistical packages automatically report the p -value based on the F statistic in Example 5.1.

As a final remark, the statistics in Theorems 5.3 and 5.4 are called the Wald-type statistics.

5.3 Prediction based on pivotal quantities

Practitioners use OLS not only to infer β but also to predict future outcomes. For the pair of future data (x_{n+1}, y_{n+1}) , we observe only x_{n+1} and want to predict y_{n+1} based on (X, Y) and x_{n+1} . Assume a stable relationship between y_{n+1} and x_{n+1} , that is,

$$y_{n+1} \sim N(x_{n+1}^T \beta, \sigma^2)$$

with the same (β, σ^2) .

First, we can predict the mean of y_{n+1} which is $x_{n+1}^T \beta$. It is just a one-dimensional linear function of β , so the theory in Theorem 5.3 is directly applicable. A natural unbiased predictor is $x_{n+1}^T \hat{\beta}$ with $1 - \alpha$ level prediction interval

$$x_{n+1}^T \hat{\beta} \pm t_{1-\alpha/2, n-p} \hat{\sigma} e_{x_{n+1}}.$$

Second, we can predict y_{n+1} itself, which is a random variable. We can still use $x_{n+1}^T \hat{\beta}$ as a natural unbiased predictor but need to modify the prediction interval. Because $y_{n+1} \perp\!\!\!\perp \hat{\beta}$, we have

$$y_{n+1} - x_{n+1}^T \hat{\beta} \sim N\{0, \sigma^2 + \sigma^2 x_{n+1}^T (X^T X)^{-1} x_{n+1}\},$$

and therefore

$$\frac{y_{n+1} - x_{n+1}^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 x_{n+1}^T (X^T X)^{-1} x_{n+1}}} = \frac{y_{n+1} - x_{n+1}^T \hat{\beta}}{\sqrt{\sigma^2 + \sigma^2 x_{n+1}^T (X^T X)^{-1} x_{n+1}}} / \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \\ \sim \frac{N(0, 1)}{\sqrt{\chi_{n-p}^2/(n-p)}},$$

where $N(0, 1)$ and χ_{n-p}^2 denote independent standard Normal and χ_{n-p}^2 random variables, respectively. Therefore,

$$\frac{y_{n+1} - x_{n+1}^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 x_{n+1}^T (X^T X)^{-1} x_{n+1}}} \sim t_{n-p}$$

is a pivotal quantity. Define the squared prediction error as

$$\begin{aligned} \hat{\text{pe}}_{x_{n+1}}^2 &= \hat{\sigma}^2 + \hat{\sigma}^2 x_{n+1}^T (X^T X)^{-1} x_{n+1} \\ &= \hat{\sigma}^2 \left\{ 1 + n^{-1} x_{n+1}^T \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1} x_{n+1} \right\}, \end{aligned}$$

which has two components. The first one has magnitude close to σ^2 , which is of constant order. The second one has a magnitude that is decreasing in n , if $n^{-1} \sum_{i=1}^n x_i x_i^T$ converges to a finite limit with large n . Therefore, the first component dominates the second one with large n , which results in the main difference between predicting the mean of y_{n+1} and predicting y_{n+1} itself. Using the notation $\hat{\text{pe}}_{x_{n+1}}$, we can construct the following $1 - \alpha$ level prediction interval:

$$x_{n+1}^T \hat{\beta} \pm t_{1-\alpha/2, n-p} \hat{\text{pe}}_{x_{n+1}}.$$

5.4 Examples

Below I illustrate the theory in this chapter with two classic datasets.

5.4.1 Univariate regression

Revisiting Galton's data, we have the following result:

```
> GaltonFamilies = read.table("GaltonFamilies.txt", header = TRUE)
>
> ## OLS fit by the "lm" function
> galton_fit = lm(childHeight ~ midparentHeight,
+                 data = GaltonFamilies)
>
> ## OLS coefficients and inference
> summary(galton_fit)$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.6362405	4.2651074	5.307308	1.390930e-07
midparentHeight	0.6373609	0.0616076	10.345491	8.053865e-24

With the fitted line, we can predict `childHeight` at different values of `midparentHeight`. In the `predict` function, if we specify `interval = "confidence"`, it gives the *confidence* intervals for the means of the new outcomes; if we specify `interval = "prediction"`, it gives the *prediction* intervals for the new outcomes themselves.

```
> ## predictions: confidence and prediction intervals
> new_mph = seq(60, 80, by = 0.5)
> new_data = data.frame(midparentHeight = new_mph)
> new_ci = predict(galton_fit, new_data,
+                 interval = "confidence")
> new_pi = predict(galton_fit, new_data,
+                 interval = "prediction")
> head(round(new_ci, 2))
```

```

      fit   lwr   upr
1 60.88 59.74 62.01
2 61.20 60.12 62.27
3 61.52 60.50 62.53
4 61.83 60.88 62.79
5 62.15 61.25 63.05
6 62.47 61.63 63.31
> head(round(new_pi, 2))
      fit   lwr   upr
1 60.88 54.13 67.63
2 61.20 54.45 67.94
3 61.52 54.78 68.25
4 61.83 55.11 68.56
5 62.15 55.44 68.87
6 62.47 55.76 69.18

```

Figure 5.1 plots the fitted line as well as the confidence intervals and prediction intervals at level 95%.

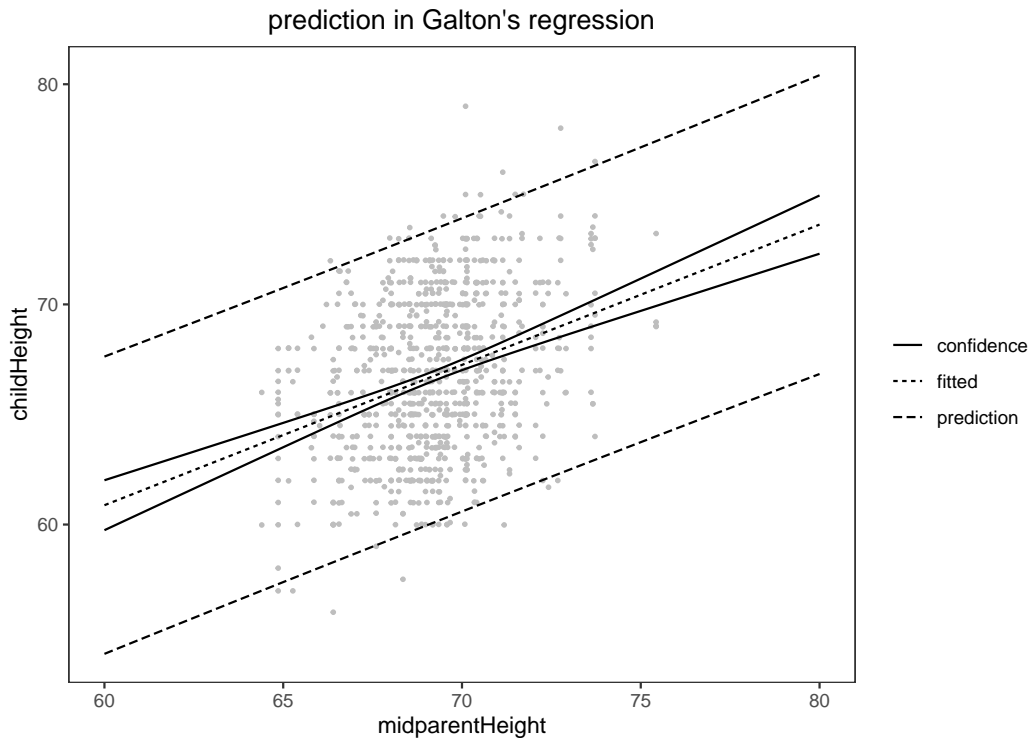


FIGURE 5.1: Prediction in Galton's regression

5.4.2 Anscombe's Quartet: the importance of graphical diagnostics

Anscombe (1973) used four simple datasets to illustrate the importance of graphical diagnostics in linear regression. His datasets are in `anscombe` in the R package `datasets`: `x1` and `y1` constitute the first dataset, and so on.

```

> library(datasets)
> ## Anscombe's Quartet
> anscombe

```

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

The four datasets have similar sample moments.

```
> ## mean of x
> c(mean(anscombe$x1),
+   mean(anscombe$x2),
+   mean(anscombe$x3),
+   mean(anscombe$x4))
[1] 9 9 9 9
> ## variance of x
> c(var(anscombe$x1),
+   var(anscombe$x2),
+   var(anscombe$x3),
+   var(anscombe$x4))
[1] 11 11 11 11
> ## mean of y
> c(mean(anscombe$y1),
+   mean(anscombe$y2),
+   mean(anscombe$y3),
+   mean(anscombe$y4))
[1] 7.500909 7.500909 7.500000 7.500909
> ## variance of y
> c(var(anscombe$y1),
+   var(anscombe$y2),
+   var(anscombe$y3),
+   var(anscombe$y4))
[1] 4.127269 4.127629 4.122620 4.123249
```

The results based on linear regression are almost identical.

```
> ols1 = lm(y1 ~ x1, data = anscombe)
> summary(ols1)

Call:
lm(formula = y1 ~ x1, data = anscombe)

Residuals:
    Min       1Q   Median       3Q      Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001      1.1247   2.667  0.02573 *
x1             0.5001      0.1179   4.241  0.00217 **

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,    Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217

> ols2 = lm(y2 ~ x2, data = anscombe)
> summary(ols2)

Call:
```

```
lm(formula = y2 ~ x2, data = anscombe)

Residuals:
    Min       1Q   Median       3Q      Max
-1.9009 -0.7609  0.1291  0.9491  1.2691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2              0.500      0.118   4.239  0.00218 **

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6662,    Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179

> ols3 = lm(y3 ~ x3, data = anscombe)
> summary(ols3)

Call:
lm(formula = y3 ~ x3, data = anscombe)

Residuals:
    Min       1Q   Median       3Q      Max
-1.1586 -0.6146 -0.2303  0.1540  3.2411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.0025      1.1245   2.670  0.02562 *
x3              0.4997      0.1179   4.239  0.00218 **

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6663,    Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176

> ols4 = lm(y4 ~ x4, data = anscombe)
> summary(ols4)

Call:
lm(formula = y4 ~ x4, data = anscombe)

Residuals:
    Min       1Q   Median       3Q      Max
-1.751 -0.831  0.000  0.809  1.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.0017      1.1239   2.671  0.02559 *
x4              0.4999      0.1178   4.243  0.00216 **

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,    Adjusted R-squared:  0.6297
F-statistic:   18 on 1 and 9 DF,  p-value: 0.002165
```

However, the scatter plots of the datasets in Figure 5.2 reveal fundamental differences between the datasets. The first dataset seems ideal for linear regression. The second dataset shows a quadratic form of y versus x , and therefore, the linear model is misspecified. The third dataset shows a linear trend of y versus x , but an outlier has severely distorted the slope of the linear line. The fourth dataset is supported on only two values of x and thus may suffer from severe extrapolation.

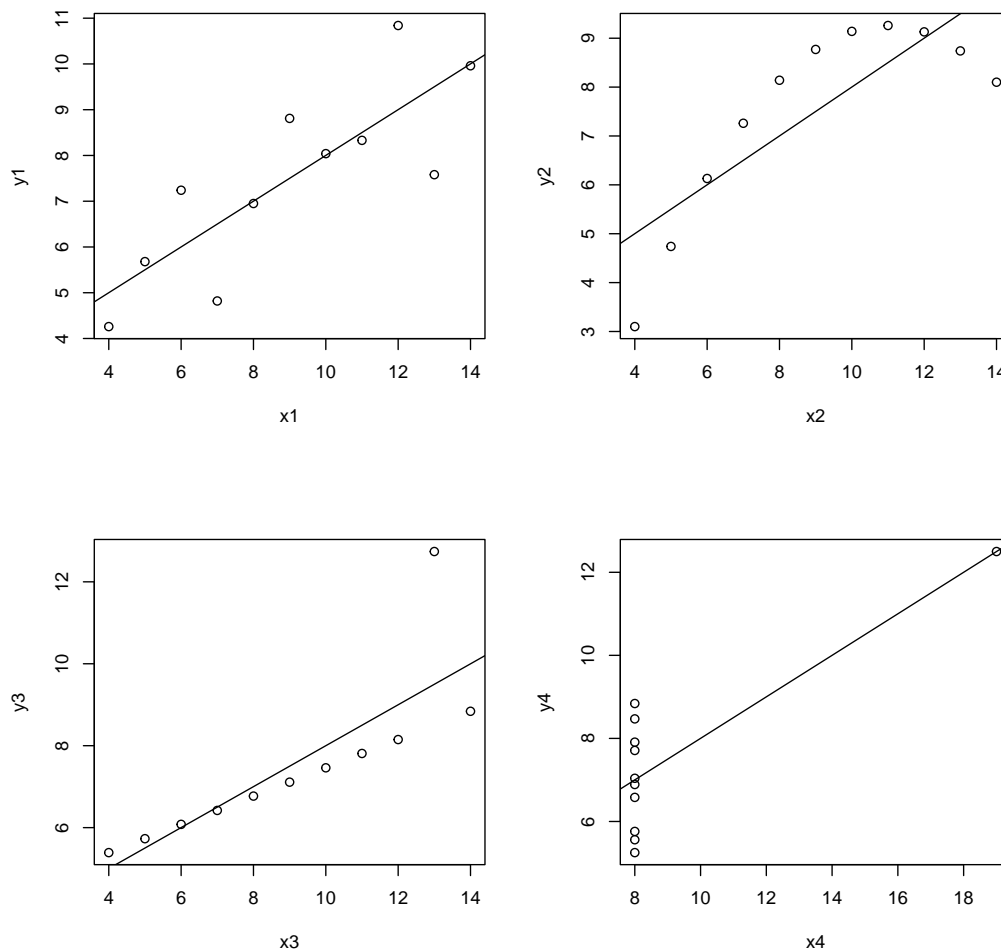


FIGURE 5.2: Anscombe's Quartet: scatter plots

5.4.3 Multivariate regression

The R package `Matching` contains an experimental dataset `lalonge` from LaLonde (1986). Units were randomly assigned to a job training program, with `treat` being the treatment indicator. The outcome `re78` is the real earnings in the year 1978, and other variables are pretreatment covariates. From the simple OLS, the treatment has a significant positive effect, whereas none of the covariates are predictive of the outcome.

```
> library("Matching")
> data(lalonge)
> lalonge_fit = lm(re78 ~ ., data = lalonge)
> summary(lalonge_fit)
```

```
Call:
lm(formula = re78 ~ ., data = lalonge)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-9612	-4355	-1572	3054	53119

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.567e+02	3.522e+03	0.073	0.94193
age	5.357e+01	4.581e+01	1.170	0.24284
educ	4.008e+02	2.288e+02	1.751	0.08058 .
black	-2.037e+03	1.174e+03	-1.736	0.08331 .
hisp	4.258e+02	1.565e+03	0.272	0.78562
married	-1.463e+02	8.823e+02	-0.166	0.86835
nodegr	-1.518e+01	1.006e+03	-0.015	0.98797
re74	1.234e-01	8.784e-02	1.405	0.16079
re75	1.974e-02	1.503e-01	0.131	0.89554
u74	1.380e+03	1.188e+03	1.162	0.24590
u75	-1.071e+03	1.025e+03	-1.045	0.29651
treat	1.671e+03	6.411e+02	2.606	0.00948 **

Residual standard error: 6517 on 433 degrees of freedom
 Multiple R-squared: 0.05822, Adjusted R-squared: 0.0343
 F-statistic: 2.433 on 11 and 433 DF, p-value: 0.005974

The above result shows that none of the pretreatment covariates is significant *marginally*. It is also of interest to test whether they are *jointly* significant. The result below shows that they are only weakly significant at the level 0.05 based on a joint test (the *p*-value is almost 0.05!).

```
> library("car")
> linearHypothesis(lalonde_fit,
+                  c("age=0", "educ=0", "black=0",
+                  "hisp=0", "married=0", "nodegr=0",
+                  "re74=0", "re75=0", "u74=0",
+                  "u75=0"))
Linear hypothesis test

Hypothesis:
age = 0
educ = 0
black = 0
hisp = 0
married = 0
nodegr = 0
re74 = 0
re75 = 0
u74 = 0
u75 = 0

Model 1: restricted model
Model 2: re78 ~ age + educ + black + hisp + married + nodegr + re74 +
re75 + u74 + u75 + treat

   Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     443 1.9178e+10
2     433 1.8389e+10  10 788799023 1.8574 0.04929 *
```

Below I create two pseudo datasets: one with all units assigned to the treatment, and the other with all units assigned to the control, fixing all the pretreatment covariates. The predicted outcomes are the *counterfactual outcomes* under the treatment and control. I further calculate their means and verify that their difference equals the OLS coefficient of *treat*.

```
> new_treat      = lalonde
> new_treat$treat = 1
```

```

> predict_lalonde1 = predict(lalonde_fit, new_treat,
+                           interval = "none")
> new_control      = lalonde
> new_control$treat = 0
> predict_lalonde0 = predict(lalonde_fit, new_control,
+                           interval = "none")
> mean(predict_lalonde1)
[1] 6276.91
> mean(predict_lalonde0)
[1] 4606.201
>
> mean(predict_lalonde1) - mean(predict_lalonde0)
[1] 1670.709

```

5.5 Homework problems

5.1 Maximum likelihood estimator and OLS

Under the Normal linear model, prove that the maximum likelihood estimator for β is the OLS estimator, but the maximum likelihood estimator for σ^2 is $\tilde{\sigma}^2 = \text{RSS}/n$. Compare the mean squared errors of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ for estimating σ^2 .

Remark: The definitions of the mean squared errors are $E\{(\hat{\sigma}^2 - \sigma^2)^2\}$ and $E\{(\tilde{\sigma}^2 - \sigma^2)^2\}$.

5.2 Optimal estimation of the variance

This problem extends Problem 4.4.

Under the Normal linear model, calculate $\text{var}(\hat{\sigma}_A^2)$ and prove that $A = I_n$ minimizes $\text{var}(\hat{\sigma}_A^2)$.

Remark: The minimizer of $\text{var}(\hat{\sigma}_A^2)$ is not unique. For instance, $A = (I_n - H)^+$ also minimizes $\text{var}(\hat{\sigma}_A^2)$, where $+$ denotes the pseudoinverse. There are other minimizers.

5.3 Maximum likelihood estimator with Laplace errors

Assume that $y_i = x_i^T \beta + \sigma \varepsilon_i$, where the ε_i 's are IID Laplace distribution with density $f(\varepsilon) = 2^{-1} e^{-|\varepsilon|}$ ($i = 1, \dots, n$). Find the Maximum likelihood estimators of (β, σ^2) .

Remark: We will revisit this problem in Chapter 26.

5.4 Joint prediction

With multiple future data points (X_{n+1}, Y_{n+1}) where $X_{n+1} \in \mathbb{R}^{l \times p}$ and $Y_{n+1} \in \mathbb{R}^l$, construct the joint predictors and prediction region for Y_{n+1} based on (X, Y) and X_{n+1} .

As a starting point, you can assume that $l \leq p$ and the rows of X_{n+1} are linearly independent. You can then consider the case in which the rows of X_{n+1} are not linearly independent.

Remark: Assume the Normal linear model for all observations and apply Theorem B.10 in Appendix B.

5.5 Two-sample problem

1. Assume that $z_1, \dots, z_m \stackrel{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ and $w_1, \dots, w_n \stackrel{\text{iid}}{\sim} N(\mu_2, \sigma^2)$, and test $H_0 : \mu_1 = \mu_2$. Under H_0 , the t statistic with pooled variance estimator equals

$$t_{\text{equal}} = \frac{\bar{z} - \bar{w}}{\sqrt{\hat{\sigma}^2(m^{-1} + n^{-1})}}$$

where

$$\hat{\sigma}^2 = \{(m-1)S_z^2 + (n-1)S_w^2\} / (m+n-2)$$

with the sample means

$$\bar{z} = m^{-1} \sum_{i=1}^m z_i, \quad \bar{w} = n^{-1} \sum_{i=1}^n w_i,$$

and the sample variances

$$S_z^2 = (m-1)^{-1} \sum_{i=1}^m (z_i - \bar{z})^2, \quad S_w^2 = (n-1)^{-1} \sum_{i=1}^n (w_i - \bar{w})^2.$$

Prove that under H_0 , the t statistic has the following distribution:

$$t_{\text{equal}} \sim t_{m+n-2},$$

Remark: The name “equal” is motivated by the “`var.equal`” parameter of the R function `t.test`.

2. We can write the above problem as testing hypothesis $H_0 : \beta_1 = 0$ in the linear regression $Y = X\beta + \varepsilon$ with

$$Y = \begin{pmatrix} z_1 \\ \vdots \\ z_m \\ w_1 \\ \vdots \\ w_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \\ \varepsilon_{m+1} \\ \vdots \\ \varepsilon_{m+n} \end{pmatrix}.$$

Based on the Normal linear model, we can compute the t statistic for the coefficient of β_1 .

Prove that the t statistic from OLS is numerically identical to t_{equal} .

5.6 Analysis of Variance (ANOVA) with a multi-level treatment

Let x_i be the indicator vector for J treatment levels in a completely randomized experiment, for example, $x_i = e_j = (0, \dots, 1, \dots, 0)^T$ with the j th element being one if unit i receives treatment level j ($j = 1, \dots, J$). Let y_i be the outcome of unit i ($i = 1, \dots, n$). Let \mathcal{T}_j be the indices of units receiving treatment j , and let $n_j = |\mathcal{T}_j|$ be the sample size and $\bar{y}_j = n_j^{-1} \sum_{i \in \mathcal{T}_j} y_i$ be the sample mean of the outcomes under treatment j . Define $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ as the grand mean. We can test whether the treatment has any effect on the outcome by testing the null hypothesis

$$H_0 : \beta_1 = \dots = \beta_J$$

in the Normal linear model $Y = X\beta + \varepsilon$. This is a special case of testing $C\beta = 0$.

Find C and prove that the corresponding F statistic is identical to

$$F = \frac{\sum_{j=1}^J n_j (\bar{y}_j - \bar{y})^2 / (J-1)}{\sum_{j=1}^J \sum_{i \in \mathcal{T}_j} (y_i - \bar{y}_j)^2 / (n-J)} \sim F_{J-1, n-J}.$$

Remarks: (1) This is Fisher's F statistic. (2) In this linear model formulation, X does not contain a column of 1's. (3) The choice of C is not unique, but the final formula for F is. (4) You may use the Sherman–Morrison formula in Problem A.3 in Appendix A.

5.7 Confidence interval for σ^2

Based on Theorem 5.1, construct a $1 - \alpha$ level confidence interval for σ^2 .

5.8 Relationship between t and F

Prove that when C containing only one row c^T , then $T_c^2 = F_C$, where T_c is defined in Theorem 5.3 and F_C is defined in Theorem 5.4.

5.9 RSS and t -statistic in univariate OLS

Focus on univariate OLS discussed in Chapter 2: $y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{\varepsilon}_i$ ($i = 1, \dots, n$).

Prove that RSS equals

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \bar{y})^2 (1 - \hat{\rho}_{xy}^2)$$

and under the homoskedasticity assumption, the t -statistic associated with $\hat{\beta}$ equals

$$\frac{\hat{\rho}_{xy}}{\sqrt{(1 - \hat{\rho}_{xy}^2)/(n-2)}}.$$

5.10 Equivalence of the t -statistics

Consider data $(x_i, y_i)_{i=1}^n$ with scalar x_i and y_i . Run OLS fit of y_i on $(1, x_i)$ to obtain $t_{y|x}$, the t -statistic of the coefficient of x_i , under the homoskedasticity assumption. Run OLS fit of x_i on $(1, y_i)$ to obtain $t_{x|y}$, the t -statistic of the coefficient of y_i , under the homoskedasticity assumption.

Prove $t_{y|x} = t_{x|y}$.

Remark: This is a numerical result that holds without any stochastic assumptions. I give an example below.

```
> library(MASS)
> #simulate bivariate normal distribution
> xy = mvrnorm(n=100, mu=c(0, 0),
+           Sigma=matrix(c(1, 0.5, 0.5, 1), ncol=2))
> xy = as.data.frame(xy)
> colnames(xy) = c("x", "y")
> ## OLS
> reg.y.x = lm(y ~ x, data = xy)
> reg.x.y = lm(x ~ y, data = xy)
> ## compare t statistics based on homoskedastic errors
> summary(reg.y.x)$coef[2, 3]
[1] 4.470331
> summary(reg.x.y)$coef[2, 3]
[1] 4.470331
```

The equivalence of the t -statistics from the OLS fit of y on x and that of x on y demonstrates that based on OLS, the data do not contain any information about the direction of the relationship between x and y .

5.11 An application

The R package `sampleSelection` (Toomet and Henningsen, 2008) describes the dataset `RandHIE` as follows: “The RAND Health Insurance Experiment was a comprehensive study of health care cost, utilization and outcome in the United States. It is the only randomized study of health insurance, and the only study which can give definitive evidence as to the causal effects of different health insurance plans.” You can find more detailed information about other variables in this package. The main outcome of interest `lnmedd01` means the log of medical expenses.

Use OLS to investigate the relationship between the outcome and various important covariates.

Remark: The solution to this problem is not unique, but you need to justify your choice of covariates and model, and need to interpret the results.

6

Asymptotic Inference in OLS: Eicker–Huber–White (EHW) robust standard error

The results in Chapter 5 rely on the assumption of the Normal linear model, which imposes strong distributional assumptions. If we think the Normal linear model is unlikely to hold, how much should we trust the results in Chapter 5?

This chapter will show that the Normality assumption is not that crucial but the homoskedasticity assumption of the errors is. The theory will show that if the homoskedasticity assumption fails, we must modify the covariance estimator of the OLS to be

$$\hat{V}_{\text{EHW}} = (X^T X)^{-1} (X^T \hat{\Omega} X) (X^T X)^{-1} \quad (6.1)$$

with $\hat{\Omega} = \text{diag}\{\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2\}$, where the $\hat{\varepsilon}_i$'s are the residuals. The matrix \hat{V}_{EHW} is called the Eicker–Huber–White (EHW) robust covariance matrix estimator.

Before diving into the theory, I will first present some numerical examples.

6.1 Motivation

6.1.1 Numerical examples

The first one is the ideal Normal linear model:

```
> library(car)
> n      = 200
> x      = runif(n, -2, 2)
> beta   = 1
> xbeta  = x*beta
> Simu1  = replicate(5000,
+                   {y = xbeta + rnorm(n)
+                   ols.fit = lm(y ~ x)
+                   c(summary(ols.fit)$coef[2, 1:2],
+                     sqrt(hccm(ols.fit)[2, 2]))
+                   })
```

In the above, I generate outcomes from a simple linear model $y_i = x_i + \varepsilon_i$ with $\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2 = 1)$. Over 5000 replications of the data, we computed the OLS coefficient $\hat{\beta}$ of x_i and reported two standard errors. One is the standard error discussed in Chapter 5 under the Normal linear model, which is also the default choice of the `lm` function of `R`. The other one is the square root of (2,2)th element of \hat{V}_{EHW} in (6.1), which can be computed by the `hccm` function in the `R` package `car`. The (1,1) the panel of Figure 6.1 shows the histogram of the estimator and reports the standard error (se0), as well as two estimated standard

errors (se1 and se2). The distribution of $\hat{\beta}$ is symmetric and bell-shaped around the true parameter 1, and both of the two estimated standard errors are close to the true one.

To investigate the impact of Normality, I change the error terms to be IID exponential with mean 1 and variance 1.

```
> Simu2 = replicate(5000,
+                   {y = xbeta + rexp(n)
+                   ols.fit = lm(y ~ x)
+                   c(summary(ols.fit)$coef[2, 1:2],
+                     sqrt(hccm(ols.fit)[2, 2]))
+                   })
```

The (1, 2) panel of Figure 6.1 corresponds to this setting. With non-Normal errors, $\hat{\beta}$ is still symmetric and bell-shaped around the true parameter 1, and the estimated standard errors are close to the true one. So the Normality of the error terms does not seem to be a crucial assumption for the validity of the inference procedure under the Normal linear model.

I then generate errors from Normal with variance depending on x :

```
> Simu3 = replicate(5000,
+                   {y = xbeta + rnorm(n, 0, abs(x))
+                   ols.fit = lm(y ~ x)
+                   c(summary(ols.fit)$coef[2, 1:2],
+                     sqrt(hccm(ols.fit)[2, 2]))
+                   })
```

The (2, 1) panel of Figure 6.1 corresponds to this setting. With heteroskedastic Normal errors, $\hat{\beta}$ is symmetric and bell-shaped around the true parameter 1, se2 is close to se0, but se1 underestimates se0. So the heteroskedasticity of the error terms does not change the Normality of the OLS estimator dramatically, although the statistical inference discussed in Chapter 5 is invalid.

Finally, I generate heteroskedastic non-Normal errors:

```
> Simu4 = replicate(5000,
+                   {y = xbeta + runif(n, -x^2, x^2)
+                   ols.fit = lm(y ~ x)
+                   c(summary(ols.fit)$coef[2, 1:2],
+                     sqrt(hccm(ols.fit)[2, 2]))
+                   })
```

The (2, 2) panel of Figure 6.1 corresponds to this setting, which has a similar pattern as the (2, 1) panel. So the Normality of the error terms is not crucial, but the homoskedasticity is.

6.1.2 Goal of this chapter

This chapter will still impose the linearity assumption, but relax the distributional assumption on the error terms. Assume the following heteroskedastic linear model.

Assumption 6.1 (Heteroskedastic linear model) *We have*

$$y_i = x_i^T \beta + \varepsilon_i,$$

where the ε_i 's are independent with mean zero and variance σ_i^2 ($i = 1, \dots, n$). The design matrix $X = (x_1^T, \dots, x_n^T)^T$ is fixed with linearly independent column vectors, and $(\beta, \sigma_1^2, \dots, \sigma_n^2)$ are unknown parameters.

Because the error terms can have different variances, they are not IID in general under the heteroskedastic linear model. Their variances can be functions of the x_i 's, and the

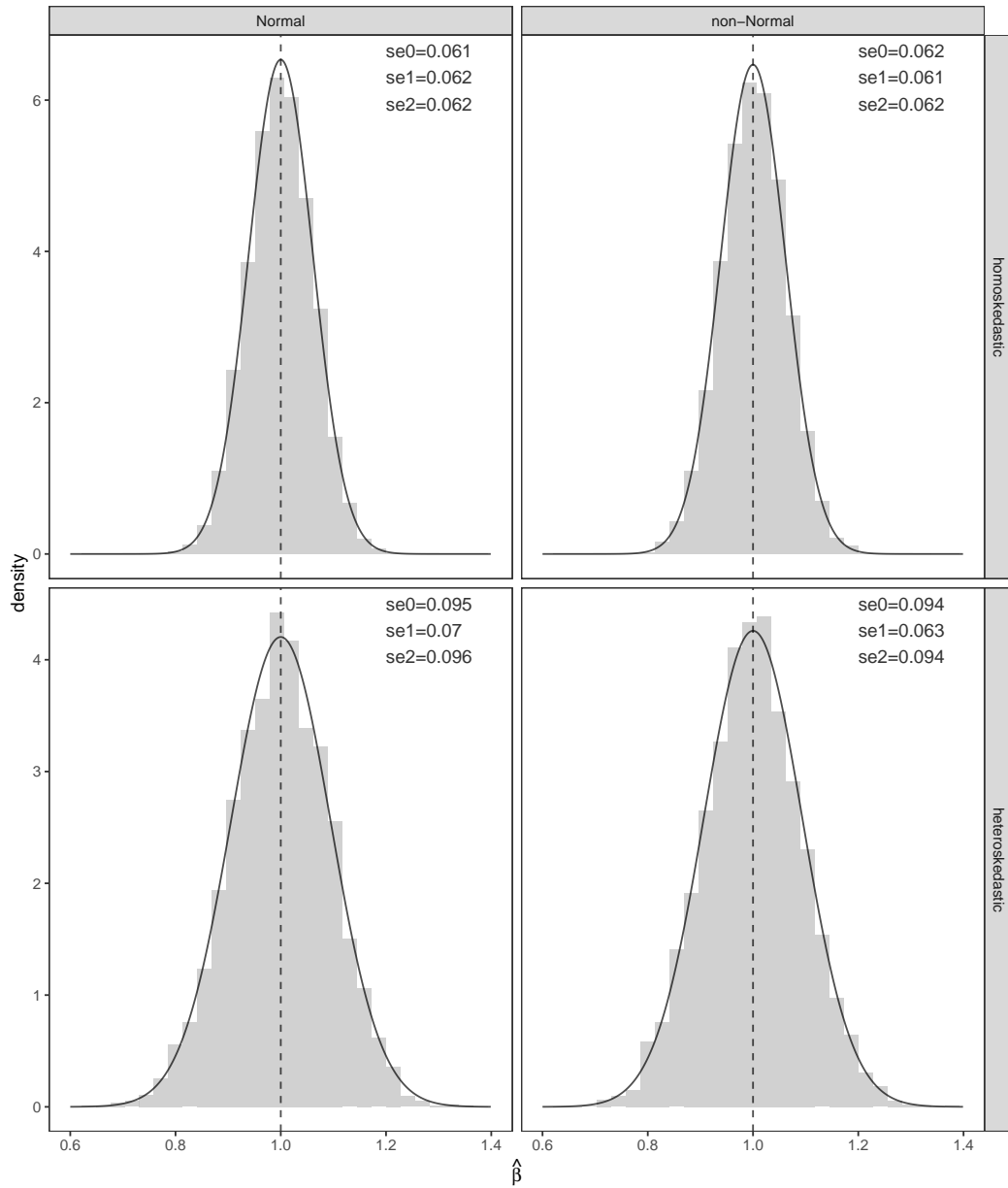


FIGURE 6.1: Simulation with 5000 replications: “se0” denotes the true standard error of $\hat{\beta}$, “se1” denotes the estimated standard error based on the homoskedasticity assumption, and “se2” denotes the Eicker–Huber–White standard error allowing for heteroskedasticity. The density curves are Normal with mean 1 and standard deviation se0.

variances σ_i^2 are n free unknown numbers. Again, treating the x_i 's as fixed is not essential, because we can condition on them if they are random.

Without imposing Normality on the error terms, we cannot determine the finite sample exact distribution of the OLS estimator. This chapter will use the asymptotic analysis, assuming that the sample size n is large so that certain laws of large numbers and central limit theorems (CLTs) hold.

The asymptotic analysis later will show that if the error terms are homoskedastic, i.e., $\sigma_i^2 = \sigma^2$ for all $i = 1, \dots, n$, we can still trust the statistical inference discussed in Chapter 5 based on the Normal linear model as long as the CLT for the OLS estimator holds as $n \rightarrow \infty$. If the error terms are heteroskedastic, i.e., their variances are different, we must modify the standard error as the so-called Eicker–Huber–White (EHW) heteroskedasticity robust standard error introduced in (6.1). I will give the technical details below. If you are unfamiliar with the asymptotic analysis, please first review the basics in Appendix C.

6.2 Consistency of OLS

Under the heteroskedastic linear model, the OLS estimator $\hat{\beta}$ is still unbiased for β because the error terms have mean zero. Moreover, we can show that it is consistent for β with large n and some regularity conditions. We start with a lemma.

Lemma 6.1 *Under Assumption 6.1, the OLS estimator has the representation $\hat{\beta} - \beta = B_n^{-1}\xi_n$, where*

$$\begin{aligned} B_n &= n^{-1} \sum_{i=1}^n x_i x_i^T, \\ \xi_n &= n^{-1} \sum_{i=1}^n x_i \varepsilon_i. \end{aligned}$$

Proof of Lemma 6.1: Since $y_i = x_i^T \beta + \varepsilon_i$, we have

$$\begin{aligned} \hat{\beta} &= B_n^{-1} n^{-1} \sum_{i=1}^n x_i y_i \\ &= B_n^{-1} n^{-1} \sum_{i=1}^n x_i (x_i^T \beta + \varepsilon_i) \\ &= B_n^{-1} B_n \beta + B_n^{-1} n^{-1} \sum_{i=1}^n x_i \varepsilon_i \\ &= \beta + B_n^{-1} \xi_n. \end{aligned}$$

Therefore, $\hat{\beta} - \beta = B_n^{-1} \xi_n$. □

In the representation of Lemma 6.1, B_n is fixed and ξ_n is random. Since $E(\xi_n) = 0$, we

know that $E(\hat{\beta}) = \beta$, so the OLS estimator is unbiased. Moreover,

$$\begin{aligned} \text{cov}(\xi_n) &= \text{cov}\left(n^{-1} \sum_{i=1}^n x_i \varepsilon_i\right) \\ &= n^{-2} \sum_{i=1}^n \sigma_i^2 x_i x_i^\top \\ &= M_n/n, \end{aligned}$$

where

$$M_n = n^{-1} \sum_{i=1}^n \sigma_i^2 x_i x_i^\top.$$

So the covariance of the OLS estimator is

$$\text{cov}(\hat{\beta}) = n^{-1} B_n^{-1} M_n B_n^{-1}.$$

It has a sandwich form, justifying the choice of notation B_n for the “bread matrix” and M_n for the “meat matrix.” Vegetarians can read M_n as the “middle matrix.”

Intuitively, if B_n and M_n have finite limits, then the covariance of $\hat{\beta}$ shrinks to zero with large n , implying that $\hat{\beta}$ will concentrate near its mean β . This is the idea of consistency, formally stated below.

Assumption 6.2 $B_n \rightarrow B$ and $M_n \rightarrow M$ where B and M are finite with B invertible.

Theorem 6.1 Under Assumptions 6.1 and 6.2, we have $\hat{\beta} \rightarrow \beta$ in probability.

Proof of Theorem 6.1: We only need to show that $\xi_n \rightarrow 0$ in probability. It has mean zero and covariance matrix M_n/n , so it converges to zero in probability using Proposition C.4 in Appendix C. \square

6.3 Asymptotic Normality of the OLS estimator

Intuitively, ξ_n is the sample average of some independent terms, and therefore, the classic Lindberg–Feller theorem (see Proposition C.8 in Appendix C) guarantees that it enjoys a CLT under some regularity conditions. Consequently, $\hat{\beta}$ also enjoys a CLT with mean β and covariance matrix $n^{-1} B_n^{-1} M_n B_n^{-1}$. The asymptotic results in this chapter require rather tedious regularity conditions. I give them for generality, and they hold automatically if we are willing to assume that the covariates and error terms are all bounded by a constant not depending on n . These general conditions are basically moment conditions required by the law of large numbers and CLT. You do not have to pay too much attention to the conditions when you first read this chapter or you focus on applied statistics.

The CLT relies on an additional condition on a higher-order moment

$$d_{2+\delta,n} = n^{-1} \sum_{i=1}^n \|x_i\|^{2+\delta} E(|\varepsilon_i|^{2+\delta}).$$

Theorem 6.2 Under Assumptions 6.1 and 6.2, if there exist a $\delta > 0$ and $C > 0$ not depending on n such that $d_{2+\delta,n} \leq C$, then

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, B^{-1} M B^{-1})$$

in distribution.

Proof of Theorem 6.2: The key is to show the CLT for ξ_n , and the CLT for $\hat{\beta}$ holds due to the Slutsky's Theorem; see Appendix C for a review. Define

$$z_{n,i} = n^{-1/2} x_i \varepsilon_i, \quad (i = 1, \dots, n)$$

with mean zero and finite covariance, and we need to verify the two conditions required by the Lindeberg–Feller CLT stated as Proposition C.8 in Appendix C.

First, the Lyapunov condition holds because

$$\begin{aligned} \sum_{i=1}^n E(\|z_{n,i}\|^{2+\delta}) &= \sum_{i=1}^n E\left(n^{-(2+\delta)/2} \|x_i\|^{2+\delta} |\varepsilon_i|^{2+\delta}\right) \\ &= n^{-\delta/2} \times n^{-1} \sum_{i=1}^n \|x_i\|^{2+\delta} E(|\varepsilon_i|^{2+\delta}) \\ &= n^{-\delta/2} \times d_{2+\delta,n} \\ &\rightarrow 0, \end{aligned}$$

by the assumption that $d_{2+\delta,n}$ is bounded by a constant C .

Second,

$$\begin{aligned} \sum_{i=1}^n \text{cov}(z_{n,i}) &= n^{-1} \sum_{i=1}^n \sigma_i^2 x_i x_i^\top \\ &= M_n \\ &\rightarrow M. \end{aligned}$$

So the Lindberg–Feller CLT implies that $n^{-1/2} \sum_{i=1}^n x_i \varepsilon_i = \sum_{i=1}^n z_{n,i} \rightarrow N(0, M)$ in distribution. \square

6.4 Eicker–Huber–White standard error

6.4.1 Sandwich variance estimator

The CLT in Theorem 6.2 shows that

$$\hat{\beta} \overset{a}{\sim} N(\beta, n^{-1} B^{-1} M B^{-1}),$$

where $\overset{a}{\sim}$ denotes “approximation in distribution.” However, the asymptotic covariance is unknown, and we need to use the data to construct a reasonable estimator for it to conduct statistical inference. It is relatively easy to replace B with its sample analog B_n , but

$$\tilde{M}_n = n^{-1} \sum_{i=1}^n \varepsilon_i^2 x_i x_i^\top$$

as the sample analog for M is not directly useful because the error terms are unknown either. It is natural to use $\hat{\varepsilon}_i^2$ to replace ε_i^2 to obtain the following estimator for M :

$$\hat{M}_n = n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i^\top.$$

Although each $\hat{\varepsilon}_i^2$ is a poor estimator for σ_i^2 , the sample average \hat{M}_n turns out to be well-behaved with large n and the regularity conditions below.

Theorem 6.3 Under Assumptions 6.1 and 6.2, we have $\hat{M}_n \rightarrow M$ in probability if

$$n^{-1} \sum_{i=1}^n \text{var}(\varepsilon_i^2) x_{ij_1}^2 x_{ij_2}^2, \quad n^{-1} \sum_{i=1}^n |x_{ij_1} x_{ij_2} x_{ij_3} x_{ij_4}|, \quad n^{-1} \sum_{i=1}^n \sigma_i^2 x_{ij_1}^2 x_{ij_2}^2 x_{ij_3}^2 \quad (6.2)$$

are bounded from above by a constant C not depending on n for any $j_1, j_2, j_3, j_4 = 1, \dots, p$.

Proof of Theorem 6.3: Assumption 6.2 ensures that $\hat{\beta} \rightarrow \beta$ in probability by Theorem 6.1. Markov's inequality and the boundedness of the first term in (6.2) ensure that $\hat{M}_n - M_n \rightarrow 0$ in probability. So we only need to show that $\hat{M}_n - \tilde{M}_n \rightarrow 0$ in probability. The (j_1, j_2) th element of their difference is

$$\begin{aligned} (\hat{M}_n - \tilde{M}_n)_{j_1, j_2} &= n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 x_{i, j_1} x_{i, j_2} - n^{-1} \sum_{i=1}^n \varepsilon_i^2 x_{i, j_1} x_{i, j_2} \\ &= n^{-1} \sum_{i=1}^n \left[\left(\varepsilon_i + x_i^\top \beta - x_i^\top \hat{\beta} \right)^2 - \varepsilon_i^2 \right] x_{i, j_1} x_{i, j_2} \\ &= n^{-1} \sum_{i=1}^n \left[\left(x_i^\top \beta - x_i^\top \hat{\beta} \right)^2 + 2\varepsilon_i \left(x_i^\top \beta - x_i^\top \hat{\beta} \right) \right] x_{i, j_1} x_{i, j_2} \\ &= \text{I} + 2 \cdot \text{II}, \end{aligned}$$

where

$$\text{I} = (\beta - \hat{\beta})^\top n^{-1} \sum_{i=1}^n x_i x_i^\top x_{i, j_1} x_{i, j_2} (\beta - \hat{\beta}), \quad (6.3)$$

$$\text{II} = (\beta - \hat{\beta})^\top n^{-1} \sum_{i=1}^n x_i x_{i, j_1} x_{i, j_2} \varepsilon_i. \quad (6.4)$$

The $(\hat{M}_n - \tilde{M}_n)_{j_1, j_2}$ converges to 0 in probability because both term I and term II converge to 0 in probability. I will show these two facts below.

First, $\text{I} \rightarrow 0$ in probability because $\beta - \hat{\beta} \rightarrow 0$ in probability and $|n^{-1} \sum_{i=1}^n x_i x_i^\top x_{i, j_1} x_{i, j_2}|$ is bounded by the assumption in (6.2).

Second, $\text{II} \rightarrow 0$ in probability because $\beta - \hat{\beta} \rightarrow 0$ in probability and $n^{-1} \sum_{i=1}^n x_i x_{i, j_1} x_{i, j_2} \varepsilon_i \rightarrow 0$ in probability due to¹

$$\begin{aligned} E(n^{-1} \sum_{i=1}^n x_i x_{i, j_1} x_{i, j_2} \varepsilon_i) &= 0, \\ \text{cov}(n^{-1} \sum_{i=1}^n x_i x_{i, j_1} x_{i, j_2} \varepsilon_i) &= n^{-2} \sum_{i=1}^n \sigma_i^2 x_i x_i^\top x_{i, j_1}^2 x_{i, j_2}^2 \rightarrow 0, \end{aligned}$$

and Markov's inequality (see Proposition C.4). \square

The final variance estimator for $\hat{\beta}$ is

$$\hat{V}_{\text{EHW}} = n^{-1} \left(n^{-1} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i^\top \right) \left(n^{-1} \sum_{i=1}^n x_i x_i^\top \right)^{-1},$$

¹Technically, we only need this term to be bounded in probability to ensure that $\text{II} \rightarrow 0$ in probability. A weaker condition is that $n^{-2} \sum_{i=1}^n \sigma_i^2 x_{ij_1}^2 x_{ij_2}^2 x_{ij_3}^2$ is bounded by a constant. Nevertheless, I invoke a stronger condition because the sample mean is easier to interpret.

which is called the Eicker–Huber–White (EHW) heteroskedasticity robust covariance matrix. In matrix form, it equals

$$\hat{V}_{\text{EHW}} = (X^T X)^{-1} (X^T \hat{\Omega} X) (X^T X)^{-1},$$

with $\hat{\Omega} = \text{diag}\{\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2\}$, which was introduced in (6.1) at the beginning of this chapter. Eicker (1967) first proposed to use \hat{V}_{EHW} . White (1980a) popularized it in economics, which has been influential in empirical research. Related estimators appeared in many other contexts of statistics. Cox (1961) and Huber (1967) discussed the sandwich variance in the context of misspecified parametric models; see Appendix D.2. Fuller (1975) proposed a more general form of \hat{V}_{EHW} in the context of survey sampling. The square root of the diagonal terms of \hat{V}_{EHW} , denoted by $\hat{\text{se}}_{\text{EHW},j}$ ($j = 1, \dots, p$), are called the heteroskedasticity-consistent standard errors, heteroskedasticity-robust standard errors, White standard errors, Huber–White standard errors, or Eicker–Huber–White standard errors, among many other names.

We can conduct statistical inference based on Normal approximations. For example, we can test linear hypotheses based on

$$\hat{\beta} \stackrel{\text{a}}{\sim} N(\beta, \hat{V}_{\text{EHW}}),$$

and in particular, we can infer each element of the coefficient based on

$$\hat{\beta}_j \stackrel{\text{a}}{\sim} N(\beta_j, \hat{\text{se}}_{\text{EHW},j}^2).$$

6.4.2 Other heteroskedasticity-consistent (HC) standard errors

Statistical inference based on the EHW standard error relaxes the parametric assumptions of the Normal linear model. However, its validity relies strongly on the asymptotic argument. In finite samples, it can have poor behavior. Since White (1980a) published his paper, several modifications of \hat{V}_{EHW} appeared aiming for better finite-sample properties. I summarize some of them below. They all rely on the h_{ii} 's, which are the diagonal elements of the projection matrix H and called the *leverage scores*. Define

$$\hat{V}_{\text{EHW},k} = n^{-1} \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n \hat{\varepsilon}_{i,k}^2 x_i x_i^T \right) \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1},$$

where

$$\hat{\varepsilon}_{i,k} = \begin{cases} \hat{\varepsilon}_i, & (k = 0, \text{ HC0}); \\ \hat{\varepsilon}_i \sqrt{\frac{n}{n-p}}, & (k = 1, \text{ HC1}); \\ \hat{\varepsilon}_i / \sqrt{1 - h_{ii}}, & (k = 2, \text{ HC2}); \\ \hat{\varepsilon}_i / (1 - h_{ii}), & (k = 3, \text{ HC3}); \\ \hat{\varepsilon}_i / (1 - h_{ii})^{\min\{2, nh_{ii}/(2p)\}}, & (k = 4, \text{ HC4}). \end{cases}$$

The HC1 correction is similar to the degrees of freedom correction in the OLS covariance estimator. The HC2 correction was motivated by the unbiasedness of covariance when the error terms have the same variance; see Problem 6.8 for more details. The HC3 correction was motivated by a method called *jackknife*, which will be discussed in Chapter 11. This version appeared even earlier than White (1980a); see Miller (1974), Hinkley (1977), and Reeds (1978). See MacKinnon and White (1985), Long and Ervin (2000) and Cribari-Neto (2004) for reviews. Based on simulation studies, Long and Ervin (2000) recommended HC3.

6.4.3 Special case with homoskedasticity

As an important special case with $\sigma_i^2 = \sigma^2$ for all $i = 1, \dots, n$, we have

$$M_n = \sigma^2 n^{-1} \sum_{i=1}^n x_i x_i^T = \sigma^2 B_n,$$

which simplifies the covariance of $\hat{\beta}$ to $\text{cov}(\hat{\beta}) = \sigma^2 B_n^{-1}/n$, and the asymptotic Normality to $\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, \sigma^2 B^{-1})$ in distribution. We have shown that under the Gauss–Markov model, $\hat{\sigma}^2 = (n - p)^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2$ is unbiased for σ^2 . Moreover, $\hat{\sigma}^2$ is consistent for σ^2 under the same condition as Theorem 6.1, justifying the use of

$$\hat{V} = \hat{\sigma}^2 \left(\sum_{i=1}^n x_i x_i^T \right) = \hat{\sigma}^2 (X^T X)^{-1}$$

as the covariance estimator. So under homoskedasticity, we can conduct statistical inference based on the following approximate Normality:

$$\hat{\beta} \stackrel{a}{\sim} N(\beta, \hat{\sigma}^2 (X^T X)^{-1}).$$

It is slightly different from the inference based on t and F distributions. But with large n , the difference is very small.

I will end this section with a formal result on the consistency of $\hat{\sigma}^2$.

Theorem 6.4 *Under Assumptions 6.1 and 6.2, we have $\hat{\sigma}^2 \rightarrow \sigma^2$ in probability if $\sigma_i^2 = \sigma^2 < \infty$ for all $i = 1, \dots, n$, and $n^{-1} \sum_{i=1}^n \text{var}(\varepsilon_i^2)$ is bounded above by a constant not depending on n .*

Proof of Theorem 6.4: Using Markov's inequality, we can show that $n^{-1} \sum_{i=1}^n \varepsilon_i^2 \rightarrow \sigma^2$ in probability. In addition, $n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2$ has the same probability limit as $\hat{\sigma}^2$. So we only need to show that $n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 - n^{-1} \sum_{i=1}^n \varepsilon_i^2 \rightarrow 0$ in probability. Their difference is

$$\begin{aligned} & n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 - n^{-1} \sum_{i=1}^n \varepsilon_i^2 \\ &= n^{-1} \sum_{i=1}^n \left\{ \left(\varepsilon_i + x_i^T \beta - x_i^T \hat{\beta} \right)^2 - \varepsilon_i^2 \right\} \\ &= n^{-1} \sum_{i=1}^n \left\{ \left(x_i^T \beta - x_i^T \hat{\beta} \right)^2 + 2 \left(x_i^T \beta - x_i^T \hat{\beta} \right) \varepsilon_i \right\} \\ &= (\beta - \hat{\beta})^T n^{-1} \sum_{i=1}^n x_i x_i^T (\beta - \hat{\beta}) + 2(\beta - \hat{\beta})^T n^{-1} \sum_{i=1}^n x_i \varepsilon_i \\ &= -(\beta - \hat{\beta})^T n^{-1} \sum_{i=1}^n x_i x_i^T (\beta - \hat{\beta}), \end{aligned}$$

where the last step follows from Lemma 6.1. So the difference converges to zero in probability because $\hat{\beta} - \beta \rightarrow 0$ in probability by Theorem 6.1 and $B_n \rightarrow B$ by Assumption 6.2. \square

6.5 Examples

I use three examples to compare various standard errors for the regression coefficients. The `car` package contains the `hccm` function that implements the EHW standard errors.

```
> library("car")
```

6.5.1 LaLonde experimental data

First, I revisit the `lalonde` data, which were analyzed in Chapter 5.4.3. In the following analysis, different standard errors give similar t statistics. Only `treat` is significant, but none of the other pretreatment covariates are significant.

```
> library("Matching")
> data(lalonde)
> ols.fit = lm(re78 ~ ., data = lalonde)
> ols.fit.hc0 = sqrt(diag(hccm(ols.fit, type = "hc0")))
> ols.fit.hc1 = sqrt(diag(hccm(ols.fit, type = "hc1")))
> ols.fit.hc2 = sqrt(diag(hccm(ols.fit, type = "hc2")))
> ols.fit.hc3 = sqrt(diag(hccm(ols.fit, type = "hc3")))
> ols.fit.hc4 = sqrt(diag(hccm(ols.fit, type = "hc4")))
> ols.fit.coef = summary(ols.fit)$coef
> tvalues = ols.fit.coef[,1]/
+   cbind(ols.fit.coef[,2], ols.fit.hc0, ols.fit.hc1,
+         ols.fit.hc2, ols.fit.hc3, ols.fit.hc4)
> colnames(tvalues) = c("ols", "hc0", "hc1", "hc2", "hc3", "hc4")
> round(tvalues, 2)
```

	ols	hc0	hc1	hc2	hc3	hc4
(Intercept)	0.07	0.07	0.07	0.07	0.07	0.07
age	1.17	1.29	1.28	1.27	1.25	1.25
educ	1.75	2.03	2.00	1.99	1.94	1.92
black	-1.74	-2.00	-1.97	-1.95	-1.91	-1.91
hisp	0.27	0.30	0.30	0.30	0.29	0.29
married	-0.17	-0.17	-0.17	-0.17	-0.16	-0.16
nodegr	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01
re74	1.40	0.98	0.96	0.92	0.87	0.77
re75	0.13	0.14	0.14	0.13	0.13	0.12
u74	1.16	0.89	0.88	0.87	0.85	0.83
u75	-1.05	-0.76	-0.75	-0.75	-0.74	-0.74
treat	2.61	2.49	2.46	2.45	2.41	2.40

6.5.2 Data from King and Roberts (2015)

The following example uses the data from King and Roberts (2015). The outcome variable is the multilateral aid flows, and the covariates include log population, log population squared, gross domestic product, former colony status, distance from the Western world, political freedom, military expenditures, arms imports, and the indicators for the years. Different standard errors give very different t statistics for some coefficients.

```
> library(foreign)
> dat = read.dta("isq.dta")
> dat = na.omit(dat[,c("multish", "lnpop", "lnpopsq",
+                     "lngdp", "lncolony", "lndist",
+                     "freedom", "militexp", "arms",
+                     "year83", "year86", "year89", "year92")])
> ols.fit = lm(multish ~ lnpop + lnpopsq + lngdp + lncolony
+               + lndist + freedom + militexp + arms
+               + year83 + year86 + year89 + year92, data=dat)
```

```

> ols.fit.hc0 = sqrt(diag(hccm(ols.fit, type = "hc0")))
> ols.fit.hc1 = sqrt(diag(hccm(ols.fit, type = "hc1")))
> ols.fit.hc2 = sqrt(diag(hccm(ols.fit, type = "hc2")))
> ols.fit.hc3 = sqrt(diag(hccm(ols.fit, type = "hc3")))
> ols.fit.hc4 = sqrt(diag(hccm(ols.fit, type = "hc4")))
> ols.fit.coef = summary(ols.fit)$coef
> tvalues = ols.fit.coef[,1]/
+   cbind(ols.fit.coef[,2], ols.fit.hc0, ols.fit.hc1,
+         ols.fit.hc2, ols.fit.hc3, ols.fit.hc4)
> colnames(tvalues) = c("ols", "hc0", "hc1", "hc2", "hc3", "hc4")
> round(tvalues, 2)

```

	ols	hc0	hc1	hc2	hc3	hc4
(Intercept)	7.40	4.60	4.54	4.43	4.27	4.14
lnpop	-8.25	-4.46	-4.40	-4.30	-4.14	-4.01
lnpopsq	9.56	4.79	4.72	4.61	4.44	4.31
lngdp	-6.39	-6.14	-6.06	-6.01	-5.88	-5.86
lncolony	4.70	4.75	4.69	4.64	4.53	4.47
lndist	-0.14	-0.16	-0.16	-0.16	-0.15	-0.16
freedom	2.25	1.80	1.78	1.75	1.69	1.65
militexp	0.51	0.59	0.59	0.57	0.55	0.52
arms	1.34	1.17	1.15	1.10	1.03	0.91
year83	1.05	0.85	0.84	0.83	0.80	0.79
year86	0.35	0.40	0.39	0.39	0.38	0.38
year89	0.70	0.81	0.80	0.80	0.78	0.79
year92	0.31	0.40	0.40	0.40	0.39	0.40

However, if we apply the log transformation on the outcome, then all standard errors give similar t statistics.

```

> ols.fit = lm(log(multish + 1) ~ lnpop + lnpopsq + lngdp + lncolony
+           + lndist + freedom + militexp + arms
+           + year83 + year86 + year89 + year92, data=dat)
> ols.fit.hc0 = sqrt(diag(hccm(ols.fit, type = "hc0")))
> ols.fit.hc1 = sqrt(diag(hccm(ols.fit, type = "hc1")))
> ols.fit.hc2 = sqrt(diag(hccm(ols.fit, type = "hc2")))
> ols.fit.hc3 = sqrt(diag(hccm(ols.fit, type = "hc3")))
> ols.fit.hc4 = sqrt(diag(hccm(ols.fit, type = "hc4")))
> ols.fit.coef = summary(ols.fit)$coef
> tvalues = ols.fit.coef[,1]/
+   cbind(ols.fit.coef[,2], ols.fit.hc0, ols.fit.hc1,
+         ols.fit.hc2, ols.fit.hc3, ols.fit.hc4)
> colnames(tvalues) = c("ols", "hc0", "hc1", "hc2", "hc3", "hc4")
> round(tvalues, 2)

```

	ols	hc0	hc1	hc2	hc3	hc4
(Intercept)	2.96	2.81	2.77	2.72	2.63	2.53
lnpop	-2.87	-2.63	-2.60	-2.54	-2.45	-2.35
lnpopsq	4.21	3.72	3.67	3.59	3.46	3.32
lngdp	-8.02	-7.49	-7.38	-7.38	-7.27	-7.33
lncolony	6.31	6.19	6.11	6.08	5.97	5.95
lndist	-0.16	-0.14	-0.14	-0.14	-0.14	-0.14
freedom	1.47	1.53	1.51	1.50	1.47	1.46
militexp	-0.32	-0.32	-0.31	-0.31	-0.30	-0.29
arms	1.27	1.12	1.10	1.05	0.98	0.86
year83	0.10	0.10	0.10	0.10	0.10	0.10
year86	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14
year89	0.46	0.45	0.44	0.44	0.44	0.44
year92	0.03	0.03	0.03	0.03	0.03	0.03

In general, the difference between the OLS and EHW standard errors may be due to the heteroskedasticity or the poor approximation of the linear model. The above two analyses based on the original and transformed outcomes suggest that the linear approximation works better for the log-transformed outcome. We will discuss the issues of model misspecification and transformation in Chapters 12 and 16, respectively.

6.5.3 Boston housing data

I also re-analyze the classic Boston housing data (Harrison Jr and Rubinfeld, 1978). The outcome variable is the median value of owner-occupied homes in US dollars 1000, and the covariates include per capita crime rate by town, the proportion of residential land zoned for lots over 25,000 square feet, the proportion of non-retail business acres per town, etc. You can find more details in the R package `mlbench`. In this example, different standard errors give very different t statistics.

```
> library("mlbench")
> data(BostonHousing)
> ols.fit = lm(medv ~ ., data = BostonHousing)
> summary(ols.fit)
```

Call:
lm(formula = medv ~ ., data = BostonHousing)

Residuals:

Min	1Q	Median	3Q	Max
-15.595	-2.730	-0.518	1.777	26.199

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.646e+01	5.103e+00	7.144	3.28e-12	***
crim	-1.080e-01	3.286e-02	-3.287	0.001087	**
zn	4.642e-02	1.373e-02	3.382	0.000778	***
indus	2.056e-02	6.150e-02	0.334	0.738288	
chas1	2.687e+00	8.616e-01	3.118	0.001925	**
nox	-1.777e+01	3.820e+00	-4.651	4.25e-06	***
rm	3.810e+00	4.179e-01	9.116	< 2e-16	***
age	6.922e-04	1.321e-02	0.052	0.958229	
dis	-1.476e+00	1.995e-01	-7.398	6.01e-13	***
rad	3.060e-01	6.635e-02	4.613	5.07e-06	***
tax	-1.233e-02	3.760e-03	-3.280	0.001112	**
ptratio	-9.527e-01	1.308e-01	-7.283	1.31e-12	***
b	9.312e-03	2.686e-03	3.467	0.000573	***
lstat	-5.248e-01	5.072e-02	-10.347	< 2e-16	***

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared: 0.7406, Adjusted R-squared: 0.7338
F-statistic: 108.1 on 13 and 492 DF, p-value: < 2.2e-16

```
>
> ols.fit.hc0 = sqrt(diag(hccm(ols.fit, type = "hc0")))
> ols.fit.hc1 = sqrt(diag(hccm(ols.fit, type = "hc1")))
> ols.fit.hc2 = sqrt(diag(hccm(ols.fit, type = "hc2")))
> ols.fit.hc3 = sqrt(diag(hccm(ols.fit, type = "hc3")))
> ols.fit.hc4 = sqrt(diag(hccm(ols.fit, type = "hc4")))
> ols.fit.coef = summary(ols.fit)$coef
> tvalues = ols.fit.coef[,1]/
+   cbind(ols.fit.coef[,2], ols.fit.hc0, ols.fit.hc1,
+         ols.fit.hc2, ols.fit.hc3, ols.fit.hc4)
> colnames(tvalues) = c("ols", "hc0", "hc1", "hc2", "hc3", "hc4")
> round(tvalues, 2)
```

	ols	hc0	hc1	hc2	hc3	hc4
(Intercept)	7.14	4.62	4.56	4.48	4.33	4.25
crim	-3.29	-3.78	-3.73	-3.48	-3.17	-2.58
zn	3.38	3.42	3.37	3.35	3.27	3.28
indus	0.33	0.41	0.41	0.41	0.40	0.40
chas1	3.12	2.11	2.08	2.05	2.00	2.00
nox	-4.65	-4.76	-4.69	-4.64	-4.53	-4.52
rm	9.12	4.57	4.51	4.43	4.28	4.18
age	0.05	0.04	0.04	0.04	0.04	0.04

```

dis          -7.40 -6.97 -6.87 -6.81 -6.66 -6.66
rad           4.61  5.05  4.98  4.91  4.76  4.65
tax          -3.28 -4.65 -4.58 -4.54 -4.43 -4.42
ptratio      -7.28 -8.23 -8.11 -8.06 -7.89 -7.93
b             3.47  3.53  3.48  3.44  3.34  3.30
lstat        -10.35 -5.34 -5.27 -5.18 -5.01 -4.93

```

The log transformation of the outcome does not remove the discrepancy among the standard errors. In this example, heteroskedasticity seems an important problem.

```

> ols.fit = lm(log(medv) ~ ., data = BostonHousing)
> summary(ols.fit)

```

Call:

```
lm(formula = log(medv) ~ ., data = BostonHousing)
```

Residuals:

```

      Min       1Q   Median       3Q      Max
-0.73361 -0.09747 -0.01657  0.09629  0.86435

```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.1020423   0.2042726  20.081 < 2e-16 ***
crim         -0.0102715   0.0013155  -7.808 3.52e-14 ***
zn           0.0011725   0.0005495   2.134 0.033349 *
indus        0.0024668   0.0024614   1.002 0.316755
chas1        0.1008876   0.0344859   2.925 0.003598 **
nox          -0.7783993   0.1528902  -5.091 5.07e-07 ***
rm           0.0908331   0.0167280   5.430 8.87e-08 ***
age          0.0002106   0.0005287   0.398 0.690567
dis          -0.0490873   0.0079834  -6.149 1.62e-09 ***
rad           0.0142673   0.0026556   5.373 1.20e-07 ***
tax          -0.0006258   0.0001505  -4.157 3.80e-05 ***
ptratio      -0.0382715   0.0052365  -7.309 1.10e-12 ***
b            0.0004136   0.0001075   3.847 0.000135 ***
lstat        -0.0290355   0.0020299 -14.304 < 2e-16 ***

```

```

Residual standard error: 0.1899 on 492 degrees of freedom
Multiple R-squared:  0.7896,    Adjusted R-squared:  0.7841
F-statistic: 142.1 on 13 and 492 DF,  p-value: < 2.2e-16

```

```

>
> ols.fit.hc0 = sqrt(diag(hccm(ols.fit, type = "hc0")))
> ols.fit.hc1 = sqrt(diag(hccm(ols.fit, type = "hc1")))
> ols.fit.hc2 = sqrt(diag(hccm(ols.fit, type = "hc2")))
> ols.fit.hc3 = sqrt(diag(hccm(ols.fit, type = "hc3")))
> ols.fit.hc4 = sqrt(diag(hccm(ols.fit, type = "hc4")))
> ols.fit.coef =summary(ols.fit)$coef
> tvalues = ols.fit.coef[,1]/
+   cbind(ols.fit.coef[,2], ols.fit.hc0, ols.fit.hc1,
+         ols.fit.hc2, ols.fit.hc3, ols.fit.hc4)
> colnames(tvalues) = c("ols", "hc0", "hc1", "hc2", "hc3", "hc4")
> round(tvalues, 2)

```

	ols	hc0	hc1	hc2	hc3	hc4
(Intercept)	20.08	14.29	14.09	13.86	13.43	13.13
crim	-7.81	-5.31	-5.24	-4.85	-4.39	-3.56
zn	2.13	2.68	2.64	2.62	2.56	2.56
indus	1.00	1.46	1.44	1.43	1.40	1.41
chas1	2.93	2.69	2.66	2.62	2.56	2.56
nox	-5.09	-4.79	-4.72	-4.67	-4.56	-4.54
rm	5.43	3.31	3.26	3.20	3.10	3.02
age	0.40	0.33	0.32	0.32	0.31	0.31
dis	-6.15	-6.12	-6.03	-5.98	-5.84	-5.82
rad	5.37	5.23	5.16	5.05	4.87	4.67
tax	-4.16	-5.05	-4.98	-4.90	-4.76	-4.69

ptratio	-7.31	-8.84	-8.72	-8.67	-8.51	-8.55
b	3.85	2.80	2.76	2.72	2.65	2.59
lstat	-14.30	-7.86	-7.75	-7.63	-7.40	-7.28

6.6 Final remarks

The beauty of the asymptotic analysis and the EHW standard error is that they hold under weak parametric assumptions on the error terms. We do not need to modify the OLS estimator but only need to modify the covariance estimator. However, this framework has limitations.

- (L1) The proofs are based on limiting theorems that require the sample size to be infinity. We are often unsure whether the sample size is large enough for a particular application we have.
 - (L2) The EHW standard errors can be severely biased and have large variability in finite samples. Problem 6.8 shows that the HC2 correction is unbiased for the true covariance matrix of $\hat{\beta}$ under the Gauss–Markov model. However, no such result exists for the heteroskedastic linear model.
 - (L3) Under the heteroskedastic linear model, the Gauss–Markov theorem does not hold, so the OLS can be inefficient. We will discuss possible improvements in Chapter 19.
 - (L4) Unlike Section 5.3, we cannot create any reasonable prediction intervals for a future observation y_{n+1} based on (X, Y, x_{n+1}) since its variance σ_{n+1}^2 is fundamentally unknown without further assumptions. Chapter 12.6 will discuss the problem of prediction under a modified statistical framework.
-

6.7 Homework problems

6.1 Testing linear hypotheses under heteroskedasticity

Under the heteroskedastic linear model, how do we test the hypotheses

$$H_0 : c^T \beta = 0,$$

for $c \in \mathbb{R}^p$, and

$$H_0 : C\beta = 0$$

for $C \in \mathbb{R}^{l \times p}$ with l linearly independent rows?

6.2 Two-sample problem continued

This problem extends Problem 5.5.

1. Assume that z_1, \dots, z_m are IID with mean μ_1 and variance σ_1^2 , and w_1, \dots, w_n are IID with mean μ_2 and variance σ_2^2 , and test $H_0 : \mu_1 = \mu_2$.

Prove that under H_0 , the following t statistic has an asymptotically Normal distribution:

$$t_{\text{unequal}} = \frac{\bar{z} - \bar{w}}{\sqrt{S_z^2/m + S_w^2/n}} \rightarrow N(0, 1)$$

in distribution.

Remark: The name “unequal” is motivated by the “var.equal” parameter of the R function `t.test`.

2. We can write the above problem as testing hypothesis $H_0 : \beta_1 = 0$ in the heteroskedastic linear regression. Based on the EHW standard error, we can compute the t statistic.

Prove that t_{unequal} is numerically identical to the t statistic based on the EHW robust standard error with the HC2 correction.

6.3 ANOVA with heteroskedasticity

This problem extends Problem 5.6.

Assume $y_i \mid i \in \mathcal{T}_j$ has mean β_j and variance σ_j^2 , which can be rewritten as a linear model without the Normality and homoskedasticity. In the process of solving Problem 5.6, you have derived the estimator of the covariance matrix of the OLS estimator under homoskedasticity. Find the HC0 and HC2 versions of the EHW covariance matrix. Which covariance matrices do you recommend, and why?

6.4 Invariance of the EHW covariance estimator

Theorem 6.5 below extends Theorem 3.3 in Problem 3.4. Prove Theorem 6.5.

Theorem 6.5 *If we transform X to $\tilde{X} = X\Gamma$ where Γ is a $p \times p$ non-degenerate matrix, the OLS fit changes from*

$$Y = X\hat{\beta} + \hat{\varepsilon}$$

to

$$Y = \tilde{X}\tilde{\beta} + \tilde{\varepsilon},$$

and the associated EHW covariance estimator changes from \hat{V}_{EHW} to \tilde{V}_{EHW} .

Then

$$\hat{V} = \Gamma\tilde{V}\Gamma^T,$$

and the above result holds for HCj ($j = 0, 1, 2, 3, 4$). The relationship also holds for the covariance estimator assuming homoskedasticity.

Remark: You can use the results in Problems 3.4 and 3.5.

6.5 Breakdown of the equivalence of the t -statistics based on the EHW standard error

This problem parallels Problem 5.10.

Consider data $(x_i, y_i)_{i=1}^n$, where both x_i and y_i are scalars. Run OLS fit of y_i on $(1, x_i)$ to obtain $t_{y|x}$, the t -statistic of the coefficient of x_i , based on the EHW standard error. Run OLS fit of x_i on $(1, y_i)$ to obtain $t_{x|y}$, the t -statistic of the coefficient of y_i , based on the EHW standard error.

Give a counterexample with $t_{y|x} \neq t_{x|y}$.

6.6 Empirical comparison of the standard errors

Long and Ervin (2000) reviewed and compared several commonly-used standard errors in OLS. Redo their simulation and replicate their Figures 1–4. They specified more details of their covariate generating process in a technical report (Long and Ervin, 1998).

6.7 Robust standard error in practice

King and Roberts (2015) gave three examples where the EHW standard errors differ from the OLS standard error. I have replicated one example in Section 6.5.2. Replicate another one using linear regression although the original analysis used Poisson regression. You can find the datasets used by King and Roberts (2015) at Harvard Dataverse (<https://dataverse.harvard.edu/>).

Remark: You may encounter the issue of $h_{ii} = 1$ for some observations. How would you deal with it? See Chapter 11 for discussions of h_{ii} .

6.8 Unbiased sandwich variance estimator under the Gauss–Markov model

Under the Gauss–Markov model with $\sigma_i^2 = \sigma^2$ for $i = 1, \dots, n$, show that the HC0 version of \hat{V}_{EHW} is biased but the HC2 version of \tilde{V}_{EHW} is unbiased for $\text{cov}(\hat{\beta})$.

Part III

Interpretation of Ordinary Least Squares Based on Partial Regressions



Frisch–Waugh–Lovell Theorem

The Frisch–Waugh–Lovell (FWL) Theorem is a powerful theorem about OLS. It allows us to reduce complicated, multi-dimensional OLS to simpler, often one-dimensional OLS. It helps to interpret the OLS coefficients. Moreover, it is a theoretical tool to derive many other results about OLS. This chapter introduces the FWL Theorem, and Chapter 8 will discuss its applications.

7.1 Long and short regressions

If we partition X and β as

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

where $X_1 \in \mathbb{R}^{n \times k}$, $X_2 \in \mathbb{R}^{n \times l}$, $\beta_1 \in \mathbb{R}^k$ and $\beta_2 \in \mathbb{R}^l$, then we can consider the *long regression*

$$\begin{aligned} Y &= X\hat{\beta} + \hat{\varepsilon} \\ &= \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} + \hat{\varepsilon} \\ &= X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}, \end{aligned}$$

and the *short regression*

$$Y = X_2\tilde{\beta}_2 + \tilde{\varepsilon},$$

where $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ and $\tilde{\beta}_2$ are the OLS coefficients, and $\hat{\varepsilon} = Y - X\hat{\beta}$ and $\tilde{\varepsilon} = Y - X_2\tilde{\beta}_2$ are the residual vectors from the long and short regressions, respectively. These two regressions are of great interest in practice. For example, we can ask the following questions:

- (Q1) if the true β_1 is zero, then what is the consequence of including X_1 in the long regression?
- (Q2) if the true β_1 is not zero, then what is the consequence of omitting X_1 in the short regression?
- (Q3) what is the difference between $\hat{\beta}_2$ and $\tilde{\beta}_2$? Both of them are measures of the “impact” of X_2 on Y . Then why are they different? Does their difference give us any information about β_1 ?

Many problems in statistics are related to the long and short regressions. We will discuss some applications in Chapter 8 and give a related result in Chapter 9.

7.2 FWL Theorem for the regression coefficients

Theorem 7.1 below helps to answer the questions in (Q1)–(Q3).

Theorem 7.1 (FWL Theorem) *The OLS estimator for β_2 in the short regression is $\tilde{\beta}_2 = (X_2^T X_2)^{-1} X_2^T Y$, and the OLS estimator for β_2 in the long regression has the following equivalent forms*

$$\hat{\beta}_2 = [(X^T X)^{-1} X^T Y]_{\text{last } l \text{ elements}} \quad (7.1)$$

$$= \{X_2^T (I_n - H_1) X_2\}^{-1} X_2^T (I_n - H_1) Y \quad \text{where } H_1 = X_1 (X_1^T X_1)^{-1} X_1^T \quad (7.2)$$

$$= (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y \quad \text{where } \tilde{X}_2 = (I_n - H_1) X_2 \quad (7.3)$$

$$= (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T \tilde{Y} \quad \text{where } \tilde{Y} = (I_n - H_1) Y. \quad (7.4)$$

Theorem 7.1 is often called the Frisch–Waugh–Lovell (FWL) Theorem in econometrics (Frisch and Waugh, 1933; Lovell, 1963), although its equivalent forms were also known in classic statistics.¹

Before proving Theorem 7.1, I will first discuss its meanings and interpretations. Equation (7.1) follows from the definition of the OLS coefficient. The matrix $I_n - H_1$ in equation (7.2) is the projection matrix onto the space orthogonal to the column space of X_1 . Equation (7.3) states that $\hat{\beta}_2$ equals the OLS coefficient of Y on $\tilde{X}_2 = (I_n - H_1) X_2$, which is the residual matrix from the column-wise OLS fit of X_2 on X_1 .² So $\hat{\beta}_2$ measures the “impact” of X_2 on Y after “adjusting” for the impact of X_1 , that is, it measures the partial or pure “impact” of X_2 on Y . Equation (7.4) is a slight modification of Equation (7.3), stating that $\hat{\beta}_2$ equals the OLS coefficient of \tilde{Y} on \tilde{X}_2 , where $\tilde{Y} = (I_n - H_1) Y$ is the residual vector from the OLS fit of Y on X_1 . From (7.3) and (7.4), it is not crucial to residualize Y , but it is crucial to residualize X_2 .

The forms (7.3) and (7.4) suggest the interpretation of $\hat{\beta}_2$ as the “impact” of X_2 on Y holding X_1 constant, or in an econometric term, the “impact” of X_2 on Y *ceteris paribus*. Marshall (1890) used the Latin phrase *ceteris paribus*. Its English meaning is “with other conditions remaining the same.” However, the algebraic meaning of the FWL Theorem is that the OLS coefficient of a variable equals the *partial regression* coefficient based on the residuals. Therefore, taking the Latin phrase too seriously may be problematic because Theorem 7.1 is a pure algebraic result without any distributional assumptions. We cannot hold X_1 constant using pure linear algebra. Sometimes, we can manipulate the value of X_1 in an experimental setting, but this relies on the assumption of the data-collecting process.

There are many ways to prove Theorem 7.1. Below I first take a detour to give an unnecessarily complicated proof because some intermediate steps will be useful for later parts of the book. I will then give a simpler proof, which requires a deep understanding of OLS as a linear projection.

The first proof relies on the following lemma.

Lemma 7.1 *The inverse of $X^T X$ is*

$$(X^T X)^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

¹Professor Alan Agresti gave me the reference of Yule (1907).

²See Problem 3.8 for more details.

where

$$\begin{aligned} S_{11} &= (X_1^T X_1)^{-1} + (X_1^T X_1)^{-1} X_1^T X_2 (\tilde{X}_2^T \tilde{X}_2)^{-1} X_2^T X_1 (X_1^T X_1)^{-1}, \\ S_{12} &= -(X_1^T X_1)^{-1} X_1^T X_2 (\tilde{X}_2^T \tilde{X}_2)^{-1}, \\ S_{21} &= S_{12}^T, \\ S_{22} &= (\tilde{X}_2^T \tilde{X}_2)^{-1}, \end{aligned}$$

with $\tilde{X}_2 = (I_n - H_1)X_2$ and $H_1 = X_1(X_1^T X_1)^{-1}X_1^T$ defined in Theorem 7.1.

I leave the proof of Lemma 7.1 as Problem 7.1. With Lemma 7.1, we can easily prove Theorem 7.1.

Proof of Theorem 7.1: (Version 1) The OLS coefficient is

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X^T X)^{-1} X^T Y = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} X_1^T Y \\ X_2^T Y \end{pmatrix}.$$

Then using Lemma 7.1, we can simplify $\hat{\beta}_2$ as

$$\begin{aligned} \hat{\beta}_2 &= S_{21} X_1^T Y + S_{22} X_2^T Y \\ &= -(\tilde{X}_2^T \tilde{X}_2)^{-1} X_2^T X_1 (X_1^T X_1)^{-1} X_1^T Y + (\tilde{X}_2^T \tilde{X}_2)^{-1} X_2^T Y \\ &= -(\tilde{X}_2^T \tilde{X}_2)^{-1} X_2^T H_1 Y + (\tilde{X}_2^T \tilde{X}_2)^{-1} X_2^T Y \\ &= (\tilde{X}_2^T \tilde{X}_2)^{-1} X_2^T (I_n - H_1) Y \end{aligned} \tag{7.5}$$

$$= (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y. \tag{7.6}$$

Equation (7.5) is the form (7.2), and Equation (7.6) is the form (7.3). Because we also have $X_2^T (I_n - H_1) Y = X_2^T (I_n - H_1)^2 Y = \tilde{X}_2^T \tilde{Y}$, we can write $\hat{\beta}_2$ as $\hat{\beta}_2 = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T \tilde{Y}$, giving the form (7.4). \square

The second proof does not invert the block matrix of $X^T X$ directly.

Proof of Theorem 7.1: (Version 2) First, the OLS decomposition $Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{\varepsilon}$ satisfies $X^T \hat{\varepsilon} = (X_1, X_2)^T \hat{\varepsilon} = 0$, which implies

$$X_1^T \hat{\varepsilon} = 0, \quad X_2^T \hat{\varepsilon} = 0.$$

Second, multiplying $I_n - H_1$ on both sides of the OLS decomposition $Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{\varepsilon}$, we have

$$(I_n - H_1)Y = (I_n - H_1)X_1 \hat{\beta}_1 + (I_n - H_1)X_2 \hat{\beta}_2 + (I_n - H_1)\hat{\varepsilon},$$

which reduces to

$$(I_n - H_1)Y = (I_n - H_1)X_2 \hat{\beta}_2 + \hat{\varepsilon} \tag{7.7}$$

because $(I_n - H_1)X_1 = 0$ and $(I_n - H_1)\hat{\varepsilon} = \hat{\varepsilon} - H_1 \hat{\varepsilon} = \hat{\varepsilon} - X_1(X_1^T X_1)^{-1}X_1^T \hat{\varepsilon} = \hat{\varepsilon}$.

Third, multiplying X_2^T on both sides of (7.7), we have

$$X_2^T (I_n - H_1)Y = X_2^T (I_n - H_1)X_2 \hat{\beta}_2$$

because $X_2^T \hat{\varepsilon} = 0$. The FWL Theorem follows immediately with $\hat{\beta}_2 = (X_2^T (I_n - H_1)X_2)^{-1} X_2^T (I_n - H_1)Y$.

A subtle issue in this proof is to verify that $X_2^T (I_n - H_1)X_2$ is invertible. It is relatively easy to show that matrix $X_2^T (I_n - H_1)X_2$ is positive semi-definite. To show it has rank l , we only need to show that

$$u_2^T X_2^T (I_n - H_1)X_2 u_2 = 0 \text{ implies } u_2 = 0.$$

We have $u_2^T X_2^T (I_n - H_1) X_2 u_2 = \|(I_n - H_1) X_2 u_2\|^2 = 0$, so $(I_n - H_1) X_2 u_2 = 0$, which further implies $X_2 u_2 \in \mathcal{C}(X_1)$ by Proposition 3.1. That is, $X_2 u_2 = X_1 u_1$ for some u_1 . So $X_1 u_1 - X_2 u_2 = 0$. Since the columns of X are linearly independent, we must have $u_1 = 0$ and $u_2 = 0$. \square

I will end this section with two byproducts of the FWL Theorem. First, \tilde{X}_2 is the residual matrix from the OLS fit of X_2 on X_1 . It is an $n \times l$ matrix with linearly independent columns as shown in the proof of Theorem 7.1 (Version 2) and induces a projection matrix

$$\tilde{H}_2 = \tilde{X}_2 (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T.$$

This projection matrix is closely related to the projection matrices induced by X and X_1 as shown in the following lemma.

Lemma 7.2 *We have*

$$H_1 \tilde{H}_2 = \tilde{H}_2 H_1 = 0 \quad (7.8)$$

and

$$H = H_1 + \tilde{H}_2. \quad (7.9)$$

Lemma 7.2 is purely algebraic. I leave the proof as Problem 7.3. The identities in (7.8) imply that the column space of \tilde{X}_2 is orthogonal to the column space of X_1 . The identity in (7.9) has a clear geometric interpretation. For any vector $v \in \mathbb{R}^n$, we have $Hv = H_1 v + \tilde{H}_2 v$, so the projection of v onto the column space of X equals the summation of the projection of v onto the column space of X_1 and the projection of v onto the column space of \tilde{X}_2 . Importantly, $H \neq H_1 + H_2$ in general.

Second, we can obtain $\hat{\beta}_2$ from (7.3) or (7.4), which corresponds to the partial regression of Y on \tilde{X}_2 or the partial regression of \tilde{Y} on \tilde{X}_2 . However, subtle issues arise with the residuals. Corollary 7.1 below states that the residual vector from the second partial regression equals the residual vector from the full regression. The conclusion does not hold if we only residualize X_2 . See Problem 7.2. Therefore, to ensure the residual vectors are the same, it is important to residualize both Y and X_2 .

Corollary 7.1 *We have $\hat{\varepsilon} = \hat{e}$, where $\hat{\varepsilon}$ is the residual vector from the OLS fit of Y on X and \hat{e} is the residual vector from the OLS fit of \tilde{Y} on \tilde{X}_2 , respectively.*

Proof of Corollary 7.1: We have $\hat{\varepsilon} = (I - H)Y$ and

$$\hat{e} = (I - \tilde{H}_2)\tilde{Y} = (I - \tilde{H}_2)(I - H_1)Y.$$

It suffices to show that $I - H = (I - \tilde{H}_2)(I - H_1)$, or, equivalently, $I - H = I - H_1 - \tilde{H}_2 + \tilde{H}_2 H_1$. This holds due to Lemma 7.2. \square

7.3 FWL Theorem for standard errors

Based on the OLS fit of Y on X , we have two estimated covariances for the second component $\hat{\beta}_2$: \hat{V} assuming homoskedasticity and \hat{V}_{EHW} allowing for heteroskedasticity.

The FWL Theorem demonstrates that we can obtain $\hat{\beta}_2$ from the OLS fit of \tilde{Y} on \tilde{X}_2 . Then based on this partial regression, we have two estimated covariances for $\hat{\beta}_2$: \tilde{V} assuming homoskedasticity and \tilde{V}_{EHW} allowing for heteroskedasticity.

Theorem 7.2 below establishes the equivalence between the estimated covariances from the long and partial regressions.

Theorem 7.2 $(n - k - l)\hat{V} = (n - l)\tilde{V}$ and $\hat{V}_{\text{EHW}} = \tilde{V}_{\text{EHW}}$.

Theorem 7.1 is well known for a long time but Theorem 7.2 is less well known. Theorem 7.2 is a numeric result that does not depend on the statistical assumptions. Lovell (1963) hinted at the first identity in Theorem 7.2, and Ding (2021a) proved Theorem 7.2. In Theorem 7.2, the equivalence of the EHW covariances only holds for the original version, and it breaks down for other modified versions discussed in Chapter 6.4.2.

Proof of Theorem 7.2: By Corollary 7.1, the full regression and partial regression have the same residual vector, denoted by $\hat{\varepsilon}$. Therefore, $\hat{\Omega}_{\text{EHW}} = \tilde{\Omega}_{\text{EHW}} = \text{diag}\{\hat{\varepsilon}^2\}$ in the EHW covariance estimators.

Based on the full regression, define $\hat{\sigma}^2 = \|\hat{\varepsilon}\|_2^2/(n - k - l)$. Then \hat{V} equals the $(2, 2)$ th block of $\hat{\sigma}^2(X^T X)^{-1}$, and \hat{V}_{EHW} equals the $(2, 2)$ th block of $(X^T X)^{-1} X^T \hat{\Omega}_{\text{EHW}} X (X^T X)^{-1}$.

Based on the partial regression, define $\tilde{\sigma}^2 = \|\hat{\varepsilon}\|_2^2/(n - l)$. Then $\tilde{V} = \tilde{\sigma}^2(\tilde{X}_2^T \tilde{X}_2)^{-1}$ and $\tilde{V}_{\text{EHW}} = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T \tilde{\Omega}_{\text{EHW}} \tilde{X}_2 (\tilde{X}_2^T \tilde{X}_2)^{-1}$.

Let $\hat{\sigma}^2 = \|\hat{\varepsilon}\|_2^2/(n - k - l)$ and $\tilde{\sigma}^2 = \|\hat{\varepsilon}\|_2^2/(n - l)$ be the common variance estimators. They are identical up to the degrees of freedom correction. Under homoskedasticity, the covariance estimator for $\hat{\beta}_2$ is the $(2, 2)$ th block of $\hat{\sigma}^2(X^T X)^{-1}$, that is, $\hat{\sigma}^2 S_{22} = \hat{\sigma}^2(\tilde{X}_2^T \tilde{X}_2)^{-1}$ by Lemma 7.1, which is identical to the covariance estimator for $\tilde{\beta}_2$ up to the degrees of freedom correction.

The EHW covariance estimator from the full regression is the $(2, 2)$ block of $\hat{A} \hat{\Omega}_{\text{EHW}} \hat{A}^T$, where

$$\begin{aligned} \hat{A} &= (X^T X)^{-1} X^T \\ &= \begin{pmatrix} * & * \\ -(\tilde{X}_2^T \tilde{X}_2)^{-1} X_2^T X_1 (X_1^T X_1)^{-1} & (\tilde{X}_2^T \tilde{X}_2)^{-1} \end{pmatrix} \begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix} \\ &= \begin{pmatrix} * \\ -(\tilde{X}_2^T \tilde{X}_2)^{-1} X_2^T H_1 + (\tilde{X}_2^T \tilde{X}_2)^{-1} X_2^T \end{pmatrix} \\ &= \begin{pmatrix} * \\ (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T \end{pmatrix}, \end{aligned}$$

by Lemma 7.1. I omit the $*$ terms because they do not affect the final calculation. Define $\tilde{A}_2 = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T$, and then

$$\hat{V}_{\text{EHW}} = \tilde{A}_2 \hat{\Omega}_{\text{EHW}} \tilde{A}_2^T = \tilde{A}_2 \tilde{\Omega}_{\text{EHW}} \tilde{A}_2^T,$$

which equals the EHW covariance estimator \tilde{V}_{EHW} from the partial regression. \square

7.4 Gram–Schmidt orthogonalization, QR decomposition, and computation of OLS

When the regressors are orthogonal, the coefficients from the long and short regressions are identical, which simplifies the calculation and theoretical discussion.

Corollary 7.2 *If $X_1^T X_2 = 0$, i.e., the columns of X_1 and X_2 are orthogonal, then $\tilde{X}_2 = X_2$ and $\hat{\beta}_2 = \tilde{\beta}_2$.*

Proof of Corollary 7.2: We can directly prove Corollary 7.2 by verifying that $X^T X$ is block diagonal.

Alternatively, Corollary 7.2 follows from

$$\tilde{X}_2 = (I_n - H_1)X_2 = X_2 - X_1(X_1^T X_1)^{-1}X_1^T X_2 = X_2,$$

and Theorem 7.1. \square

The simple fact of Corollary 7.2 motivates us to orthogonalize the columns of the covariate matrix X , which in turn gives the famous QR decomposition in linear algebra. Interestingly, the `lm` function in `R` uses the QR decomposition to compute the OLS estimator. To facilitate the discussion, I will use the notation

$$\hat{\beta}_{V_2|V_1} V_1$$

as the linear projection of the vector $V_2 \in \mathbb{R}^n$ onto the vector $V_1 \in \mathbb{R}^n$, where $\hat{\beta}_{V_2|V_1} = V_2^T V_1 / V_1^T V_1$. This is from the univariate OLS of V_2 on V_1 (recall Chapter 2.2).

With a slight abuse of notation, partition the covariate matrix into column vectors $X = (X_1, \dots, X_p)$. The goal is to find orthogonal vectors (U_1, \dots, U_p) that generate the same column space as X . Start with

$$X_1 = U_1.$$

Regress X_2 on U_1 to obtain the fitted and residual vector

$$X_2 = \hat{\beta}_{X_2|U_1} U_1 + U_2;$$

by OLS, U_1 and U_2 must be orthogonal. Regress X_3 on (U_1, U_2) to obtain the fitted and residual vector

$$X_3 = \hat{\beta}_{X_3|U_1} U_1 + \hat{\beta}_{X_3|U_2} U_2 + U_3;$$

by Corollary 7.2, the OLS reduces to two univariate OLS by $U_1 \perp U_2$ and ensures that U_3 is orthogonal to both U_1 and U_2 . This justifies the notation $\hat{\beta}_{X_3|U_1}$ and $\hat{\beta}_{X_3|U_2}$. Continue this procedure to the last column vector:

$$X_p = \sum_{j=1}^{p-1} \hat{\beta}_{X_p|U_j} U_j + U_p;$$

by OLS, U_p is orthogonal to all U_j ($j = 1, \dots, p-1$). We further normalize the U vectors to have unit length:

$$Q_j = U_j / \|U_j\|, \quad (j = 1, \dots, p).$$

The whole process is called the Gram–Schmidt orthogonalization, which is essentially the sequential OLS fits. This process generates an $n \times p$ matrix with orthonormal column vectors

$$Q = (Q_1, \dots, Q_p).$$

More interestingly, the column vectors of X and Q can linearly represent each other because

$$\begin{aligned} X &= (X_1, \dots, X_p) \\ &= (U_1, \dots, U_p) \begin{pmatrix} 1 & \hat{\beta}_{X_2|U_1} & \hat{\beta}_{X_3|U_1} & \cdots & \hat{\beta}_{X_p|U_1} \\ 0 & 1 & \hat{\beta}_{X_3|U_2} & \cdots & \hat{\beta}_{X_p|U_2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \\ &= Q \text{diag}\{\|U_j\|\}_{j=1}^p \begin{pmatrix} 1 & \hat{\beta}_{X_2|U_1} & \hat{\beta}_{X_3|U_1} & \cdots & \hat{\beta}_{X_p|U_1} \\ 0 & 1 & \hat{\beta}_{X_3|U_2} & \cdots & \hat{\beta}_{X_p|U_2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}. \end{aligned}$$

We can verify that the product of the second and the third matrix is an upper triangular matrix, denoted by R . By definition, the j th diagonal element of R equals $\|U_j\|$, and the (j, j') th element of R equals $\|U_j\|\hat{\beta}_{X_{j'}|U_j}$ for $j' > j$. Therefore, we can decompose X as

$$X = QR$$

where Q is an $n \times p$ matrix with orthonormal columns and R is a $p \times p$ upper triangular matrix. This is called the QR decomposition of X .

Most software packages, for example, R, do not calculate the inverse of $X^T X$ directly. Instead, they first find the QR decomposition of $X = QR$. Since the Normal equation simplifies to

$$\begin{aligned} X^T X \hat{\beta} &= X^T Y, \\ R^T Q^T Q R \hat{\beta} &= R^T Q^T Y, \\ R \hat{\beta} &= Q^T Y, \end{aligned}$$

they then backsolve the last linear equation since R is upper triangular.

In R, the `qr` function returns the QR decomposition of a matrix.

```
> X = matrix(rnorm(7*3), 7, 3)
> X
      [,1]      [,2]      [,3]
[1,] -0.57231223  0.1196325  0.8087505
[2,] -1.76090225  1.0627631  1.8170361
[3,] -0.04144281 -0.2904749 -1.8372247
[4,] -0.37627821  0.4476932 -0.9629320
[5,] -1.40848027  0.2735408 -0.8047917
[6,]  1.84878518  0.7290005  1.2688929
[7,]  0.06432856  0.2256284  0.3972229
> qrX = qr(X)
> qr.Q(qrX)
      [,1]      [,2]      [,3]
[1,] -0.19100878 -0.03460617  0.30340481
[2,] -0.58769981 -0.60442928  0.23753900
[3,] -0.01383151  0.21191991 -0.55839928
[4,] -0.12558257 -0.28728403 -0.62864750
[5,] -0.47007924 -0.07020076 -0.36640938
[6,]  0.61703067 -0.68778411 -0.09999859
[7,]  0.02146961 -0.16748246  0.01605493
> qr.R(qrX)
      [,1]      [,2]      [,3]
[1,]  2.996261 -0.3735673  0.0937788
[2,]  0.000000 -1.3950642 -2.1217223
[3,]  0.000000  0.0000000  2.4826186
```

If we specify `qr = TRUE` in the `lm` function, it will also return the QR decomposition of the covariate matrix.³

```
> Y = rnorm(7)
> lmfit = lm(Y ~ 0 + X, qr = TRUE)
> qr.Q(lmfit$qr)
      [,1]      [,2]      [,3]
[1,] -0.43535054 -0.25679823 -0.65480400
[2,] -0.47091275 -0.13639459  0.14746444
[3,]  0.66494532  0.07725435 -0.39436265
[4,]  0.21136347 -0.78814737  0.34611820
[5,] -0.04493356 -0.40413829 -0.01156273
[6,] -0.28046504  0.10655755 -0.16193251
```

³The “0 +” in the code below forces the OLS to exclude the constant term.

```
[7,] 0.14561808 -0.33708219 -0.49780561
> qr.R(lmfit$qr)
      X1      X2      X3
1 3.190035 -0.6964269 1.8693260
2 0.000000 2.0719787 1.9210212
3 0.000000 0.0000000 -0.9261921
```

7.5 Homework problems

7.1 Inverse of a block matrix

Prove Lemma 7.1 and the following alternative form:

$$(X^T X)^{-1} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix},$$

where $H_2 = X_2(X_2^T X_2)^{-1} X_2^T$, $\tilde{X}_1 = (I_n - H_2)X_1$, and

$$\begin{aligned} Q_{11} &= (\tilde{X}_1^T \tilde{X}_1)^{-1}, \\ Q_{12} &= -(\tilde{X}_1^T \tilde{X}_1)^{-1} \tilde{X}_1^T X_2 (X_2^T X_2)^{-1}, \\ Q_{21} &= Q_{12}^T, \\ Q_{22} &= (X_2^T X_2)^{-1} + (X_2^T X_2)^{-1} X_2^T X_1 (\tilde{X}_1^T \tilde{X}_1)^{-1} X_1^T X_2 (X_2^T X_2)^{-1}. \end{aligned}$$

Remark: Use the formula in Problem A.3.

7.2 Residuals in the FWL Theorem

Give an example in which the residual vector from the partial regression of Y on \tilde{X}_2 does not equal to the residual vector from the full regression.

7.3 Projection matrices

Prove Lemma 7.2.

Remark: Use Lemma 7.1.

7.4 FWL Theorem and leverage scores

Consider the partitioned regression $Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{\varepsilon}$. To obtain the coefficient $\hat{\beta}_2$, we can run two OLS fits:

(R1) regress X_2 on X_1 to obtain the residual \tilde{X}_2 ;

(R2) regress Y on \tilde{X}_2 to obtain the coefficient, which equals $\hat{\beta}_2$ by the FWL Theorem.

Although partial regression (R2) can recover the OLS coefficient, the leverage scores from (R2) are not the same as those from the long regression. Prove that the summation of the corresponding leverage scores from (R1) and (R2) equals the leverage scores from the long regression.

Remark: The leverage scores are the diagonal elements of the hat matrix from OLS fits. Chapter 6 before mentioned them and Chapter 11 later will discuss them in more detail.

7.5 Another invariance property of the OLS coefficient

Partition the covariate matrix as $X = (X_1, X_2)$ where $X_1 \in \mathbb{R}^{n \times k}$ and $X_2 \in \mathbb{R}^{n \times l}$. Given any $A \in \mathbb{R}^{k \times l}$, define $\tilde{X}_2 = X_2 - X_1 A$. Fit two OLS:

$$Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{\varepsilon}$$

and

$$Y = X_1 \tilde{\beta}_1 + \tilde{X}_2 \tilde{\beta}_2 + \tilde{\varepsilon}.$$

Prove that

$$\hat{\beta}_2 = \tilde{\beta}_2, \quad \hat{\varepsilon} = \tilde{\varepsilon}.$$

Remark: You can use the result in Problem 3.4 to prove the result in this problem. As a special case, if we choose $A = (X_1^T X_1)^{-1} X_1^T X_2$ to be the coefficient matrix of the OLS fit of X_2 on X_1 , then the above result ensures that $\hat{\beta}_2$ from the OLS fit of Y on X_1 and X_2 equals $\tilde{\beta}_2$ from the OLS fit of Y on X_1 and $(I_n - H_1)X_2$, which is coherent with the FWL Theorem since $X_1^T(I_n - H_1)X_2 = 0$.

7.6 Alternative formula for the EHW standard error

Consider the partition regression $Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{\varepsilon}$ with X_1 is an $n \times (p-1)$ matrix and X_2 is an n dimensional vector. So $\hat{\beta}_2$ is a scalar, and the (p, p) th element of \hat{V}_{EHW} equals $\hat{\text{se}}_{\text{EHW}, 2}^2$, the squared EHW standard error for $\hat{\beta}_2$.

Define

$$\tilde{X}_2 = (I_n - H_1)X_2 = \begin{pmatrix} \tilde{x}_{12} \\ \vdots \\ \tilde{x}_{n2} \end{pmatrix}.$$

Prove that under Assumption 6.1, we have

$$\begin{aligned} \text{var}(\hat{\beta}_2) &= \sum_{i=1}^n w_i \sigma_i^2, \\ \hat{\text{se}}_{\text{EHW}, 2}^2 &= \sum_{i=1}^n w_i \hat{\varepsilon}_i^2 \end{aligned}$$

where

$$w_i = \frac{\tilde{x}_{i2}^2}{(\sum_{i=1}^n \tilde{x}_{i2}^2)^2}.$$

Remark: You can use Theorems 7.1 and 7.2 to prove the result. The original formula of the EHW covariance matrix has a complex form. However, using the FWL theorems, we can simplify each of the squared EHW standard errors as a weighted average of the squared residuals, or, equivalently, a simple quadratic form of the residual vector.

7.7 A counterexample to the Gauss–Markov Theorem

The Gauss–Markov Theorem does not hold under the heteroskedastic linear model. This problem gives a counterexample in a simple linear model.

Assume $y_i = \beta x_i + \varepsilon_i$ without the intercept and with potentially different $\text{var}(\varepsilon_i) = \sigma_i^2$ across $i = 1, \dots, n$. Consider two OLS estimators: the first OLS estimator does not contain the intercept $\hat{\beta} = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2$; the second OLS estimator contains the intercept $\tilde{\beta} = \sum_{i=1}^n (x_i - \bar{x}) y_i / \sum_{i=1}^n (x_i - \bar{x})^2$ even though the true linear model does not contain the intercept.

The Gauss-Markov Theorem ensures that if $\sigma_i^2 = \sigma^2$ for all i 's, then the variance of $\hat{\beta}$ is smaller than or equal to the variance of $\tilde{\beta}$. However, it does not hold when σ_i^2 's vary.

Give a counterexample in which the variance of $\hat{\beta}$ is larger than the variance of $\tilde{\beta}$.

7.8 QR decomposition of X and the computation of OLS

Verify that the R matrix equals

$$R = \begin{pmatrix} Q_1^T X_1 & Q_1^T X_2 & \cdots & Q_1^T X_p \\ 0 & Q_2^T X_2 & \cdots & Q_2^T X_p \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_p^T X_p \end{pmatrix}.$$

Based on the QR decomposition of X , prove that

$$H = QQ^T,$$

and the its (i, i) the diagonal element h_{ii} equals the squared length of the i -th row of Q .

7.9 Uniqueness of the QR decomposition

Prove that if X has linearly independent column vectors, the QR decomposition must be unique. That is, if $X = QR = Q_1 R_1$ where Q and Q_1 have orthonormal columns and R and R_1 are upper triangular, then we must have

$$Q = Q_1, \quad R = R_1.$$

Applications of the Frisch–Waugh–Lovell Theorem

The Frisch–Waugh–Lovell (FWL) theorem has many applications. I will highlight some of them in this chapter.

8.1 Centering regressors

8.1.1 Intercept and centering regressors

As a special case, partition the covariate matrix into $X = (X_1, X_2)$ with $X_1 = 1_n$. This is the usual case including the constant as the first regressor. The projection matrix

$$H_1 = 1_n(1_n^T 1_n)^{-1} 1_n^T = n^{-1} 1_n 1_n^T = \begin{pmatrix} n^{-1} & \cdots & n^{-1} \\ \vdots & & \vdots \\ n^{-1} & \cdots & n^{-1} \end{pmatrix} \equiv A_n$$

contains n^{-1} 's as its elements, and

$$C_n = I_n - n^{-1} 1_n 1_n^T$$

is the projection matrix orthogonal to 1_n . The matrices A_n and C_n have convenient properties:

(P1) Multiplying any vector by A_n is equivalent to obtaining the average of its components.

(P2) Multiplying any vector by C_n is equivalent to centering that vector.

For example,

$$A_n Y = \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} = \bar{y} 1_n,$$

and

$$C_n Y = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}.$$

More generally, multiplying any matrix by A_n is equivalent to averaging each column, and multiplying any matrix by C_n is equivalent to centering each column of that matrix. For

example,

$$A_n X_2 = \begin{pmatrix} \bar{x}_2^T \\ \vdots \\ \bar{x}_2^T \end{pmatrix} = 1_n \bar{x}_2^T,$$

and

$$C_n X_2 = \begin{pmatrix} x_{12}^T - \bar{x}_2^T \\ \vdots \\ x_{n2}^T - \bar{x}_2^T \end{pmatrix},$$

where X_2 contains row vectors $x_{12}^T, \dots, x_{n2}^T$ with average $\bar{x}_2 = n^{-1} \sum_{i=1}^n x_{i2}$. The FWL Theorem implies that the coefficient of X_2 in the OLS fit of Y on $(1_n, X_2)$ equals the coefficient of $C_n X_2$ in the OLS fit of $C_n Y$ on $C_n X_2$, that is, the OLS fit of the centered response vector on the column-wise centered X_2 . An immediate consequence is that if each column is centered in the design matrix, then to obtain the OLS coefficients, it does not matter whether to include the column 1_n or not.

The centering matrix C_n has another property:

(P3) The quadratic form of C_n equals the sample variance multiplied by $n - 1$.

For example,

$$\begin{aligned} Y^T C_n Y &= Y^T C_n^T C_n Y \\ &= (y_1 - \bar{y}, \dots, y_n - \bar{y}) \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= (n - 1) \hat{\sigma}_y^2, \end{aligned}$$

where $\hat{\sigma}_y^2$ is the sample variance of the outcomes. For an $n \times p$ matrix X ,

$$\begin{aligned} X^T C_n X &= \begin{pmatrix} X_1^T \\ \vdots \\ X_p^T \end{pmatrix} C_n \begin{pmatrix} X_1 & \cdots & X_p \end{pmatrix} \\ &= \begin{pmatrix} X_1^T C_n X_1 & \cdots & X_1^T C_n X_p \\ \vdots & & \vdots \\ X_p^T C_n X_1 & \cdots & X_p^T C_n X_p \end{pmatrix} \\ &= (n - 1) \begin{pmatrix} \hat{\sigma}_{11} & \cdots & \hat{\sigma}_{1p} \\ \vdots & & \vdots \\ \hat{\sigma}_{p1} & \cdots & \hat{\sigma}_{pp} \end{pmatrix}, \end{aligned}$$

where

$$\hat{\sigma}_{j_1 j_2} = (n - 1)^{-1} \sum_{i=1}^n (x_{ij_1} - \bar{x}_{\cdot j_1})(x_{ij_2} - \bar{x}_{\cdot j_2})$$

is the sample covariance between X_{j_1} and X_{j_2} . So $(n - 1)^{-1} X^T C_n X$ equals the sample covariance matrix of X . For these reason, I choose the notation C_n with “C” for both “centering” and “covariance.”

8.1.2 Dummy variables and centering regressors within groups

As another important special case, X_1 contains the dummies for a discrete variable, for example, the indicators for different treatment levels or groups. See Example 3.2 for the background. With k groups, X_1 can take the following two forms:

$$X_1 = \begin{pmatrix} 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & & 1 \\ 1 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}_{n \times k} \quad \text{or} \quad X_1 = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{n \times k}, \quad (8.1)$$

where the first form of X_1 contains 1_n and $k - 1$ dummy variables, and the second form of X_1 contains k dummy variables. In both forms of X_1 , the observations are sorted according to the group indicators. If we regress Y on X_1 , the residual vector is

$$Y - \begin{pmatrix} \bar{y}_{[1]} \\ \vdots \\ \bar{y}_{[1]} \\ \vdots \\ \bar{y}_{[k]} \\ \vdots \\ \bar{y}_{[k]} \end{pmatrix}, \quad (8.2)$$

where $\bar{y}_{[1]}, \dots, \bar{y}_{[k]}$ are the averages of the outcomes within groups $1, \dots, k$. Effectively, we center Y by group-specific means. Similarly, if we regress X_2 on X_1 , we center each column of X_2 by the group-specific means. Let Y^c and X_2^c be the centered response vector and design matrix. The FWL Theorem implies that the OLS coefficient of X_2 in the long regression of Y on (X_1, X_2) equals the OLS coefficient of X_2^c in the partial regression of Y^c on X_2^c . When k is large, running the OLS with centered variables can reduce the computational cost. See Problem 8.4 for an application in econometrics.

8.2 Partial correlation coefficient and Simpson's paradox

The sample Pearson correlation coefficient between n observations of two scalars $(x_i, y_i)_{i=1}^n$,

$$\hat{\rho}_{yx} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

measures the linear relationship between x and y . How do we measure the linear relationship between x and y after controlling for some other variables $w \in \mathbb{R}^{k-1}$? Intuitively, we can

measure it with the sample Pearson correlation coefficient based on the residuals from the following two OLS fits:

- (R1) run OLS of Y on $(1, W)$ and obtain residual vector $\hat{\varepsilon}_y$ and residual sum of squares RSS_y ;
- (R2) run OLS of X on $(1, W)$ and obtain residual vector $\hat{\varepsilon}_x$ and residual sum of squares RSS_x .

With $\hat{\varepsilon}_y$ and $\hat{\varepsilon}_x$, we can define the sample partial correlation coefficient between x and y given w as

$$\hat{\rho}_{yx|w} = \frac{\sum_{i=1}^n \hat{\varepsilon}_{x,i} \hat{\varepsilon}_{y,i}}{\sqrt{\sum_{i=1}^n \hat{\varepsilon}_{x,i}^2} \sqrt{\sum_{i=1}^n \hat{\varepsilon}_{y,i}^2}}. \quad (8.3)$$

In (8.3), we do not center the residuals because they have zero sample means due to the inclusions of the intercepts in the OLS fits (RR1) and (RR2). The sample partial correlation coefficient determines the coefficient of $\hat{\varepsilon}_x$ in the OLS fit of $\hat{\varepsilon}_y$ on $\hat{\varepsilon}_x$:

$$\hat{\beta}_{yx|w} = \frac{\sum_{i=1}^n \hat{\varepsilon}_{x,i} \hat{\varepsilon}_{y,i}}{\sum_{i=1}^n \hat{\varepsilon}_{x,i}^2} = \hat{\rho}_{yx|w} \sqrt{\frac{\sum_{i=1}^n \hat{\varepsilon}_{y,i}^2}{\sum_{i=1}^n \hat{\varepsilon}_{x,i}^2}} = \hat{\rho}_{yx|w} \frac{\hat{\sigma}_{y|w}}{\hat{\sigma}_{x|w}}, \quad (8.4)$$

where $\hat{\sigma}_{y|w}^2 = \text{RSS}_y/(n-k)$ and $\hat{\sigma}_{x|w}^2 = \text{RSS}_x/(n-k)$ are the variance estimators based on regressions (RR1) and (RR2) motivated by the Gauss–Markov model. Based on the FWL Theorem, $\hat{\beta}_{yx|w}$ equals the OLS coefficient of X in the long regression of Y on $(1, X, W)$. Therefore, (8.4) is the Galtonian formula for multiple regression, which is analogous to that for univariate regression (1.1).

To investigate the relationship between y and x , different researchers may run different regressions. One may run OLS of Y on $(1, X, W)$, and the other may run OLS of Y on $(1, X, W')$, where W' is a subset of W . Let $\hat{\beta}_{yx|w}$ be the coefficient of X in the first regression, and let $\hat{\beta}_{yx|w'}$ be the coefficient of X in the second regression. Mathematically, it is possible that these two coefficients have different signs, which is called *Simpson's paradox*.¹ It is a paradox because we expect both coefficients to measure the “impact” of X on Y . Because these two coefficients have the same signs as the partial correlation coefficients $\hat{\rho}_{yx|w}$ and $\hat{\rho}_{yx|w'}$, Simpson's paradox is equivalent to

$$\hat{\rho}_{yx|w} \hat{\rho}_{yx|w'} < 0.$$

To simplify the presentation, we discuss the special case with w being a scalar and w' being an empty set. Simpson's paradox is then equivalent to

$$\hat{\rho}_{yx|w} \hat{\rho}_{yx} < 0.$$

Theorem 8.1 below gives an expression that links $\hat{\rho}_{yx|w}$ and $\hat{\rho}_{yx}$.

Theorem 8.1 For $Y, X, W \in \mathbb{R}^n$, we have

$$\hat{\rho}_{yx|w} = \frac{\hat{\rho}_{yx} - \hat{\rho}_{yw} \hat{\rho}_{xw}}{\sqrt{1 - \hat{\rho}_{yw}^2} \sqrt{1 - \hat{\rho}_{xw}^2}}.$$

The proof of Theorem 8.1 is purely algebraic, so I leave it as Problem 8.8. Theorem 8.1 states that we can obtain the sample partial correlation coefficient based on the three pairwise correlation coefficients. Figure 8.1 illustrates the interplay among three variables. In

¹The usual form of Simpson's paradox is in terms of a $2 \times 2 \times 2$ table with all binary variables. Here we focus on its continuous version.

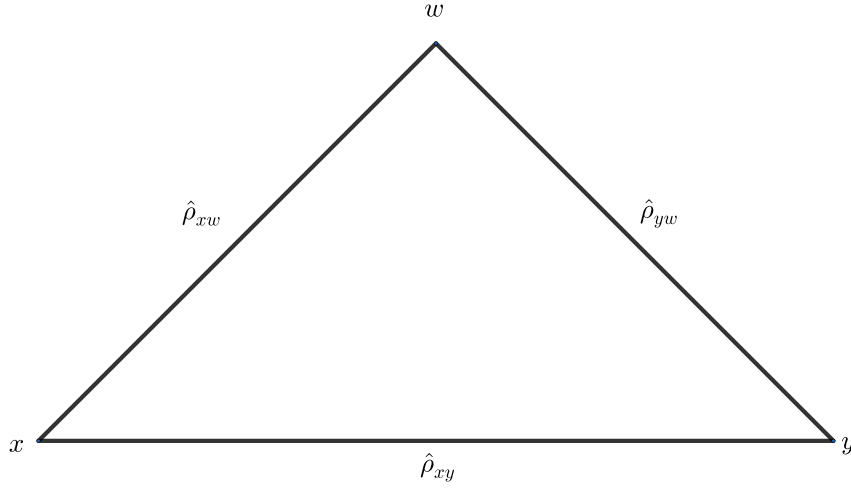


FIGURE 8.1: Correlations among three variables

particular, the correlation between x and y is due to two “pathways”: the one acting through w and the one acting independent of w . The first path way is related to the product term $\hat{\rho}_{yw}\hat{\rho}_{xw}$, and the second pathway is related to $\hat{\rho}_{yx|w}$. This gives some intuition for Theorem 8.1.

Based on data $(y_i, x_i, w_i)_{i=1}^n$, we can compute the sample correlation matrix

$$\begin{pmatrix} 1 & \hat{\rho}_{yx} & \hat{\rho}_{yw} \\ \hat{\rho}_{xy} & 1 & \hat{\rho}_{xw} \\ \hat{\rho}_{wy} & \hat{\rho}_{wx} & 1 \end{pmatrix},$$

which is symmetric and positive semi-definite. Simpson's paradox happens if and only if

$$\hat{\rho}_{yx}(\hat{\rho}_{yx} - \hat{\rho}_{yw}\hat{\rho}_{xw}) < 0,$$

which is equivalent to

$$\hat{\rho}_{yx}^2 < \hat{\rho}_{yx}\hat{\rho}_{yw}\hat{\rho}_{xw}.$$

We can observe Simpson's Paradox in the following simulation.

```
> n = 1000
> w = rbinom(n, 1, 0.5)
> x1 = rnorm(n, -1, 1)
> x0 = rnorm(n, 2, 1)
> x = ifelse(w, x1, x0)
> y = x + 6*w + rnorm(n)
> fit.xw = lm(y ~ x + w)$coef
> fit.x = lm(y ~ x)$coef
> fit.xw
(Intercept)          x          w
 0.05655442  0.97969907  5.92517072
> fit.x
(Intercept)          x
 3.6422978  -0.3743368
```

Because w is binary, we can plot (x, y) in each group of $w = 1$ and $w = 0$ in Figure 8.2. In both groups, y and x are positively associated with positive regression coefficients; but in the pooled data, y and x are negatively associated with a negative regression coefficient.

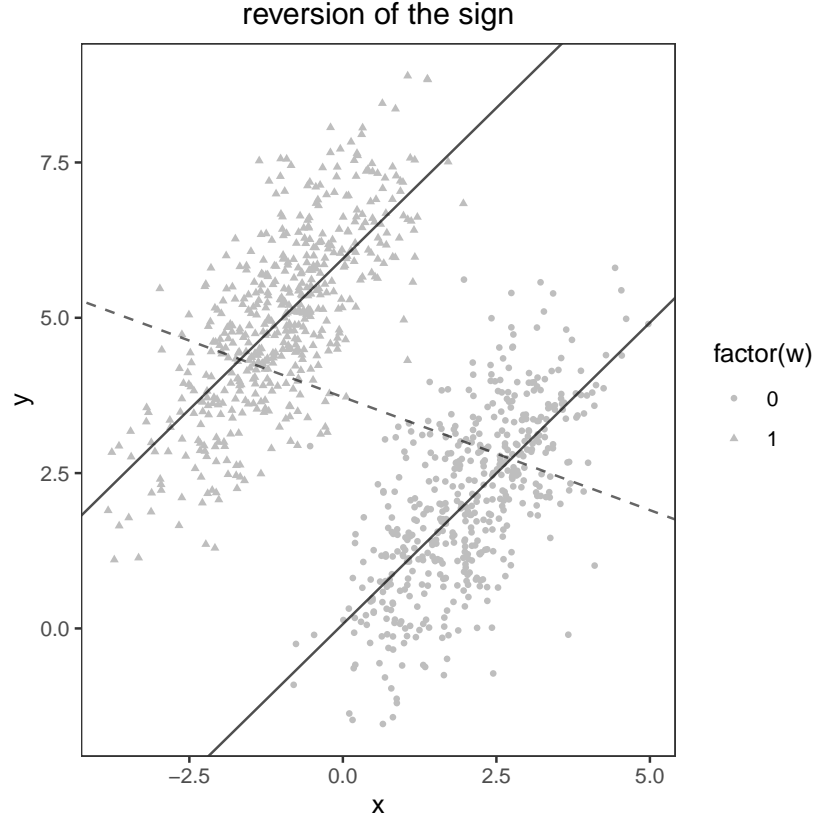


FIGURE 8.2: An Example of Simpson's Paradox. The two solid regression lines are fitted separately using the data from two groups, and the dash regression line is fitted using the pooled data.

8.3 Hypothesis testing and analysis of variance

Partition X and β as

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

where $X_1 \in \mathbb{R}^{n \times k}$, $X_2 \in \mathbb{R}^{n \times l}$, $\beta_1 \in \mathbb{R}^k$ and $\beta_2 \in \mathbb{R}^l$. We are often interested in testing

$$H_0 : \beta_2 = 0$$

in the long regression

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon, \quad (8.5)$$

where $\varepsilon \sim N(0, \sigma^2 I_n)$. If H_0 holds, then X_2 is redundant and a short regression suffices:

$$Y = X_1\beta + \varepsilon. \quad (8.6)$$

This is a special case of testing $C\beta = 0$ with

$$C = \begin{pmatrix} 0_{l \times k} & I_{l \times l} \end{pmatrix}.$$

As discussed in Chapter 5, we can use

$$\hat{\beta}_2 \sim N(0, \sigma^2 S_{22})$$

with $S_{22} = (\tilde{X}_2^T \tilde{X}_2)^{-1}$ being the $(2, 2)$ th block of $(X^T X)^{-1}$ by Lemma 7.1, to construct the Wald-type statistic for hypothesis testing:

$$\begin{aligned} F_{\text{Wald}} &= \frac{\hat{\beta}_2^T (S_{22})^{-1} \hat{\beta}_2}{l \hat{\sigma}^2} \\ &= \frac{\hat{\beta}_2^T \tilde{X}_2^T \tilde{X}_2 \hat{\beta}_2}{l \hat{\sigma}^2} \\ &\sim F_{l, n-p}. \end{aligned}$$

Now I will discuss testing H_0 from an alternative perspective based on comparing the residual sum of squares in the long regression (8.5) and the short regression (8.6). This technique is called the analysis of variance (ANOVA), pioneered by R. A. Fisher in the design and analysis of experiments. Intuitively, if $\beta_2 = 0$, then the residual vectors from the long regression (8.5) and the short regression (8.6) should not be “too different.” However, with the error term ε , these residuals are random, then the key is to quantify the magnitude of the difference. Define

$$\text{RSS}_{\text{long}} = Y^T (I_n - H) Y$$

and

$$\text{RSS}_{\text{short}} = Y^T (I_n - H_1) Y$$

as the residual sums of squares from the long and short regressions, respectively. By the definition of OLS, it must be true that

$$\text{RSS}_{\text{long}} \leq \text{RSS}_{\text{short}}$$

and therefore,

$$\text{RSS}_{\text{short}} - \text{RSS}_{\text{long}} = Y^T (H - H_1) Y \geq 0. \quad (8.7)$$

To understand the magnitude of the change in the residual sum of squares, we can standardize the above difference and define

$$F_{\text{ANOVA}} = \frac{(\text{RSS}_{\text{short}} - \text{RSS}_{\text{long}})/l}{\text{RSS}_{\text{long}}/(n-p)},$$

In the definition of the above statistic, l and $n-p$ are the degrees of freedom to make the mathematics more elegant, but they do not change the discussion fundamentally. The denominator of F_{ANOVA} is $\hat{\sigma}^2$, so we can also write it as

$$F_{\text{ANOVA}} = \frac{\text{RSS}_{\text{short}} - \text{RSS}_{\text{long}}}{l \hat{\sigma}^2}. \quad (8.8)$$

Theorem 8.2 below states that these two perspectives yield an identical test statistic.

Theorem 8.2 *Under Assumption 5.1, if $\beta_2 = 0$, then $F_{\text{ANOVA}} \sim F_{l, n-p}$. In fact, $F_{\text{ANOVA}} = F_{\text{Wald}}$, which is a numerical result without Assumption 5.1.*

I divide the proof into two parts. The first part derives the exact distribution of F_{ANOVA} under the Normal linear model. It relies on the following lemma on the basic properties of the projection matrices. I relegate its proof to Problem 8.11.

Lemma 8.1 *We have*

$$HX_1 = X_1, \quad HX_2 = X_2, \quad HH_1 = H_1, \quad H_1H = H_1$$

Moreover, $H - H_1$ is a projection matrix of rank $p - k = l$, $I_n - H$ is a projection matrix of rank $n - p$, and they are orthogonal:

$$(H - H_1)(I_n - H) = 0. \quad (8.9)$$

Proof of Theorem 8.2 (Part I):

The residual vector from the long regression equals

$$\hat{\varepsilon} = (I_n - H)Y = (I_n - H)(X\beta + \varepsilon) = (I_n - H)\varepsilon,$$

so the residual sum of squares equals

$$\text{RSS}_{\text{long}} = \hat{\varepsilon}^T \hat{\varepsilon} = \varepsilon^T (I_n - H) \varepsilon.$$

Since $\beta_2 = 0$, the residual vector from the short regression equals

$$\tilde{\varepsilon} = (I_n - H_1)Y = (I_n - H_1)(X_1\beta_1 + \varepsilon) = (I_n - H_1)\varepsilon,$$

so the residual sum of squares equals

$$\text{RSS}_{\text{short}} = \tilde{\varepsilon}^T \tilde{\varepsilon} = \varepsilon^T (I_n - H_1) \varepsilon.$$

Let $\varepsilon_0 = \varepsilon/\sigma \sim N(0, I_n)$ be a standard Normal random vector, then we can write F_{ANOVA} as

$$\begin{aligned} F_{\text{ANOVA}} &= \frac{\varepsilon^T (H - H_1) \varepsilon / l}{\varepsilon^T (I_n - H) \varepsilon / (n - p)} \\ &= \frac{\varepsilon_0^T (H - H_1) \varepsilon_0 / l}{\varepsilon_0^T (I_n - H) \varepsilon_0 / (n - p)} \\ &= \frac{\|(H - H_1)\varepsilon_0\|^2 / l}{\|(I_n - H)\varepsilon_0\|^2 / (n - p)}. \end{aligned} \quad (8.10)$$

We have the following joint Normality using the basic fact (8.9):

$$\begin{aligned} \begin{pmatrix} (H - H_1)\varepsilon_0 \\ (I_n - H)\varepsilon_0 \end{pmatrix} &= \begin{pmatrix} H - H_1 \\ I_n - H \end{pmatrix} \varepsilon_0 \\ &\sim N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} H - H_1 & 0 \\ 0 & I_n - H \end{pmatrix} \right\}. \end{aligned}$$

So $(H - H_1)\varepsilon_0$ and $(I_n - H)\varepsilon_0$ are Normal with mean zero and two projection matrices $H - H_1$ and $I_n - H$ as covariances, respectively, and moreover, they are independent. These imply that their squared lengths are chi-squared (by Theorem B.10 in Appendix B):

$$\begin{aligned} \|(H - H_1)\varepsilon_0\|^2 &\sim \chi_l^2, \\ \|(I_n - H)\varepsilon_0\|^2 &\sim \chi_{n-p}^2, \end{aligned}$$

and they are independent. These facts, coupled with (8.10), imply that $F_{\text{ANOVA}} \sim F_{l, n-p}$. \square

The second part demonstrates that $F_{\text{ANOVA}} = F_{\text{Wald}}$ without assuming the Normal linear model, which gives an indirect proof for the exact distribution of F_{ANOVA} under the Normal linear model.

Proof of Theorem 8.2 (Part II): Using the FWL Theorem that $\hat{\beta}_2 = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y$, we can rewrite F_{Wald} as

$$\begin{aligned} F_{\text{Wald}} &= \frac{Y^T \tilde{X}_2 (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T \tilde{X}_2 (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y}{l\hat{\sigma}^2} \\ &= \frac{Y^T \tilde{X}_2 (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y}{l\hat{\sigma}^2} \\ &= \frac{Y^T \tilde{H}_2 Y}{l\hat{\sigma}^2}, \end{aligned} \quad (8.11)$$

recalling that $\tilde{H}_2 = \tilde{X}_2 (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T$ is the projection matrix onto the column space of \tilde{X}_2 . Therefore, $F_{\text{ANOVA}} = F_{\text{Wald}}$ follows from the basic identity $H - H_1 = \tilde{H}_2$ ensured by Lemma 7.2. \square

We can use the `anova` function in `R` to compute the F statistic and the p -value. Below I revisit the `lalonge` data, first analyzed in Chapter 5.4.3 of this book. The result is identical as in Section 5.4.3.

```
> library("Matching")
> data(lalonge)
> lalonge_full = lm(re78 ~ ., data = lalonge)
> lalonge_treat = lm(re78 ~ treat, data = lalonge)
> anova(lalonge_treat, lalonge_full)
Analysis of Variance Table

Model 1: re78 ~ treat
Model 2: re78 ~ age + educ + black + hisp + married + nodegr + re74 +
  re75 + u74 + u75 + treat
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1      443 1.9178e+10
2      433 1.8389e+10 10  788799023  1.8574 0.04929 *
```

In fact, we can conduct an analysis of variance in a sequence of models. For example, we can supplement the above analysis with a model containing only the intercept. The function `anova` works for a sequence of nested models with increasing complexities.

```
> lalonge1 = lm(re78 ~ 1, data = lalonge)
> anova(lalonge1, lalonge_treat, lalonge_full)
Analysis of Variance Table

Model 1: re78 ~ 1
Model 2: re78 ~ treat
Model 3: re78 ~ age + educ + black + hisp + married + nodegr + re74 +
  re75 + u74 + u75 + treat
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1      444 1.9526e+10
2      443 1.9178e+10  1 348013456  8.1946 0.004405 **
3      433 1.8389e+10 10  788799023  1.8574 0.049286 *
```

Overall, the treatment variable is significantly related to the outcome, but none of the pretreatment covariates are.

8.4 Homework problems

8.1 FWL Theorem with an intercept

Theorem 8.3 below extends Theorem 7.1 and highlights the inclusion of the intercept in OLS. Prove Theorem 8.3.

With the intercept in OLS, partition X and β as

$$X = \begin{pmatrix} 1_n & X_1 & X_2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix},$$

where $X_1 \in \mathbb{R}^{n \times k}$, $X_2 \in \mathbb{R}^{n \times l}$, $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}^k$ and $\beta_2 \in \mathbb{R}^l$. Consider the *long regression*

$$\begin{aligned} Y &= X\hat{\beta} + \hat{\varepsilon} \\ &= \begin{pmatrix} 1_n & X_1 & X_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} + \hat{\varepsilon} \\ &= 1_n\hat{\beta}_0 + X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}, \end{aligned}$$

and the *short regression*

$$Y = 1_n\tilde{\beta}_0 + X_2\tilde{\beta}_2 + \tilde{\varepsilon},$$

where $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ and $\begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_2 \end{pmatrix}$ are the OLS coefficients, and $\hat{\varepsilon}$ and $\tilde{\varepsilon}$ are the residual vectors from the long and short regressions, respectively.

Theorem 8.3 *The OLS estimator for β_2 in the long regression equals the coefficient of \tilde{X}_2 in the OLS fit of Y on $(1_n, \tilde{X}_2)$, where \tilde{X}_2 is the residual matrix of the column-wise OLS fit of X_2 on $(1_n, X_1)$, and also equals the coefficient of \tilde{X}_2 in the OLS fit of \tilde{Y} on $(1_n, \tilde{X}_2)$, where \tilde{Y} is the residual vector of the OLS fit of Y on $(1_n, X_1)$.*

8.2 General centering

Verify (8.2).

8.3 Two-way centering of a matrix

Given $X \in \mathbb{R}^{n \times p}$, show that all rows and columns of $C_n X C_p$ have mean 0, where $C_n = I_n - n^{-1}1_n 1_n^T$ and $C_p = I_p - p^{-1}1_p 1_p^T$.

8.4 Linear fixed-effects regression for panel data

Assume that we have data over n units and T time points: $\{y_{it} \in \mathbb{R}, x_{it} \in \mathbb{R}^p : i = 1, \dots, n; t = 1, \dots, T\}$. Prove that the solution of the problem

$$(\hat{\alpha}_1, \dots, \hat{\alpha}_n, \hat{\beta}) = \arg \min_{a_1, \dots, a_n, b \in \mathbb{R}^p} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - a_i - x_{it}^T b)^2$$

is $\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{x}_{i\cdot}^T \hat{\beta}$, with

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^n \sum_{t=1}^T (\dot{y}_{it} - \dot{x}_{it}^T b)^2$$

and

$$\begin{aligned}\bar{y}_{i\cdot} &= T^{-1} \sum_{t=1}^T y_{it}, & \dot{y}_{it} &= y_{it} - \bar{y}_{i\cdot}, \\ \bar{x}_{i\cdot} &= T^{-1} \sum_{t=1}^T x_{it}, & \dot{x}_{it} &= x_{it} - \bar{x}_{i\cdot}.\end{aligned}$$

Remark: In econometrics, cross-section data contain observations over units at one time point, whereas panel data contain observations over units and time. The OLS problem is motivated by the following linear fixed-effects panel data model:

$$y_{it} = \alpha_i + x_{it}^T \beta + \varepsilon_{it},$$

where the α_i 's and β are the unknown parameters. The total number of parameters grows with n , so solving the original OLS problem directly can be computationally inefficient.

8.5 Linear two-way fixed-effects regression for panel data

This problem extends Problem 8.4.

Assume that we have data over n units and T time points: $\{y_{it} \in \mathbb{R}, x_{it} \in \mathbb{R}^p : i = 1, \dots, n; t = 1, \dots, T\}$. Prove that the solution of the problem

$$\begin{aligned}(\hat{\mu}, \hat{\alpha}_1, \dots, \hat{\alpha}_n, \hat{\gamma}_1, \dots, \hat{\gamma}_T, \hat{\beta} \in \mathbb{R}^p) &= \arg \min_{\mu, \alpha_1, \dots, \alpha_n, r_1, \dots, r_T, b} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \mu - \alpha_i - r_t - x_{it}^T b)^2 \\ &\text{such that } \sum_{i=1}^n \alpha_i = \sum_{t=1}^T r_t = 0\end{aligned}$$

is $\hat{\mu} = \bar{y}_{\cdot\cdot} - \bar{x}_{\cdot\cdot}^T \hat{\beta}$, $\hat{\alpha}_i = (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) - (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^T \hat{\beta}$, and $\hat{\gamma}_t = (\bar{y}_{\cdot t} - \bar{y}_{\cdot\cdot}) - (\bar{x}_{\cdot t} - \bar{x}_{\cdot\cdot})^T \hat{\beta}$, with

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^n \sum_{t=1}^T (\ddot{y}_{it} - \ddot{x}_{it}^T b)^2$$

and

$$\begin{aligned}\bar{y}_{i\cdot} &= T^{-1} \sum_{t=1}^T y_{it}, & \bar{y}_{\cdot t} &= n^{-1} \sum_{i=1}^n y_{it}, & \bar{y}_{\cdot\cdot} &= (nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T y_{it}, & \ddot{y}_{it} &= y_{it} - \bar{y}_{i\cdot} - \bar{y}_{\cdot t} + \bar{y}_{\cdot\cdot}, \\ \bar{x}_{i\cdot} &= T^{-1} \sum_{t=1}^T x_{it}, & \bar{x}_{\cdot t} &= n^{-1} \sum_{i=1}^n x_{it}, & \bar{x}_{\cdot\cdot} &= (nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T x_{it}, & \ddot{x}_{it} &= x_{it} - \bar{x}_{i\cdot} - \bar{x}_{\cdot t} + \bar{x}_{\cdot\cdot}.\end{aligned}$$

Remark: The OLS problem is motivated by the following linear two-way fixed-effects panel data model:

$$y_{it} = \mu + \alpha_i + \gamma_t + x_{it}^T \beta + \varepsilon_{it},$$

where the μ , α_i 's, γ_t 's and β are the unknown parameters. The total number of parameters grows with n and T , so solving the original OLS problem directly can be computationally inefficient.

Before solving this problem, you can first consider a simpler problem without covariates:

$$\begin{aligned}(\hat{\mu}, \hat{\alpha}_1, \dots, \hat{\alpha}_n, \hat{\gamma}_1, \dots, \hat{\gamma}_T) &= \arg \min_{\mu, \alpha_1, \dots, \alpha_n, r_1, \dots, r_T} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \mu - \alpha_i - r_t)^2 \\ &\text{such that } \sum_{i=1}^n \alpha_i = \sum_{t=1}^T r_t = 0\end{aligned}$$

which is called the balanced two-way analysis of variance (ANOVA) model in statistics.

8.6 *t*-statistic in multivariate OLS

This problem extends Problem 5.9.

Focus on multivariate OLS discussed in Chapter 5: $y_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2^T x_{i2} + \hat{\varepsilon}_i$ ($i = 1, \dots, n$), where x_{i1} is a scalar and x_{i2} can be a vector. Show that under homoskedasticity, the *t*-statistic associated with $\hat{\beta}_1$ equals

$$\frac{\hat{\rho}_{y x_1 | x_2}}{\sqrt{(1 - \hat{\rho}_{y x_1 | x_2}^2)/(n - p)}},$$

where p is the total number of regressors and $\hat{\rho}_{y x_1 | x_2}$ is the sample partial correlation coefficient between y and x_1 given x_2 .

Remark: Frank (2000) applied this formula to study causal inference.

8.7 Equivalence of the *t*-statistics in multivariate OLS

This problem extends Problems 5.10 and 6.5.

Consider data $(x_{i1}, x_{i2}, y_i)_{i=1}^n$, where both x_{i1} and y_i are scalars and x_{i2} can be a vector. Run OLS fit of y_i on $(1, x_{i1}, x_{i2})$ to obtain $t_{y|x_1, x_2}$, the *t*-statistic of the coefficient of x_{i1} , under the homoskedasticity assumption. Run OLS fit of x_{i1} on $(1, y_i, x_{i2})$ to obtain $t_{x_1|y, x_2}$, the *t*-statistic of the coefficient of y_i , under the homoskedasticity assumption.

Show $t_{y|x_1, x_2} = t_{x_1|y, x_2}$. Give a counterexample in which the numerical equivalence of the *t*-statistics breaks down based on the EHW standard error.

8.8 Formula of the partial correlation coefficient

Prove Theorem 8.1 based on the definition in (8.3).

8.9 Examples of Simpson's Paradox

Give three numerical examples of (Y, X, W) that cause Simpson's Paradox. Report the mean and covariance matrix for each example.

8.10 Simpson's Paradox in reality

Find a real-life dataset with Simpson's Paradox.

8.11 Basic properties of projection matrices

Prove Lemma 8.1.

8.12 Correlation of the regression coefficients

1. Regress Y on $(1_n, X_1, X_2)$ where X_1 and X_2 are two n -vectors with positive sample Pearson correlation $\hat{\rho}_{x_1 x_2} > 0$.

Prove that the corresponding OLS coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are negatively correlated under the Gauss–Markov model of Y on $(1_n, X_1, X_2)$.

2. Regress Y on $(1_n, X_1, X_2, X_3)$ where X_1 and X_2 are two n -vectors and X_3 is an $n \times L$ dimensional matrix. Assume the partial correlation coefficient between X_1 and X_2 given X_3 is positive.

Prove that the corresponding OLS coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are negatively correlated under the Gauss–Markov model Y on $(1_n, X_1, X_2, X_3)$.

8.13 Inverse of sample covariance matrix and partial correlation coefficient

This is the sample version of Problem B.12 in Appendix B.

Based on $X \in \mathbb{R}^{n \times p}$, we can compute the sample covariance matrix $\hat{\Sigma}$. Denote its inverse by $\hat{\Sigma}^{-1} = (\hat{\sigma}^{jk})_{1 \leq j, k \leq p}$. Prove Theorem 8.4 below.

Theorem 8.4 For any pair $j \neq k$, we have

$$\hat{\sigma}^{jk} = 0 \text{ if and only if } \hat{\rho}_{x_j x_k | x_{\setminus(j,k)}} = 0$$

where $\hat{\rho}_{x_j x_k | x_{\setminus(j,k)}}$ is the partial correlation coefficient of X_j and X_k given all other variables.



9

Cochran's Formula and Omitted-Variable Bias

Frisch–Waugh–Lovell (FWL) Theorem in Chapter 7 and Cochran's formula in this chapter are sister results about OLS. The FWL Theorem states the equivalence between the coefficients in the long regression and partial regressions. Cochran's formula compares the coefficients in the long and short regressions. They are both useful for interpreting the OLS coefficients.

9.1 Cochran's formula

Consider an $n \times 1$ vector Y , an $n \times k$ matrix X_1 , and an $n \times l$ matrix X_2 . Similar to the FWL Theorem, we do not impose any statistical models. We can fit the following OLS:

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}, \quad (9.1)$$

$$Y = X_2\tilde{\beta}_2 + \tilde{\varepsilon}, \quad (9.2)$$

$$X_1 = X_2\hat{\delta} + \hat{U}, \quad (9.3)$$

where $\hat{\varepsilon}, \tilde{\varepsilon}$ are the residual vectors, and \hat{U} is the residual matrix from the column-wise OLS of X_1 on X_2 . Therefore, \hat{U} is an $n \times k$ matrix.

Theorem 9.1 *Under the OLS fits (9.1)–(9.3), we have*

$$\tilde{\beta}_2 = \hat{\beta}_2 + \hat{\delta}\hat{\beta}_1.$$

This is a pure linear algebra fact similar to the FWL Theorem. It is called *Cochran's formula* in statistics. Sir David Cox (Cox, 2007) attributed the formula to Cochran (1938) although Cochran himself attributed the formula to Fisher (1925a).

Cochran's formula may seem familiar to readers who know the chain rule in calculus. In a deterministic world with scalar y, x_1, x_2 , if

$$y(x_1, x_2) = x_1\beta_1 + x_2\beta_2$$

is a function of x_1 and x_2 , and

$$x_1(x_2) = x_2\delta$$

is a function of x_2 , then the derivative of y with respect to x_2 equals

$$\begin{aligned} \frac{dy}{dx_2} &= \frac{\partial y}{\partial x_1} \frac{\partial x_1}{\partial x_2} + \frac{\partial y}{\partial x_2} \\ &= \delta\beta_1 + \beta_2. \end{aligned}$$

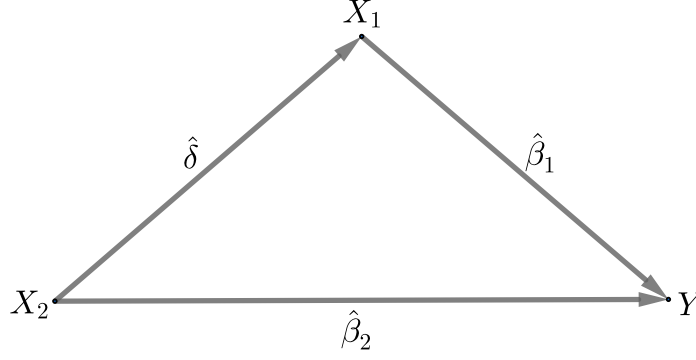


FIGURE 9.1: A diagram for Cochran's formula

But the OLS decompositions in (9.1)–(9.3) do not establish any deterministic relationships among Y and (X_1, X_2) .

In some sense, the formula in Theorem 9.1 is obvious. From the OLS fits (9.1) and (9.3), we have

$$\begin{aligned}
 Y &= X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon} \\
 &= (X_2\hat{\delta} + \hat{U})\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon} \\
 &= X_2\hat{\delta}\hat{\beta}_1 + \hat{U}\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon} \\
 &= X_2(\hat{\delta}\hat{\beta}_1 + \hat{\beta}_2) + (\hat{U}\hat{\beta}_1 + \hat{\varepsilon}).
 \end{aligned} \tag{9.4}$$

This suggests that $\tilde{\beta}_2 = \hat{\beta}_2 + \hat{\delta}\hat{\beta}_1$. The above derivation follows from simple algebraic manipulations and does not use any properties of the OLS. To prove Theorem 9.1, we need to verify that the last line is indeed the OLS fit of Y on X_2 . The proof is indeed very simple. **Proof of Theorem 9.1:** Based on the above discussion, we only need to show that (9.4) is the OLS fit of Y on X_2 , which is equivalent to show that $\hat{U}\hat{\beta}_1 + \hat{\varepsilon}$ is orthogonal to all columns of X_2 . This follows from

$$X_2^T(\hat{U}\hat{\beta}_1 + \hat{\varepsilon}) = X_2^T\hat{U}\hat{\beta}_1 + X_2^T\hat{\varepsilon} = 0,$$

because $X_2^T\hat{U} = 0$ based on the OLS fit in (9.3) and $X_2^T\hat{\varepsilon} = 0$ based on the OLS fit in (9.1). \square

Figure 9.1 illustrates Theorem 9.1. Intuitively, $\tilde{\beta}_2$ measures the total impact of X_2 on Y , which has two channels:

- (C1) $\hat{\beta}_2$ measures the impact acting directly;
- (C2) $\hat{\delta}\hat{\beta}_1$ measures the impact acting indirectly through X_1 .

This interpretation is closely related to *mediation analysis* in causal inference. See Problem 9.1 for more details. If you are interested in more discussions on mediation analysis, you can read Chapter 27 of Ding (2024) or the monograph of VanderWeele (2015).

Figure 9.1 shows the interplay among three variables. Theoretically, we can discuss a system of more than three variables which is called the *path model*. This more advanced topic is beyond the scope of this book. Wright (1921, 1934)'s initial discussion of this approach was motivated by genetic studies. Problems 9.1 and 9.2 are two examples. See Freedman (2009) for a critical textbook introduction.

9.2 Omitted-variable bias

The proof of Theorem 9.1 is very simple. However, it is one of the most insightful formulas in statistics. Econometricians often call it the *omitted-variable bias* formula because it quantifies the bias of the OLS coefficient of X_2 in the short regression omitting possibly important variables in X_1 . If the OLS coefficient from the long regression is unbiased then the OLS coefficient from the short regression has a biased term $\hat{\delta}\hat{\beta}_1$, which equals the product of the coefficient of X_2 in the OLS fit of X_1 on X_2 and the coefficient of X_1 in the long regression.

Below I will discuss a canonical example of using OLS to estimate the treatment effect in observational studies. For unit i ($i = 1, \dots, n$), let y_i be the outcome, z_i be the binary treatment indicator (1 for the treatment group and 0 for the control group) and x_i be the observed covariate vector. Practitioners often fit the following OLS:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 z_i + \tilde{\beta}_2^T x_i + \tilde{\varepsilon}_i$$

and interpret $\tilde{\beta}_1$ as the treatment effect estimate. However, observational studies may suffer from unmeasured confounding, that is, the treatment and control units differ in unobserved but important ways. In the simplest case, the above OLS may have omitted a variable u_i for each unit i , which is called a *confounder*. The oracle OLS is

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 z_i + \hat{\beta}_2^T x_i + \hat{\beta}_3 u_i + \hat{\varepsilon}_i$$

and the coefficient $\hat{\beta}_1$ is an unbiased estimator if the model with u_i is correct. With X_1 containing the values of the u_i 's and X_2 containing the values of the $(1, z_i, x_i^T)$'s, Cochran's formula implies that

$$\begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} + \hat{\beta}_3 \begin{pmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix}$$

where $(\hat{\delta}_0, \hat{\delta}_1, \hat{\delta}_2^T)^T$ is the coefficient vector in the OLS fit of u_i on $(1, z_i, x_i)$. Therefore, we can quantify the difference between the observed estimate $\tilde{\beta}_1$ and oracle estimate $\hat{\beta}_1$:

$$\tilde{\beta}_1 - \hat{\beta}_1 = \hat{\beta}_3 \hat{\delta}_1,$$

which is sometimes called the *confounding bias*.

Using the basic properties of OLS, we can show that $\hat{\delta}_1$ equals the difference in means of $e_i = u_i - \hat{\delta}_2^T x_i$ across the treatment and control groups:

$$\hat{\delta}_1 = \bar{e}_1 - \bar{e}_0,$$

where the bar and subscript jointly denote the sample mean of a particular variable within a treatment group. So

$$\tilde{\beta}_1 - \hat{\beta}_1 = \hat{\beta}_3 (\bar{e}_1 - \bar{e}_0). \quad (9.5)$$

Moreover, we can obtain a more explicit formula for $\hat{\delta}_1$:

$$\hat{\delta}_1 = \bar{u}_1 - \bar{u}_0 - \hat{\delta}_2^T (\bar{x}_1 - \bar{x}_0).$$

So

$$\tilde{\beta}_1 - \hat{\beta}_1 = \hat{\beta}_3 (\bar{u}_1 - \bar{u}_0) - \hat{\beta}_3 \hat{\delta}_2^T (\bar{x}_1 - \bar{x}_0). \quad (9.6)$$

Both (9.5) and (9.6) give some insights into the bias due to omitting an important covariate u . It is clear that the bias depends on $\hat{\beta}_3$, which quantifies the relationship between u and y . The formula (9.5) shows that the bias also depends on the imbalance in means of u across the treatment and control groups, after adjusting for the observed covariates x , that is, the imbalance in means of the residual confounding. The formula (9.6) shows a more explicit formula of the bias. The above discussion is often called *bias analysis* in epidemiology or *sensitivity analysis* in statistics and econometrics.

9.3 Homework problems

9.1 Baron–Kenny method for mediation analysis

The Baron–Kenny method, popularized by Baron and Kenny (1986), is one of the most cited methods in social science. It concerns the interplay among three variables z, m, y , after controlling for some other variables x . Let Z, M, Y be $n \times 1$ vectors representing the observed values of z, m, y , and let X be the $n \times p$ matrix representing the observations of x . The question of interest is to assess the “direct” and “indirect” effects of z on y , acting independently and through m , respectively. We do not need to define these notions precisely since we are only interested in the algebraic property below.

The Baron–Kenny method runs the OLS

$$Y = \hat{\beta}_0 \mathbf{1}_n + \hat{\beta}_1 Z + \hat{\beta}_2 M + X \hat{\beta}_3 + \hat{\varepsilon}_Y$$

and interprets $\hat{\beta}_1$ as the estimator of the “direct effect” of z on y . The “indirect effect” of z on y through m has two estimators. First, based on the OLS

$$Y = \tilde{\beta}_0 \mathbf{1}_n + \tilde{\beta}_1 Z + X \tilde{\beta}_3 + \tilde{\varepsilon}_Y,$$

define the *difference estimator* as $\tilde{\beta}_1 - \hat{\beta}_1$. Second, based on the OLS

$$M = \hat{\gamma}_0 \mathbf{1}_n + \hat{\gamma}_1 Z + X \hat{\gamma}_2 + \hat{\varepsilon}_M,$$

define the *product estimator* as $\hat{\gamma}_1 \hat{\beta}_2$. Figure 9.2 illustrates the OLS fits used in defining the estimators.

Prove that

$$\tilde{\beta}_1 - \hat{\beta}_1 = \hat{\gamma}_1 \hat{\beta}_2$$

that is, the difference estimator and product estimator are numerically identical.

9.2 A special case of path analysis

Figure 9.3 represents the order of the variables $X_1, X_2, X_3, Y \in \mathbb{R}^n$. Run the following OLS:

$$\begin{aligned} Y &= \hat{\beta}_0 \mathbf{1}_n + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\varepsilon}_Y, \\ X_3 &= \hat{\delta}_0 \mathbf{1}_n + \hat{\delta}_1 X_1 + \hat{\delta}_2 X_2 + \hat{\varepsilon}_3, \\ X_2 &= \hat{\theta}_0 \mathbf{1}_n + \hat{\theta}_1 X_1 + \hat{\varepsilon}_2, \end{aligned}$$

and

$$Y = \tilde{\beta}_0 \mathbf{1}_n + \tilde{\beta}_1 X_1 + \tilde{\varepsilon}_Y.$$

Prove that

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\theta}_1 + \hat{\beta}_3 \hat{\delta}_1 + \hat{\beta}_3 \hat{\delta}_2 \hat{\theta}_1.$$

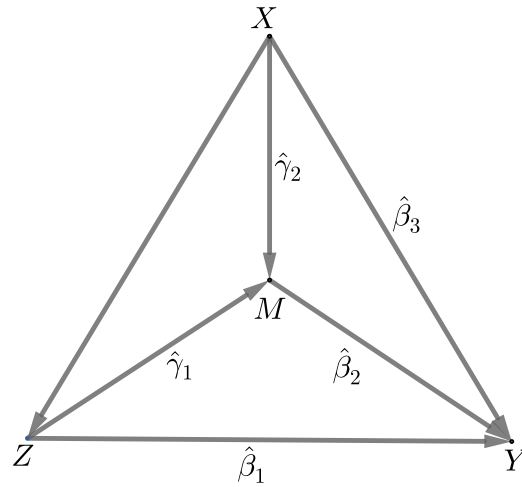


FIGURE 9.2: The graph for the Baron–Kenny method

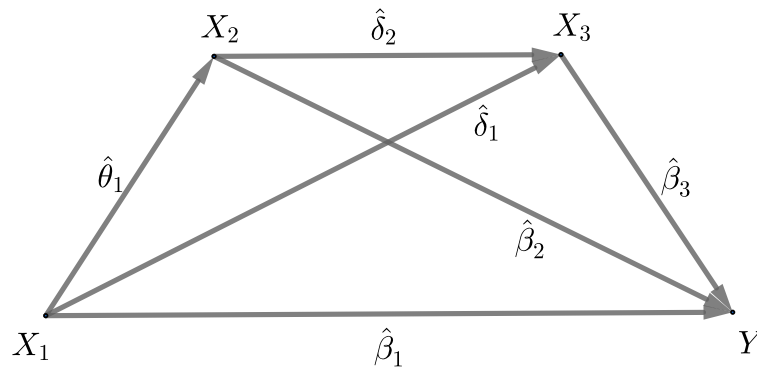


FIGURE 9.3: A path model

Remark: The OLS coefficient of X_1 in the short regression of Y on $(1_n, X_1)$ equals the summation of all the path coefficients from X_1 to Y as illustrated by Figure 9.3:

$$\begin{aligned}
 &X_1 \longrightarrow Y, \\
 &X_1 \longrightarrow X_2 \longrightarrow Y, \\
 &X_1 \longrightarrow X_3 \longrightarrow Y, \\
 &X_1 \longrightarrow X_2 \longrightarrow X_3 \longrightarrow Y.
 \end{aligned}$$

This problem is a special case of the path model, but the conclusion holds in general.

9.3 EHW in long and short regressions

Theorem 9.1 gives Cochran's formula related to the coefficients from three OLS fits. This problem concerns the covariance estimation. There are at least two ways to estimate the covariance of $\tilde{\beta}_2$ in the short regression (9.2). First, from the short regression (9.2), the

EHW covariance estimator is

$$\tilde{V}_2 = (X_2^T X_2)^{-1} X_2^T \text{diag}(\tilde{\varepsilon}^2) X_2 (X_2^T X_2)^{-1}.$$

Second, Cochran's formula implies that

$$\tilde{\beta}_2 = (\hat{\delta}, I_l) \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

is a linear transformation of the coefficient from the long regression, which further justifies the EHW covariance estimator

$$\tilde{V}_2' = (\hat{\delta}, I_l) (X^T X)^{-1} X^T \text{diag}(\tilde{\varepsilon}^2) X (X^T X)^{-1} \begin{pmatrix} \hat{\delta}^T \\ I_l \end{pmatrix}.$$

Prove that

$$\tilde{V}_2' = (X_2^T X_2)^{-1} X_2^T \text{diag}(\tilde{\varepsilon}^2) X_2 (X_2^T X_2)^{-1}.$$

Remark: This problem states that $\tilde{V}_2 \neq \tilde{V}_2'$ in general. To prove the result, you can use the result in Problem 7.1.

Moreover, based on Theorem 7.2, the EHW covariance estimator for $\hat{\beta}_2$ is

$$\hat{V}_2 = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T \text{diag}(\tilde{\varepsilon}^2) \tilde{X}_2 (\tilde{X}_2^T \tilde{X}_2)^{-1},$$

where $\tilde{X}_2 = (I_n - H_1)X_2$. It differs from \tilde{V}_2 and \tilde{V}_2' in general.

9.4 Statistical properties of under-fitted OLS

Assume that $Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$ follows the Gauss–Markov model, where $X_1 \in \mathbb{R}^{n \times k}$, $X_2 \in \mathbb{R}^{n \times l}$, and $\text{cov}(\varepsilon) = \sigma^2 I_n$. However, we only fit the OLS of Y on X_2 with coefficient $\hat{\beta}_2$ and estimated variance $\tilde{\sigma}^2$.

Prove that

$$\begin{aligned} E(\tilde{\beta}_2) &= (X_2^T X_2)^{-1} X_2^T X_1 \beta_1 + \beta_2, \\ \text{var}(\tilde{\beta}_2) &= \sigma^2 (X_2^T X_2)^{-1}, \\ E(\tilde{\sigma}^2) &= \sigma^2 + \beta_1^T X_1^T (I_n - H_2) X_1 \beta_1 / (n - l) \geq \sigma^2. \end{aligned}$$

Part IV

Model Fitting, Checking, and Misspecification



Multiple Correlation Coefficient

This chapter will introduce the R^2 , the *multiple correlation coefficient*, also called the *coefficient of determination* (Wright, 1921). It can achieve two goals:

- (G1) it extends the sample Pearson correlation coefficient between two scalars to a measure of correlation between a scalar outcome and a vector covariate;
- (G2) it measures how well multiple covariates can linearly represent an outcome.

10.1 Equivalent definitions of the multiple correlation coefficient

I start with the standard definition of R^2 between Y and X . Slightly different from other chapters, X in this chapter excludes the column of 1's, so now X is an $n \times (p - 1)$ matrix. Based on the OLS of Y on $(1_n, X)$, we define

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (10.1)$$

You may wonder why the numerator of (10.1) does not use $\bar{\hat{y}} = n^{-1} \sum_{i=1}^n \hat{y}_i$. This is because the average of the fitted values equals the average of the original observed outcomes, i.e., $\bar{\hat{y}} = \bar{y}$, if we include 1_n in the OLS; see Problem 10.1. With scaling factor $(n - 1)^{-1}$, the denominator of R^2 is the sample variance of the outcomes, and the numerator of R^2 is the sample variance of the fitted values. We can verify the following decomposition:

Lemma 10.1 *Based on the OLS of Y on $(1_n, X)$, we have the following variance decomposition:*

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

I leave the proof of Lemma 10.1 as Problem 10.2. Lemma 10.1 states that the total sum of squares $\sum_{i=1}^n (y_i - \bar{y})^2$ equals the regression sum of squares $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ plus the residual sum of squares (RSS) $\sum_{i=1}^n (y_i - \hat{y}_i)^2$. From Lemma 10.1, R^2 must be lie within the interval $[0, 1]$ which measures the proportion of the regression sum of squares in the total sum of squares. An immediate consequence of Lemma 10.1 is that

$$\text{RSS} = (1 - R^2) \times \sum_{i=1}^n (y_i - \bar{y})^2.$$

We can also verify that R^2 is the squared sample Pearson correlation coefficient between Y and \hat{Y} .

Theorem 10.1 We have $R^2 = \hat{\rho}_{y\hat{y}}^2$, where

$$\hat{\rho}_{y\hat{y}} = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}}. \quad (10.2)$$

I leave the proof of Theorem 10.1 to Problem 10.3. Theorem 10.1 states that the multiple correlation coefficient equals the squared Pearson correlation coefficient between the observed outcome y_i and its fitted value \hat{y}_i . Although the sample Pearson correlation coefficient can be positive or negative, R^2 is always non-negative. Geometrically, R^2 equals the squared cosine of the angle between the centered vectors $Y - \bar{y}1_n$ and $\hat{Y} - \bar{y}1_n$; see Appendix A.1 for the geometric interpretation of the Pearson correlation coefficient.

In terms of long and short regressions, we can partition the design matrix into 1_n and X , then the OLS fit of the long regression is

$$Y = 1_n \hat{\beta}_0 + X \hat{\beta} + \hat{\varepsilon}, \quad (10.3)$$

and the OLS fit of the short regression is

$$Y = 1_n \tilde{\beta}_0 + \tilde{\varepsilon}, \quad (10.4)$$

with $\tilde{\beta}_0 = \bar{y}$. The total sum of squares is the residual sum of squares from the short regression so by Lemma 10.1, R^2 also equals

$$R^2 = \frac{\text{RSS}_{\text{short}} - \text{RSS}_{\text{long}}}{\text{RSS}_{\text{short}}}. \quad (10.5)$$

10.2 The multiple correlation coefficient and F statistic

Under the Normal linear model

$$Y = 1_n \beta_0 + X \beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \quad (10.6)$$

we can use the F statistic to test whether $\beta = 0$:

$$F = \frac{(\text{RSS}_{\text{short}} - \text{RSS}_{\text{long}})/(p-1)}{\text{RSS}_{\text{long}}/(n-p)}.$$

This F statistic is a monotone function of R^2 in (10.5) as shown in Theorem 10.2 below. Most standard software packages report both F and R^2 .

Theorem 10.2 We have

$$F = \frac{n-p}{p-1} \times \frac{R^2}{1-R^2}.$$

I leave the proof of Theorem 10.2 as Problem 10.4. Theorem 10.2 is a numeric result without assuming that model (10.6) is correct. Under the Normal linear model, we can derive the exact distribution of R^2 .

Corollary 10.1 Under the Normal linear model (10.6), if $\beta = 0$, then

$$R^2 \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right).$$

Proof of Corollary 10.1: By definition, the F statistic can be represented as

$$F = \frac{\chi_{p-1}^2/(p-1)}{\chi_{n-p}^2/(n-p)},$$

where χ_{p-1}^2 and χ_{n-p}^2 denote independent χ_{p-1}^2 and χ_{n-p}^2 random variables, respectively. Using Theorem 10.2, we have

$$\frac{R^2}{1-R^2} = F \times \frac{p-1}{n-p} = \frac{\chi_{p-1}^2}{\chi_{n-p}^2},$$

which implies

$$R^2 = \frac{\chi_{p-1}^2}{\chi_{p-1}^2 + \chi_{n-p}^2}.$$

Because $\chi_{p-1}^2 \sim \text{Gamma}(\frac{p-1}{2}, \frac{1}{2})$ and $\chi_{n-p}^2 \sim \text{Gamma}(\frac{n-p}{2}, \frac{1}{2})$ by Proposition B.1 in Appendix B, we have

$$R^2 = \frac{\text{Gamma}(\frac{p-1}{2}, \frac{1}{2})}{\text{Gamma}(\frac{p-1}{2}, \frac{1}{2}) + \text{Gamma}(\frac{n-p}{2}, \frac{1}{2})}$$

where $\text{Gamma}(\frac{p-1}{2}, \frac{1}{2})$ and $\text{Gamma}(\frac{n-p}{2}, \frac{1}{2})$ denote independent Gamma random variables. The R^2 follows the Beta distribution by the Beta–Gamma duality in Theorem B.1 in Appendix B. \square

10.3 Numerical examples

I first revisit the LaLonde data to verify Theorems 10.1 and 10.2 numerically.

```
> library("Matching")
> data(lalonde)
> ols.fit = lm(re78 ~ ., y = TRUE, data = lalonde)
> ols.summary = summary(ols.fit)
> r2 = ols.summary$r.squared
> all.equal(r2, (cor(ols.fit$y, ols.fit$fitted.values))^2,
+           check.names = FALSE)
[1] TRUE
>
> fstat = ols.summary$fstatistic
> all.equal(fstat[1], fstat[3]/fstat[2]*r2/(1-r2),
+           check.names = FALSE)
[1] TRUE
```

I then revisit the data from King and Roberts (2015) to verify Theorems 10.1 and 10.2 numerically.

```
> library(foreign)
> dat = read.dta("isq.dta")
> dat = na.omit(dat[,c("multish", "lnpop", "lnpopsq",
+                     "lngdp", "lncolony", "lndist",
+                     "freedom", "militexp", "arms",
+                     "year83", "year86", "year89", "year92")])
>
> ols.fit = lm(log(multish + 1) ~ lnpop + lnpopsq + lngdp + lncolony
+             + lndist + freedom + militexp + arms
```

```

+           + year83 + year86 + year89 + year92,
+           y = TRUE, data=dat)
> ols.summary = summary(ols.fit)
> r2 = ols.summary$r.squared
> all.equal(r2, (cor(ols.fit$y, ols.fit$fitted.values))^2,
+           check.names = FALSE)
[1] TRUE
>
> fstat = ols.summary$fstatistic
> all.equal(fstat[1], fstat[3]/fstat[2]*r2/(1-r2),
+           check.names = FALSE)
[1] TRUE

```

10.4 Homework problems

10.1 Average outcome equals average fitted values

Prove that from the OLS of Y on $(1_n, X)$, we have $\bar{\hat{y}} = \bar{y}$.

10.2 Variance decomposition

Prove Lemma 10.1.

10.3 R^2 and the sample Pearson correlation coefficient

Prove Theorem 10.1.

10.4 F and R^2

Prove Theorem 10.2.

10.5 Exact distribution of $\hat{\rho}$

Assume the Normal linear model $y_i = \alpha + \beta x_i + \varepsilon_i$ with a univariate x_i with $\beta = 0$ and ε_i 's IID $N(0, \sigma^2)$. Find the exact distribution of $\hat{\rho}_{xy}$.

10.6 Partial R^2

The form (10.5) of R^2 is well defined in more general long and short regressions:

$$Y = 1_n \hat{\beta}_0 + X \hat{\beta} + W \hat{\gamma} + \hat{\varepsilon}_Y$$

and

$$Y = 1_n \tilde{\beta}_0 + W \tilde{\gamma} + \tilde{\varepsilon}_Y$$

where X is an $n \times k$ matrix and W is an $n \times l$ matrix. Define the *partial* R^2 between Y and X given W as

$$R_{Y.X|W}^2 = \frac{\text{RSS}(Y \sim 1_n + W) - \text{RSS}(Y \sim 1_n + X + W)}{\text{RSS}(Y \sim 1_n + W)}$$

which spells out the formulas of the long and short regressions. This is an intuitive measure of the multiple correlation between Y and X after controlling for W . The following properties make this intuition more explicit.

1. The partial R^2 equals

$$R_{Y.X|W}^2 = \frac{R_{Y.XW}^2 - R_{Y.W}^2}{1 - R_{Y.W}^2},$$

where $R_{Y.XW}^2$ is the multiple correlation between Y and (X, W) , and $R_{Y.W}^2$ is the multiple correlation between Y and W .

2. The partial R^2 equals the R^2 between $\tilde{\varepsilon}_Y$ and $\tilde{\varepsilon}_X$:

$$R_{Y.X|W}^2 = R_{\tilde{\varepsilon}_Y.\tilde{\varepsilon}_X}^2,$$

where $\tilde{\varepsilon}_X$ is the residual matrix from the OLS fit of X on $(1_n, W)$.

Prove the above two results.

Do the following two results hold?

$$\begin{aligned} R_{Y.XW}^2 &= R_{Y.W}^2 + R_{Y.X|W}^2, \\ R_{Y.XW}^2 &= R_{Y.W|X}^2 + R_{Y.X|W}^2. \end{aligned}$$

For each result, give a proof if it is correct, or give a counterexample if it is incorrect in general.

10.7 Omitted-variable bias in terms of the partial R^2

Revisit Section 9.2 on the following three OLS fits. The first one involves only the observed variables:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 z_i + \tilde{\beta}_2^T x_i + \tilde{\varepsilon}_i,$$

and the second and third ones involve the unobserved u :

$$\begin{aligned} y_i &= \hat{\beta}_0 + \hat{\beta}_1 z_i + \hat{\beta}_2^T x_i + \hat{\beta}_3 u_i + \hat{\varepsilon}_i, \\ u_i &= \hat{\delta}_0 + \hat{\delta}_1 z_i + \hat{\delta}_2^T x_i + \hat{v}_i. \end{aligned}$$

The omitted-variable bias formula states that

$$\tilde{\beta}_1 - \hat{\beta}_1 = \hat{\beta}_3 \hat{\delta}_1.$$

This formula is simple but may be difficult to interpret since u is unobserved and its scale is unclear to researchers.

Prove the following formula:

$$|\tilde{\beta}_1 - \hat{\beta}_1|^2 = R_{Y.U|ZX}^2 \times \frac{R_{Z.U|X}^2}{1 - R_{Z.U|X}^2} \times \frac{\text{RSS}(Y \sim 1_n + Z + X)}{\text{RSS}(Z \sim 1_n + X)}. \quad (10.7)$$

Remark: Cinelli and Hazlett (2020) suggested the partial R^2 parametrization for the omitted-variable bias formula. The formula (10.7) has three factors:

- (F1) the first factor depends on the unknown *sensitivity parameters* $R_{Y.U|ZX}^2$, which measures the confounder-outcome relationship;
- (F2) the second factor depends on the unknown *sensitivity parameters* $R_{Z.U|X}^2$, which measures the treatment-confounder relationship;
- (F3) the third factor equals the ratio of the two residual sums of squares based on the observed data, which does not depend on the unmeasured confounder.

The beauty of (10.7) is that the partial R^2 parameters $R_{Y.U|ZX}^2$ and $R_{Z.U|X}^2$ do not depend on the unknown scale of u . Zhang and Ding (2022) derived more general omitted-variables bias formulas based on partial R^2 .



11

Leverage Scores and Leave-One-Out Formulas

This chapter will discuss two related topics: leverage scores and leave-one-out formulas. We have seen leverage scores in previous chapters, which are defined as the diagonal elements of the hat matrix $H = X(X^T X)^{-1} X^T$ from OLS. This chapter will present more properties of the leverage scores. Intuitively, the leverage score of unit i measures whether covariate x_i is an outlier among other covariates. Consequently, units with larger leverage scores will have larger impact on the final OLS outputs, which can be quantified by various leave-one-out formulas proved in this chapter.

Leave-one-out is a general idea in statistics and machine learning. It is a basic idea to avoid overfitting. It applies to all statistical models. The beauty of OLS is that we can derive leave-one-out formulas explicitly. Those formulas not only allow for fast computation but also provide insights into the properties of OLS.

From an exploratory data analysis perspective, we can use the idea of leave-one-out to assess the impact of individual observations, or equivalently, to assess the *stability* of the results with respect to deleting individual observations (Yu and Kumbier, 2020). In the linear model, plotting the leverage scores can be viewed as a shortcut to the leave-one-out.

This chapter will provide more details for the above high-level statements.

11.1 Leverage scores

The hat matrix $H = X(X^T X)^{-1} X^T$ is a key matrix for OLS. Because

$$H = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} (X^T X)^{-1} (x_1 \quad \cdots \quad x_n),$$

its (i, j) th element equals $h_{ij} = x_i^T (X^T X)^{-1} x_j$. In this chapter, we will pay special attention to its diagonal elements

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \quad (i = 1, \dots, n)$$

often called the *leverage scores*, which play important roles in many discussions later.

11.1.1 The average leverage score equals p/n

Because H is a projection matrix of rank p , we have

$$\sum_{i=1}^n h_{ii} = \text{trace}(H) = \text{rank}(H) = p,$$

which implies that

$$n^{-1} \sum_{i=1}^n h_{ii} = p/n,$$

i.e., the average of the leverage scores equals p/n . Therefore, the maximum of the leverage scores must be larger than or equal to p/n :

$$\max_{1 \leq i \leq n} h_{ii} \geq n^{-1} \sum_{i=1}^n h_{ii} = p/n.$$

As p/n becomes larger, we will observe more extreme leverage scores.

11.1.2 The leverage scores are all bounded between 0 and 1

Because $H = H^2$ and $H = H^T$, we have

$$\begin{aligned} h_{ii} &= \sum_{j=1}^n h_{ij} h_{ji} \\ &= \sum_{j=1}^n h_{ij}^2 \\ &= h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \\ &\geq h_{ii}^2, \end{aligned}$$

which implies

$$h_{ii} \in [0, 1],$$

i.e., each leverage score is bounded between 0 and 1.¹

11.1.3 The i th leverage score measures the impact of the i th observation in prediction

Because $\hat{Y} = HY$, we have

$$\hat{y}_i = \sum_{j=1}^n h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j,$$

which implies that

$$\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii}.$$

So h_{ii} measures the contribution of y_i in the predicted value \hat{y}_i . In general, we do not want the contribution of y_i in predicting itself to be too large, because this means we do not borrow enough information from other observations, making the prediction very noisy. This is also clear from the variance of the predicted value $\hat{y}_i = x_i^T \hat{\beta}$ under the Gauss–Markov model:²

$$\begin{aligned} \text{var}(\hat{y}_i) &= x_i^T \text{cov}(\hat{\beta}) x_i \\ &= \sigma^2 x_i^T (X^T X)^{-1} x_i \\ &= \sigma^2 h_{ii}. \end{aligned}$$

¹This also follows from Theorem A.4 in Appendix A since the eigenvalues of H are 0 and 1.

²We have already proved a more general result on the covariance matrix of \hat{Y} in Theorem 4.2.

So the variance of \hat{y}_i increases with h_{ii} .

11.1.4 The i th leverage score measures whether x_i is an outlier compared with other covariates

The h_{ii} measures whether observation i is an outlier based on its covariate value, that is, whether x_i is far from the center of the data. Partition the design matrix as $X = \begin{pmatrix} 1_n & X_2 \end{pmatrix}$ with $H_1 = n^{-1}1_n1_n^T$. The covariates X_2 has sample mean $\bar{x}_2 = n^{-1}\sum_{i=1}^n x_{i2}$ and sample covariance

$$\begin{aligned} S &= (n-1)^{-1} \sum_{i=1}^n (x_{i2} - \bar{x}_2)(x_{i2} - \bar{x}_2)^T \\ &= (n-1)^{-1} X_2^T (I_n - H_1) X_2. \end{aligned}$$

The sample Mahalanobis distance between x_{i2} and the center \bar{x}_2 is

$$D_i^2 = (x_{i2} - \bar{x}_2)^T S^{-1} (x_{i2} - \bar{x}_2).$$

Theorem 11.1 below shows that h_{ii} is a monotone function of D_i^2 :

Theorem 11.1 *Assume that we include the intercept in OLS. We have*

$$h_{ii} = \frac{1}{n} + \frac{D_i^2}{n-1}, \quad (11.1)$$

so $h_{ii} \geq 1/n$.

Proof of Theorem 11.1: The definition of D_i^2 implies that it is the (i, i) th element of the following matrix:

$$\begin{aligned} & \begin{pmatrix} x_{12} - \bar{x}_2 \\ \vdots \\ x_{n2} - \bar{x}_2 \end{pmatrix}^T S^{-1} \begin{pmatrix} x_{12} - \bar{x}_2 & \cdots & x_{n2} - \bar{x}_2 \end{pmatrix} \\ &= (I_n - H_1) X_2 \{ (n-1)^{-1} X_2^T (I_n - H_1) X_2 \}^{-1} X_2^T (I_n - H_1) \\ &= (n-1) \tilde{X}_2 (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T \\ &= (n-1) \tilde{H}_2 \\ &= (n-1)(H - H_1), \end{aligned}$$

recalling that $\tilde{X}_2 = (I_n - H_1) X_2$, $\tilde{H}_2 = \tilde{X}_2 (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T$, and $H = H_1 + \tilde{H}_2$ by Lemma 7.2. Therefore,

$$D_i^2 = (n-1)(h_{ii} - 1/n)$$

which implies (11.1). \square

11.1.5 Other properties of the leverage scores

Another advanced result on the leverage scores is due to Huber (1973). He proved that in the linear model with non-Normal IID ε_i with mean 0 and variance $\sigma^2 < \infty$, all linear combinations of the OLS coefficient are asymptotically Normal if and only if the maximum leverage score converges to 0. This is a very elegant asymptotic result on the OLS coefficient. I give more details in Appendix C as an application of the Lindeberg–Feller CLT.

These are the basic properties of the leverage scores. Chatterjee and Hadi (1988) provided an in-depth discussion of the properties of the leverage scores. We will see their roles frequently in later parts of this chapter.

11.2 Leave-one-out formulas

To measure the impact of the i th observation on the final OLS estimator, a natural approach is to delete the i th row from the full data

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

and check how much the OLS estimator changes. Let

$$X_{[-i]} = \begin{pmatrix} x_1^T \\ \vdots \\ x_{i-1}^T \\ x_{i+1}^T \\ \vdots \\ x_n^T \end{pmatrix}, \quad Y_{[-i]} = \begin{pmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{pmatrix}$$

denote the leave- i -out data, and define

$$\hat{\beta}_{[-i]} = (X_{[-i]}^T X_{[-i]})^{-1} X_{[-i]}^T Y_{[-i]} \quad (11.2)$$

as the corresponding OLS estimator. We can fit n OLS by deleting the i th row ($i = 1, \dots, n$). However, this is computationally intensive especially when n is large. Theorem 11.2 shows that we need only to fit OLS once and then compute all leave-one-out coefficients explicitly.

Theorem 11.2 *Recall that $\hat{\beta}$ is the full data OLS, $\hat{\varepsilon}_i$ is the residual and h_{ii} is the leverage score for the i th observation. We have*

$$\hat{\beta}_{[-i]} = \hat{\beta} - (1 - h_{ii})^{-1} (X^T X)^{-1} x_i \hat{\varepsilon}_i$$

if $X^T X$ is invertible and $h_{ii} \neq 1$.

The condition that $X^T X$ is invertible ensures the full OLS has a unique solution. The additional condition $h_{ii} \neq 1$ ensures that the leave- i -out OLS has a unique solution; see Problem 11.10.

Proof of Theorem 11.2: From (11.2), we need to invert

$$X_{[-i]}^T X_{[-i]} = \sum_{i' \neq i} x_{i'} x_{i'}^T = X^T X - x_i x_i^T$$

and calculate

$$X_{[-i]}^T Y_{[-i]} = \sum_{i' \neq i} x_{i'} y_{i'} = X^T Y - x_i y_i,$$

which are the original $X^T X$ and $X^T Y$ without the contribution of the i th observation. Using the following Sherman–Morrison formula in Problem A.3:

$$(A + uv^T)^{-1} = A^{-1} - (1 + v^T A^{-1} u)^{-1} A^{-1} u v^T A^{-1}$$

with $A = X^T X$, $u = x_i$, and $v = -x_i$ we can invert $X_{[-i]}^T X_{[-i]}$ as³

$$\begin{aligned} (X_{[-i]}^T X_{[-i]})^{-1} &= (X^T X)^{-1} + \{1 - x_i^T (X^T X)^{-1} x_i\}^{-1} (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} \\ &= (X^T X)^{-1} + (1 - h_{ii})^{-1} (X^T X)^{-1} x_i x_i^T (X^T X)^{-1}. \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{\beta}_{[-i]} &= (X_{[-i]}^T X_{[-i]})^{-1} X_{[-i]}^T Y_{[-i]} \\ &= \left\{ (X^T X)^{-1} + (1 - h_{ii})^{-1} (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} \right\} (X^T Y - x_i y_i) \\ &= (X^T X)^{-1} X^T Y \\ &\quad - (X^T X)^{-1} x_i y_i \\ &\quad + (1 - h_{ii})^{-1} (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} X^T Y \\ &\quad - (1 - h_{ii})^{-1} (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} x_i y_i \\ &= \hat{\beta} - (X^T X)^{-1} x_i y_i + (1 - h_{ii})^{-1} (X^T X)^{-1} x_i x_i^T \hat{\beta} - h_{ii} (1 - h_{ii})^{-1} (X^T X)^{-1} x_i y_i \\ &= \hat{\beta} - (1 - h_{ii})^{-1} (X^T X)^{-1} x_i y_i + (1 - h_{ii})^{-1} (X^T X)^{-1} x_i \hat{y}_i \\ &= \hat{\beta} - (1 - h_{ii})^{-1} (X^T X)^{-1} x_i \hat{\varepsilon}_i. \end{aligned}$$

□

With the leave- i -out OLS estimator $\hat{\beta}_{[-i]}$, we can define the predicted residual

$$\hat{\varepsilon}_{[-i]} = y_i - x_i^T \hat{\beta}_{[-i]},$$

which is different from the original residual $\hat{\varepsilon}_i$. The predicted residual based on leave-one-out can better measure the performance of the prediction because it mimics the real problem of predicting a future observation. In contrast, the original residual based on the full data $\hat{\varepsilon}_i = y_i - x_i^T \hat{\beta}$ gives an overly optimistic measure of the performance of the prediction. This is related to the overfitting issue discussed in Chapter 13. Under the Gauss–Markov model, Theorem 4.2 implies that the original residual has mean zero and variance

$$\text{var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}), \quad (11.3)$$

and we can show that the predicted residual has mean zero and variance

$$\text{var}(\hat{\varepsilon}_{[-i]}) = \text{var}(y_i - x_i^T \hat{\beta}_{[-i]}) = \sigma^2 + \sigma^2 x_i^T (X_{[-i]}^T X_{[-i]})^{-1} x_i. \quad (11.4)$$

Theorem 11.3 below further simplifies the predicted residual and its variance.

Theorem 11.3 Assume $h_{ii} \neq 1$. We have

$$\hat{\varepsilon}_{[-i]} = \hat{\varepsilon}_i / (1 - h_{ii}),$$

and under Assumption 4.1, we have

$$\text{var}(\hat{\varepsilon}_{[-i]}) = \sigma^2 / (1 - h_{ii}). \quad (11.5)$$

Proof of Theorem 11.3: By definition and Theorem 11.2, we have

$$\begin{aligned} \hat{\varepsilon}_{[-i]} &= y_i - x_i^T \hat{\beta}_{[-i]} \\ &= y_i - x_i^T \left\{ \hat{\beta} - (1 - h_{ii})^{-1} (X^T X)^{-1} x_i \hat{\varepsilon}_i \right\} \\ &= y_i - x_i^T \hat{\beta} + (1 - h_{ii})^{-1} x_i^T (X^T X)^{-1} x_i \hat{\varepsilon}_i \\ &= \hat{\varepsilon}_i + h_{ii} (1 - h_{ii})^{-1} \hat{\varepsilon}_i \\ &= \hat{\varepsilon}_i / (1 - h_{ii}). \end{aligned} \quad (11.6)$$

³See Problem A.6 in Appendix A for a related linear algebra result.

Combining (11.3) and (11.6), we can derive its variance:

$$\text{var}(\hat{\varepsilon}_{[-i]}) = \text{var}(\hat{\varepsilon}_i)/(1 - h_{ii})^2 = \sigma^2(1 - h_{ii})/(1 - h_{ii})^2 = \sigma^2/(1 - h_{ii}).$$

□

Comparing formulas (11.4) and (11.5), we obtain that

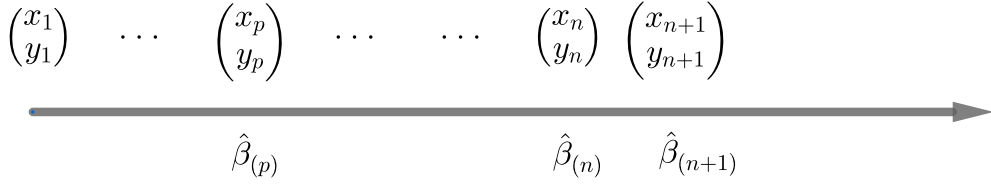
$$1 + x_i^T (X_{[-i]}^T X_{[-i]})^{-1} x_i = (1 - h_{ii})^{-1} = \{1 - x_i^T (X^T X)^{-1} x_i\}^{-1}.$$

This is not an obvious linear algebra identity, but it follows immediately from the two ways of calculating the variance of the predicted residual.

11.3 Applications of the leave-one-out formulas

11.3.1 Gauss updating formula

Consider an online setting in which the data points come sequentially as illustrated by the figure below:



In this setting, we can update the OLS estimator step by step: based on the first n data points $(x_i, y_i)_{i=1}^n$, we calculate the OLS estimator $\hat{\beta}_{(n)}$, and with an additional data point (x_{n+1}, y_{n+1}) , we update the OLS estimator as $\hat{\beta}_{(n+1)}$. These two OLS estimators are closely related as shown in Theorem 11.4 below.

Theorem 11.4 *Let $X_{(n)}$ be the design matrix and $Y_{(n)}$ be the outcome vector for the first n observations. We have*

$$\hat{\beta}_{(n+1)} = \hat{\beta}_{(n)} + \gamma_{(n+1)} \hat{\varepsilon}_{[n+1]},$$

where $\gamma_{(n+1)} = (X_{(n+1)}^T X_{(n+1)})^{-1} x_{n+1}$ and $\hat{\varepsilon}_{[n+1]} = y_{n+1} - x_{n+1}^T \hat{\beta}_{(n)}$ is the predicted residual of the $(n+1)$ th outcome based on the OLS of the first n observations.

Proof of Theorem 11.4: This is the reverse form of the leave-one-out formula in Theorem 11.2. We can view the first $n+1$ data points as the full data, and $\hat{\beta}_{(n)}$ as the OLS estimator leaving the $(n+1)$ th observation out. Applying Theorem 11.2, we have

$$\begin{aligned} \hat{\beta}_{(n)} &= \hat{\beta}_{(n+1)} - (X_{(n+1)}^T X_{(n+1)})^{-1} x_{n+1} \frac{\hat{\varepsilon}_{n+1}}{1 - h_{n+1, n+1}} \\ &= \hat{\beta}_{(n+1)} - \gamma_{(n+1)} \hat{\varepsilon}_{[n+1]}, \end{aligned}$$

where $\hat{\varepsilon}_{n+1}$ is the $(n+1)$ th residual based on the full data OLS, and the $(n+1)$ th predicted residual equals $\hat{\varepsilon}_{[n+1]} = \hat{\varepsilon}_{n+1}/(1 - h_{n+1, n+1})$ based on Theorem 11.3. □

Theorem 11.4 shows that to obtain $\hat{\beta}_{(n+1)}$ from $\hat{\beta}_{(n)}$, the adjustment depends on the predicted residual $\hat{\varepsilon}_{[n+1]}$. If we have a perfect prediction of the $(n+1)$ th observation based on $\hat{\beta}_{(n)}$, then we do not need to make any adjustment to obtain $\hat{\beta}_{(n+1)}$; if the predicted residual is large, then we need to make a large adjustment.

Theorem 11.4 suggests an algorithm for sequentially computing the OLS estimators. But it gives a formula that involves inverting $X_{(n+1)}^T X_{(n+1)}$ at each step. Using the Sherman–Morrison formula in Problem A.3 for updating the inverse of $X_{(n+1)}^T X_{(n+1)}$ based on the inverse of $X_{(n)}^T X_{(n)}$, we have an even simpler algorithm below:

(G1) Start with $V_{(n)} = (X_{(n)}^T X_{(n)})^{-1}$ and $\hat{\beta}_{(n)}$.

(G2) Update

$$V_{(n+1)} = V_{(n)} - \{1 + x_{n+1}^T V_{(n)} x_{n+1}\}^{-1} V_{(n)} x_{n+1} x_{n+1}^T V_{(n)}.$$

(G3) Calculate $\gamma_{(n+1)} = V_{(n+1)} x_{n+1}$ and $\hat{\varepsilon}_{[n+1]} = y_{n+1} - x_{n+1}^T \hat{\beta}_{(n)}$.

(G4) Update $\hat{\beta}_{(n+1)} = \hat{\beta}_{(n)} + \gamma_{(n+1)} \hat{\varepsilon}_{[n+1]}$.

11.3.2 Outlier detection based on residuals

Under the Normal linear model $Y = X\beta + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2 I_n)$, we know some basic probabilistic properties of the residual vector:

$$E(\hat{\varepsilon}) = 0, \quad \text{var}(\hat{\varepsilon}) = \sigma^2(I_n - H).$$

At the same time, the residual vector is computable based on the data. So it is sensible to check whether these properties of the residual vector are plausible based on data, which in turn serves as modeling checking for the Normal linear model.

The first quantity is the standardized residual

$$\text{standr}_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}.$$

We may hope that it has mean 0 and variance 1. However, because of the dependence between $\hat{\varepsilon}_i$ and $\hat{\sigma}^2$, it is not easy to find the exact distribution of standr_i .

The second quantity is the studentized residual based on the predicted residual:

$$\text{studr}_i = \frac{\hat{\varepsilon}_{[-i]}}{\sqrt{\hat{\sigma}_{[-i]}^2/(1 - h_{ii})}} = \frac{y_i - x_i^T \hat{\beta}_{[-i]}}{\sqrt{\hat{\sigma}_{[-i]}^2/(1 - h_{ii})}},$$

where $\hat{\beta}_{[-i]}$ and $\hat{\sigma}_{[-i]}^2$ are the estimates of the coefficient and variance based on the leave- i -out OLS. Because $(y_i, \hat{\beta}_{[-i]}, \hat{\sigma}_{[-i]}^2)$ are mutually independent under the Normal linear model, we can show that

$$\text{studr}_i \sim t_{n-p-1}. \quad (11.7)$$

I leave the rigorous proof of (11.7) as Problem 11.2. Because we know the distribution of studr_i , we can compare it with the quantiles of the t distribution.

The third quantity is Cook's distance (Cook, 1977):

$$\begin{aligned} \text{cook}_i &= (\hat{\beta}_{[-i]} - \hat{\beta})^T X^T X (\hat{\beta}_{[-i]} - \hat{\beta}) / (p\hat{\sigma}^2) \\ &= (X\hat{\beta}_{[-i]} - X\hat{\beta})^T (X\hat{\beta}_{[-i]} - X\hat{\beta}) / (p\hat{\sigma}^2), \end{aligned}$$

where the first form measures the change of the OLS estimator and the second form measures the change in the predicted values based on leaving- i -out. It has a slightly different motivation, but eventually, it is related to the previous two residuals due to the leave-one-out formulas.

Theorem 11.5 *Cook's distance is related to the standardized residual via:*

$$\text{cook}_i = \text{standr}_i^2 \times \frac{h_{ii}}{p(1 - h_{ii})}.$$

I leave the proof of Theorem 11.5 as Problem 11.4.

I will end this subsection with two examples. The first one is simulated. I generate data from a univariate Normal linear model without outliers. I then use `hatvalues`, `r.standard`, `r.student` and `cooks.distance` to an `lm.object` to calculate the leverage scores, standardized residuals, studentized residuals, and Cook's distances. Their plots are in the first column of Figure 11.1.

```
n = 100
x = seq(0, 1, length = n)
y = 1 + 3*x + rnorm(n)
lmmod = lm(y ~ x)
hatvalues(lmmod)
rstandard(lmmod)
rstudent(lmmod)
cooks.distance(lmmod)
```

If I add 8 to the outcome of the last observation, the plots change to the second column of Figure 11.1. If I add 8 to the 50th observation, the plots change to the last column of Figure 11.1. Both visually show the outliers. In this example, the three residual plots give qualitatively the same pattern, so the choice among them does not matter much. In general cases, we may prefer studr_i because it has a known distribution under the Normal linear model.

The second one is a further analysis of the Lalonde data. Based on the plots in Figure 11.2, there are indeed some outliers in the data. It is worth investigating them more carefully.

Although the outliers detection methods above are very classic, they are rarely implemented in modern data analyses. They are simple and useful diagnostics. I recommend using them at least as a part of the exploratory data analysis.

11.3.3 Jackknife

Jackknife is a general strategy for bias reduction and variance estimation proposed by Quenouille (1949, 1956) and popularized by Tukey (1958). Based on independent data (Z_1, \dots, Z_n) , how to estimate the variance of a general estimator $\hat{\theta}(Z_1, \dots, Z_n)$ of the parameter θ ? Define $\hat{\theta}_{[-i]}$ as the estimator without observation i , and the pseudo-value as

$$\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{[-i]}.$$

The jackknife point estimator is $\hat{\theta}_j = n^{-1} \sum_{i=1}^n \tilde{\theta}_i$, and the jackknife variance estimator is

$$\hat{V}_j = \frac{1}{n(n-1)} \sum_{i=1}^n (\tilde{\theta}_i - \hat{\theta}_j)(\tilde{\theta}_i - \hat{\theta}_j)^T.$$

Where does the idea of jackknife come from? Honestly, I do not know. However, as a sanity check, we can test whether it makes sense in OLS. See details below.

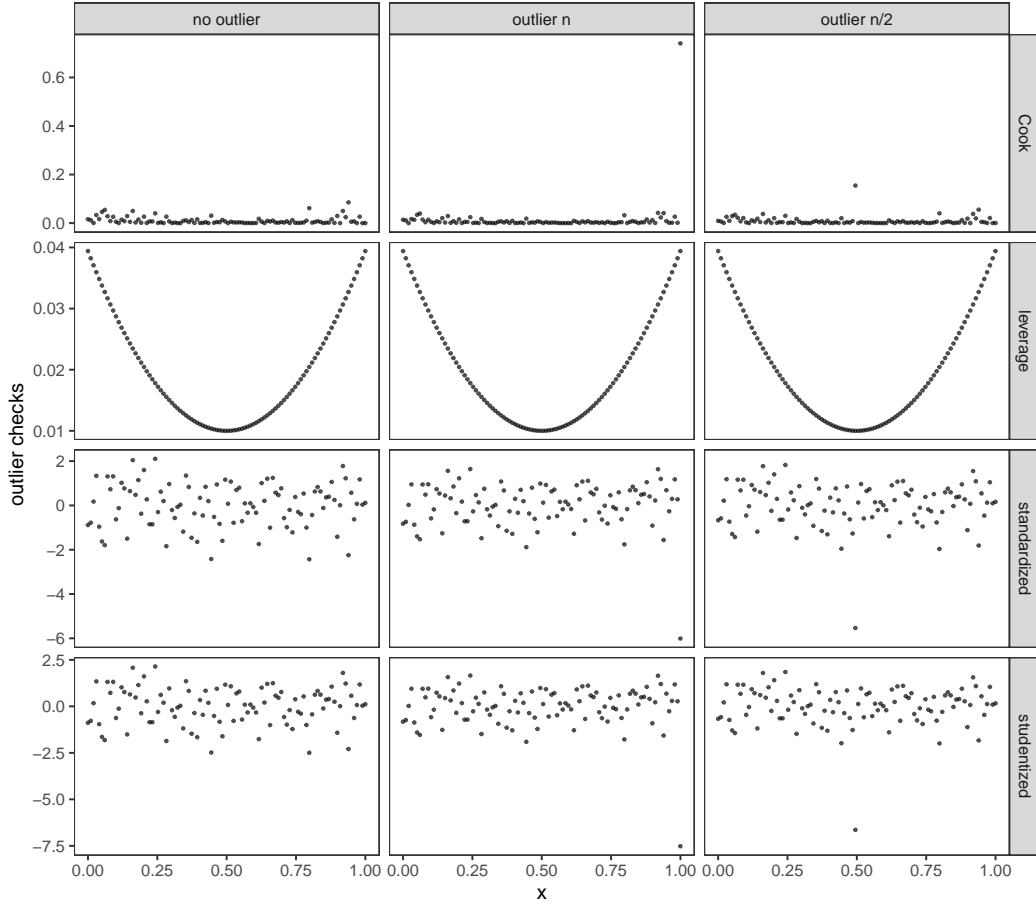


FIGURE 11.1: Outlier detections

We have already shown that the OLS coefficient is unbiased and derived several variance estimators for it. Here we focus on the jackknife in OLS using the leave-one-out formula for the coefficient. The pseudo-value is

$$\begin{aligned}
 \tilde{\beta}_i &= n\hat{\beta} - (n-1)\hat{\beta}_{[-i]} \\
 &= n\hat{\beta} - (n-1) \left\{ \hat{\beta} - (1-h_{ii})^{-1}(X^T X)^{-1}x_i\hat{\varepsilon}_i \right\} \\
 &= \hat{\beta} + (n-1)(1-h_{ii})^{-1}(X^T X)^{-1}x_i\hat{\varepsilon}_i.
 \end{aligned}$$

The jackknife point estimator is

$$\hat{\beta}_J = \hat{\beta} + \frac{n-1}{n} \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n x_i \frac{\hat{\varepsilon}_i}{1-h_{ii}} \right).$$

It is a little unfortunate that the jackknife point estimator is not identical to the OLS estimator, which is BLUE under the Gauss–Markov model. We can show that $E(\hat{\beta}_J) = \beta$ and it is a linear estimator. So the Gauss–Markov theorem ensures that $\text{cov}(\hat{\beta}_J) \succeq \text{cov}(\hat{\beta})$. (Readers, make sure these two sentences make sense to you; see Problem 11.6.)

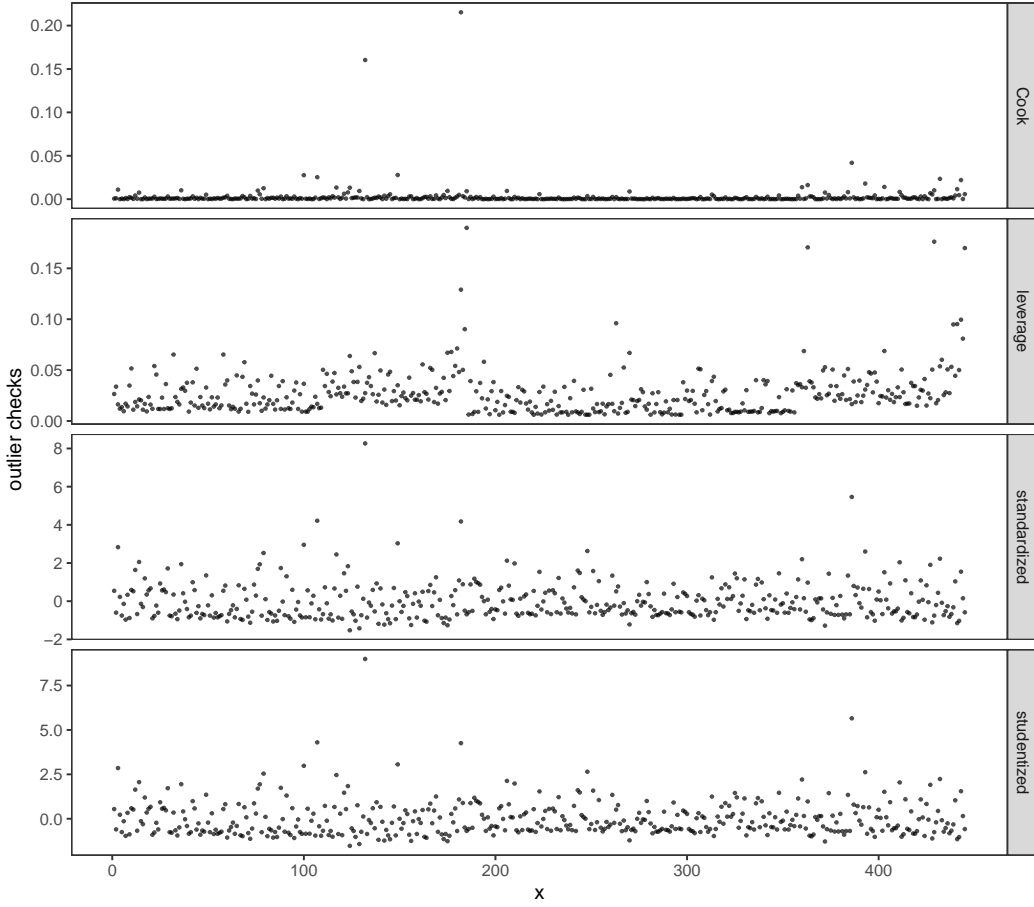


FIGURE 11.2: Outlier detections in the LaLonde data

Nevertheless, the difference between $\hat{\beta}_j$ and $\hat{\beta}$ is quite small. I omit their difference in the following derivation. Assuming that $\hat{\beta}_j \cong \hat{\beta}$, we can continue to calculate the approximate jackknife variance estimator:

$$\begin{aligned} \hat{V}_j &\cong \frac{1}{n(n-1)} \sum_{i=1}^n (\tilde{\beta}_i - \hat{\beta})(\tilde{\beta}_i - \hat{\beta})^T \\ &= \frac{n-1}{n} (X^T X)^{-1} \sum_{i=1}^n \left(\frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right)^2 x_i x_i^T (X^T X)^{-1}, \end{aligned}$$

which is almost identical to the HC3 form of the EHW covariance matrix introduced in Chapter 6.4.2. Miller (1974) first analyzed the jackknife in OLS but dismissed it immediately. Hinkley (1977) modified the original jackknife and proposed a version that is identical to HC1, and Wu (1986) proposed some further modifications and proposed a version that is identical to HC2. Weber (1986) made connections between EHW and jackknife standard errors. However, Long and Ervin (2000)'s finite-sample simulation seems to favor the original jackknife or HC3.

11.4 Homework problems

11.1 Implementing the Gauss updating formula

Implement the algorithm in (G1)–(G4), and try it on simulated data.

11.2 The distribution of the studentized residual

Prove (11.7).

11.3 Leave-one-out coefficient

Prove

$$\hat{\beta} = \sum_{i=1}^n w_i \hat{\beta}_{[-i]},$$

and find the weights w_i 's. Prove that they are positive and sum to one. Does $\hat{\beta} = n^{-1} \sum_{i=1}^n \hat{\beta}_{[-i]}$ hold in general?

11.4 Cook's distance and the standardized residual

Prove Theorem 11.5.

11.5 The relationship between the standardized and studentized residual

Prove Theorem 11.6 below.

Theorem 11.6 *We have*

1. $(n-p-1)\hat{\sigma}_{[-i]}^2 = (n-p)\hat{\sigma}^2 - \hat{\varepsilon}_i^2/(1-h_{ii})$.
2. *There is a monotone relationship between the standardized and studentized residual:*

$$studr_i = standr_i \sqrt{\frac{n-p-1}{n-p-standr_i^2}}.$$

Remark: The right-hand side of the first result must be nonnegative so $\sum_{k=1}^n \hat{\varepsilon}_k^2 \geq \hat{\varepsilon}_i^2/(1-h_{ii})$, which implies the following inequality:

$$h_{ii} + \frac{\hat{\varepsilon}_i^2}{\sum_{k=1}^n \hat{\varepsilon}_k^2} \leq 1.$$

From this inequality, if $h_{ii} = 1$ then $\hat{\varepsilon}_i = 0$ which further implies that $h_{ij} = 0$ for all $j \neq i$.

11.6 More on the jackknife estimator in OLS

Prove that under the Gauss–Markov model, $E(\hat{\beta}_j) = \beta$ and $\text{cov}(\hat{\beta}_j) \succeq \text{cov}(\hat{\beta})$.

11.7 Subset and full-data OLS coefficients

Leaving one observation out, we have the OLS coefficient formula in Theorem 11.2. Leave a subset of observations out, we have the OLS coefficient formula below. Partition the covariate matrix and outcome vector based on a subset S of $\{1, \dots, n\}$:

$$X = \begin{pmatrix} X_S \\ X_{\setminus S} \end{pmatrix}, \quad Y = \begin{pmatrix} Y_S \\ Y_{\setminus S} \end{pmatrix}.$$

Without using the observations in set S , we have the OLS coefficient

$$\hat{\beta}_{\setminus S} = (X_{\setminus S}^T X_{\setminus S})^{-1} X_{\setminus S}^T Y_{\setminus S}.$$

The corresponding leave- S -out residual vector is

$$\hat{\varepsilon}_{\setminus S} = Y_S - X_S \hat{\beta}_{\setminus S}.$$

Theorem 11.7 below facilitates the computation of many $\hat{\beta}_{\setminus S}$'s and $\hat{\varepsilon}_{\setminus S}$'s simultaneously, without running the OLS for each S . It relies crucially on the subvector of the residuals

$$\hat{\varepsilon}_S = Y_S - X_S \hat{\beta}$$

and the submatrix of H

$$H_{SS} = X_S (X^T X)^{-1} X_S^T$$

corresponding to the set S . Prove Theorem 11.7.

Theorem 11.7 Assume $X^T X$ and $I - H_{SS}$ are both invertible, where I is the identity matrix with the same dimension as H_{SS} . Recall that $\hat{\beta}$ is the full data OLS. We have

$$\hat{\beta}_{\setminus S} = \hat{\beta} - (X^T X)^{-1} X_S^T (I - H_{SS})^{-1} \hat{\varepsilon}_S$$

and

$$\hat{\varepsilon}_{\setminus S} = (I - H_{SS})^{-1} \hat{\varepsilon}_S.$$

11.8 Gauss updating formula with batches of data

This problem extends the discussion in Chapter 11.3.1. Assume the data come in batches, with $X_b \in \mathbb{R}^{n_b \times p}$ and $Y_b \in \mathbb{R}^{n_b}$ for $b = 1, \dots, B, B+1$. Derive the Gauss-type updating formula and design an algorithm to compute the OLS coefficients efficiently.

11.9 Bounding the leverage score

With the intercept included in the OLS, Theorem 11.1 shows $n^{-1} \leq h_{ii} \leq 1$ for all $i = 1, \dots, n$. Prove the following stronger result:

$$n^{-1} \leq h_{ii} \leq s_i^{-1}$$

where s_i is the number of rows that are identical to x_i or $-x_i$.

11.10 More on the leverage score

Prove Theorem 11.8 below.

Theorem 11.8 $\det(X_{[-i]}^T X_{[-i]}) = (1 - h_{ii}) \det(X^T X)$.

Remark: If $h_{ii} = 1$, then $X_{[-i]}^T X_{[-i]}$ is degenerate with determinant 0. Therefore, if we delete an observation i with leverage score 1, the columns of the covariate matrix become linearly dependent.

12

Population Ordinary Least Squares and Misspecified Linear Model

Previous chapters assume fixed X and random Y . We can also view each observation (x_i, y_i) as IID draws from a population and discuss population OLS. The population OLS allows us to achieve the following goals:

- (G1) We can view the OLS from the level of random variables instead of data points.
- (G2) The population OLS facilitates the discussion of the properties of misspecified linear models.
- (G3) The population OLS motivates a prediction procedure called *conformal prediction* without imposing any distributional assumptions.

12.1 Population OLS

Assume that $(x_i, y_i)_{i=1}^n$ are IID with the same distribution as (x, y) , where $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$. Below I will use (x, y) to denote a general observation, dropping the subscript i for simplicity. Let the joint distribution be $F(x, y)$, and $E(\cdot)$, $\text{var}(\cdot)$, and $\text{cov}(\cdot)$ be the expectation, variance, and covariance under this joint distribution. We want to use x to predict y . Theorem 12.1 below states that the conditional expectation $E(y | x)$ is the best predictor in terms of the mean squared error.

Theorem 12.1 *For any function $m(x)$, we have the decomposition*

$$E \left[\{y - m(x)\}^2 \right] = E \left[\{E(y | x) - m(x)\}^2 \right] + E\{\text{var}(y | x)\}, \quad (12.1)$$

provided the existence of the moments in (12.1). The decomposition (12.1) implies

$$E(y | x) = \arg \min_{m(\cdot)} E \left[\{y - m(x)\}^2 \right]$$

with the minimum value equaling $E\{\text{var}(y | x)\}$, the expectation of the conditional variance of y given x .

Theorem 12.1 is well known in probability theory. I relegate its proof as Problem 12.1. We have finite data points, but the function $E(y | x)$ lies in an infinite dimensional space. Nonparametric estimation of $E(y | x)$ is generally a hard problem, especially with a multidimensional x . As a starting point, we often use a linear function of x to approximate $E(y | x)$ and define the population OLS coefficient as

$$\beta = \arg \min_{b \in \mathbb{R}^p} \mathcal{L}(b),$$

where

$$\begin{aligned}\mathcal{L}(b) &= E \{ (y - x^T b)^2 \} \\ &= E \{ y^2 + b^T x x^T b - 2y x^T b \} \\ &= E(y^2) + b^T E(x x^T) b - 2E(y x^T) b\end{aligned}$$

is a quadratic function of b . From the first-order condition, we have

$$\frac{\partial \mathcal{L}(b)}{\partial b} \Big|_{b=\beta} = 2E(x x^T) \beta - 2E(y x) = 0$$

based on Proposition A.9 in Appendix A, so

$$\beta = \{E(x x^T)\}^{-1} E(y x), \quad (12.2)$$

if $E(x x^T)$ is non-degenerate. From the second-order condition,

$$\frac{\partial^2 \mathcal{L}(b)}{\partial b \partial b^T} = 2E(x x^T)$$

is positive definite. So β is the unique minimizer of $\mathcal{L}(b)$.

The above derivation shows that $x^T \beta$ is the best linear predictor. Theorem 12.2 below states that $x^T \beta$ is the best linear approximation to the possibly nonlinear conditional mean $E(y | x)$.

Theorem 12.2 *If $E(x x^T)$ is non-degenerate, then*

$$\begin{aligned}\beta &= \arg \min_{b \in \mathbb{R}^p} E \left[\{E(y | x) - x^T b\}^2 \right] \\ &= \{E(x x^T)\}^{-1} E \{x E(y | x)\}.\end{aligned}$$

As a special case, when the covariate “ x ” in the above formulas contains 1 and a scalar x , the OLS coefficient has the following form.

Corollary 12.1 *For scalar x and y , we have*

$$(\alpha, \beta) = \arg \min_{a, b} E(y - a - bx)^2,$$

where $\alpha = E(y) - E(x)\beta$ and

$$\beta = \frac{\text{cov}(x, y)}{\text{var}(x)} = \rho_{xy} \sqrt{\frac{\text{var}(y)}{\text{var}(x)}}$$

Recall that

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}}$$

is the population Pearson correlation coefficient. So Corollary 12.1 gives the population version of the Galtonian formula. I leave the proofs of Theorem 12.2 and Corollary 12.1 as Problems 12.2 and 12.3.

With β , we can define

$$\varepsilon = y - x^T \beta \quad (12.3)$$

as the population residual. Because we usually do not use the upper-case letter E for ε , this notation may cause confusion with previous discussion on OLS, where ε denotes the vector of the error terms. Here ε is a scalar in (12.3). By the definition of β , we can verify

$$E(x\varepsilon) = E\{x(y - x^\top\beta)\} = E(xy) - E(xx^\top)\beta = 0. \quad (12.4)$$

If we include 1 as a component of x , then $E(\varepsilon) = 0$, which, coupled with (12.4), implies $\text{cov}(x, \varepsilon) = 0$. So with an intercept in β , the mean of the population residual must be zero, and it is uncorrelated with other covariates by construction.

We can also rewrite (12.3) as

$$y = x^\top\beta + \varepsilon, \quad (12.5)$$

which holds by the definition of the population OLS coefficient and residual without any modeling assumption. We call (12.5) with (12.2) and (12.3) the *population OLS decomposition*.

12.2 Population FWL Theorem and Cochran's formula

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 . With

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

define the population OLS of y on x_1 and x_2 as

$$y = \beta^\top x = \beta_1^\top x_1 + \beta_2^\top x_2 + \varepsilon. \quad (12.6)$$

To aid the interpretation of the population OLS coefficient, we have the following population FWL theorem.

Theorem 12.3 (Population FWL Theorem) *Consider the population OLS decomposition (12.6). Define \tilde{x}_2 as the residual of the component-wise¹ population OLS of x_2 on x_1 :*

$$\tilde{x}_2 = x_2 - E(x_2 x_1^\top) E(x_1 x_1^\top)^{-1} x_1.$$

Define \tilde{y} as the residual of the population OLS of y on x_1 :

$$\tilde{y} = y - E(y x_1^\top) E(x_1 x_1^\top)^{-1} x_1.$$

Then the coefficient β_2 has the following equivalent forms:

$$\beta_2 = [E(xx^\top)^{-1} E(xy)]_{\text{the last } l \text{ components}} \quad (12.7)$$

$$= E(\tilde{x}_2 \tilde{x}_2^\top)^{-1} E(\tilde{x}_2 y) \quad (12.8)$$

$$= E(\tilde{x}_2 \tilde{x}_2^\top)^{-1} E(\tilde{x}_2 \tilde{y}). \quad (12.9)$$

The form (12.7) is the definition of β_2 from the population OLS (12.6). The form (12.8) states that β_2 equals the population OLS coefficient of y on \tilde{x}_2 . The form (12.9) states that β_2 equals the population OLS coefficient of \tilde{y} on \tilde{x}_2 . Angrist and Pischke (2008) provide a special case of Theorem 12.3 with $l = 1$.

¹If x_2 is a vector, we run population OLS of each component of x_2 on x_1 to obtain the residual. Vectorize the residuals to obtain \tilde{x}_2 .

Proof of Theorem 12.3: Introduce the notation:

$$E \left(\begin{pmatrix} y \\ x_1 \\ x_2 \end{pmatrix} \begin{pmatrix} y & x_1^T & x_2^T \end{pmatrix} \right) = \begin{pmatrix} E(y^2) & E(yx_1^T) & E(yx_2^T) \\ E(x_1y) & E(x_1x_1^T) & E(x_1x_2^T) \\ E(x_2y) & E(x_2x_1^T) & E(x_2x_2^T) \end{pmatrix} = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} & \Sigma_{02} \\ \Sigma_{10} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{20} & \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

By definition of the population OLS of y on x_1 and x_2 , we have

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{10} \\ \Sigma_{20} \end{pmatrix}.$$

Use the first form of the inverse of 2×2 block matrix in Problem A.3 to obtain

$$\begin{aligned} \beta_2 &= \begin{pmatrix} * & * \\ -\Sigma_{22|1}^{-1} \Sigma_{21} \Sigma_{11}^{-1} & \Sigma_{22|1}^{-1} \end{pmatrix} \begin{pmatrix} \Sigma_{10} \\ \Sigma_{20} \end{pmatrix} \\ &= -\Sigma_{22|1}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{10} + \Sigma_{22|1}^{-1} \Sigma_{20} \\ &= \Sigma_{22|1}^{-1} (\Sigma_{20} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{10}), \end{aligned}$$

where $\Sigma_{22|1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$ and $*$ signifies unimportant terms.

Now we consider the population OLS coefficient of y on \tilde{x}_2 :

$$\begin{aligned} &E(\tilde{x}_2 \tilde{x}_2^T)^{-1} E(\tilde{x}_2 y) \\ &= [E\{(x_2 - \Sigma_{21} \Sigma_{11}^{-1} x_1)(x_2^T - x_1^T \Sigma_{11}^{-1} \Sigma_{12})\}]^{-1} E\{(x_2 - \Sigma_{21} \Sigma_{11}^{-1} x_1)y\} \\ &= (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\Sigma_{20} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{10}) \\ &= \Sigma_{22|1}^{-1} (\Sigma_{20} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{10}) \\ &= \beta_2. \end{aligned}$$

Finally, we consider the population OLS coefficient of \tilde{y} on \tilde{x}_2 :

$$\begin{aligned} &E(\tilde{x}_2 \tilde{x}_2^T)^{-1} E(\tilde{x}_2 \tilde{y}) \\ &= [E\{(x_2 - \Sigma_{21} \Sigma_{11}^{-1} x_1)(x_2^T - x_1^T \Sigma_{11}^{-1} \Sigma_{12})\}]^{-1} E\{(x_2 - \Sigma_{21} \Sigma_{11}^{-1} x_1)(y - x_1^T \Sigma_{11}^{-1} \Sigma_{10})\} \\ &= (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\Sigma_{20} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{10}) \\ &= \Sigma_{22|1}^{-1} (\Sigma_{20} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{10}) \\ &= \beta_2. \end{aligned}$$

□

We also have a population version of Cochran's formula. We have the following OLS decompositions of random variables

$$y = \beta_1^T x_1 + \beta_2^T x_2 + \varepsilon, \quad (12.10)$$

$$y = \tilde{\beta}_2^T x_2 + \tilde{\varepsilon}, \quad (12.11)$$

$$x_1 = \delta^T x_2 + u. \quad (12.12)$$

Equation (12.10) defines the population long regression, and Equation (12.11) defines the population short regression. In Equation (12.12), δ is a $l \times k$ matrix because it is the OLS decomposition of a vector on a vector. We can view (12.12) as OLS decomposition of each component of x_{i1} on x_{i2} . Theorem 12.4 below states the population version of Cochran's formula.

Theorem 12.4 (population Cochran's formula) *Based on (12.10)–(12.12), we have*

$$\tilde{\beta}_2 = \beta_2 + \delta \beta_1.$$

I relegate the proof of Theorem 12.4 as Problem 12.5.

12.3 Population R^2 and partial correlation coefficient

Exclude 1 from x and assume $x \in \mathbb{R}^{p-1}$ in this subsection. Assume that covariates and outcome are centered with mean zero and covariance matrix

$$\text{cov} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \sigma_y^2 \end{pmatrix}.$$

There are multiple equivalent definitions of R^2 . I start with the following definition

$$R^2 = \frac{\Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}}{\sigma_y^2}, \quad (12.13)$$

and will give several equivalent definitions below. Let β be the population OLS coefficient of y on x , and $\hat{y} = x^T \beta$ be the best linear predictor.

Theorem 12.5 *The R^2 defined in (12.13) has the following equivalent forms:*

$$R^2 = \frac{\text{var}(\hat{y})}{\text{var}(y)} \quad (12.14)$$

$$= \max_{b \in \mathbb{R}^{p-1}} \rho^2(y, x^T b) \quad (12.15)$$

$$= \rho^2(y, \hat{y}). \quad (12.16)$$

The form (12.14) states that R^2 equals the ratio of the variance of the best linear predictor of y and the variance of y itself. The form (12.15) states that R^2 equals the maximum value of the squared Pearson correlation coefficient between y and a linear combination of x , over all possible linear combinations of x . The form (12.16) states that R^2 equals the squared Pearson correlation coefficient between y and the best linear predictor of y .

Proof of Theorem 12.5: I first prove (12.14). Because of the centering of x , we have $\beta = \Sigma_{xx}^{-1} \Sigma_{xy}$ and

$$\begin{aligned} \text{var}(\hat{y}) &= \beta^T \Sigma_{xx} \beta \\ &= \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xx} \Sigma_{xx}^{-1} \Sigma_{xy} \\ &= \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}. \end{aligned}$$

Therefore, $\text{var}(\hat{y})/\text{var}(y) = R^2$.

I then prove (12.15). We have

$$\rho^2(y, x^T b) = \frac{\text{cov}^2(y, x^T b)}{\text{var}(y) \text{var}(x^T b)} = \frac{b^T \Sigma_{xy} \Sigma_{yx} b}{\sigma_y^2 \times b^T \Sigma_{xx} b}.$$

Define $\gamma = \Sigma_{xx}^{1/2} b$ and $b = \Sigma_{xx}^{-1/2} \gamma$ such that b and γ have one-to-one mapping. So the maximum value of

$$\sigma_y^2 \times \rho^2(y, x^T b) = \frac{\gamma^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yx} \Sigma_{xx}^{-1/2} \gamma}{\gamma^T \gamma}$$

equals the maximum eigenvalue of $\Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yx} \Sigma_{xx}^{-1/2}$, attained when γ equals the corresponding eigenvector; see Theorem A.4 in Appendix A. The matrix $\Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yx} \Sigma_{xx}^{-1/2}$ is

positive semi-definite and has rank one due to Proposition A.3 in Appendix A, so it has exactly one non-zero eigenvalue which must equal its trace. So

$$\begin{aligned}
 \max_{b \in \mathbb{R}^{p-1}} \rho^2(y, x^T b) &= \sigma_y^{-2} \text{trace}(\Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yx} \Sigma_{xx}^{-1/2}) \\
 &= \sigma_y^{-2} \text{trace}(\Sigma_{xy} \Sigma_{yx} \Sigma_{xx}^{-1/2} \Sigma_{xx}^{-1/2}) \\
 &= \sigma_y^{-2} \text{trace}(\Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}) \\
 &= \sigma_y^{-2} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \\
 &= R^2.
 \end{aligned}$$

I finally prove (12.16). We have

$$\begin{aligned}
 \rho^2(y, \hat{y}) &= \frac{\text{cov}(y, \hat{y})^2}{\text{var}(y) \text{var}(\hat{y})} \\
 &= \frac{\beta^T \Sigma_{xy} \Sigma_{yx} \beta}{\sigma_y^2 \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}},
 \end{aligned}$$

by the results in the proofs of (12.15) and (12.15). Use the formula $\beta = \Sigma_{xx}^{-1} \Sigma_{xy}$ to further simplify the above expression as

$$\begin{aligned}
 \rho^2(y, \hat{y}) &= \frac{\Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}}{\sigma_y^2 \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}} \\
 &= \frac{\Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}}{\sigma_y^2} \\
 &= R^2.
 \end{aligned}$$

□

We can also define population partial correlation and R^2 . For scalar y and x with another scalar or vector w , we can define the population OLS decomposition based on $(1, w)$ as

$$y = \hat{y} + \tilde{y}, \quad x = \hat{x} + \tilde{x}, \quad (12.17)$$

where

$$\tilde{y} = \{y - E(y)\} - \{w - E(w)\}^T \beta_y, \quad \tilde{x} = \{x - E(x)\} - \{w - E(w)\}^T \beta_x,$$

with β_y and β_x being the coefficients of w in these population OLS. We then define the population partial correlation coefficient as

$$\rho_{yx|w} = \rho_{\tilde{y}\tilde{x}}.$$

If the marginal correlation and partial correlation have different signs, then we have Simpson's paradox at the population level.

With a scalar w , we have a more explicit formula below.

Theorem 12.6 *For scalar (y, x, w) , we have*

$$\rho_{yx|w} = \frac{\rho_{yx} - \rho_{xw} \rho_{yw}}{\sqrt{1 - \rho_{xw}^2} \sqrt{1 - \rho_{yw}^2}}.$$

I relegate the proof of Theorem 12.6 as Problem 12.7.

We can also extend the definition of $\rho_{yx|w}$ to the partial R^2 with a scalar y and possibly vector x and w . The population OLS decompositions (12.17) still hold in the general case. Then we can define the partial R^2 between y and x given w as the R^2 between \tilde{y} and \tilde{x} :

$$R_{y.x|w}^2 = R_{\tilde{y}\tilde{x}}^2.$$

12.4 Inference for the population OLS

12.4.1 Inference with the Eicker–Huber–White standard errors

Based on the IID data $(x_i, y_i)_{i=1}^n$, we can obtain the moment estimator for the population OLS coefficient

$$\hat{\beta} = \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n x_i y_i \right),$$

and the residuals $\hat{\varepsilon}_i = y_i - x_i^T \hat{\beta}$. Assume finite fourth moments of (x, y) . We can use the law of large numbers to show that

$$\begin{aligned} n^{-1} \sum_{i=1}^n x_i x_i^T &\rightarrow E(xx^T), \\ n^{-1} \sum_{i=1}^n x_i y_i &\rightarrow E(xy), \end{aligned}$$

so $\hat{\beta} \rightarrow \beta$ in probability. We can use the central limit theorem (CLT) to show that $n^{-1/2} \sum_{i=1}^n x_i \varepsilon_i \rightarrow N(0, M)$ in distribution, where $M = E(\varepsilon^2 x x^T)$, so

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, V) \quad (12.18)$$

in distribution, where $V = B^{-1} M B^{-1}$ with $B = E(xx^T)$. The moment estimator for the asymptotic variance of $\hat{\beta}$ is again the Eicker–Huber–White (EHW) robust covariance estimator:

$$\hat{V}_{\text{EHW}} = n^{-1} \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i^T \right) \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1}. \quad (12.19)$$

Following the almost the same proof of Theorem 6.3, we can show that \hat{V}_{EHW} is consistent for the asymptotic covariance V . I summarize the formal results below, with the proof relegated as Problem 12.4.

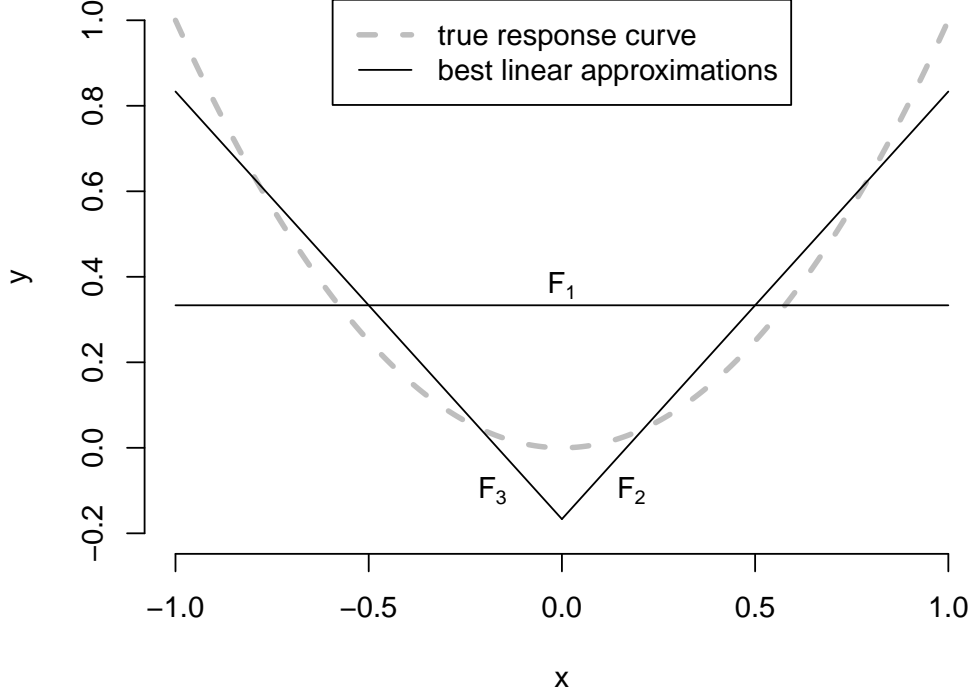
Theorem 12.7 Assume that $(x_i, y_i)_{i=1}^n \stackrel{\text{IID}}{\sim} (x, y)$ with $E(\|x\|^4) < \infty$ and $E(y^4) < \infty$. We have (12.18) and $n\hat{V}_{\text{EHW}} \rightarrow V$ in probability.

So the EHW standard error is not only robust to the heteroskedasticity of the errors but also robust to the misspecification of the linear model (Huber, 1967; White, 1980b; Angrist and Pischke, 2008; Buja et al., 2019a). Of course, when the linear model is wrong, we need to modify the interpretation of β : it is the coefficient of x in the best linear prediction of y or the best linear approximation of the conditional mean function $E(y | x)$.

12.5 To model or not to model?

12.5.1 Population OLS and the restricted mean model

I start with a simple example. In the following calculation, I will use the fact that the k th moment of a Uniform(0, 1) random variable equals $1/(k+1)$; see Problem B.1.

FIGURE 12.1: Best linear approximations correspond to three different distributions of x .

Example 12.1 Assume that $x \sim F(x)$, $\varepsilon \sim N(0, 1)$, $x \perp \varepsilon$, and $y = x^2 + \varepsilon$.

1. If $x \sim F_1(x)$ is $\text{Uniform}(-1, 1)$, then the best linear approximation is $1/3 + 0 \cdot x$. We can see this result from

$$\beta(F_1) = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{\text{cov}(x, x^2)}{\text{var}(x)} = \frac{E(x^3)}{\text{var}(x)} = 0,$$

and $\alpha(F_1) = E(y) = E(x^2) = 1/3$.

2. If $x \sim F_2(x)$ is $\text{Uniform}(0, 1)$, then the best linear approximation is $-1/6 + x$. We can see this result from

$$\beta(F_2) = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{\text{cov}(x, x^2)}{\text{var}(x)} = \frac{E(x^3) - E(x)E(x^2)}{E(x^2) - \{E(x)\}^2} = \frac{1/4 - 1/2 \times 1/3}{1/3 - (1/2)^2} = 1,$$

and $\alpha(F_2) = E(y) - \beta E(x) = E(x^2) - E(x) = 1/3 - 1/2 = -1/6$

3. If $x \sim F_3(x)$ is $\text{Uniform}(-1, 0)$, then the best linear approximation is $-1/6 - x$. This result holds by symmetry.

Figure 12.1 shows the true conditional mean function x^2 and the best linear approximations. As highlighted in the notation above, the best linear approximation depends on the distribution of x .

From Example 12.1, we can see that the best linear approximation depends on the distribution of X . This complicates the interpretation of β from the population OLS decomposition (Buja et al., 2019a). Consequently, this can cause problems for *external validity* because β will be different in a future environment with different distribution of X . Sims (2010, page 66) pointed this out as a critique of Angrist and Pischke (2008).

To ensure the stability of β across different environments, we often invoke the following *restricted mean model*.

Assumption 12.1 (restricted mean model) *Assume*

$$E(y | x) = x^T \beta \quad (12.20)$$

or, equivalently,

$$y = x^T \beta + \varepsilon, \quad E(\varepsilon | x) = 0.$$

Assumption 12.1 restricts the conditional mean of y given x to be linear in x , justifying the name “restricted mean model.” Nevertheless, Assumption 12.1 imposes weak assumptions on the distribution of y or ε . Under Assumption 12.1, the population OLS coefficient equals the true parameter in the restricted mean model:

$$\begin{aligned} \{E(xx^T)\}^{-1} E(xy) &= \{E(xx^T)\}^{-1} E\{xE(y | x)\} \\ &= \{E(xx^T)\}^{-1} E(xx^T \beta) \\ &= \beta. \end{aligned}$$

Moreover, the population OLS coefficient does not depend on the distribution of x . The above asymptotic inference applies to this model too.

Freedman (1981) distinguished two types of OLS as shown in Figure 12.2:

- (M1) the *regression model*, as shown in left-hand side of Figure 12.2;
- (M2) the *correlation model*, as shown in right-hand side of Figure 12.2.

In the regression model, we first generate x and ε under some restrictions, for example, $E(\varepsilon | x) = 0$, and then generate the outcome based on $y = x^T \beta + \varepsilon$, a linear function of x with error ε . In the correlation model, we start with a pair (x, y) , then decompose y into the best linear predictor $x^T \beta$ and the leftover residual ε . Compare the subtle difference between the ε in the regression model and the correlation model. The regression model requires $E(\varepsilon | x) = 0$, whereas the correlation model ensures $E(\varepsilon x) = 0$. The regression model imposes a stronger assumption because $E(\varepsilon | x) = 0$ implies

$$E(\varepsilon x) = E\{E(\varepsilon x | x)\} = E\{E(\varepsilon | x)x\} = 0.$$

12.5.2 More on residual plots

Most standard statistical theory for inference assumes a correctly specified linear model (e.g., Gauss–Markov model, Normal linear model, or restricted mean model). However, the corresponding inferential procedures are often criticized since it is challenging to ensure that the model is correctly specified. Alternatively, we can argue that without assuming the linear model, the OLS estimator is consistent for the coefficient in the best linear approximation of the conditional mean function $E(y | x)$, which is often a meaningful quantify even the linear model is misspecified. This can be misleading. Example 12.1 shows that the best linear approximation can be a bad approximation to a nonlinear conditional mean function, and it depends on the distribution of the covariates.

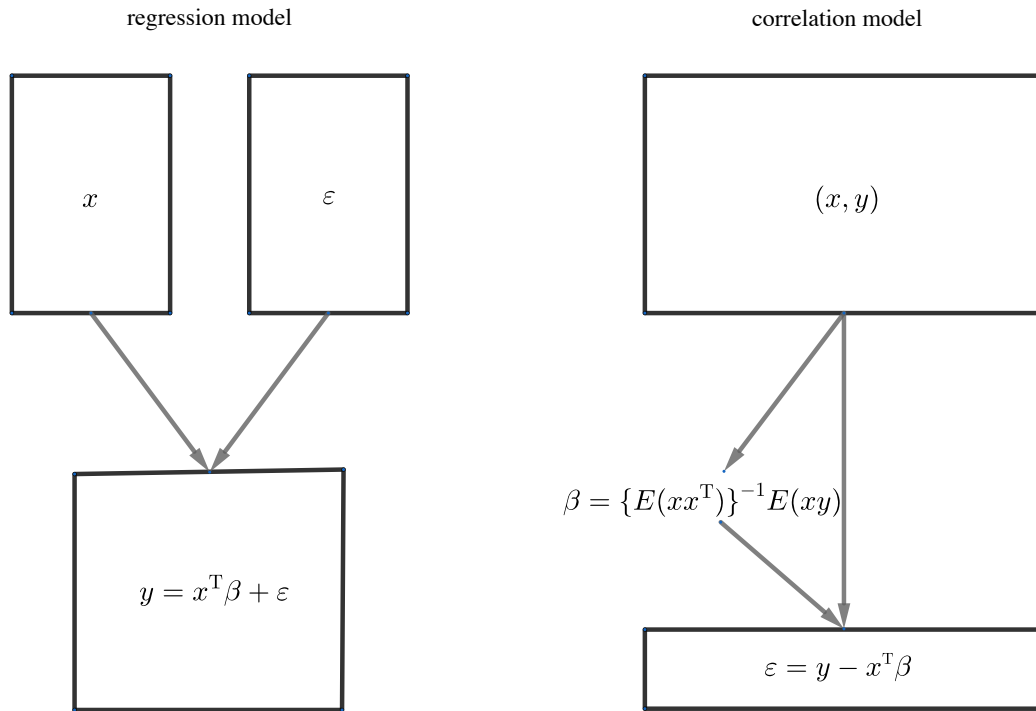


FIGURE 12.2: Freedman's classification of OLS

A classic statistical approach is to check whether the residual $\hat{\varepsilon}_i$ has any nonlinear trend with respect to the covariates. With only a few covariates, we can plot the residual against each covariate; with many covariates, we can plot the residual against the fitted value \hat{y}_i . Figure 12.3 gives four examples. In these examples, the covariates are the same:

```
n = 200
x1 = rexp(n)
x2 = runif(n)
```

The outcome models differ:

- (Y1) linear homoskedastic: $y = x1 + x2 + \text{rnorm}(n)$;
- (Y2) linear heteroskedastic: $y = x1 + x2 + \text{rnorm}(n, 0, x1+x2)$;
- (Y3) quadratic homoskedastic: $y = x1^2 + x2^2 + \text{rnorm}(n)$;
- (Y4) quadratic heteroskedastic: $y = x1^2 + x2^2 + \text{rnorm}(n, 0, x1+x2)$.

In the last two outcome models (Y3) and (Y4), the residuals indeed show some nonlinear relationship with the covariates and the fitted value. This suggests that the linear function can be a poor approximation to the true conditional mean function.

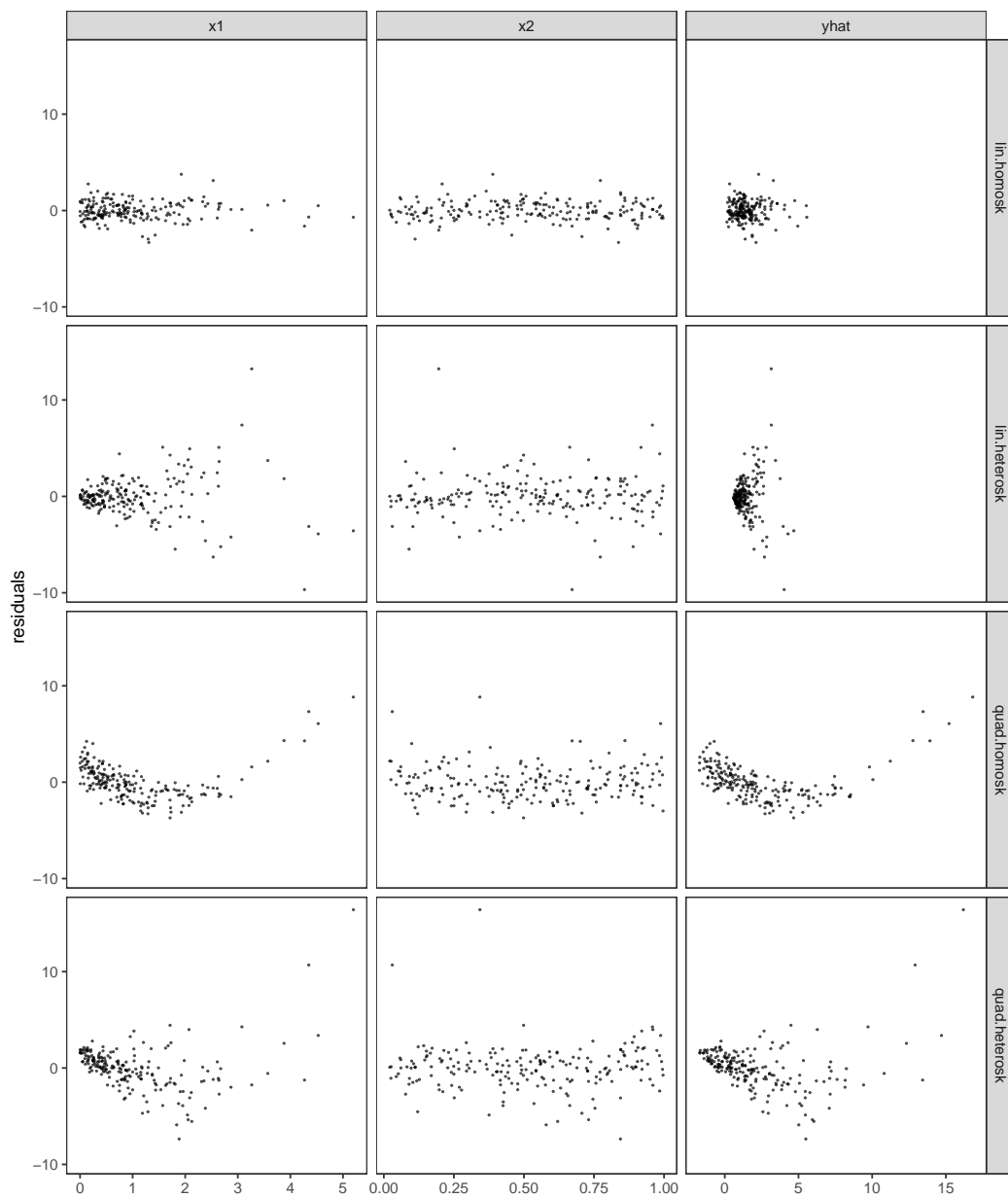


FIGURE 12.3: Residual plots

12.6 Conformal prediction based on exchangeability

Chapter 5.3 discusses the prediction of a future outcome y_{n+1} based on x_{n+1} and (X, Y) . It requires the Normal linear model assumption. Chapter 6 relaxes the Normality assumption on the error term in statistical inference but does not discuss prediction. Under the heteroskedastic linear model assumption in Chapter 6, we can predict the mean $E(y_{n+1}) = x_{n+1}^T \beta$ by $x_{n+1}^T \hat{\beta}$ with asymptotic standard error $(x_{n+1}^T \hat{V}_{\text{EHW}} x_{n+1})^{1/2}$, where \hat{V}_{EHW} is the EHW covariance matrix for the OLS coefficient. However, it is fundamentally challenging to predict y_{n+1} itself since the heteroskedastic linear model allows it to have a completely unknown variance σ_{n+1}^2 .

Under the population OLS formulation, it seems even more challenging to predict the future outcome since we do not even assume that the linear model is correctly specified. In particular, $x_{n+1}^T \hat{\beta}$ does not have the same mean as y_{n+1} in general. Perhaps surprisingly, we can construct a prediction interval for y_{n+1} based on x_{n+1} and (X, Y) using an idea called *conformal prediction* (Vovk et al., 2005; Lei et al., 2018). It leverages the *exchangeability*² of the data points

$$(x_1, y_1), \dots, (x_n, y_n), (x_{n+1}, y_{n+1}).$$

Pretending that we know the value $y_{n+1} = y^*$, we can fit OLS using $n + 1$ data points and obtain residuals

$$\hat{\varepsilon}_i(y^*) = y_i - x_i^T \hat{\beta}(y^*), \quad (i = 1, \dots, n + 1)$$

where we emphasize the dependence of the OLS coefficient and residuals on the unknown y^* . The absolute values of the residuals $|\hat{\varepsilon}_i(y^*)|$'s are also exchangeable, so the rank of $|\hat{\varepsilon}_{n+1}(y^*)|$, denoted by

$$\hat{R}_{n+1}(y^*) = 1 + \sum_{i=1}^n 1\{|\hat{\varepsilon}_i(y^*)| \leq |\hat{\varepsilon}_{n+1}(y^*)|\},$$

must have a uniform distribution over $\{1, 2, \dots, n, n + 1\}$, a known distribution not depending on anything else. It is a pivotal quantity satisfying

$$\text{pr} \left\{ \hat{R}_{n+1}(y^*) \leq \lceil (1 - \alpha)(n + 1) \rceil \right\} \geq 1 - \alpha. \quad (12.21)$$

Equivalently, this is a statement linking the unknown quantity y^* and observed data, so it gives a confidence set for y^* at level $1 - \alpha$. In practice, we can use a grid search to solve for the inequality (12.21) involving y^* .

Below we evaluate the leave-one-out prediction with the Boston housing data.

```
library("mlbench")
data(BostonHousing)
attach(BostonHousing)
n = dim(BostonHousing)[1]
p = dim(BostonHousing)[2] - 1
ymin = min(medv)
ymax = max(medv)
grid.y = seq(ymin - 30, ymax + 30, 0.1)
BostonHousing = BostonHousing[order(medv), ]
detach(BostonHousing)
```

²Exchangeability is a technical term in probability and statistics. Random elements z_1, \dots, z_n are exchangeable if $(z_{\pi(1)}, \dots, z_{\pi(n)})$ have the same distribution as (z_1, \dots, z_n) , where $\pi(1), \dots, \pi(n)$ is a permutation of the integers $1, \dots, n$. In other words, a set of random elements are exchangeable if their joint distribution does not change under re-ordering. IID random elements are exchangeable.

```

ols.fit.full = lm(medv ~ ., data = BostonHousing,
                  x = TRUE, y = TRUE, qr = TRUE)
beta        = ols.fit.full$coef
e.sigma      = summary(ols.fit.full)$sigma
X            = ols.fit.full$x
Y           = ols.fit.full$y
X.QR        = ols.fit.full$qr
X.Q         = qr.Q(X.QR)
X.R         = qr.R(X.QR)
Gram.inv     = solve(t(X.R)%*%X.R)
hatmat       = X.Q%*%t(X.Q)
resmat       = diag(n) - hatmat
leverage     = diag(hatmat)
Resvec       = ols.fit.full$residuals

cvt = qt(0.975, df = n-p-1)
cvr = ceiling(0.95*(n+1))

loo.pred = matrix(0, n, 5)
loo.cov  = matrix(0, n, 2)
for(i in 1:n)
{
  beta.i = beta - Gram.inv%*%X[i, ]*Resvec[i]/(1-leverage[i])
  e.sigma.i = sqrt(e.sigma^2*(n - p) -
                  (Resvec[i])^2/(1 - leverage[i]))/
              sqrt(n - p - 1)
  pred.i = sum(X[i, ]*beta.i)
  lower.i = pred.i - cvt*e.sigma.i/sqrt(1 - leverage[i])
  upper.i = pred.i + cvt*e.sigma.i/sqrt(1 - leverage[i])
  loo.pred[i, 1:3] = c(pred.i, lower.i, upper.i)
  loo.cov[i, 1] = findInterval(Y[i], c(lower.i, upper.i))

  grid.r = sapply(grid.y,
                  FUN = function(y){
                    Res = Resvec + resmat[, i]*(y - Y[i])
                    rank(abs(Res))[i]
                  })
  Cinterval = range(grid.y[grid.r<=cvr])
  loo.pred[i, 4:5] = Cinterval
  loo.cov[i, 2] = findInterval(Y[i], Cinterval)
}

```

In the above code, I use the QR decomposition to compute $X^T X$ and H . Moreover, the calculations of `lower.i`, `upper.i`, and `Res` use some tricks to avoid fitting n OLS. I relegate the justification of them to Problem 12.10.

The variable `loo.pred` has five columns corresponding to the point predictors, lower and upper intervals based on the Normal linear model and conformal prediction.

```

> colnames(loo.pred) = c("point", "G.l", "G.u", "c.l", "c.u")
> head(loo.pred)
      point      G.l      G.u      c.l      c.u
[1,]  6.633514 -2.941532 16.208559 -3.5 16.7
[2,]  8.806641 -1.349367 18.962649 -2.6 20.1
[3,] 12.044154  2.608290 21.480018  2.2 21.8
[4,] 11.025253  1.565152 20.485355  1.2 21.0
[5,] -5.181154 -14.819041  4.456733 -15.0  4.9
[6,]  8.324114 -1.382910 18.031138 -2.0 18.8

```

Figure 12.4 plots the observed outcomes and the prediction intervals for the 20 observations with the outcomes at the bottom, middle, and top. The Normal and conformal intervals are almost indistinguishable. For the observations with the highest outcome, the

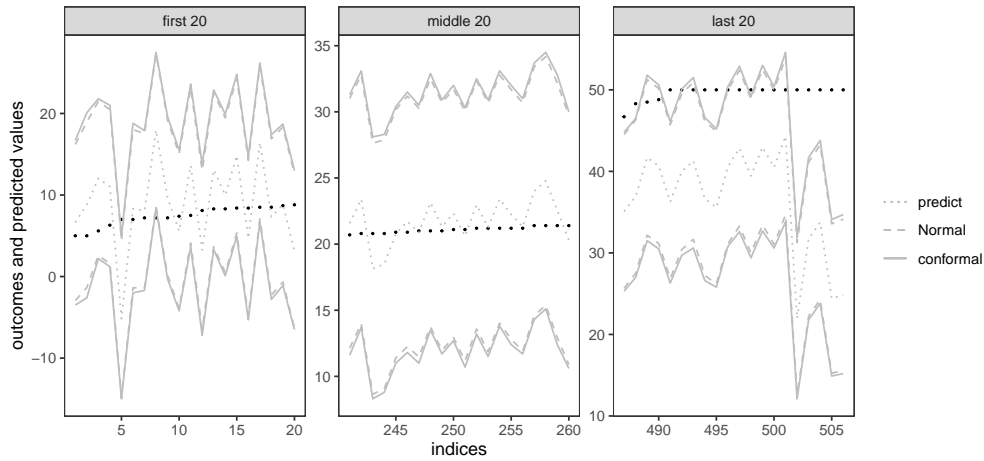


FIGURE 12.4: Leave-one-out prediction intervals based on the Boston housing data. The left, middle and right panels are for the 20 observations with the outcomes at the bottom, middle, and top, respectively. Each panel shows the original outcomes, predicted outcomes, as well as the prediction intervals based on the Normal linear model and conformal prediction.

predictions are quite poor. Surprisingly, the overall coverage rates across observations are close to 95% for both methods.

```
> apply(loo.cov==1, 2, mean)
[1] 0.9486166 0.9525692
```

Figure 12.5 compares the lengths of the two prediction intervals. Although the conformal prediction intervals are slightly wider than the Normal prediction interval, the differences are rather small, with the ratio of the length above 0.96.

If you read the above argument for conformal prediction again, you will realize that the whole argument does not rely on using OLS as the predictor. You can replace OLS with an arbitrary predictor, without harming the theoretical guarantee of the conformal prediction interval. Conformal prediction is a powerful idea for using black-box predictors while maintaining confidence interval guarantees. See Angelopoulos and Bates (2023) for its recent developments and applications.

Nevertheless, the confidence interval guarantees based on conformal prediction differ from that based on the Normal linear model. In particular, the conformal prediction interval covers y_{n+1} with probability larger than or equal to $1 - \alpha$, *averaged* over the randomness of the past data and future x_{n+1} , whereas the prediction interval based on the Normal linear model covers y_{n+1} with probability larger than or equal to $1 - \alpha$, *conditional* on all observed covariates. Therefore, the magic of conformal prediction comes from modifying the statistical model and the theoretical guarantees. Some practitioners may argue that the conditional coverage guarantee is more relevant than the average coverage guarantee. In those cases, the conformal prediction interval should be used with caution.

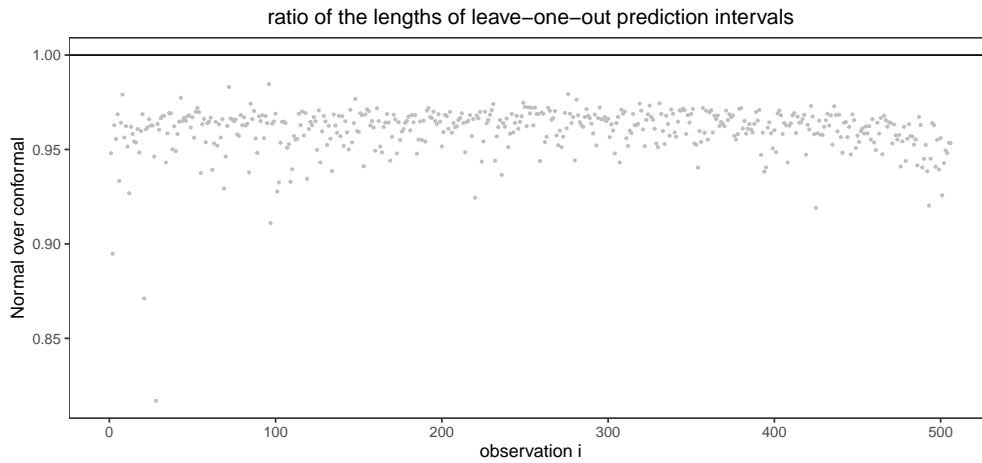


FIGURE 12.5: Boston housing data

12.7 Homework problems

12.1 Conditional mean

Prove Theorem 12.1.

12.2 Best linear approximation

Prove Theorem 12.2.

Remark: It is similar to Problem 8.8.

12.3 Univariate population OLS

Prove Corollary 12.1.

12.4 Asymptotics for the population OLS

Prove Theorem 12.7.

12.5 Population Cochran's formula

Prove Theorem 12.4.

12.6 Canonical correlation analysis (CCA)

Assume that (x, y) , where $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^k$, has the joint non-degenerate covariance matrix:

$$\begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}.$$

1. Find the best linear combinations (α, β) that give the maximum Pearson correlation coefficient:

$$(\alpha, \beta) = \arg \max_{a \in \mathbb{R}^k, b \in \mathbb{R}^p} \rho(y^T a, x^T b).$$

Note that you need to detail the steps in calculating (α, β) based on the covariance matrix above.

2. Define the maximum value as $\text{CC}(x, y)$. Prove that $\text{CC}(x, y) \geq 0$, and $\text{CC}(x, y) = 0$ if $x \perp\!\!\!\perp y$.

Remark: The maximum value $\text{CC}(x, y)$ is called the canonical correlation between x and y . We can also define partial canonical correlation between x and y given w .

12.7 Population partial correlation coefficient

Prove Theorem 12.6.

12.8 Independence and correlation

With scalar random variables x and y , it is well known that if $x \perp\!\!\!\perp y$, then $\rho_{yx} = 0$. However, with another random variable w , if $x \perp\!\!\!\perp y \mid w$, then $\rho_{yx|w} = 0$ may not hold.

Give a counterexample in which $x \perp\!\!\!\perp y \mid w$ but $\rho_{yx|w} \neq 0$.

Remark: With scalar random variables x and y , if $x \perp\!\!\!\perp y$, then we have $\text{cov}(y, x) = 0$, which implies

$$\rho_{yx} = \frac{\text{cov}(y, x)}{\sqrt{\text{var}(y)\text{var}(x)}} = 0.$$

With another random variable w , if $x \perp\!\!\!\perp y \mid w$, then we have $\text{cov}(y, x \mid w) = 0$, which, however, does not imply $\rho_{yx|w} = 0$ because $\rho_{yx|w}$ is not defined as

$$\frac{\text{cov}(y, x \mid w)}{\sqrt{\text{var}(y \mid w)\text{var}(x \mid w)}}.$$

Shah and Peters (2020) had related discussion on this issue.

12.9 Best linear approximation of a cubic curve

Assume that $x \sim N(0, 1)$, $\varepsilon \sim N(0, \sigma^2)$, $x \perp\!\!\!\perp \varepsilon$, and $y = x^3 + \varepsilon$. Find the best linear approximation of $E(y \mid x)$ based on $(1, x)$. Plot both $E(y \mid x)$ and its best linear approximation together and compare them.

12.10 Leave-one-out formula in conformal prediction

Justify the calculations of `lower.i`, `upper.i`, and `Res` in Section 12.6.

12.11 Conformal prediction for multiple outcomes

Assuming exchangeability of

$$(x_1, y_1), \dots, (x_n, y_n), (x_{n+1}, y_{n+1}), \dots, (x_{n+k}, y_{n+k}).$$

Propose a method to construct joint conformal prediction regions for $(y_{n+1}, \dots, y_{n+k})$ based on (X, Y) and $(x_{n+1}, \dots, x_{n+k})$.

12.12 Cox's theorem

Cox (1960) considered the data-generating process

$$x_1 \longrightarrow x_2 \longrightarrow y$$

under the following linear models: for $i = 1, \dots, n$, we have

$$x_{i2} = \alpha_0 + \alpha_1 x_{i1} + \eta_i$$

and

$$y_i = \beta_0 + \beta_1 x_{i2} + \varepsilon_i$$

where η_i has mean 0 and variance σ_η^2 , ε_i has mean 0 and variance σ_ε^2 , and the η_i s and ε_i s are independent. The linear model implies

$$y_i = (\beta_0 + \beta_1 \alpha_0) + (\beta_1 \alpha_1) x_{i1} + (\varepsilon_i + \beta_1 \eta_i)$$

where $\varepsilon_i + \beta_1 \eta_i$ s are independent with mean 0 and variance $\sigma_\varepsilon^2 + \beta_1^2 \sigma_\eta^2$.

Therefore, we have two ways to estimate $\beta_1 \alpha_1$:

- (W1) the first estimator is $\hat{\gamma}_1$, the OLS estimator of the y_i 's on the x_{i1} 's with the intercept;
- (W2) the second estimator is $\hat{\alpha}_1 \hat{\beta}_1$, the product of the OLS estimator of the x_{i2} 's on the x_{i1} 's with the intercept and that of the y_i 's on the x_{i2} 's with the intercept.

Cox (1960) reported Theorem 12.8 below. Prove Theorem 12.8.

Theorem 12.8 Let $X_1 = (x_{11}, \dots, x_{n1})$. We have

$$\text{var}(\hat{\alpha}_1 \hat{\beta}_1 \mid X_1) \leq \text{var}(\hat{\gamma}_1 \mid X_1),$$

and more precisely,

$$\text{var}(\hat{\gamma}_1 \mid X_1) = \frac{\sigma_\varepsilon^2 + \beta_1^2 \sigma_\eta^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}$$

and

$$\text{var}(\hat{\alpha}_1 \hat{\beta}_1 \mid X_1) = \frac{\sigma_\varepsilon^2 E(\hat{\rho}_{12}^2 \mid X_1) + \beta_1^2 \sigma_\eta^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}$$

where $\hat{\rho}_{12} \in [-1, 1]$ is the sample Pearson correlation coefficient between the x_{i1} 's and the x_{i2} 's.

Remark: If we further assume that the error terms are Normal, then $\hat{\alpha}_1 \hat{\beta}_1$ is the maximum likelihood estimator for $\alpha_1 \beta_1$. Therefore, the asymptotic optimality theory for the maximum likelihood estimator justifies the superiority of $\hat{\alpha}_1 \hat{\beta}_1$ over $\hat{\gamma}_1$. Theorem 12.8 provides a stronger finite-sample result without assuming the Normality of the error terms.

12.13 Measurement error and Frisch's bounds

- Given scalar random variables x and y , we can obtain the population OLS coefficient (α, β) of y on $(1, x)$. However, x and y may be measured with errors, that is, we observe $x^* = x + u$ and $y^* = y + v$, where u and v are mean zero error terms satisfying $u \perp\!\!\!\perp v$ and $(u, v) \perp\!\!\!\perp (x, y)$. We can obtain the population OLS coefficient (α^*, β^*) of y^* on $(1, x^*)$ and the population OLS coefficient (a^*, b^*) of x^* on $(1, y^*)$.

Prove that if $\beta = 0$ then $\beta^* = b^* = 0$; if $\beta \neq 0$ then

$$|\beta^*| \leq |\beta| \leq 1/|b^*|.$$

2. Given scalar random variables x, y and a random vector w , we can obtain the population OLS coefficient (α, β, γ) of y on $(1, x, w)$. When x and y are measured with error as above with mean zero errors satisfying $u \perp\!\!\!\perp v$ and $(u, v) \perp\!\!\!\perp (x, y, w)$, we can obtain the population OLS coefficient $(\alpha^*, \beta^*, \gamma^*)$ of y on $(1, x^*, w)$, and the population OLS coefficient (a^*, b^*, c^*) of x^* on $(1, y^*, w)$.

Prove that the same result holds as in the first part of the problem.

Remark: Tamer (2010) reviewed Frisch (1934)'s upper and lower bounds for the univariate OLS coefficient based on the two OLS coefficients of the observables. The second part of the problem extends the result to the multivariate OLS with a covariate subject to measurement error. The lower bound is well documented in most books on measurement errors, but the upper bound is much less well known.

12.14 A three-way decomposition

The main text of this chapter focuses on the two-way decomposition of the outcome: $y = x^T \beta + \varepsilon$, where β is the population OLS coefficient and ε is the population OLS residual. However, $x^T \beta$ is only the best linear approximation to the true conditional mean function $\mu(x) = E(y | x)$. This suggests the following three-way decomposition of the outcome:

$$y = x^T \beta + \{\mu(x) - x^T \beta\} + \{y - \mu(x)\},$$

which must hold without any assumptions. Introduce the notation for the linear term

$$\hat{y} = x^T \beta,$$

the notation for the approximation error:

$$\delta = \mu(x) - x^T \beta,$$

and the notation for the “ideal residual”:

$$e = y - \mu(x).$$

Then we can decompose the outcome as

$$y = \hat{y} + \delta + e$$

and the population OLS residual as

$$\varepsilon = \{\mu(x) - x^T \beta\} + \{y - \mu(x)\} = \delta + e.$$

1. Prove

$$E(\hat{y}e | x) = 0, \quad E(\delta e | x) = 0$$

and

$$E(\hat{y}e) = 0, \quad E(\delta e) = 0, \quad E(\hat{y}\delta) = 0.$$

Prove

$$E(\varepsilon^2) = E(\delta^2) + E(e^2).$$

2. Introduce an intermediate quantity between the population OLS coefficient β and the OLS coefficient $\hat{\beta}$:

$$\tilde{\beta} = \left(n^{-1} \sum_{i=1}^n x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n x_i \mu(x_i) \right).$$

Equation (12.18) states that $\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, B^{-1}MB^{-1})$ in distribution, where $B = E(xx^T)$ and $M = E(\varepsilon^2 xx^T)$.

Prove that

$$\text{cov}(\hat{\beta} - \beta) = \text{cov}(\hat{\beta} - \tilde{\beta}) + \text{cov}(\tilde{\beta} - \beta),$$

and moreover,

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \tilde{\beta}) &\rightarrow N(0, B^{-1}M_1B^{-1}), \\ \sqrt{n}(\tilde{\beta} - \beta) &\rightarrow N(0, B^{-1}M_2B^{-1})\end{aligned}$$

in distribution, where

$$\begin{aligned}M_1 &= E(e^2 xx^T), \\ M_2 &= E(\delta^2 xx^T)\end{aligned}$$

Verify that $M = M_1 + M_2$.

Remark: To prove the result, you may find the law of total covariance formula in (B.4) helpful. We can also write M_1 as $M_1 = E\{\text{var}(y | x)xx^T\}$. So the meat matrix M has two sources of uncertainty, one is from the conditional variance of y given x , and the other is from the approximation error.



Part V

Overfitting, Regularization, and Model Selection



Perils of Overfitting

Previous chapters assume that the covariate matrix X is given and the linear model, correctly specified or not, is also given. Although including useless covariates in the linear model results in less precise estimators, this problem is not severe when the total number of covariates is small compared with the sample size. In many modern applications, however, the number of covariates can be large compared with the sample size. Sometimes, it can be a nonignorable fraction of the sample size; sometimes, it can even be larger than the sample size. For instance, modern DNA sequencing technology often generates covariates of millions of dimensions, which is much larger than the usual sample size under study. In these applications, the theory in previous chapters is inadequate. This chapter introduces an important notion in statistics: overfitting.

13.1 David Freedman's simulation

Freedman (1983) used a simple simulation to illustrate the problem with a large number of covariates. He simulated data from the following Normal linear model $Y = X\beta + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2 I_n)$ and $\beta = (\mu, 0, \dots, 0)^T$. He then computed the sample R^2 . Since the covariates do not explain any variability of the outcome at all in the true model, we would expect R^2 to be extremely small over repeated sampling. However, he showed, via both simulation and theory, that R^2 is surprisingly large when p/n is not close to 0.

Figure 13.1 shows the results from Freedman's simulation setting with $n = 100$ and $p = 50$, over 1000 replications. The (1,1)th subfigure shows the histogram of the R^2 , which centers around 0.5. This can be easily explained by the exact distribution of R^2 proved in Corollary 10.1:

$$R^2 \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right),$$

with the density shown in the (1,1)th and (1,2)th subfigure of Figure 13.1. Based on the formulas in Proposition B.4, the beta distribution above has mean

$$E(R^2) = \frac{\frac{p-1}{2}}{\frac{p-1}{2} + \frac{n-p}{2}} = \frac{p-1}{n-1}$$

and variance

$$\begin{aligned} \text{var}(R^2) &= \frac{\frac{p-1}{2} \times \frac{n-p}{2}}{\left(\frac{p-1}{2} + \frac{n-p}{2}\right)^2 \left(\frac{p-1}{2} + \frac{n-p}{2} + 1\right)} \\ &= \frac{2(p-1)(n-p)}{(n-1)^2(n+1)}. \end{aligned}$$

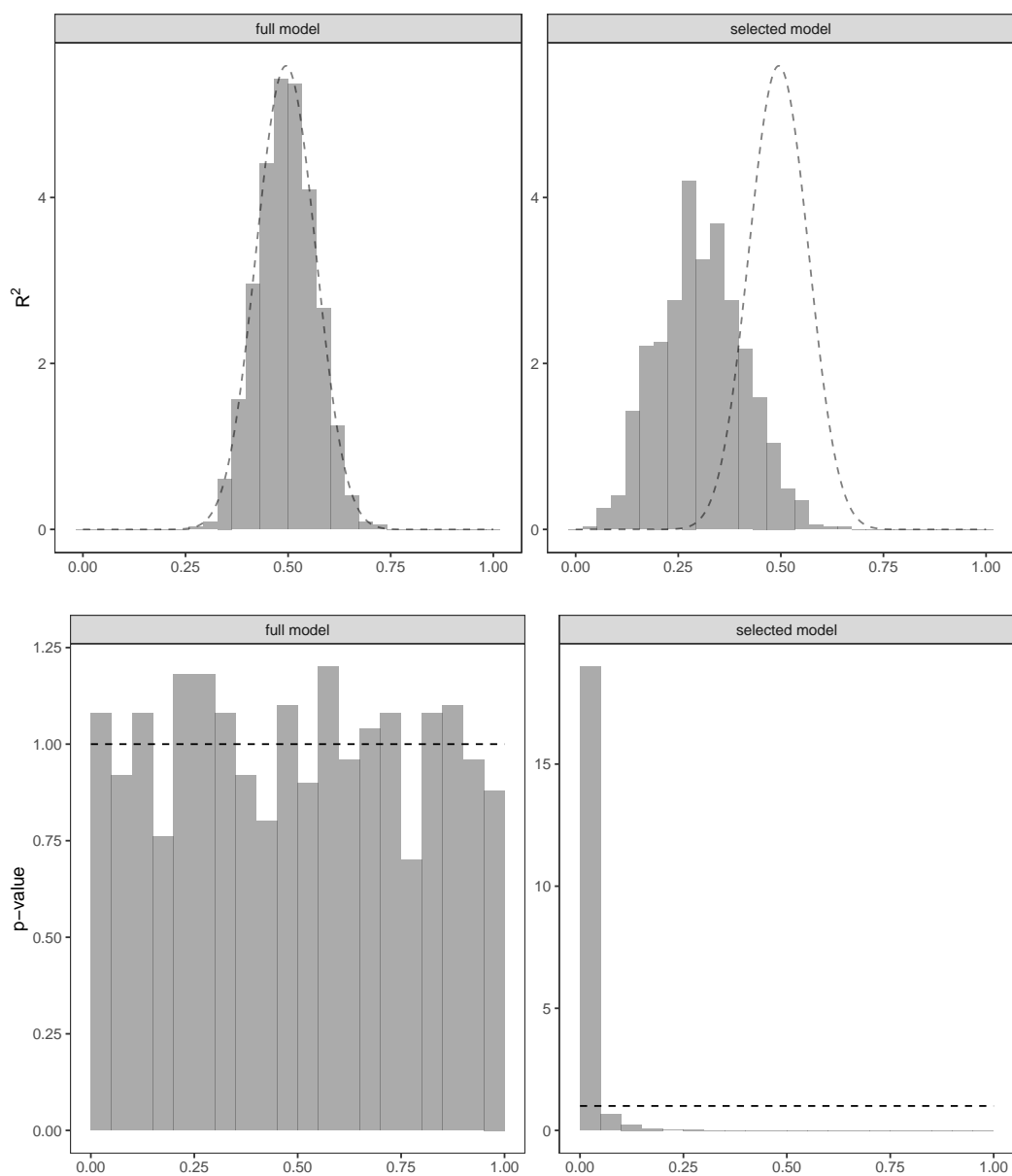


FIGURE 13.1: Freedman's simulation. The first row shows the histograms of the R^2 s, and the second row shows the histograms of the p -values in testing that all coefficients are 0. The first column corresponds to the full model without testing, and the second column corresponds to the selected model with testing at the significance level 0.25.

When $p/n \rightarrow 0$, we have

$$E(R^2) \rightarrow 0, \quad \text{var}(R^2) \rightarrow 0,$$

so Markov's inequality implies that $R^2 \rightarrow 0$ in probability (see Proposition C.4 for a related result). However, when $p/n \rightarrow \gamma \in (0, 1)$, we have

$$E(R^2) \rightarrow \gamma, \quad \text{var}(R^2) \rightarrow 0,$$

so Markov's inequality implies that $R^2 \rightarrow \gamma$ in probability. This means that when p has the same order as n , the sample R^2 is close to the ratio p/n even though there is no association between the covariates and the outcome in the true data-generating process. In Freedman's simulation, $\gamma = 0.5$ so R^2 is close to 0.5.

The (1, 2)th subfigure shows the histogram of the R^2 based on a model selection first step by dropping all covariates with p -values larger than 0.25. The R^2 in the (1, 2)th subfigure are slightly smaller but still centered around 0.37. The joint F test based on the selected model does not generate uniform p -values in the (2, 2)th subfigure, in contrast to the uniform p -values in the (2, 1)th subfigure. With a model selection step, statistical inference becomes much more complicated. This is a topic called *selective inference*, which is beyond the scope of this book.

The above simulation and calculation give an important warning: we cannot over-interpret the sample R^2 , because it can be too optimistic about model fitting. In many empirical research, R^2 is at most 0.1 with a large number of covariates, making us wonder whether those researchers are just chasing the noise rather than the signal. So we do not trust R^2 as a model-fitting measure with a large number of covariates. In general, R^2 cannot avoid overfitting, and we must modify it for model selection.

13.2 Variance inflation factor

Theorem 13.1 below quantifies the potential problem of including too many covariates in OLS. It introduces the notion of variance inflation factor (VIF).

Theorem 13.1 *Consider a fixed covariate matrix $X = (X_1, \dots, X_p) \in \mathbb{R}^p$. Let $\hat{\beta}_j$ be the coefficient of X_j of the OLS fit of Y on $(1_n, X_j : j \in S)$, where S is a subset of $\{1, \dots, p\}$. Under the model $y_i = f(x_i) + \varepsilon_i$ with an unknown (and possibly nonlinear) $f(\cdot)$ and the ε_i 's uncorrelated with mean zero and variance σ^2 , the variance of $\hat{\beta}_j$ equals*

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \times \frac{1}{1 - R_j^2},$$

where R_j^2 is the sample R^2 from the OLS fit of X_j on 1_n and all other covariates in $\{X_j : j \in S\}$.

Theorem 13.1 does not even assume that the true mean function is linear. It states that the variance of $\hat{\beta}_j$ has two multiplicative components. If we run a short regression of Y on 1_n and $X_j = (x_{1j}, \dots, x_{nj})^T$, the coefficient equals

$$\tilde{\beta}_j = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j) y_i}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

where $\bar{x}_j = n^{-1} \sum_{i=1}^n x_{ij}$. It has variance

$$\begin{aligned} \text{var}(\tilde{\beta}_j) &= \text{var} \left\{ \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j) y_i}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \right\} \\ &= \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \sigma^2}{\left\{ \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right\}^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}. \end{aligned}$$

So the first component is the variance of the OLS coefficient in the short regression. The second component $1/(1 - R_j^2)$ is called the VIF. The VIF indeed inflates the variance of $\tilde{\beta}_j$, and the more covariates are added into the long regression, the larger the variance inflation factor is. In R, the `car` package provides the function `vif` to compute the VIF for each covariate.

The proof of Theorem 13.1 below is based on the FWL Theorem in Chapter 7.

Proof of Theorem 13.1: Let $\tilde{X}_j = (\tilde{x}_{1j}, \dots, \tilde{x}_{nj})^T$ be the residual vector from the OLS fit of X_j on 1_n and all other covariates in $\{X_j : j \in S\}$, which have sample mean 0. The FWL Theorem implies that

$$\hat{\beta}_j = \frac{\sum_{i=1}^n \tilde{x}_{ij} y_i}{\sum_{i=1}^n \tilde{x}_{ij}^2},$$

which has variance

$$\text{var}(\hat{\beta}_j) = \frac{\sum_{i=1}^n \tilde{x}_{ij}^2 \sigma^2}{\left\{ \sum_{i=1}^n \tilde{x}_{ij}^2 \right\}^2} = \frac{\sigma^2}{\sum_{i=1}^n \tilde{x}_{ij}^2}. \quad (13.1)$$

Because $\sum_{i=1}^n \tilde{x}_{ij}^2$ is the residual sum of squares from the OLS of X_j on 1_n and all other covariates in $\{X_j : j \in S\}$, it is related to R_j^2 via

$$R_j^2 = 1 - \frac{\sum_{i=1}^n \tilde{x}_{ij}^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

or, equivalently,

$$\sum_{i=1}^n \tilde{x}_{ij}^2 = (1 - R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2. \quad (13.2)$$

Combining (13.1) and (13.2) gives Theorem 13.1. \square

13.3 Bias-variance trade-off

Theorem 13.1 above characterizes the variance of the OLS coefficient, but it does not characterize its bias. In general, a more complex model is closer to the true mean function $f(x_i)$, and can then reduce the bias of approximating the mean function. However, Theorem 13.1 implies that a more complex model results in larger variances of the OLS coefficients. So we face a bias-variance trade-off.

Consider a simple case where the true data-generating process is linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_s x_{is} + \varepsilon_i. \quad (13.3)$$

Ideally, we want to use the model (13.3) with exactly s covariates. In practice, we may not know which covariates to include in the OLS. If we underfit the data using a short regression with $q < s$:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_q x_{iq} + \tilde{\varepsilon}_i, \quad (i = 1, \dots, n) \quad (13.4)$$

then the OLS coefficients are biased. If we increase the complexity of the model to overfit the data using a long regression with $p > s$:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} + \hat{\varepsilon}_i, \quad (i = 1, \dots, n) \quad (13.5)$$

then the OLS coefficients are unbiased. Theorem 13.1, however, shows that the OLS coefficients from the under-fitted model (13.4) have smaller variances than those from the overfitted model (13.5).

Example 13.1 below illustrates the idea of overfitting under an ideal Normal linear model.

Example 13.1 *In general, we have a sequence of models with increasing complexity. For simplicity, we consider nested models containing 1_n and covariates*

$$\{X_1\} \subset \{X_1, X_2\} \subset \cdots \subset \{X_1, \dots, X_p\}$$

in the following simulation setting. The true linear model is $y_i = x_i^\top \beta + \varepsilon_i$, $\varepsilon_i \stackrel{\text{IID}}{\sim} N(0, 1)$ with $p = 40$ but only the first 10 covariates have non-zero coefficients 1 and all other covariates have coefficients 0. We generate two datasets: both have sample size $n = 200$, all covariates have IID $N(0, 1)$ entries, and the error terms are IID. We use the first dataset to fit the OLS and thus call it the “training dataset.” We use the second dataset to assess the performance of the fitted OLS from the training dataset, and thus call it the “testing dataset.”¹ Figure 13.2 plots the residual sum of squares against the number of covariates in the training and testing datasets. By definition of OLS, the residual sum of squares decreases with the number of covariates in the training dataset, but it first decreases and then increases in the testing dataset with minimum value attained at 10, the number of covariates in the true data generating process.

Example 13.2 below illustrates the idea of overfitting under a nonlinear model. Even though the true mean function is nonlinear, we still use OLS with polynomials of covariates to approximate the truth.²

Example 13.2 *The true nonlinear model is $y_i = \sin(2\pi x_i) + \varepsilon_i$, $\varepsilon_i \stackrel{\text{IID}}{\sim} N(0, 1)$ with the x_i ’s equally spaced in $[0, 1]$ and the error terms are IID. The training and testing datasets both have sample sizes $n = 200$. Figure 13.3 plots the residual sum of squares against the order of the polynomial in the OLS fit*

$$y_i = \sum_{j=0}^{p-1} \beta_j x_i^j + \varepsilon_i.$$

By the definition of OLS, the residual sum of squares decreases with the order of polynomials in the training dataset, but it achieves the minimum near $p = 5$ in the testing dataset. We can show that the residual sum of squares decreases to zero with $p = n$ in the training dataset; see Problem 13.7. However, it is larger than that under $p = 5$ in the testing dataset.

¹Splitting a dataset into the *training dataset* and the *testing dataset* is a standard tool to assess the out-of-sample performance of proposed methods. It is important in statistics and machine learning.

²A celebrated theorem due to Weierstrass states that on a bounded interval, any continuous function can be approximated arbitrarily well by a polynomial function:

Theorem 13.2 (Weierstrass’s theorem) *Suppose f is a continuous function defined on the interval $[a, b]$. For every $\varepsilon > 0$, there exists a polynomial p such that for all $x \in [a, b]$, we have $|f(x) - p(x)| < \varepsilon$.*

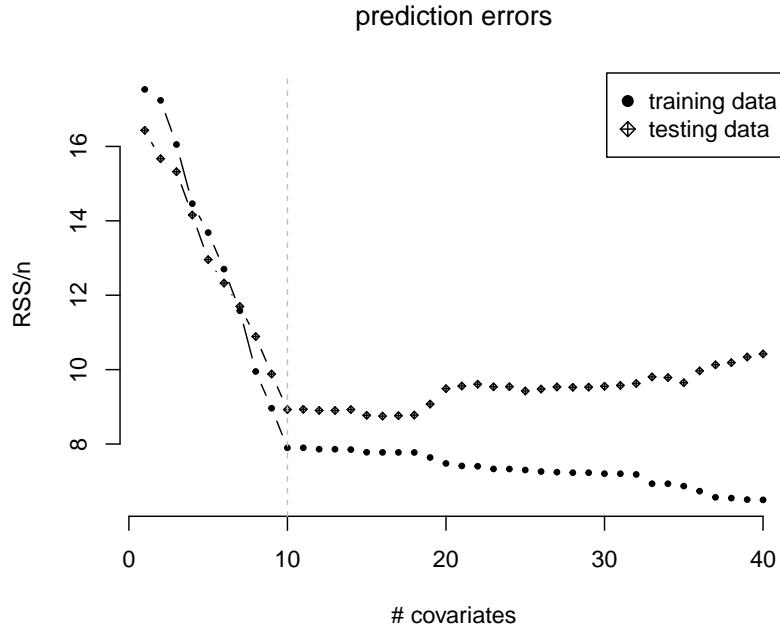


FIGURE 13.2: Training and testing errors: linear mean function

13.4 Model selection criteria

With a large number of covariates $X_1, \dots, X_{\bar{p}}$, we want to select a model that has the best performance for prediction. In total, we have $2^{\bar{p}}$ possible models. Which one is the best? What is the criterion for the best model? Practitioners often use the linear model for multiple purposes. A dominant criterion is the prediction performance of the linear model in a new dataset (Yu and Kumbier, 2020). However, we do not have the new dataset yet in the statistical modeling stage. So we need to find criteria that are good proxies for the prediction performance.

13.4.1 RSS, R-squared and adjusted R-squared

The first obvious criterion is the residual sum of squares (RSS), which, however, is not a good criterion because it favors the largest model. The sample R^2 has the same problem of favoring the largest model. Most model selection criteria are in some sense modifications of RSS or R^2 .

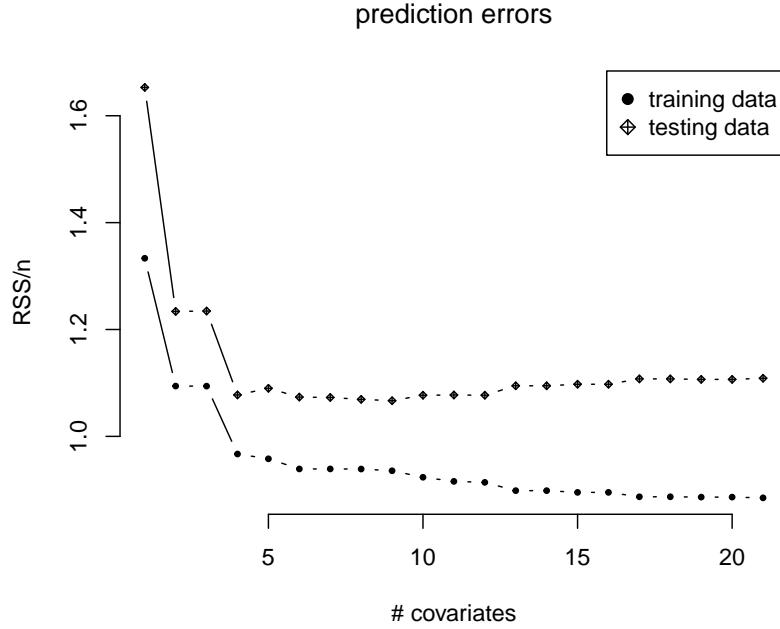


FIGURE 13.3: Training and testing errors: nonlinear mean function

The adjusted R^2 takes into account the complexity of the model:

$$\begin{aligned}
 \bar{R}^2 &= 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} \\
 &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p)}{\sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)} \\
 &= 1 - \frac{n - 1}{n - p} (1 - R^2).
 \end{aligned}$$

So based on \bar{R}^2 , the best model has the smallest $\hat{\sigma}^2$, the estimator for the variance of the error term in the Gauss–Markov model. Theorem 13.3 below shows that \bar{R}^2 is closely related to the F statistic in testing two nested Normal linear models.

Theorem 13.3 *Consider the setting of Chapter 8.3. Test two nested Normal linear models:*

$$Y = X_1\beta_1 + \varepsilon$$

versus

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

or, equivalently, test $\beta_2 = 0$. We can use the standard F statistic defined in Chapter 8.3, and we can also compare the adjusted R^2 's from these two models: \bar{R}_1^2 and \bar{R}_2^2 .

They are related in the following sense: $F > 1$ if and only if $\bar{R}_1^2 < \bar{R}_2^2$.

I leave the proof of Theorem 13.3 as Problem 13.3. From Theorem 13.3, \bar{R}^2 does not necessarily favor the largest model. However, \bar{R}^2 still favors unnecessarily large models compared with the usual hypothesis testing based on the Normal linear model because the

mean of F is approximately 1, but the upper quantile of F is much larger than 1 (for example, the 95% quantile of $F_{1,n-p}$ is larger than 3.8, and the 95% quantile of $F_{2,n-p}$ is larger than 2.9).

13.4.2 Information criteria

Taking into account the model complexity, we can find the model with the smallest AIC or BIC, defined as

$$\begin{aligned}\text{AIC} &= n \log \frac{\text{RSS}}{n} + 2p, \\ \text{BIC} &= n \log \frac{\text{RSS}}{n} + p \log n,\end{aligned}$$

with full names “Akaike’s information criterion” and “Bayes information criterion,” respectively. The theoretical derivations of AIC and BIC are beyond the scope of this book.

AIC and BIC are both monotone functions of the RSS penalized by the number of parameters p in the model. The penalty in BIC is larger, so it favors smaller models than AIC. Shao (1997)’s results suggested that BIC can consistently select the true model if the linear model is correctly specified, but AIC can select the model that minimizes the prediction error if the linear model is misspecified. In most statistical practice, the linear model assumption cannot be justified, so we recommend using AIC.

13.4.3 Cross-validation (CV)

We can use the leave-one-out cross-validation based on the predicted residual:

$$\text{PRESS} = \sum_{i=1}^n \hat{\varepsilon}_{[-i]}^2,$$

which is called the predicted residual error sum of squares (PRESS) statistic. By Theorem 11.3, it equals

$$\text{PRESS} = \sum_{i=1}^n \frac{\hat{\varepsilon}_i^2}{(1 - h_{ii})^2}, \quad (13.6)$$

and therefore, it depends not only on the residuals but also on the leverage scores.

Because the average value of h_{ii} is $n^{-1} \sum_{i=1}^n h_{ii} = p/n$, we can approximate PRESS by the generalized cross-validation (GCV) criterion:

$$\begin{aligned}\text{GCV} &= \sum_{i=1}^n \frac{\hat{\varepsilon}_i^2}{(1 - p/n)^2} \\ &= \text{RSS} \times \left(1 - \frac{p}{n}\right)^{-2}.\end{aligned}$$

When $p/n \approx 0$, we have³

$$\begin{aligned}\log \text{GCV} &= \log \text{RSS} - 2 \log \left(1 - \frac{p}{n}\right) \\ &\approx \log \text{RSS} + \frac{2p}{n} \\ &= \text{AIC}/n + \log n,\end{aligned}$$

³The approximation is due to the Taylor expansion $\log(1+x) = x - x^2/2 + x^3/3 - \dots \approx x$.

so GCV is approximately equivalent to AIC with small p/n . With large p/n , they may have large differences.

GCV is not crucial for OLS, because it is easy to compute PRESS. However, it is much more useful in other models where we need to fit the data n times to compute PRESS. For a general model without simple leave-one-out formulas, it is computationally intensive to obtain PRESS. The K -fold cross-validation (K -CV) is computationally more attractive. The best model has the smallest K -CV, computed as follows:

1. randomly shuffle the observations;
2. split the data into K folds;
3. for each fold k , use all other folds as the training data, and compute the predicted errors on fold k ($k = 1, \dots, K$);
4. sum up the prediction errors across K folds, denoted by K -CV.

When $K = 3$, we split the data into 3 folds. Run OLS to obtain a fitted function with folds 2, 3 and use it to predict on fold 1, yielding prediction error r_1 ; run OLS with folds 1, 3 and predict on fold 2, yielding prediction error r_2 ; run OLS with folds 1, 2 and predict on fold 3, yielding prediction error r_3 . The total prediction error is $r = r_1 + r_2 + r_3$. We want to select a model that minimizes r . Usually, practitioners choose $K = 5$ or 10, but this can depend on the computational resource.

13.5 Best subset and forward/backward selection

Given a model selection criterion, we can select the best model.

For a small \bar{p} , we can enumerate all $2^{\bar{p}}$ models. The function `regsubsets` in the `R` package `leaps` implements this.⁴ Figure 13.4 shows the results of the best subset selection in two applications. RSS always favors the largest model. BIC favors slightly smaller model than AIC.

⁴Note that this function uses a definition of BIC that differs from the above definition by a constant, but this does not change the model selection result.

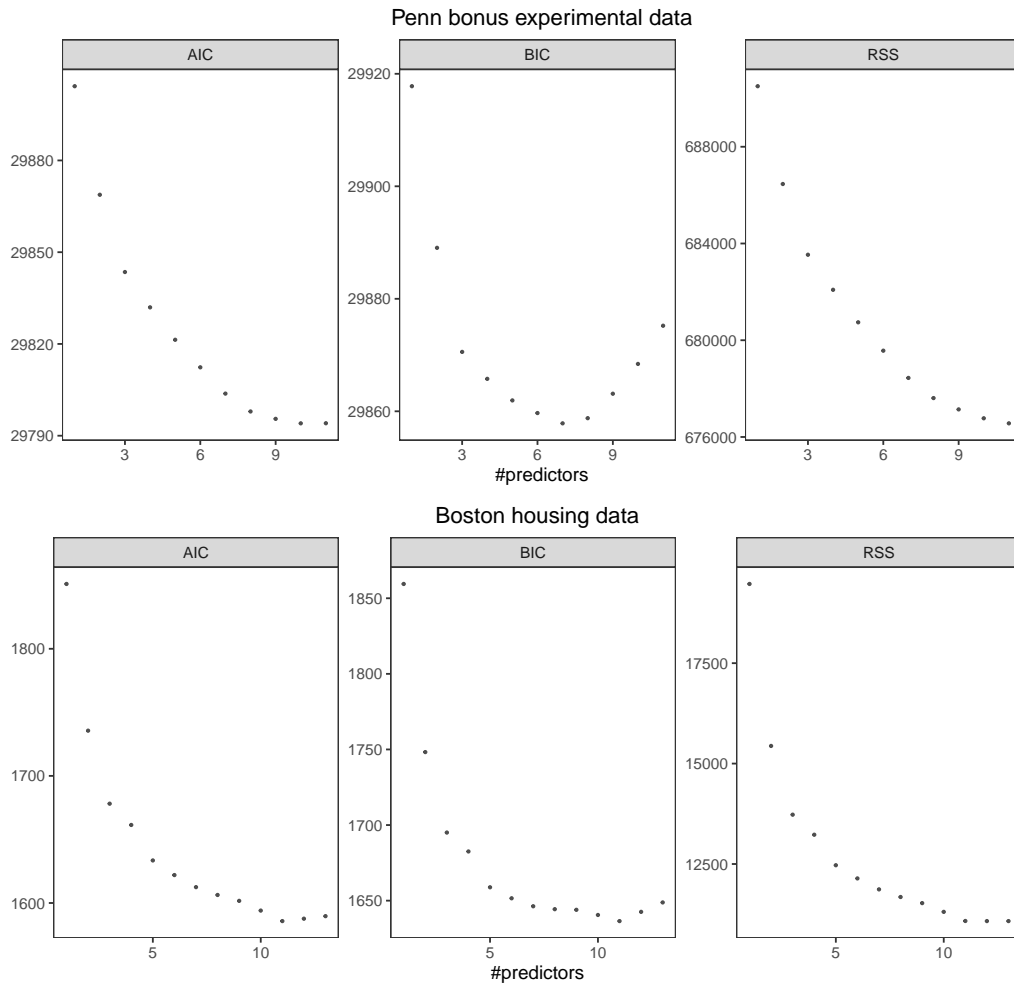


FIGURE 13.4: Best subset selection

For large \bar{p} , we can use forward or backward regressions. Forward regression starts with a model with only the intercept. In step one, it finds the best covariate among the \bar{p} candidates based on the prespecified criterion. In step two, it keeps this covariate in the model and finds the next best covariate among the remaining $\bar{p} - 1$ candidates. It proceeds by adding the next best covariate one by one.

The backward regression does the opposite. It starts with the full model with all \bar{p} covariates. In step one, it drops the worst covariate among the \bar{p} candidates based on the prespecified criterion. In step two, it drops the next worst covariate among the remaining $\bar{p} - 1$ candidates. It proceeds by dropping the next worst covariate one by one.

Both methods generate a sequence of models, and select the best one based on the prespecified criterion. Forward regression works in the case with $p \geq n$ but it stops at step $n - 1$; backward regression works only in the case with $p < n$. The functions `step` or `stepAIC` in the `MASS` package implement them.

13.6 Homework problems

13.1 Inflation and deflation of the estimated variance

This problem extends Theorem 13.1 to the estimated variance.

The covariate matrix X has columns $1_n, X_1, \dots, X_p$. Compare the coefficient of X_1 in the following long and short regressions:

$$Y = \hat{\beta}_0 1_n + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p + \hat{\varepsilon},$$

and

$$Y = \tilde{\beta}_0 1_n + \tilde{\beta}_1 X_1 + \dots + \tilde{\beta}_q X_q + \tilde{\varepsilon},$$

where $q < p$. Under the condition in Theorem 13.1,

$$\frac{\text{var}(\hat{\beta}_1)}{\text{var}(\tilde{\beta}_1)} = \frac{1 - R_{X_1.X_2 \dots X_q}^2}{1 - R_{X_1.X_2 \dots X_p}^2} \geq 1,$$

recalling that $R_{U.V}^2$ denotes the R^2 of U on V . Now we compare the corresponding estimated variances $\hat{\text{var}}(\hat{\beta}_1)$ and $\hat{\text{var}}(\tilde{\beta}_1)$ based on homoskedasticity.

1. Prove that

$$\frac{\hat{\text{var}}(\hat{\beta}_1)}{\hat{\text{var}}(\tilde{\beta}_1)} = \frac{1 - R_{Y.X_1 \dots X_p}^2}{1 - R_{Y.X_1 \dots X_q}^2} \times \frac{1 - R_{X_1.X_2 \dots X_q}^2}{1 - R_{X_1.X_2 \dots X_p}^2} \times \frac{n - q - 1}{n - p - 1}.$$

2. Using the definition of the partial R^2 in Problem 10.6, prove that

$$\frac{\hat{\text{var}}(\hat{\beta}_1)}{\hat{\text{var}}(\tilde{\beta}_1)} = \frac{1 - R_{Y.X_{q+1} \dots X_p | X_1 \dots X_q}^2}{1 - R_{X_1.X_{q+1} \dots X_p | X_2 \dots X_q}^2} \times \frac{n - q - 1}{n - p - 1}.$$

Remark: The first result shows that the ratio of the estimated variances has three factors:

- (F1) the first factor corresponds to the R^2 's of the outcome on the covariates,
- (F2) the second factor equals the ratio of the true variances $\text{var}(\hat{\beta}_1)/\text{var}(\tilde{\beta}_1)$,
- (F3) the third factor corresponds to the degrees of freedom correction.

The first factor deflates the estimated variance since the R^2 increases with more covariates included in the regression, and the second and the third factors inflate the estimated variance. Overall, whether adding more covariates inflate or deflate the estimated variance depends on the interplay of the three factors. The answer is not as definite as Theorem 13.1.

The variance inflation result in Theorem 13.1 sometimes causes confusion. It only concerns the variance. When we view some covariates as random, then the bias term can also contribute to the variance of the OLS estimator. In this case, we should interpret Theorem 13.1 with caution. See Ding (2021b) for a related discussion.

13.2 Inflation and deflation of the variance under heteroskedasticity

Revisit Section 13.2. Relax the condition in Theorem 13.1 as $\text{var}(\varepsilon_i) = \sigma_i^2$ to allow for heteroskedasticity with possibly different variances of the error terms. Give a counterexample in which

$$\text{var}(\hat{\beta}_j) > \text{var}(\tilde{\beta}_j)$$

for some j .

Remark: First derive the formulas of $\text{var}(\hat{\beta}_j)$ and $\text{var}(\tilde{\beta}_j)$ under heteroskedasticity, and then give a numerical example by specifying the X matrix and the σ_i^2 's.

13.3 Equivalence of F and \bar{R}^2

Prove Theorem 13.3.

13.4 Using PRESS to construct an unbiased estimator for σ^2

Prove that

$$\hat{\sigma}_{\text{PRESS}}^2 = \frac{\text{PRESS}}{\sum_{i=1}^n (1 - h_{ii})^{-1}}$$

is unbiased for σ^2 under the Gauss–Markov model in Assumption 4.1, recalling PRESS in (13.6) and the leverage score h_{ii} of unit i .

Remark: Theorem 4.3 shows that $\hat{\sigma}^2 = \text{RSS}/(n - p)$ is unbiased for σ^2 under the Gauss–Markov model. RSS is the “in-sample” residual sum of squares, whereas PRESS is the “leave-one-out” residual sum of squares. The estimator $\hat{\sigma}^2$ is standard, whereas $\hat{\sigma}_{\text{PRESS}}^2$ appeared in Shen et al. (2023).

13.5 Simulation with misspecified linear models

Replicate the simulation in Example 13.1 with correlated covariates and an outcome model with quadratic terms of covariates.

13.6 Best subset selection in `lalonae` data

Produce the figure similar to the ones in Figure 13.4 based on the `lalonae` data in the `Matching` package. Report the selected model based on AIC, BIC, PRESS, and GCV.

13.7 Perfect polynomial

Prove that given distinct x_i ($i = 1, \dots, n$) within $[0, 1]$ and any y_i ($i = 1, \dots, n$), we can always find an $(n - 1)$ th order polynomial

$$p_n(x) = \sum_{j=0}^{n-1} b_j x^j$$

such that

$$p_n(x_i) = y_i, \quad (i = 1, \dots, n).$$

Remark: Use the formula in (A.4) in Appendix A.

14

Ridge Regression

The OLS estimator has many nice properties. For example, Chapter 4 shows that it is BLUE under the Gauss–Markov model, and Chapter 5 shows that it follows a Normal distribution that allows for finite-sample exact inference under the Normal linear model. However, OLS has problems with the columns of X are highly correlated. This issue becomes salient when the number of covariates p is large compared with the sample size n . In particular, when $p > n$, the OLS estimator is not unique because $X^T X$ is not invertible. This chapter will first discuss these issues and then introduce ridge regression as a modification of OLS.

14.1 Introduction to ridge regression

The first motivation of ridge regression is straightforward from the linear algebra perspective. From the formula

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

if the columns of X are highly correlated, then $X^T X$ will be nearly singular; more extremely, if the number of covariates is larger than the sample size, then $X^T X$ has a rank smaller than or equal to n and thus is not invertible. So numerically, the OLS estimator can be unstable due to inverting $X^T X$. Because $X^T X$ must be positive semi-definite, its smallest eigenvalue determines whether it is invertible or not. Hoerl and Kennard (1970) proposed the following ridge estimator as a modification of OLS:

$$\hat{\beta}^{\text{ridge}}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T Y, \quad (\lambda > 0) \quad (14.1)$$

which involves a positive tuning parameter λ . Because the smallest eigenvalue of $X^T X + \lambda I_p$ is larger than or equal to $\lambda > 0$, the ridge estimator is always well defined.

Now I turn to the second equivalent motivation. The OLS estimator minimizes the residual sum of squares

$$\text{RSS}(b_0, b_1, \dots, b_p) = \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip})^2.$$

From Theorem 13.1 on the variance inflation factor, the variances of the OLS estimators increase with additional covariates included in the regression, leading to unnecessarily large estimators by chance. To avoid large OLS coefficients, we can penalize the residual sum of squares criterion with the squared length of the coefficients¹ and use

$$\hat{\beta}^{\text{ridge}}(\lambda) = \arg \min_{b_0, b_1, \dots, b_p} \left\{ \text{RSS}(b_0, b_1, \dots, b_p) + \lambda \sum_{j=1}^p b_j^2 \right\}. \quad (14.2)$$

¹This is also called the Tikhonov regularization (Tikhonov, 1943). See Bickel and Li (2006) for a review of the idea of regularization in statistics.

Again in (14.2), λ is a tuning parameter that ranges from zero to infinity. We first discuss the ridge estimator with a fixed λ and then discuss how to choose it. When $\lambda = 0$, it reduces to OLS; when $\lambda = \infty$, all coefficients must be zero except that $\hat{\beta}_0^{\text{ridge}}(\infty) = \bar{y}$. With $\lambda \in (0, \infty)$, the ridge coefficients are generally smaller than the OLS coefficients, and the penalty shrinks the OLS coefficients toward zero. So the parameter λ controls the magnitudes of the coefficients or the “complexity” of the model. In (14.2), we only penalize the slope parameters not the intercept.

As an equivalent form of (14.2), we can also define the ridge estimator as

$$\begin{aligned} \hat{\beta}^{\text{ridge}}(t) &= \arg \min_{b_0, b_1, \dots, b_p} \text{RSS}(b_0, b_1, \dots, b_p) \\ \text{such that } &\sum_{j=1}^p b_j^2 \leq t. \end{aligned} \quad (14.3)$$

Definitions (14.2) and (14.3) are equivalent because for a given λ , we can always find a t such that the solutions from (14.2) and (14.3) are identical. In fact, the corresponding t and λ satisfy $t = \|\hat{\beta}^{\text{ridge}}(\lambda)\|^2$.

However, the ridge estimator has an obvious problem: it is not invariant to linear transformations of X . In particular, it is not equivalent under different scaling of the covariates. Intuitively, the b_j 's depend on the scale of X_j 's, but the penalty term $\sum_{j=1}^p b_j^2$ puts equal weight on each coefficient. A convention in practice is to standardize the covariates before applying the ridge estimator.²

Condition 14.1 (standardization) *The covariates satisfy*

$$n^{-1} \sum_{i=1}^n x_{ij} = 0, \quad n^{-1} \sum_{i=1}^n x_{ij}^2 = 1, \quad (j = 1, \dots, p)$$

and the outcome satisfy $\bar{y} = 0$.

With all covariates centered at zero, the ridge estimator for the intercept, given any values of the slopes and the tuning parameter λ , equals $\hat{\beta}_0^{\text{ridge}} = \bar{y}$. So if we center the outcomes at mean zero, then we can drop the intercept in the ridge estimators defined in (14.2) and (14.3).

For descriptive simplicity, I will assume Condition 14.1 and call it *standardization* from now on. Condition 14.1 allows us to drop the intercept. Using the matrix form, the ridge estimator minimizes

$$(Y - Xb)^T(Y - Xb) + \lambda b^T b,$$

which is a quadratic function of b . From the first order condition, we have

$$-2X^T\{Y - X\hat{\beta}^{\text{ridge}}(\lambda)\} + 2\lambda\hat{\beta}^{\text{ridge}}(\lambda) = 0.$$

Solve the above linear equation of $\hat{\beta}^{\text{ridge}}(\lambda)$ to obtain

$$\hat{\beta}^{\text{ridge}}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T Y,$$

which coincides with the definition in (14.1). We also have the second-order condition:

$$2X^T X + 2\lambda I_p \succ 0,$$

²I choose this standardization because it is the default choice in the function `lm.ridge` in the R package MASS. In practical data analysis, the covariates may have concrete meanings. In those cases, you may not want to scale the covariates in the way as Condition 14.1. However, the discussion below does not rely on the choice of scaling although it requires centering the covariates and outcome.

is positive definite with any $\lambda > 0$, which verifies that the ridge estimator is indeed the minimizer. The predicted vector is

$$\begin{aligned}\hat{Y}^{\text{ridge}}(\lambda) &= X\hat{\beta}^{\text{ridge}}(\lambda) \\ &= X(X^T X + \lambda I_p)^{-1} X^T Y \\ &= H(\lambda)Y,\end{aligned}$$

where

$$H(\lambda) = X(X^T X + \lambda I_p)^{-1} X^T$$

is the hat matrix for ridge regression. When $\lambda = 0$, it reduces to the hat matrix for the OLS; when $\lambda > 0$, it is not a projection matrix because $\{H(\lambda)\}^2 \neq H(\lambda)$.

14.2 Ridge regression via the SVD of the covariate matrix

I will focus on the case with $n \geq p$ and relegate the discussion of the case with $n \leq p$ to Problem 14.12. To facilitate the presentation, I will use the singular value decomposition (SVD) of the covariate matrix:

$$X = UDV^T$$

where $U \in \mathbb{R}^{n \times p}$ has orthonormal columns such that $U^T U = I_p$, $V \in \mathbb{R}^{p \times p}$ is an orthogonal matrix with $VV^T = V^T V = I_p$, and $D \in \mathbb{R}^{p \times p}$ is a diagonal matrix consisting of the singular values. Figure 14.1 illustrates the SVD. For readers who are not familiar with SVD, please review Appendix A before reading the remaining parts of this chapter.

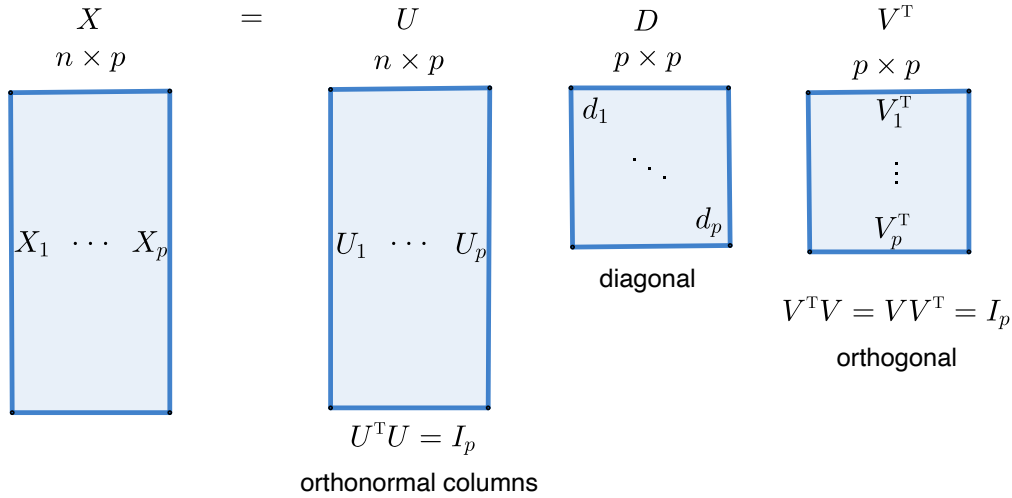


FIGURE 14.1: SVD of X , when $n > p$

The SVD of X implies the eigen-decomposition of $X^T X$:

$$X^T X = V D^2 V^T$$

with eigenvectors V_j being the column vectors of V and eigenvalues d_j^2 being the squared singular values. Lemma 14.1 below is crucial for simplifying the theory and computation.

Lemma 14.1 *The ridge coefficient equals*

$$\hat{\beta}^{\text{ridge}}(\lambda) = V \text{diag} \left(\frac{d_j}{d_j^2 + \lambda} \right) U^T Y,$$

where the diagonal matrix is $p \times p$.

Proof of Lemma 14.1: The ridge coefficient equals

$$\begin{aligned} \hat{\beta}^{\text{ridge}}(\lambda) &= (X^T X + \lambda I_p)^{-1} X^T Y \\ &= (V D U^T U D V^T + \lambda I_p)^{-1} V D U^T Y \\ &= V(D^2 + \lambda I_p)^{-1} V^T V D U^T Y \\ &= V(D^2 + \lambda I_p)^{-1} D U^T Y \\ &= V \text{diag} \left(\frac{d_j}{d_j^2 + \lambda} \right) U^T Y. \end{aligned}$$

□

14.3 Statistical properties

The Gauss–Markov theorem shows that the OLS estimator is BLUE under the Gauss–Markov model: $Y = X\beta + \varepsilon$, where ε has mean zero and covariance $\sigma^2 I_n$. Then in what sense, can ridge regression improve OLS? I will discuss the statistical properties of the ridge estimator under the Gauss–Markov model.

Based on Lemma 14.1, we can calculate the mean of the ridge estimator:

$$\begin{aligned} E\{\hat{\beta}^{\text{ridge}}(\lambda)\} &= V \text{diag} \left(\frac{d_j}{d_j^2 + \lambda} \right) U^T X \beta \\ &= V \text{diag} \left(\frac{d_j}{d_j^2 + \lambda} \right) U^T U D V^T \beta \\ &= V \text{diag} \left(\frac{d_j^2}{d_j^2 + \lambda} \right) V^T \beta, \end{aligned}$$

which does not equal β in general. So the ridge estimator is biased. We can also calculate the covariance matrix of the ridge estimator:

$$\begin{aligned} \text{cov}\{\hat{\beta}^{\text{ridge}}(\lambda)\} &= \sigma^2 V \text{diag} \left(\frac{d_j}{d_j^2 + \lambda} \right) U^T U \text{diag} \left(\frac{d_j}{d_j^2 + \lambda} \right) V^T \\ &= \sigma^2 V \text{diag} \left(\frac{d_j^2}{(d_j^2 + \lambda)^2} \right) V^T. \end{aligned}$$

The mean squared error (MSE) is a measure capturing the bias-variance trade-off:

$$\text{MSE}(\lambda) = E \left[\left\{ \hat{\beta}^{\text{ridge}}(\lambda) - \beta \right\}^T \left\{ \hat{\beta}^{\text{ridge}}(\lambda) - \beta \right\} \right].$$

Using Theorem B.8 on the expectation of a quadratic form, we have

$$\text{MSE}(\lambda) = \underbrace{[E\{\hat{\beta}^{\text{ridge}}(\lambda)\} - \beta]^T [E\{\hat{\beta}^{\text{ridge}}(\lambda)\} - \beta]}_{C_1} + \underbrace{\text{trace}[\text{cov}\{\hat{\beta}^{\text{ridge}}(\lambda)\}]}_{C_2}.$$

Theorem 14.1 below simplifies C_1 and C_2 .

Theorem 14.1 *Under Assumption 4.1, the ridge estimator satisfies*

$$C_1 = \lambda^2 \sum_{j=1}^p \frac{\gamma_j^2}{(d_j^2 + \lambda)^2}, \quad (14.4)$$

where $\gamma = V^T \beta = (\gamma_1, \dots, \gamma_p)^T$ has the j th coordinate $\gamma_j = V_j^T \beta$, and

$$C_2 = \sigma^2 \sum_{j=1}^p \frac{d_j^2}{(d_j^2 + \lambda)^2}. \quad (14.5)$$

Proof of Theorem 14.1: First, $\hat{\beta}^{\text{ridge}}(\lambda)$ is biased for estimating β :

$$E\{\hat{\beta}^{\text{ridge}}(\lambda)\} - \beta = \left\{ V \text{diag} \left(\frac{d_j^2}{d_j^2 + \lambda} \right) V^T - I_p \right\} \beta.$$

Therefore,

$$\begin{aligned} C_1 &= \beta^T \left\{ V \text{diag} \left(\frac{d_j^2}{d_j^2 + \lambda} \right) V^T - I_p \right\}^2 \beta \\ &= \beta^T V \text{diag} \left(\frac{\lambda^2}{(d_j^2 + \lambda)^2} \right) V^T \beta \\ &= \gamma^T \text{diag} \left(\frac{\lambda^2}{(d_j^2 + \lambda)^2} \right) \gamma \\ &= \lambda^2 \sum_{j=1}^p \frac{\gamma_j^2}{(d_j^2 + \lambda)^2}. \end{aligned}$$

Second, we have

$$\begin{aligned} C_2 &= \sigma^2 \text{trace} \left(V \text{diag} \left(\frac{d_j^2}{(d_j^2 + \lambda)^2} \right) V^T \right) \\ &= \sigma^2 \text{trace} \left(\text{diag} \left(\frac{d_j^2}{(d_j^2 + \lambda)^2} \right) \right) \\ &= \sigma^2 \sum_{j=1}^p \frac{d_j^2}{(d_j^2 + \lambda)^2}. \end{aligned}$$

□

Theorem 14.1 shows the bias-variance trade-off for the ridge estimator. The MSE is

$$\begin{aligned} \text{MSE}(\lambda) &= C_1 + C_2 \\ &= \lambda^2 \sum_{j=1}^p \frac{\gamma_j^2}{(d_j^2 + \lambda)^2} + \sigma^2 \sum_{j=1}^p \frac{d_j^2}{(d_j^2 + \lambda)^2}. \end{aligned}$$

When $\lambda = 0$, the ridge estimator reduces to the OLS estimator: the bias is zero and the variance $\sigma^2 \sum_{j=1}^p d_j^{-2}$ dominates. When $\lambda = \infty$, the ridge estimator reduces to zero: the bias $\sum_{j=1}^p \gamma_j^2$ dominates and the variance is zero. As we increase λ from zero, the bias increases and the variance decreases. So we face a bias-variance trade-off.

14.4 Selection of the tuning parameter

14.4.1 Based on parameter estimation

For parameter estimation, we want to choose the λ that minimizes the MSE. So the optimal λ must satisfy the following first-order condition:

$$\frac{\partial \text{MSE}(\lambda)}{\partial \lambda} = 2 \sum_{j=1}^p \gamma_j^2 \frac{\lambda}{d_j^2 + \lambda} \frac{d_j^2 + \lambda - \lambda}{(d_j^2 + \lambda)^2} - 2\sigma^2 \sum_{j=1}^p \frac{d_j^2}{(d_j^2 + \lambda)^3} = 0$$

which is equivalent to

$$\lambda \sum_{j=1}^p \frac{\gamma_j^2 d_j^2}{(d_j^2 + \lambda)^3} = \sigma^2 \sum_{j=1}^p \frac{d_j^2}{(d_j^2 + \lambda)^3}. \quad (14.6)$$

However, (14.6) is not directly useful because we do not know γ and σ^2 . Three methods below try to solve (14.6) approximately.

Dempster et al. (1977) used OLS to construct an unbiased estimator $\hat{\sigma}^2$ and $\hat{\gamma} = V^T \hat{\beta}$, and then solve λ from

$$\lambda \sum_{j=1}^p \frac{\hat{\gamma}_j^2 d_j^2}{(d_j^2 + \lambda)^3} = \hat{\sigma}^2 \sum_{j=1}^p \frac{d_j^2}{(d_j^2 + \lambda)^3},$$

which is a nonlinear equation of λ .

Hoerl et al. (1975) assumed that $X^T X = I_p$. Then $d_j^2 = 1$ ($j = 1, \dots, p$) and $\gamma = \beta$, and solve λ from

$$\lambda \sum_{j=1}^p \frac{\hat{\beta}_j^2}{(1 + \lambda)^3} = \hat{\sigma}^2 \sum_{j=1}^p \frac{1}{(1 + \lambda)^3},$$

resulting in

$$\lambda_{\text{HKB}} = p\hat{\sigma}^2 / \|\hat{\beta}\|^2.$$

Lawless (1976) used

$$\lambda_{\text{LW}} = p\hat{\sigma}^2 / \hat{\beta}^T D^2 \hat{\beta}$$

to weight the β_j 's based on the eigenvalues of $X^T X$.

But all these methods require estimating (β, σ^2) . If the initial OLS estimator is not reliable, then these estimates of λ are unlikely to be reliable. None of these methods work for the case with $p > n$.

14.4.2 Based on prediction

For prediction, we need slightly different criteria. Without estimating (β, σ^2) , we can use the leave-one-out cross-validation. The leave-one-out formula for the ridge below is similar to that for OLS.

Theorem 14.2 Define $\hat{\beta}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T Y$ as the ridge coefficient (dropping the superscript “ridge” for simplicity), $\hat{\varepsilon}(\lambda) = Y - X\hat{\beta}(\lambda)$ as the residual vector using the full data, and $h_{ii}(\lambda) = x_i^T (X^T X + \lambda I_p)^{-1} x_i$ as the (i, i) th diagonal element of $H(\lambda) = X(X^T X + \lambda I_p)^{-1} X^T$. Define $\hat{\beta}_{[-i]}(\lambda)$ as the ridge coefficient without observation i , and $\hat{\varepsilon}_{[-i]}(\lambda) = y_i - x_i^T \hat{\beta}_{[-i]}(\lambda)$ as the predicted residual. The leave-one-out formulas for ridge regression are

$$\hat{\beta}_{[-i]}(\lambda) = \hat{\beta}(\lambda) - \{1 - h_{ii}(\lambda)\}^{-1} (X^T X + \lambda I_p)^{-1} x_i \hat{\varepsilon}_i(\lambda)$$

and

$$\hat{\varepsilon}_{[-i]}(\lambda) = \hat{\varepsilon}_i(\lambda) / \{1 - h_{ii}(\lambda)\}.$$

I leave the proof of Theorem 14.2 as Problem 14.5.

By Theorem 14.2, the PRESS statistic for ridge is

$$\text{PRESS}(\lambda) = \sum_{i=1}^n \{\hat{\varepsilon}_{[-i]}(\lambda)\}^2 = \sum_{i=1}^n \frac{\{\hat{\varepsilon}_i(\lambda)\}^2}{\{1 - h_{ii}(\lambda)\}^2}.$$

Golub et al. (1979) proposed the GCV criterion to simplify the calculation of the PRESS statistic by replacing $h_{ii}(\lambda)$ with their average value $n^{-1} \text{trace}\{H(\lambda)\}$:

$$\text{GCV}(\lambda) = \frac{\sum_{i=1}^n \{\hat{\varepsilon}_i(\lambda)\}^2}{[1 - n^{-1} \text{trace}\{H(\lambda)\}]^2}.$$

In the R package MASS, the function `lm.ridge` implements the ridge regression, `kHKB` and `kLW` report two estimators for λ , and `gcv` contains the GCV values for a sequence of λ .

14.5 Computation of ridge regression

Lemma 14.1 gives the ridge coefficients. So the predicted vector equals

$$\begin{aligned} \hat{Y}(\lambda) &= X\hat{\beta}^{\text{ridge}}(\lambda) \\ &= UDV^T V \text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right) U^T Y \\ &= U D \text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right) U^T Y \\ &= U \text{diag}\left(\frac{d_j^2}{d_j^2 + \lambda}\right) U^T Y, \end{aligned}$$

and the hat matrix equals

$$H(\lambda) = U \text{diag}\left(\frac{d_j^2}{d_j^2 + \lambda}\right) U^T.$$

These formulas allow us to compute the ridge coefficient and predictor vector for many values of λ without inverting each $X^T X + \lambda I_p$. We have similar formulas for the case with $n < p$; see Problem 14.12.

A subtle point is due to the standardization of the covariates of the outcome. In `R`, the `lm.ridge` function first computes the ridge coefficient based on the standardized covariates and outcome, and then transforms them back to the original scale. Let $\bar{x}_1, \dots, \bar{x}_p, \bar{y}$ be the means of the covariates and outcome, and let $\text{sd}_j = \{n^{-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2\}^{1/2}$ ($j = 1, \dots, p$) be the standard deviation of the covariates which are report as `scales` in the output of `lm.ridge`. From the ridge coefficients $\{\hat{\beta}_1^{\text{ridge}}(\lambda), \dots, \hat{\beta}_p^{\text{ridge}}(\lambda)\}$ based on the standardized variables, we can obtain the predicted values based on the original variables as

$$\hat{y}_i(\lambda) - \bar{y} = \hat{\beta}_1^{\text{ridge}}(\lambda)(x_{i1} - \bar{x}_1)/\text{sd}_1 + \dots + \hat{\beta}_p^{\text{ridge}}(\lambda)(x_{ip} - \bar{x}_p)/\text{sd}_p$$

or, equivalently,

$$\hat{y}_i(\lambda) = \hat{\alpha}^{\text{ridge}}(\lambda) + \hat{\beta}_1^{\text{ridge}}(\lambda)/\text{sd}_1 \times x_{i1} + \dots + \hat{\beta}_p^{\text{ridge}}(\lambda)/\text{sd}_p \times x_{ip}$$

where

$$\hat{\alpha}^{\text{ridge}}(\lambda) = \bar{y} - \hat{\beta}_1^{\text{ridge}}(\lambda)\bar{x}_1/\text{sd}_1 - \dots - \hat{\beta}_p^{\text{ridge}}(\lambda)\bar{x}_p/\text{sd}_p.$$

14.6 Numerical examples

We can use the following numerical example to illustrate the bias-variance trade-off in selecting λ in the ridge.

14.6.1 Uncorrelated covariates

I first simulate data from a Normal linear model with uncorrelated covariates.

```
library(MASS)
n = 200
p = 100
beta = rep(1/sqrt(p), p)
sig = 1/2
X = matrix(rnorm(n*p), n, p)
X = scale(X)
X = X*sqrt(n/(n-1))
Y = as.vector(X*beta + rnorm(n, 0, sig))
```

The following code calculates the theoretical bias, variance, and mean squared error, reported in the (1,1)th panel of Figure 14.2.

```
eigenxx = eigen(t(X)%*%X)
xis = eigenxx$values
gammas = t(eigenxx$vectors)%*%beta

lambda.seq = seq(0, 70, 0.01)
bias2.seq = lambda.seq
var.seq = lambda.seq
mse.seq = lambda.seq
for(i in 1:length(lambda.seq))
{
  ll = lambda.seq[i]
  bias2.seq[i] = ll^2*sum(gammas^2/(xis + ll)^2)
  var.seq[i] = sig^2*sum(xis/(xis + ll)^2)
  mse.seq[i] = bias2.seq[i] + var.seq[i]
}

y.min = min(bias2.seq, var.seq, mse.seq)
```

```

y.max = max(bias2.seq, var.seq, mse.seq)
par(mfrow = c(2, 2))
plot(bias2.seq ~ lambda.seq, type = "l",
     ylim = c(y.min, y.max),
     xlab = expression(lambda), main = "",
     ylab = "bias-variance tradeoff",
     lty = 2, bty = "n")
lines(var.seq ~ lambda.seq, lty = 3)
lines(mse.seq ~ lambda.seq, lwd = 3, lty = 1)
abline(v = lambda.seq[which.min(mse.seq)],
       lty = 1, col = "grey")
legend("topright", c("bias", "variance", "mse"),
       lty = c(2, 3, 1), lwd = c(1, 1, 4), bty = "n")

```

The (1,1)th panel also reported the λ 's based on different approaches.

```

ridge.fit = lm.ridge(Y ~ X, lambda = lambda.seq)
abline(v = lambda.seq[which.min(ridge.fit$GCV)],
       lty = 2, col = "grey")
abline(v = ridge.fit$kHKB, lty = 3, col = "grey")
abline(v = ridge.fit$kLW, lty = 4, col = "grey")
legend("bottomright",
       c("MSE", "GCV", "HKB", "LW"),
       lty = 1:4, col = "grey", bty = "n")

```

I also calculate the prediction error of the ridge estimator in the testing dataset, which follows the same data-generating process as the training dataset. The (1,2)th panel of Figure 14.2 shows its relationship with λ . Overall, GCV, HKB, and LW are similar, but the λ selected by the MSE criterion is the worst for prediction.

```

X.new = matrix(rnorm(n*p), n, p)
X.new = scale(X.new)
X.new = X.new*matrix(sqrt(n/(n-1)), n, p)
Y.new = as.vector(X.new%*%beta + rnorm(n, 0, sig))
predict.error = Y.new - X.new%*%ridge.fit$coef
predict.mse = apply(predict.error^2, 2, mean)
plot(predict.mse ~ lambda.seq, type = "l",
     xlab = expression(lambda),
     ylab = "predicted MSE", bty = "n")
abline(v = lambda.seq[which.min(mse.seq)],
       lty = 1, col = "grey")
abline(v = lambda.seq[which.min(ridge.fit$GCV)],
       lty = 2, col = "grey")
abline(v = ridge.fit$kHKB, lty = 3, col = "grey")
abline(v = ridge.fit$kLW, lty = 4, col = "grey")
legend("bottomright",
       c("MSE", "GCV", "HKB", "LW"),
       lty = 1:4, col = "grey", bty = "n")

mtext("independent covariates", side = 1,
      line = -58, outer = TRUE, font.main = 1, cex=1.5)

```

14.6.2 Correlated covariates

I then simulate data from a Normal linear model with correlated covariates.

```

n = 200
p = 100
beta = rep(1/sqrt(p), p)
sig = 1/2
## correlated Normals
X = matrix(rnorm(n*p), n, p) + rnorm(n, 0, 0.5)
## standardize the covariates

```

```

X = scale(X)
X = X*matrix(sqrt(n/(n-1)), n, p)
Y = as.vector(X%*%beta + rnorm(n, 0, sig))

```

The second row of Figure 14.2 shows the bias-variance trade-off. Overall, GCV works the best for selecting λ for prediction.

14.7 Further comments on OLS, ridge, and PCA

The SVD of X is closely related to the *principal component analysis* (PCA). Assume that the columns of X are centered, so $X^T X = V D^2 V^T$ is proportional to the sample covariance matrix of X . Assume $d_1 \geq d_2 \geq \dots$. PCA tries to find linear combinations of the covariate x_i that contain maximal information. For a vector $v \in \mathbb{R}^p$, the linear combination $v^T x_i$ has sample variance proportional to

$$Q(v) = v^T X^T X v.$$

If we multiply v by a constant c , the above sample variance will change by the factor c^2 . So a meaningful criterion is to maximize $Q(v)$ such that $\|v\| = 1$. This is exactly the setting of Theorem A.3. The maximum value equals d_1^2 which is achieved by V_1 , the first column of V . We call

$$XV_1 = \begin{pmatrix} x_1^T V_1 \\ \vdots \\ x_n^T V_1 \end{pmatrix}$$

the first principal component of X . Similar to Theorem A.3, we can further maximize $Q(v)$ such that $\|v\| = 1$ and $v \perp V_1$, yielding the maximum value d_2^2 which is achieved by V_2 . We call XV_2 the second principal component of X . By induction, we can define all the p principal components, stacked in the following $n \times p$ matrix:

$$(XV_1, \dots, XV_p) = XV = UDV^T V = UD.$$

So UD in the SVD decomposition contains the principal components of X . Since D is a diagonal matrix that only changes the scales of the columns of U , we also call $U = (U_1, \dots, U_p)$ the principal components of X . They are orthogonal since $U^T U = I_p$.

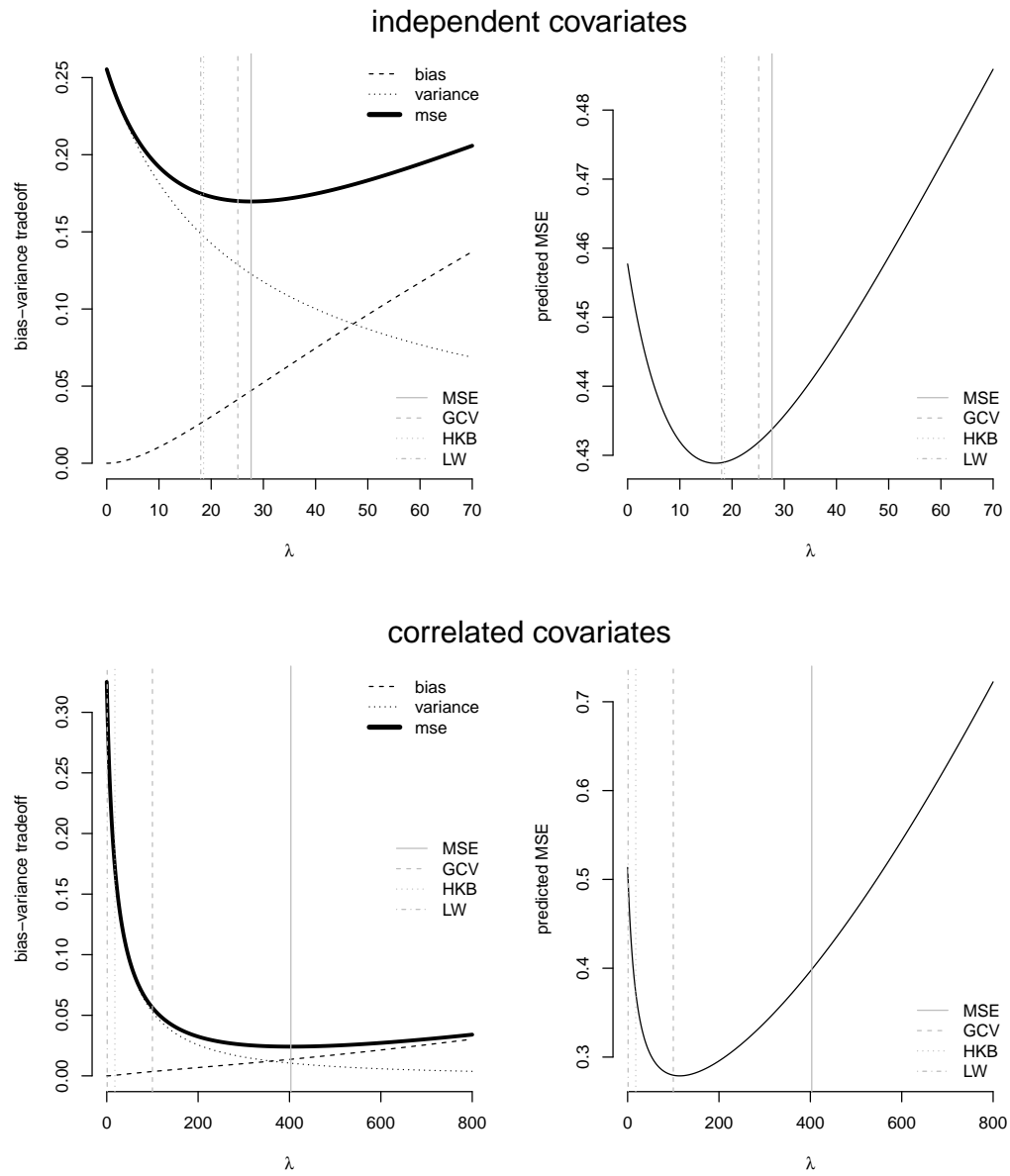
Section 14.5 shows that the ridge estimator yields the predicted value

$$\begin{aligned} \hat{Y}(\lambda) &= U \operatorname{diag} \left(\frac{d_j^2}{d_j^2 + \lambda} \right) U^T Y \\ &= \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda} \langle U_j, Y \rangle U_j \end{aligned}$$

where $\langle U_j, Y \rangle = U_j^T Y$ denotes the inner product of vectors U_j and Y . As a special case with $\lambda = 0$, the OLS estimator yields the predicted value

$$\begin{aligned} \hat{Y} &= U U^T Y \\ &= \sum_{j=1}^p \langle U_j, Y \rangle U_j, \end{aligned}$$

which is identical to the predicted value based on OLS of Y on the principal components U . Moreover, the principal components in U are orthogonal and have unit length, so by



Corollary 7.2, the OLS fit of Y on U is equivalent to the component-wise OLS of Y on U_j with coefficient $\langle U_j, Y \rangle$ ($j = 1, \dots, p$). So the predicted value based OLS equals a linear combination of the principal components with coefficients $\langle U_j, Y \rangle$; the predicted value based on ridge also equals a linear combination of the principal components but the coefficients are shrunk by the factors $d_j^2/(d_j^2 + \lambda)$.

When the columns of X are not linearly independent, for example, $p > n$, we cannot run OLS of Y on X or OLS of Y on U , but we can still run ridge regression. Motivated by the formulas above, another approach is to run OLS of Y on the first p^* principal components $\tilde{U} = (U_1, \dots, U_{p^*})$ with $p^* < p$. This is called the *principal component regression* (PCR). The predicted value is

$$\begin{aligned}\hat{Y}(p^*) &= (\tilde{U}^T \tilde{U})^{-1} \tilde{U}^T Y \\ &= \sum_{j=1}^{p^*} \langle U_j, Y \rangle U_j,\end{aligned}$$

which truncates the summation in the formula of \hat{Y} based on OLS. Compared with the predicted values of OLS and ridge regression, $\hat{Y}(p^*)$ effectively imposes zero weights on the principal components corresponding to small singular values. It depends on a tuning parameter p^* similar to λ in the ridge regression. Since p^* must be a positive integer and λ can be any positive real value, PCR is a discrete procedure while ridge regression is a continuous procedure.

14.8 Homework problems

14.1 Ridge coefficient as a posterior mode under a Normal prior

Assume fixed X, σ^2 and τ^2 . Prove that if

$$Y \mid \beta \sim N(X\beta, \sigma^2 I_n) \quad (14.7)$$

and

$$\beta \sim N(0, \tau^2 I_p), \quad (14.8)$$

then the mode of the posterior distribution of $\beta \mid Y$ equals $\hat{\beta}^{\text{ridge}}(\sigma^2/\tau^2)$:

$$\hat{\beta}^{\text{ridge}}(\sigma^2/\tau^2) = \arg \max_{\beta} f(\beta \mid Y)$$

where $f(\beta \mid Y)$ is the posterior density of β given Y .

Remark: In Bayesian statistics, (14.7) is the Normal linear model, whereas (14.8) is the *prior distribution* of the parameter β .

14.2 Derivative of the MSE

Prove that

$$\left. \frac{\partial \text{MSE}(\lambda)}{\partial \lambda} \right|_{\lambda=0} < 0.$$

Remark: This result ensures that the ridge estimator must have a smaller MSE than OLS in a neighborhood of $\lambda = 0$, which is coherent with the pattern in Figure 14.2.

14.3 Ridge and OLS

Prove that if X has linearly independent columns, then

$$\begin{aligned}\hat{\beta}^{\text{ridge}}(\lambda) &= (X^T X + \lambda I_p)^{-1} X^T X \hat{\beta} \\ &= V \text{diag} \left(\frac{d_j^2}{d_j^2 + \lambda} \right) V^T \hat{\beta}\end{aligned}$$

where $\hat{\beta}$ is the OLS coefficient.

14.4 Ridge as OLS with augmented data

Prove that $\hat{\beta}^{\text{ridge}}(\lambda)$ equals the OLS coefficient of \tilde{Y} on \tilde{X} with augmented data

$$\tilde{Y} = \begin{pmatrix} Y \\ 0_p \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda} I_p \end{pmatrix},$$

where \tilde{Y} is an $n + p$ dimensional vector and \tilde{X} is an $(n + p) \times p$ matrix.

Remark: The columns of \tilde{X} must be linearly independent, so the inverse of $\tilde{X}^T \tilde{X}$ always exists. This is a theoretical result of the ridge regression. It should not be used for computation especially when p is large.

14.5 Leave-one-out formulas for ridge

Prove Theorem 14.2.

Remark: You can use the result in Problem 14.4 and apply the leave-one-out formulas for OLS in Theorems 11.2 and 11.3.

14.6 Generalized ridge regression

Covariates have different importance, so it is reasonable to use different weights in the penalty term. Find the explicit formula for the ridge regression with general quadratic penalty:

$$\arg \min_{b \in \mathbb{R}^p} \{(Y - Xb)^T(Y - Xb) + \lambda b^T Q b\}$$

where Q is a $p \times p$ positive definite matrix.

14.7 Degrees of freedom of ridge regression

For a predictor \hat{Y} for Y , define the degrees of freedom of the predictor as $\sum_{i=1}^n \text{cov}(y_i, \hat{y}_i) / \sigma^2$. Calculate the degrees of freedom of ridge regression in terms of the eigenvalues of $X^T X$.

14.8 Extending the simulation in Figure 14.2

Re-run the simulation that generates Figure 14.2, and report the λ selected by Dempster et al. (1977)'s method, PRESS, and K -fold CV. Extend the simulation to the case with $p > n$.

14.9 Unification of OLS, ridge, and PCR

We can unify the predicted values of the OLS, ridge, and PCR as

$$\hat{Y} = \sum_{j=1}^p s_j \langle U_j, Y \rangle U_j,$$

where

$$s_j = \begin{cases} 1, & \text{OLS,} \\ \frac{d_j^2}{d_j^2 + \lambda}, & \text{ridge,} \\ 1(j \leq p^*), & \text{PCR.} \end{cases}$$

Based on the unified formula, show that under Assumption 4.1, we have

$$E(\hat{Y}) = \sum_{j=1}^p s_j d_j \gamma_j U_j$$

with the γ_j 's defined in Theorem 14.1, and

$$\text{cov}(\hat{Y}) = \sigma^2 \sum_{j=1}^p s_j^2 U_j U_j^T.$$

14.10 An equivalent form of ridge coefficient

Prove Theorem 14.3 below which states two equivalent forms of the ridge coefficient.

Theorem 14.3 For $\lambda > 0$, we have

$$\hat{\beta}^{\text{ridge}}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T Y = X^T (X X^T + \lambda I_n)^{-1} Y.$$

Remark: Theorem 14.3 has several interesting implications. The first form of the ridge coefficient involves inverting a $p \times p$ matrix, and it is more useful when $p < n$. The second form of the ridge coefficient involves inverting an $n \times n$ matrix, so it is more useful when $p > n$. From the second form of the ridge coefficient, we can see that the ridge estimator lies in $\mathcal{C}(X^T)$, the row space of X . That is, the ridge estimator can be written as $X^T \delta$, where $\delta = (X X^T + \lambda I_n)^{-1} Y \in \mathbb{R}^p$. This always holds but is particularly interesting in the case with $p > n$ when the row space of X is not the entire \mathbb{R}^p . Third, if $p > n$ and $X X^T$ is invertible, then we can let λ go to zero on the right-hand side, yielding

$$\hat{\beta}^{\text{ridge}}(0) = X^T (X X^T)^{-1} Y$$

which is the minimum norm OLS estimator; see Problem 18.7. Using the definition of the pseudoinverse in Appendix A, we can further show that

$$\hat{\beta}^{\text{ridge}}(0) = X^+ Y.$$

14.11 Ridge estimator as the minimum norm OLS estimator with augmented features

This problem is the dual problem of Problem 14.4.

Prove that $\hat{\beta}^{\text{ridge}}(\lambda)$ equals the first p components of the minimum norm OLS estimator of Y on \tilde{X} :

$$\tilde{X}^T (\tilde{X} \tilde{X}^T)^{-1} Y$$

with the $n \times (p + n)$ matrix

$$\tilde{X} = (X, \sqrt{\lambda} I_n).$$

Remark: Use the results in Problem 14.10.

14.12 Computation of ridge with $n < p$

When $n < p$, X has singular value decomposition $X = UDV^T$, where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix containing the singular values, $U \in \mathbb{R}^{n \times n}$ is an orthogonal matrix with $UU^T = U^T U = I_n$, and $V \in \mathbb{R}^{p \times n}$ has orthonormal columns with $V^T V = I_n$.

Prove that the ridge coefficient, the predicted value, and the hat matrix have the same form as the case with $n > p$. The only subtle difference is that the diagonal matrices have dimension $n \times n$.

Remark: The above result also ensures that Theorem 14.1 holds when $p > n$ if we modify the summation as “from $j = 1$ to n .”

14.13 Recommended reading

To celebrate the 50th anniversary of Hoerl and Kennard (1970)’s paper in *Technometrics*, the editor invited Roger W. Hoerl, the son of Art Hoerl, to review the historical aspects of the original paper, and Trevor Hastie to review the essential role of the idea of ridge regression in data science. See Hoerl (2020) and Hastie (2020).



15

Lasso

Ridge regression works well for prediction, but it may be difficult to interpret many small but non-zero ridge coefficients. Tibshirani (1996) proposed to use the lasso, the acronym for the Least Absolute Shrinkage and Selection Operator, to achieve the ambitious goal of simultaneously estimating parameters and selecting important variables in the linear regression. By changing the penalty term in ridge regression, the lasso automatically estimates some parameters as zero, dropping them out of the model and thus selecting the remaining variables as important predictors. This chapter introduces the lasso.

15.1 Introduction to the lasso

Recall the definition of the residual sum of squares

$$\text{RSS}(b_0, b_1, \dots, b_p) = \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip})^2.$$

Tibshirani (1996) defined the lasso as

$$\begin{aligned} \hat{\beta}^{\text{lasso}}(t) &= \arg \min_{b_0, b_1, \dots, b_p} \text{RSS}(b_0, b_1, \dots, b_p) \\ &\text{such that } \sum_{j=1}^p |b_j| \leq t. \end{aligned} \quad (15.1)$$

Osborne et al. (2000) studied its equivalent form

$$\hat{\beta}^{\text{lasso}}(\lambda) = \arg \min_{b_0, b_1, \dots, b_p} \left\{ \text{RSS}(b_0, b_1, \dots, b_p) + \lambda \sum_{j=1}^p |b_j| \right\}. \quad (15.2)$$

The two forms of lasso are equivalent in the sense that for a given λ in (15.2), there exists a t such that the solution for (15.1) is identical to the solution for (15.2). In particular, $t = \sum_{j=1}^p \hat{\beta}_j^{\text{lasso}}(\lambda)$. Technically, the minimizer of the lasso problem may not be unique especially when $p > n$, so the right-hand sides of the optimization problems (15.1) and (15.2) should be a set. Fortunately, even though the minimizer may not be unique, the resulting predictor is always unique. Tibshirani (2013) clarifies this issue; see Problem 15.1.

Both forms (15.1) and (15.2) are useful for understanding the lasso. We will use the form (15.1) for geometric intuition and use the form (15.2) for computation. Similar to ridge regression, the lasso is not invariant to the linear transformation of X . We proceed after standardizing the covariates and outcome as Condition 14.1. For the same reason as ridge regression, we can drop the intercept after standardization.

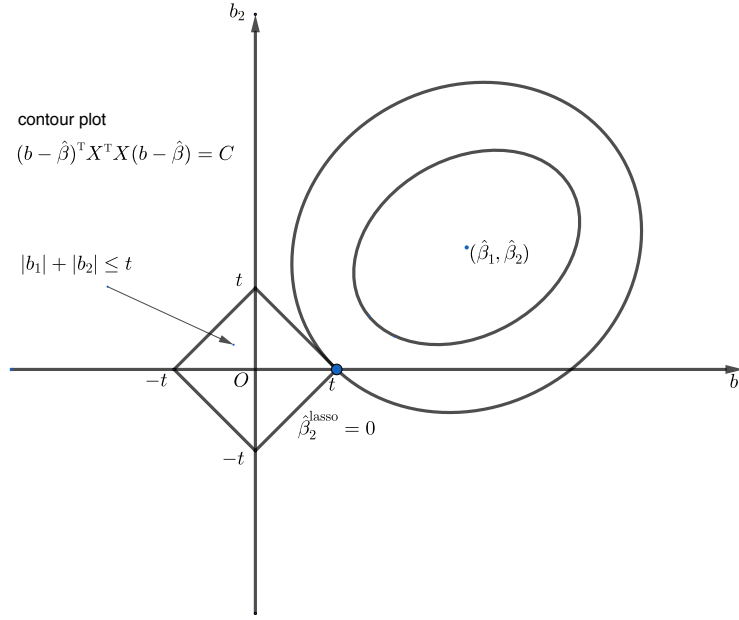


FIGURE 15.1: Lasso with a sparse solution

15.2 Comparing the lasso and ridge: a geometric perspective

The ridge and lasso are very similar: both minimize a penalized version of the residual sum of squares. They differ in the penalty term: ridge uses an L_2 penalty, i.e., the L_2 norm of the coefficient $\|b\|^2 = \sum_{j=1}^p b_j^2$, and lasso uses an L_1 penalty, i.e., the L_1 norm of the coefficient $\|b\|_1 = \sum_{j=1}^p |b_j|$. Compared to the ridge, the lasso can give sparse solutions due to the non-smooth penalty term. That is, estimators of some coefficients are exactly zero.

Focus on the form (15.1). We can gain insights from the contour plot of the residual sum of squares as a function of b . With a well-defined OLS estimator $\hat{\beta}$, Theorem 3.2 ensures

$$(Y - Xb)^T(Y - Xb) = (Y - X\hat{\beta})^T(Y - X\hat{\beta}) + (b - \hat{\beta})^T X^T X (b - \hat{\beta}),$$

which equals a constant term plus a quadratic function centered at the OLS coefficient. Without any penalty, the minimizer is of course the OLS coefficient. With the L_1 penalty, the OLS coefficient may not be in the region defined by $\sum_{j=1}^p |b_j| \leq t$. If this happens, the intersection of the contour plot of $(Y - Xb)^T(Y - Xb)$ and the border of the restriction region $\sum_{j=1}^p |b_j| \leq t$ can be at some axis. For example, Figure 15.1 shows a case with $p = 2$, and the lasso estimator hits the x-axis, resulting in a zero coefficient for the second coordinate. However, this does not mean that lasso always generates sparse solutions because sometimes the intersection of the contour plot of $(Y - Xb)^T(Y - Xb)$ and the border of the restriction region is at an edge of the region. For example, Figure 15.2 shows a case with a non-sparse lasso solution.

In contrast, the restriction region of the ridge is a circle, so the ridge solution does not hit any axis unless the original OLS coefficient is zero. Figure 15.3 shows the general ridge estimator.

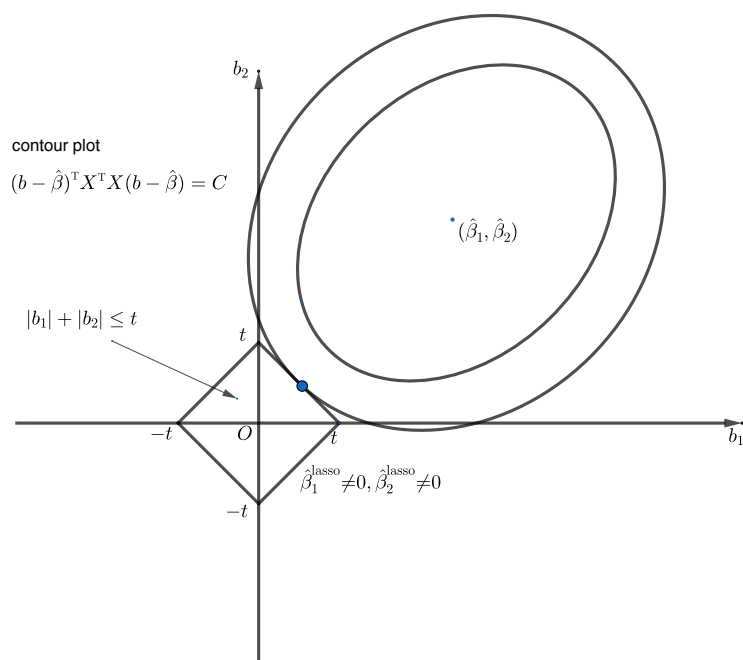


FIGURE 15.2: Lasso with a non-sparse solution

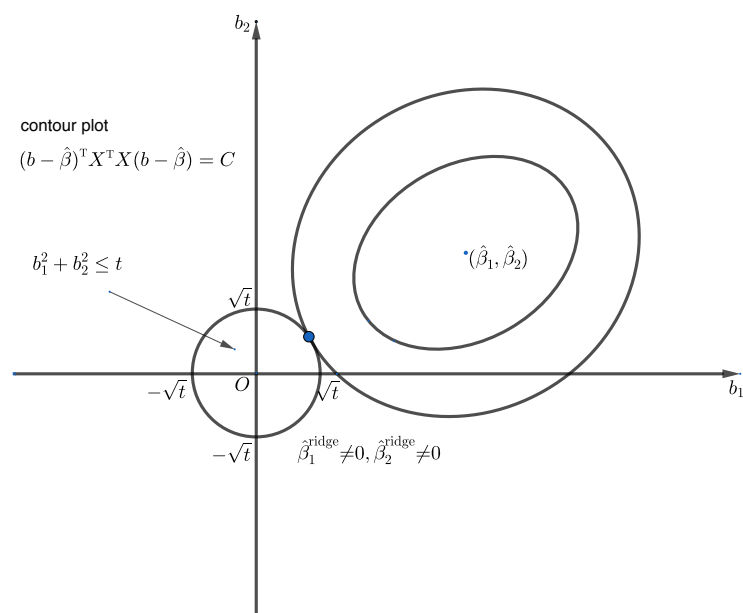


FIGURE 15.3: Ridge regression

15.3 Computing the lasso coefficients via coordinate descent

Many efficient algorithms can solve the lasso problem. The `glmnet` package in `R` uses the coordinate descent algorithm based on the form (15.2) (Friedman et al., 2007, 2010). I will first review a lemma which is the stepstone for the algorithm.

15.3.1 The soft-thresholding lemma

Let $\text{sign}(x)$ denote the sign of a real number x , which equals 1, 0, -1 if $x > 0, x = 0, x < 0$, respectively. Let $(x)_+ = \max(x, 0)$ denote the positive part of a real number x .

Lemma 15.1 *Given b_0 and $\lambda \geq 0$, we have*

$$\begin{aligned} \arg \min_{b \in \mathbb{R}} \frac{1}{2}(b - b_0)^2 + \lambda|b| &= \text{sign}(b_0) (|b_0| - \lambda)_+ \\ &= \begin{cases} b_0 - \lambda, & \text{if } b_0 \geq \lambda, \\ 0 & \text{if } -\lambda \leq b_0 \leq \lambda, \\ b_0 + \lambda & \text{if } b_0 \leq -\lambda. \end{cases} \end{aligned}$$

The solution in Lemma 15.1 is a function of b_0 and λ , and we will use the notation

$$S(b_0, \lambda) = \text{sign}(b_0) (|b_0| - \lambda)_+$$

in this chapter, where S denotes the soft-thresholding operator. For a given $\lambda > 0$, it is a function of b_0 illustrated by Figure 15.4. The proof of Lemma 15.1 is to solve the optimization problem. It is tricky since we cannot naively solve the first-order condition due to the non-smoothness of $|b|$ at 0. Nevertheless, it is only a one-dimensional optimization problem, and I relegate the proof to Problem 15.2.

15.3.2 Coordinate descent for the lasso

For a given $\lambda > 0$, we can use the following algorithm:

1. Standardize the data to satisfy Condition 14.1. So we need to solve a lasso problem without the intercept. For simplicity of derivation, we change the scale of the residual sum of squares without essentially changing the problem:¹

$$\min_{b_1, \dots, b_p} \frac{1}{2n} \sum_{i=1}^n (y_i - b_1 x_{i1} - \dots - b_p x_{ip})^2 + \lambda \sum_{j=1}^p |b_j|.$$

Initialize $\hat{\beta}$.

2. Update $\hat{\beta}_j$ given all other coefficients. Define the partial residual as $r_{ij} = y_i - \sum_{k \neq j} \hat{\beta}_k x_{ik}$. Updating $\hat{\beta}_j$ is equivalent to minimizing

$$\frac{1}{2n} \sum_{i=1}^n (r_{ij} - b_j x_{ij})^2 + \lambda |b_j|.$$

¹It will change the scale of λ . However, λ is a tuning parameter anyway, which will often be determined via cross-validation.

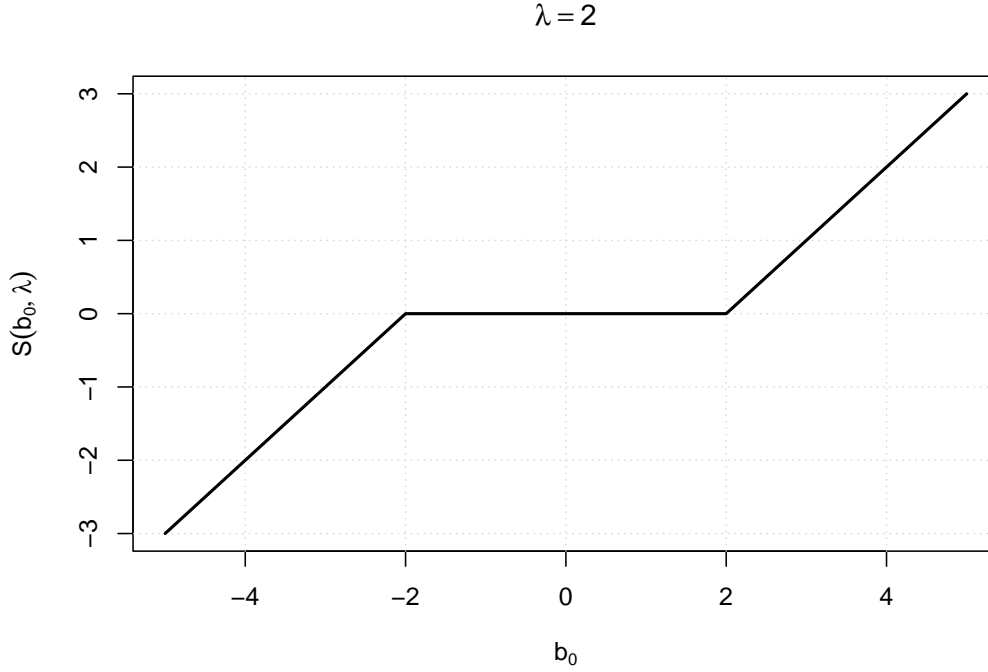


FIGURE 15.4: Soft-thresholding

Define

$$\hat{\beta}_{j,0} = \frac{\sum_{i=1}^n x_{ij} r_{ij}}{\sum_{i=1}^n x_{ij}^2} = n^{-1} \sum_{i=1}^n x_{ij} r_{ij}$$

as the OLS coefficient of the r_{ij} 's on the x_{ij} 's, so

$$\begin{aligned} \frac{1}{2n} \sum_{i=1}^n (r_{ij} - b_j x_{ij})^2 &= \frac{1}{2n} \sum_{i=1}^n (r_{ij} - \hat{\beta}_{j,0} x_{ij})^2 + \frac{1}{2n} \sum_{i=1}^n x_{ij}^2 (b_j - \hat{\beta}_{j,0})^2 \\ &= \text{constant} + \frac{1}{2} (b_j - \hat{\beta}_{j,0})^2. \end{aligned}$$

Then updating $\hat{\beta}_j$ is equivalent to minimizing $\frac{1}{2} (b_j - \hat{\beta}_{j,0})^2 + \lambda |b_j|$. Lemma 15.1 implies

$$\hat{\beta}_j = S(\hat{\beta}_{j,0}, \lambda).$$

3. Iterate until convergence.

Does the algorithm always converge? The theory of Tseng (2001) ensures it converges, but this is beyond the scope of this book. We can start with a large λ and all 0 lasso coefficients. We then gradually decrease λ , and for each λ , we apply the above algorithm. We finally select λ via K -fold cross-validation. Since we gradually decrease λ , the initial values from the last step are very close to the minimizer and the algorithm converges fairly fast.

15.4 Example: comparing OLS, ridge and lasso

In the Boston housing data, the OLS, ridge, and lasso have similar performance in out-of-sample prediction. Lasso and ridge have similar coefficients. See Figure 15.5(a).

```
> library("mlbench")
> library("glmnet")
> library("MASS")
> data(BostonHousing)
>
> ## training and testing data
> set.seed(230)
> nsample = dim(BostonHousing)[1]
> trainindex = sample(1:nsample, floor(nsample*0.9))
>
> xmatrix = model.matrix(medv ~ ., data = BostonHousing)[, -1]
> yvector = BostonHousing$medv
> dat = data.frame(yvector, xmatrix)
>
> ## linear regression
> bostonlm = lm(yvector ~ ., data = dat[trainindex, ])
> predictorerror = dat$yvector[- trainindex] -
+               predict(bostonlm, dat[- trainindex, ])
> mse.ols = sum(predictorerror^2)/length(predictorerror)
>
> ## ridge regression
> lambdas= seq(0, 5, 0.01)
> lm0 = lm.ridge(yvector ~ ., data = dat[trainindex, ],
+               lambda = lambdas)
> coefridge = coef(lm0)[which.min(lm0$GCV), ]
> predictorerrorridge = dat$yvector[- trainindex] -
+               cbind(1, xmatrix[- trainindex, ])%*%coefridge
> mse.ridge = sum(predictorerrorridge^2)/length(predictorerrorridge)
>
> ## lasso
> cvboston = cv.glmnet(x = xmatrix[trainindex, ], y = yvector[trainindex])
> coeflasso = coef(cvboston, s = "lambda.min")
> predictorerrorlasso = dat$yvector[- trainindex] -
+               cbind(1, xmatrix[- trainindex, ])%*%coeflasso
> mse.lasso = sum(predictorerrorlasso^2)/length(predictorerrorlasso)
>
> c(mse.ols, mse.ridge, mse.lasso)
[1] 29.37365 29.07174 28.88161
```

But if we artificially add 200 columns of covariates of pure noise $N(0, 1)$, then the ridge and lasso perform much better. Lasso can automatically shrink many coefficients to zero. See Figure 15.5(b).

```
> ## adding more noisy covariates
> n.noise = 200
> xnoise = matrix(rnorm(nsample*n.noise), nsample, n.noise)
> xmatrix = cbind(xmatrix, xnoise)
> dat = data.frame(yvector, xmatrix)
>
> ## linear regression
> bostonlm = lm(yvector ~ ., data = dat[trainindex, ])
> predictorerror = dat$yvector[- trainindex] -
+               predict(bostonlm, dat[- trainindex, ])
> mse.ols = sum(predictorerror^2)/length(predictorerror)
>
> ## ridge regression
> lambdas= seq(100, 150, 0.01)
```

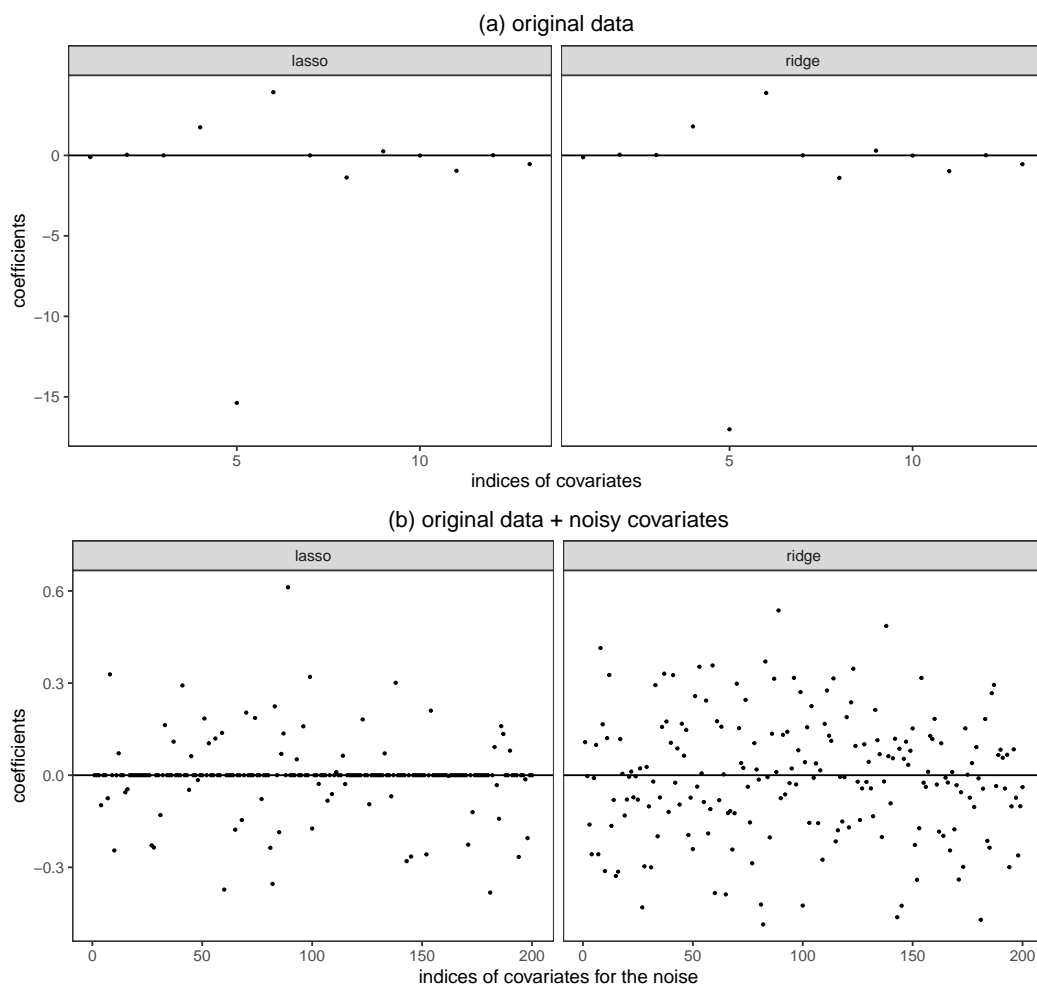


FIGURE 15.5: Comparing ridge and lasso

```

> lm0 = lm.ridge(yvector ~ ., data = dat[trainindex, ],
+               lambda = lambdas)
> coefridge = coef(lm0)[which.min(lm0$GCV), ]
> predicterrorridge = dat$yvector[- trainindex] -
+   cbind(1, xmatrix[- trainindex, ])%%coefridge
> mse.ridge = sum(predicterrorridge^2)/length(predicterrorridge)
>
>
> ## lasso
> cvboston = cv.glmnet(x = xmatrix[trainindex, ], y = yvector[trainindex])
> coeflasso = coef(cvboston, s = "lambda.min")
>
> predicterrorlasso = dat$yvector[- trainindex] -
+   cbind(1, xmatrix[- trainindex, ])%%coeflasso
> mse.lasso = sum(predicterrorlasso^2)/length(predicterrorlasso)
>
> c(mse.ols, mse.ridge, mse.lasso)
[1] 41.80376 33.33372 32.64287

```

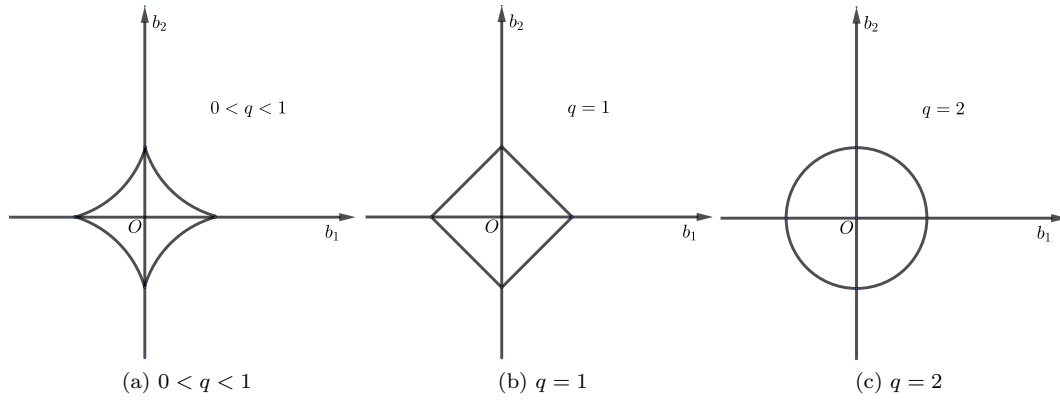


FIGURE 15.6: Shrinkage estimators

15.5 Other shrinkage estimators

15.5.1 Bridge estimator

A general class of shrinkage estimators is the bridge estimator (Frank and Friedman, 1993):

$$\hat{\beta}(\lambda) = \arg \min_{b_0, b_1, \dots, b_p} \left\{ \text{RSS}(b_0, b_1, \dots, b_p) + \lambda \sum_{j=1}^p |b_j|^q \right\},$$

or, equivalently,

$$\begin{aligned} \hat{\beta}(t) = \arg \min_{b_0, b_1, \dots, b_p} & \text{RSS}(b_0, b_1, \dots, b_p) \\ \text{such that } & \sum_{j=1}^p |b_j|^q \leq t. \end{aligned}$$

Figure 15.6 shows the constraints corresponding to different values of q .

15.5.2 Elastic net

Zou and Hastie (2005) proposed the elastic net, which combines the penalties of the lasso and ridge:

$$\hat{\beta}^{\text{enet}}(\lambda, \alpha) = \arg \min_{b_0, b_1, \dots, b_p} \left[\text{RSS}(b_0, b_1, \dots, b_p) + \lambda \sum_{j=1}^p \left\{ \alpha b_j^2 + (1 - \alpha) |b_j| \right\} \right].$$

Figure 15.7 compares the constraints corresponding to the ridge, lasso, and elastic net. Because the constraint of the elastic net is not smooth, it encourages sparse solutions in the same way as the lasso. Due to the ridge penalty, the elastic net can deal with the collinearity of the covariates better than the lasso.

Friedman et al. (2007) proposed to use the coordinate descent algorithm to solve for the elastic net estimator, and Friedman et al. (2009) implemented it in an R package `glmnet`.

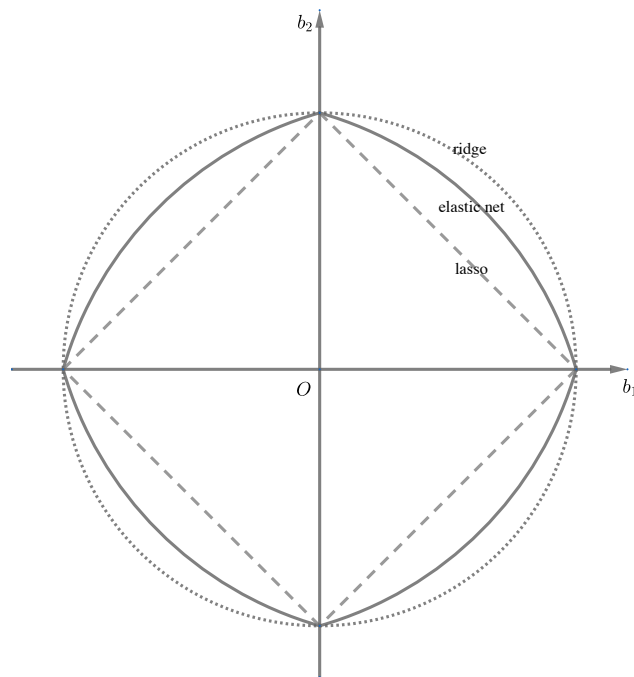


FIGURE 15.7: Comparing the ridge, lasso, and elastic net

15.6 Homework problems

15.1 Uniqueness of the lasso prediction

Consider the lasso problem:

$$\min_{b \in \mathbb{R}^p} \|Y - Xb\|^2 + \lambda \|b\|_1.$$

Prove that if $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ are two solutions, then $\alpha\hat{\beta}^{(1)} + (1-\alpha)\hat{\beta}^{(2)}$ must also be a solution, and $X\hat{\beta}^{(1)} = X\hat{\beta}^{(2)}$ must hold, for any $0 \leq \alpha \leq 1$.

Remark: The following facts may be useful for the problem. The function $\|\cdot\|^2$ is strongly convex. That is, for any v_1, v_2 and $0 < \alpha < 1$, we have

$$\|\alpha v_1 + (1-\alpha)v_2\|^2 \leq \alpha\|v_1\|^2 + (1-\alpha)\|v_2\|^2$$

and the inequality holds when $v_1 \neq v_2$. The function $\|\cdot\|$ is convex. That is, for any v_1, v_2 and $0 < \alpha < 1$, we have

$$\|\alpha v_1 + (1-\alpha)v_2\|_1 \leq \alpha\|v_1\|_1 + (1-\alpha)\|v_2\|_1.$$

Tibshirani (2013) provided detailed discussion about the uniqueness of the lasso problem.

15.2 The soft-thresholding lemma

Prove Lemma 15.1.

15.3 Penalized OLS with an orthogonal design matrix

Consider the special case with standardized and orthogonal design matrix:

$$X^T \mathbf{1}_n = 0, \quad X^T X = I_p.$$

For a fixed $\lambda \geq 0$, find the explicit formulas of the j th coordinates of the following estimators in terms of the corresponding j th coordinate of the OLS estimator $\hat{\beta}_j$ and λ ($j = 1, \dots, p$):

$$\hat{\beta}^{\text{ridge}}(\lambda) = \arg \min_{b \in \mathbb{R}^p} \{ \|Y - Xb\|^2 + \lambda \|b\|^2 \},$$

$$\hat{\beta}^{\text{lasso}}(\lambda) = \arg \min_{b \in \mathbb{R}^p} \{ \|Y - Xb\|^2 + \lambda \|b\|_1 \},$$

$$\hat{\beta}^{\text{enet}}(\lambda) = \arg \min_{b \in \mathbb{R}^p} \{ \|Y - Xb\|^2 + \lambda(\alpha \|b\|^2 + (1 - \alpha) \|b\|_1) \},$$

$$\hat{\beta}^{\text{subset}}(\lambda) = \arg \min_{b \in \mathbb{R}^p} \{ \|Y - Xb\|^2 + \lambda \|b\|_0 \},$$

where

$$\begin{aligned} \|b\|^2 &= \sum_{j=1}^p b_j^2, \\ \|b\|_1 &= \sum_{j=1}^p |b_j|, \\ \|b\|_0 &= \sum_{j=1}^p 1(b_j \neq 0). \end{aligned}$$

15.4 Standardization in the elastic net

For fixed λ and α , prove that the intercept in $\hat{\beta}^{\text{enet}}(\lambda, \alpha)$ equals zero under the standardization in Condition 14.1.

15.5 Coordinate descent for the elastic net

Give the detailed coordinate descent algorithm for the elastic net.

15.6 Reducing elastic net to lasso

Consider the following form of the elastic net:

$$\arg \min_{b \in \mathbb{R}^p} \|Y - Xb\|^2 + \lambda \{ \alpha \|b\|^2 + (1 - \alpha) \|b\|_1 \}.$$

Prove that it reduces to the following lasso:

$$\arg \min_{b \in \mathbb{R}^p} \|\tilde{Y} - \tilde{X}b\|^2 + \tilde{\lambda} \|b\|_1,$$

where

$$\tilde{Y} = \begin{pmatrix} Y \\ 0_p \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda \alpha} I_p \end{pmatrix}, \quad \tilde{\lambda} = \lambda(1 - \alpha).$$

Remark: Use the result in Problem 14.4.

15.7 Reducing lasso to iterative ridge

Based on the simple result

$$\min_{ac=b} (a^2 + c^2)/2 = |b|,$$

for scalars a, b, c , Hoff (2017) rewrote the lasso problem

$$\min_{b \in \mathbb{R}^p} \{\|Y - Xb\|^2 + \lambda \|b\|_1\}$$

as

$$\min_{u, v \in \mathbb{R}^p} \{\|Y - X(u \circ v)\|^2 + \lambda(\|u\|^2 + \|v\|^2)/2\}$$

where \circ denotes the component-wise product of vectors. Hoff (2017, Lemma 1) proved that a local minimum of the new problem must be a local minimum of the lasso problem.

Prove that the new problem can be solved based on the following iterative ridge regressions:

1. given u , we update v based on the ridge regression of Y on X_u with tuning parameter $\lambda/2$, where $X_u = X \text{diag}(u_1, \dots, u_p)$;
2. given v , we update u based on the ridge regression of Y on X_v with tuning parameter $\lambda/2$, where $X_v = X \text{diag}(v_1, \dots, v_p)$.

15.8 More noise in the Boston housing data

The Boston housing data have $n = 506$ observations. Add $p = n$ columns of covariates of random noise, and compare OLS, ridge, and lasso, as in Section 15.4. Add $p = 2n$ columns of covariates of random noise, and compare OLS, ridge, and lasso.

15.9 Recommended reading

Tibshirani (2011) gives a review of the lasso, as well as its history and recent developments. Two discussants, Professors Peter Bühlmann and Chris Holmes, make some excellent comments.



Part VI

Transformation and Weighting



16

Transformations in OLS

Transforming the outcome and covariates is fundamental in linear models. Whenever we specify a linear model $y_i = x_i^T \beta + \varepsilon_i$, we implicitly have transformed the original y and x , or at least we have chosen the scales of them. Carroll and Ruppert (1988) is a textbook on transformations in OLS. This chapter discusses some important special cases.

16.1 Transformation of the outcome

Although we can view

$$y_i = x_i^T \beta + \varepsilon_i, \quad (i = 1, \dots, n)$$

as a linear projection that works for any type of outcome $y_i \in \mathbb{R}$, the linear model works the best for continuous outcomes and especially for Normally distributed outcomes. Sometimes, the linear model can be a poor approximation of the original outcome but may perform well for certain transformations of the outcome.

16.1.1 Log transformation

With positive, especially heavy-tailed outcomes, a standard transformation is the log transformation. So we fit a linear model

$$\log y_i = x_i^T \beta + \varepsilon_i, \quad (i = 1, \dots, n).$$

The interpretation of the coefficients changes a little bit. Because

$$\frac{\partial \log \hat{y}_i}{\partial x_{ij}} = \frac{\partial \hat{y}_i}{\hat{y}_i} / \partial x_{ij} = \hat{\beta}_j,$$

we can interpret $\hat{\beta}_j$ in the following way: ceteris paribus, if x_{ij} increases by one unit, then the proportional increase in the average outcome is $\hat{\beta}_j$. In economics, $\hat{\beta}_j$ is the *semi-elasticity* of y on x_j in the model with log transformation on the outcome.

Sometimes, we may apply the log transformation on both the outcome and a certain covariate:

$$\log y_i = \beta_1 x_{i1} + \dots + \beta_j \log x_{ij} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad (i = 1, \dots, n).$$

The j th fitted coefficient becomes

$$\frac{\partial \log \hat{y}_i}{\partial \log x_{ij}} = \frac{\partial \hat{y}_i}{\hat{y}_i} / \frac{\partial x_{ij}}{x_{ij}} = \hat{\beta}_j,$$

so *ceteris paribus*, if x_{ij} increases by 1%, then the average outcome increases by $\hat{\beta}_j\%$. In economics, $\hat{\beta}_j$ is the x_j -*elasticity* of y in the model with log transformation on both the outcome and x_j .

The log transformation only works for positive variables. For a nonnegative outcome, we can modify the log transformation to $\log(y_i + 1)$.

16.1.2 Box–Cox transformation

Power transformation is another important class. The Box–Cox transformation unifies the log transformation and the power transformation:

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0. \end{cases}$$

L'Hôpital's rule implies that

$$\lim_{\lambda \rightarrow 0} \frac{y^\lambda - 1}{\lambda} = \lim_{\lambda \rightarrow 0} \frac{dy^\lambda/d\lambda}{1} = \lim_{\lambda \rightarrow 0} y^\lambda \log y = \log y,$$

so as a function of λ , $g_\lambda(y)$ is continuous at $\lambda = 0$. The log transformation is a limiting version of the power transformation. Can we choose λ based on data? Box and Cox (1964) proposed a strategy based on the maximum likelihood under the Normal linear model:

$$Y_\lambda = \begin{pmatrix} y_{\lambda 1} \\ \vdots \\ y_{\lambda n} \end{pmatrix} = \begin{pmatrix} g_\lambda(y_1) \\ \vdots \\ g_\lambda(y_n) \end{pmatrix} \sim N(X\beta, \sigma^2 I_n).$$

The density function of Y_λ is

$$f(Y_\lambda) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (Y_\lambda - X\beta)^\top (Y_\lambda - X\beta) \right\}.$$

The Jacobian of the transformation from Y to Y_λ is

$$\det \left(\frac{\partial Y_\lambda}{\partial Y} \right) = \det \begin{pmatrix} y_1^{\lambda-1} & & & \\ & y_2^{\lambda-1} & & \\ & & \ddots & \\ & & & y_n^{\lambda-1} \end{pmatrix} = \prod_{i=1}^n y_i^{\lambda-1},$$

so the density function of Y is

$$f(Y) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (Y_\lambda - X\beta)^\top (Y_\lambda - X\beta) \right\} \prod_{i=1}^n y_i^{\lambda-1}.$$

If we treat the density function of Y as a function of $(\beta, \sigma^2, \lambda)$, then it is the likelihood function, defined as $L(\beta, \sigma^2, \lambda)$. Given (σ^2, λ) , maximizing the likelihood function is equivalent to minimizing $(Y_\lambda - X\beta)^\top (Y_\lambda - X\beta)$, i.e., we can run OLS of Y_λ on X to obtain

$$\hat{\beta}(\lambda) = (X^\top X)^{-1} X^\top Y_\lambda.$$

Given λ , maximizing the likelihood function is equivalent to first obtaining $\hat{\beta}(\lambda)$ and then obtaining $\hat{\sigma}^2(\lambda) = n^{-1} Y_\lambda^\top (I_n - H) Y_\lambda$. The final step is to maximize the *profile likelihood* as a function of λ :

$$L(\hat{\beta}(\lambda), \hat{\sigma}^2(\lambda), \lambda) = \{2\pi\hat{\sigma}^2(\lambda)\}^{-n/2} \exp \left\{ -\frac{n\hat{\sigma}^2(\lambda)}{2\hat{\sigma}^2(\lambda)} \right\} \prod_{i=1}^n y_i^{\lambda-1}.$$

Dropping some constants, the log profile likelihood function of λ is

$$l_p(\lambda) = -\frac{n}{2} \log \hat{\sigma}^2(\lambda) + (\lambda - 1) \sum_{i=1}^n \log y_i.$$

The `boxcox` function in the R package `MASS` plots $l_p(\lambda)$, finds its maximizer $\hat{\lambda}$, and constructs a 95% confidence interval $[\hat{\lambda}_L, \hat{\lambda}_U]$ based on the following asymptotic pivotal quantity

$$2 \left\{ l_p(\hat{\lambda}) - l_p(\lambda) \right\} \stackrel{a}{\sim} \chi_1^2,$$

which holds by Wilks' Theorem. In practice, we often use the λ values within $[\hat{\lambda}_L, \hat{\lambda}_U]$ that have scientific meanings.

I use two datasets to illustrate the Box–Cox transformation.

Example 16.1 I use the `jobs` data in the `mediation` package (Tingley et al., 2014) to illustrate the Box–Cox transformation. The outcome of interest is `job_seek`, the level of job-search self-efficacy. The regressor `treat` is a binary indicator for whether the participant was randomly selected for the JOBS II training program. Descriptions of other covariates can be found in the R package. In this example, $\lambda = 2$ seems a plausible value. See the R code below and Figure 16.1.

```
library(MASS)
library(mediation)
par(mfrow = c(1, 3))
jobslm = lm(job_seek ~ treat + econ_hard + depress1 + sex + age + occp + marital +
             nonwhite + educ + income, data = jobs)
boxcox(jobslm, lambda = seq(1.5, 3, 0.1), plotit = TRUE)
jobslm2 = lm(I(job_seek^2) ~ treat + econ_hard + depress1 + sex + age + occp + marital +
             nonwhite + educ + income, data = jobs)
hist(jobslm$residuals, xlab = "residual", ylab = "",
     main = "job_seek", font.main = 1)
hist(jobslm2$residuals, , xlab = "residual", ylab = "",
     main = "job_seek^2", font.main = 1)
```

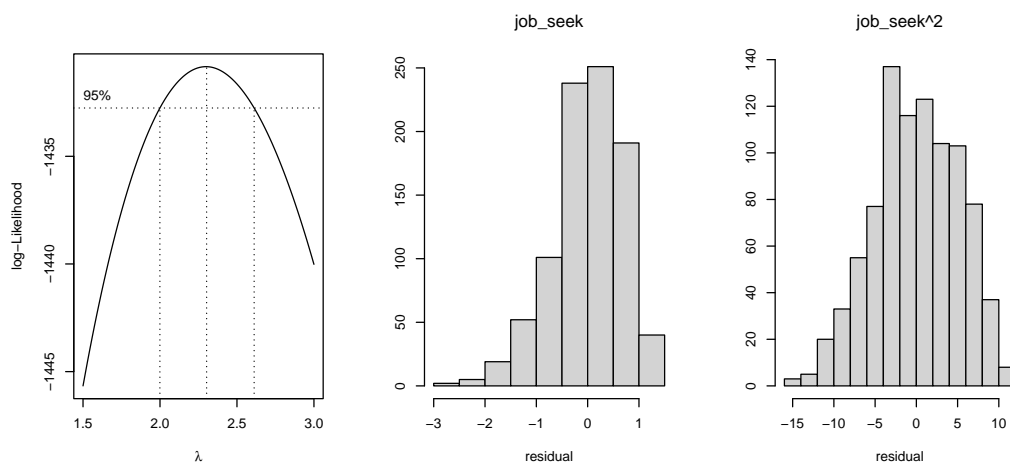


FIGURE 16.1: Box–Cox transformation in the `jobs` data

Example 16.2 I use the Penn bonus experiment data to illustrate the Box–Cox transformation. The outcome of interest is *duration*, the duration until employment. See Koenker and Xiao (2002) for descriptions of other regressors. In this example, $\lambda = 0.3$ seems a plausible value. However, the residual plot does not seem Normal, making the Box–Cox transformation not very meaningful. See the R code below and Figure 16.2.

```
penndata = read.table("pennbonus.txt")
par(mfrow = c(1, 3))
pennlm = lm(duration ~ ., data = penndata)
boxcox(pennlm, lambda = seq(0.2, 0.4, 0.05), plotit = TRUE)

pennlm.3 = lm(I(duration^(0.3)) ~ ., data = penndata)

hist(pennlm$residuals, xlab = "residual", ylab = "",
     main = "duration", font.main = 1)
hist(pennlm.3$residuals, xlab = "residual", ylab = "",
     main = "duration^0.3", font.main = 1)
```

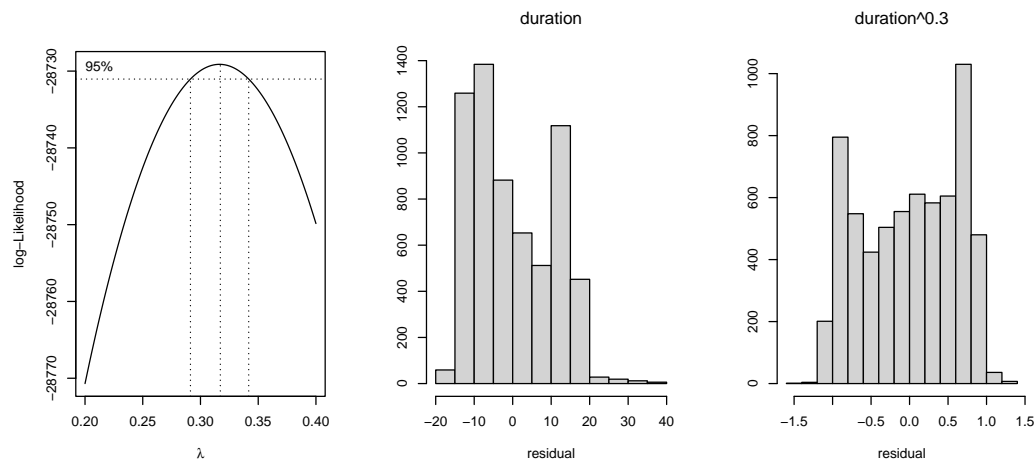


FIGURE 16.2: Box–Cox transformation in the Penn bonus experiment data

16.2 Transformation of the covariates

16.2.1 Polynomial, basis expansion, and generalized additive model

Linear approximations may not be adequate, so we can consider a polynomial specification. With one-dimensional x , we can use

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \cdots + \beta_p x_i^{p-1} + \varepsilon_i.$$

In economics, it is almost the default choice to include the quadratic term of working experience in the log wage equation. I give an example below using the data from Angrist et al. (2006). The quadratic term of *exper* is significant.

```

> library(foreign)
> census00 = read.dta("census00.dta")
> head(census00)
  age educ   logwk   perwt exper exper2 black
1  48  12 6.670576 1.0850021   30   900     0
2  42  13 6.783905 0.9666383   23   529     0
3  49  13 6.762383 1.2132297   30   900     0
4  44  13 6.302851 0.4833191   25   625     0
5  45  16 6.043386 0.9666383   23   529     0
6  43  13 5.061138 1.0850021   24   576     0
>
> census00ols1 = lm(logwk ~ educ + exper + black,
+                   data = census00)
> census00ols2 = lm(logwk ~ educ + exper + I(exper^2) + black,
+                   data = census00)
> round(summary(census00ols1)$coef, 4)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8918      0.0315 155.0540  0.0000
educ           0.1152      0.0012  99.1472  0.0000
exper          0.0002      0.0008   0.2294  0.8185
black         -0.2466      0.0085 -29.1674  0.0000
> round(summary(census00ols2)$coef, 4)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0777      0.0887  57.2254  0.0000
educ           0.1148      0.0012  97.6506  0.0000
exper         -0.0148      0.0067  -2.2013  0.0277
I(exper^2)     0.0003      0.0001   2.2425  0.0249
black         -0.2467      0.0085 -29.1732  0.0000

```

We can also include polynomial terms of more than one covariate, for example,

$$(1, x_{1i}, \dots, x_{i1}^d, x_{i2}, \dots, x_{i2}^l)$$

or

$$(1, x_{1i}, \dots, x_{i1}^d, x_{i2}, \dots, x_{i2}^l, x_{i1}x_{i2}, \dots, x_{i1}^d x_{i2}^l).$$

We can also approximate the conditional mean function of the outcome by a linear combination of some basis functions:

$$\begin{aligned}
 y_i &= f(x_i) + \varepsilon_i \\
 &\cong \sum_{j=1}^J \beta_j S_j(x_i) + \varepsilon_i,
 \end{aligned}$$

where the $S_j(x_i)$'s are basis functions. The `gam` function in the `mgcv` package uses this strategy including the automatic procedure of choosing the number of basis functions J . The following example has a sine function as the truth, and the basis expansion approximation yields reasonable performance with sample size $n = 1000$. Figure 16.3 plots both the true and estimated curves.

```

library(mgcv)
n = 1000
dat = data.frame(x <- seq(0, 1, length.out = n),
                 true <- sin(x*10),
                 y <- true + rnorm(n))
np.fit = gam(y ~ s(x), data = dat)
plot(y ~ x, data = dat, bty = "n",
     pch = 19, cex = 0.1, col = "grey")
lines(true ~ x, col = "grey")
lines(np.fit$fitted.values ~ x, lty = 2)
legend("bottomright", c("true", "estimated"),
     lty = 1:2, col = c("grey", "black"),
     bty = "n")

```

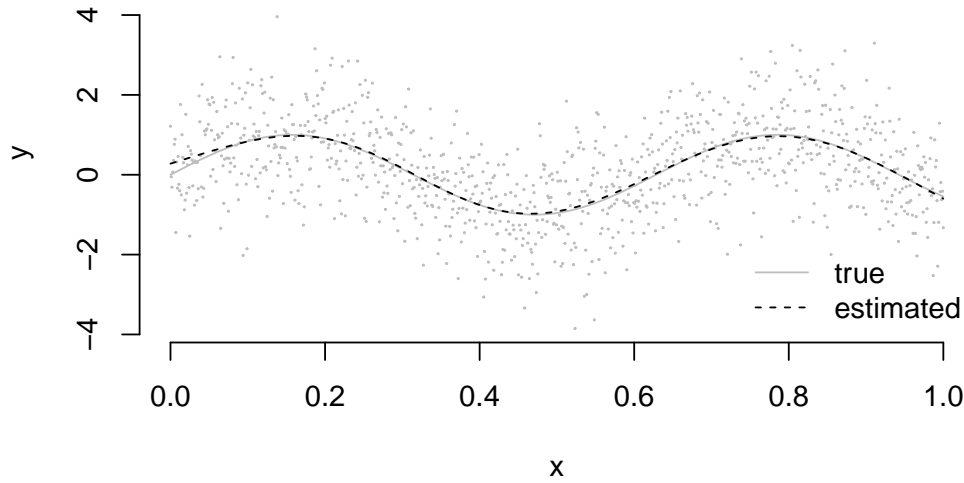


FIGURE 16.3: Nonparametric regression using the basis expansion: simulated data

The generalized additive model is an extension of the multivariate case:

$$y_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \varepsilon_i$$

$$\cong \sum_{j=1}^{J_1} \beta_{1j} S_j(x_{i1}) + \cdots + \sum_{j=1}^{J_p} \beta_{pj} S_j(x_{ip}) + \varepsilon_i.$$

The `gam` function in the `mgcv` package implements this strategy. Again I use the dataset from Angrist et al. (2006) to illustrate the procedure with nonlinearity in `educ` and `exper` shown in Figure 16.4.

```
census00gam = gam(logwk ~ s(educ) + s(exper) + black,
                  data = census00)
summary(census00gam)
par(mfrow = c(1, 2))
plot(census00gam, bty = "n")
```

See Wood (2017) for more details about the generalized additive model.

16.2.2 Regression discontinuity and regression kink

The left panel of Figure 16.5 shows an example of regression discontinuity, where the linear functions before and after a cutoff point can differ with a possible jump. A simple way to capture the two regimes of linear regression is to fit the following model:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 1(x_i > c) + \beta_4 x_i 1(x_i > c) + \varepsilon_i.$$

So

$$y_i = \begin{cases} \beta_1 + \beta_2 x_i + \varepsilon_i & x_i \leq c, \\ (\beta_1 + \beta_3) + (\beta_2 + \beta_4) x_i + \varepsilon_i, & x_i > c. \end{cases}$$

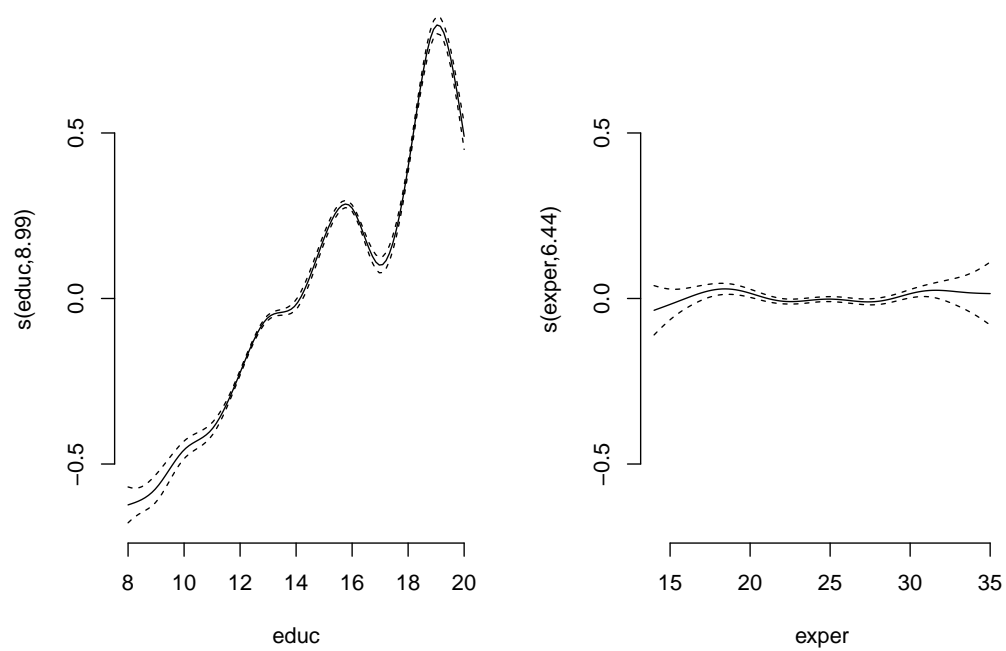


FIGURE 16.4: Generalized additive model

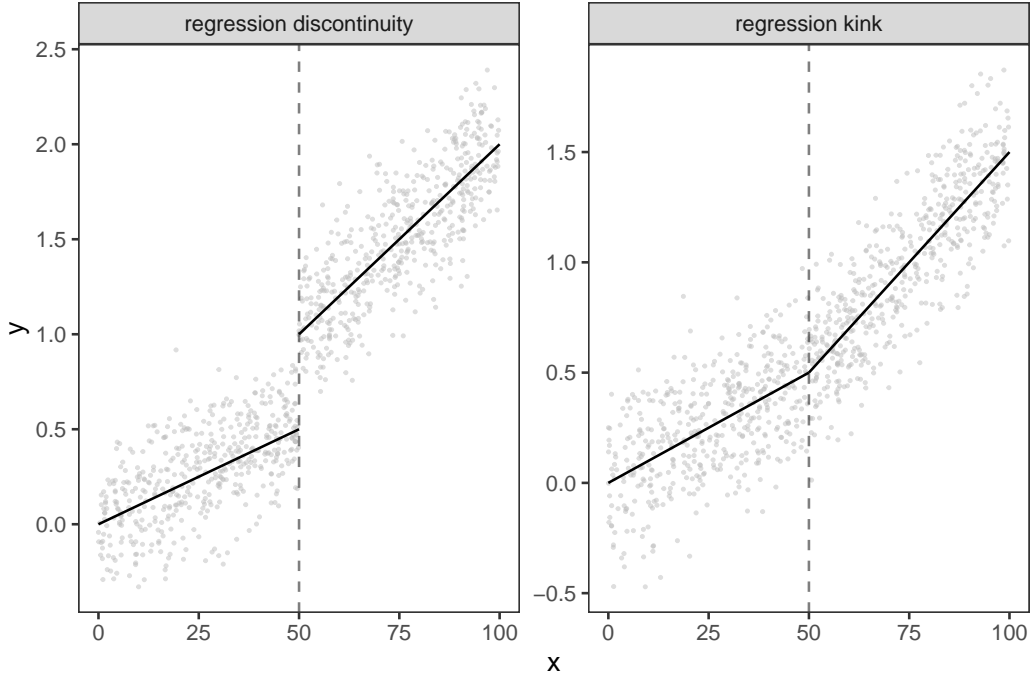


FIGURE 16.5: Regression discontinuity and kink

Testing the discontinuity at c is equivalent to testing

$$(\beta_1 + \beta_3) + (\beta_2 + \beta_4)c = \beta_1 + \beta_2c,$$

which is equivalent to

$$\beta_3 + \beta_4c = 0.$$

If we center the covariates at c , then

$$y_i = \beta_1 + \beta_2(x_i - c) + \beta_31(x_i > c) + \beta_4(x_i - c)1(x_i > c) + \varepsilon_i$$

and

$$y_i = \begin{cases} \beta_1 + \beta_2(x_i - c) + \varepsilon_i & x_i \leq c, \\ (\beta_1 + \beta_3) + (\beta_2 + \beta_4)(x_i - c) + \varepsilon_i, & x_i > c. \end{cases}$$

So testing the discontinuity at c is equivalent to testing $\beta_3 = 0$.

The right panel of Figure 16.5 shows an example of regression kink, where the linear functions before and after a cutoff point can differ but the whole regression line is continuous. A simple way to capture the two regimes of linear regression is to fit the following model:

$$y_i = \beta_1 + \beta_2R_c(x_i) + \beta_3(x_i - c) + \varepsilon_i$$

using

$$R_c(x) = \max(0, x - c) = \begin{cases} 0, & x \leq c, \\ x - c, & x > c. \end{cases}$$

So

$$y_i = \begin{cases} \beta_1 + \beta_3(x_i - c) + \varepsilon_i, & x_i \leq c, \\ \beta_1 + (\beta_2 + \beta_3)(x_i - c) + \varepsilon_i, & x_i > c. \end{cases}$$

This ensures that the mean function is continuous at c with both left and right limits equaling β_1 . Testing the kink is equivalent to testing $\beta_2 = 0$.

These regressions have many applications in economics, but I omit the economic background. Readers can find more discussions in Angrist and Pischke (2008) and Card et al. (2015).

16.3 Homework problems

16.1 Piecewise linear regression

Generate data in the same way as the example in Figure 16.3, and fit a continuous piecewise linear function with cutoff points 0, 0.2, 0.4, 0.6, 0.8, 1.



Interactions in OLS

Interaction is an important notion in applied statistics. It measures the interplay of two or more variables acting simultaneously on an outcome. Epidemiologists find that cigarette smoking and alcohol consumption both increase the risks of many cancers. Then they want to measure how cigarette smoking and alcohol consumption jointly increase the risks. That is, does cigarette smoking increase the risks of cancers more in the presence of alcohol consumption than in the absence of it? Political scientists are interested in measuring the interplay of different get-out-to-vote interventions on voting behavior.

This chapter will review many aspects of interaction in the context of linear regression. Cox (1984) and Berrington de González and Cox (2007) reviewed interaction from a statistical perspective. VanderWeele (2015) offers a textbook discussion on interaction with a focus on applications in epidemiology.

17.1 Two binary covariates interact

Start with the simplest yet nontrivial example with two binary covariates $x_{i1}, x_{i2} \in \{0, 1\}$. We can fit the OLS:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_{12} x_{i1} x_{i2} + \hat{\varepsilon}_i. \quad (17.1)$$

We can express the coefficients in terms of the means of the outcomes within four combinations of the covariates. The following proposition is an algebraic result.

Proposition 17.1 *From (17.1), we have*

$$\begin{aligned} \hat{\beta}_0 &= \bar{y}_{00}, \\ \hat{\beta}_1 &= \bar{y}_{10} - \bar{y}_{00}, \\ \hat{\beta}_2 &= \bar{y}_{01} - \bar{y}_{00}, \\ \hat{\beta}_{12} &= (\bar{y}_{11} - \bar{y}_{10}) - (\bar{y}_{01} - \bar{y}_{00}), \end{aligned}$$

where $\bar{y}_{f_1 f_2}$ is the average value of the y_i 's with $x_{i1} = f_1$ and $x_{i2} = f_2$:

$$\bar{y}_{f_1 f_2} = \frac{\sum_{i=1}^n I(x_{i1} = f_1, x_{i2} = f_2) y_i}{\sum_{i=1}^n I(x_{i1} = f_1, x_{i2} = f_2)}.$$

The proof of Proposition 17.1 is pure algebraic which is relegated to Problem 17.1. The proposition generalizes to OLS with more than two binary covariates. See Zhao and Ding (2022) for more details.

The coefficient $\hat{\beta}_{12}$ equals the difference between $\bar{y}_{11} - \bar{y}_{10}$, the effect of x_{i2} on y_i holding x_{i1} at level 1, and $\bar{y}_{01} - \bar{y}_{00}$, the effect of x_{i2} on y_i holding x_{i1} at level 0. It also equals

$$\hat{\beta}_{12} = (\bar{y}_{11} - \bar{y}_{01}) - (\bar{y}_{10} - \bar{y}_{00}),$$

that is, the difference between $\bar{y}_{11} - \bar{y}_{01}$, the effect of x_{i1} on y_i holding x_{i2} at level 1, and $\bar{y}_{10} - \bar{y}_{00}$, the effect of x_{i1} on y_i holding x_{i2} at level 0. The formula shows the symmetry of x_{i1} and x_{i2} in defining interaction.

17.2 A binary covariate interacts with a general covariate

17.2.1 Treatment effect heterogeneity

In many studies, we are interested in the effect of a binary treatment z_i on an outcome y_i , adjusting for some background covariates x_i . The covariates can play many roles in this problem. They may affect the treatment, enter the outcome model, and modify the effect of the treatment on the outcome. We can formulate the problem in terms of linear regression:

$$y_i = \beta_0 + \beta_1 z_i + \beta_2^T x_i + \beta_3^T x_i z_i + \varepsilon_i, \quad (17.2)$$

where $E(\varepsilon_i | z_i, x_i) = 0$. So

$$E(y_i | z_i = 1, x_i) = \beta_0 + \beta_1 + (\beta_2 + \beta_3)^T x_i$$

and

$$E(y_i | z_i = 0, x_i) = \beta_0 + \beta_2^T x_i,$$

which implies that

$$E(y_i | z_i = 1, x_i) - E(y_i | z_i = 0, x_i) = \beta_1 + \beta_3^T x_i.$$

The difference between two conditional expectations, $E(y_i | z_i = 1, x_i) - E(y_i | z_i = 0, x_i)$, is often called the conditional average treatment effect (CATE). Under model (17.2), the CATE is a linear function of the covariates. As long as $\beta_3 \neq 0$, we have treatment effect heterogeneity, which is also called *effect modification*. A statistical test for $\beta_3 = 0$ is straightforward based on OLS and EHW standard error.

Note that (17.2) includes the interaction of the treatment and all covariates. With prior knowledge, we may believe that the treatment effect varies with respect to a subset of covariates, or, equivalently, we may set some components of β_3 to be zero.

17.2.2 Johnson–Neyman technique

Johnson and Neyman (1936) proposed a technique to identify the region of covariates in which the conditional average treatment $\beta_1 + \beta_3^T x$ is zero. For a given x , we can test the null hypothesis that $\beta_1 + \beta_3^T x = 0$, which is a linear combination of the regression coefficients of (17.2). If we fail to reject the null hypothesis, then this x belongs to the region of zero CATE. See Rogosa (1981) for more discussions.

17.2.3 Blinder–Oaxaca decomposition

The linear regression (17.2) also applies to descriptive statistics, when z_i is a binary indicator for subgroups. For example, z_i can be a binary indicator for age, racial, or gender groups, y_i can be the log wage, and x_i can be a vector of explanatory variables such as education, experience, industry, and occupation. Sometimes, it is more insightful to write (17.2) in terms of two possibly non-parallel linear regressions:

$$y_i = \gamma_0 + \theta_0^T x_i + \varepsilon_i, \quad E(\varepsilon_i | z_i = 0, x_i) = 0 \quad (17.3)$$

for the group with $z_i = 0$, and

$$y_i = \gamma_1 + \theta_1^T x_i + \varepsilon_i, \quad E(\varepsilon_i \mid z_i = 1, x_i) = 0 \quad (17.4)$$

for the group with $z_i = 1$. Regressions (17.3) and (17.4) are just a reparametrization of (17.2) with

$$\gamma_0 = \beta_0, \quad \theta_0 = \beta_2, \quad \gamma_1 = \beta_0 + \beta_1, \quad \theta_1 = \beta_2 + \beta_3.$$

Based on (17.3) and (17.4), we can decompose the difference in the outcome means as

$$\begin{aligned} & E(y_i \mid z_i = 1) - E(y_i \mid z_i = 0) \\ &= \{\gamma_1 + \theta_1^T E(x_i \mid z_i = 1)\} - \{\gamma_0 + \theta_0^T E(x_i \mid z_i = 0)\} \\ &= \theta_0^T \{E(x_i \mid z_i = 1) - E(x_i \mid z_i = 0)\} \\ &\quad + (\theta_1 - \theta_0)^T E(x_i \mid z_i = 0) + \gamma_1 - \gamma_0 \\ &\quad + (\theta_1 - \theta_0)^T \{E(x_i \mid z_i = 1) - E(x_i \mid z_i = 0)\} \\ &= \mathcal{E} + \mathcal{C} + \mathcal{I}. \end{aligned}$$

The decomposition has three components:

(C1) The first component

$$\begin{aligned} \mathcal{E} &= \theta_0^T \{E(x_i \mid z_i = 1) - E(x_i \mid z_i = 0)\} \\ &= \beta_2^T \{E(x_i \mid z_i = 1) - E(x_i \mid z_i = 0)\} \end{aligned}$$

measures the *endowment effect*, because it is due to the difference in the covariate means across groups;

(C2) The second component

$$\begin{aligned} \mathcal{C} &= (\theta_1 - \theta_0)^T E(x_i \mid z_i = 0) + \gamma_1 - \gamma_0 \\ &= \beta_3^T E(x_i \mid z_i = 0) + \beta_1 \end{aligned}$$

measures the difference in coefficients;

(C3) The third component

$$\begin{aligned} \mathcal{I} &= (\theta_1 - \theta_0)^T \{E(x_i \mid z_i = 1) - E(x_i \mid z_i = 0)\} \\ &= \beta_3^T \{E(x_i \mid z_i = 1) - E(x_i \mid z_i = 0)\} \end{aligned}$$

measures the interaction between the endowment and coefficients.

The above decomposition is called the Blinder–Oaxaca decomposition. Jann (2008) reviews other forms of the decomposition, extending the original forms in Blinder (1973) and Oaxaca (1973).

Estimation and testing for \mathcal{E} , \mathcal{C} , and \mathcal{I} are straightforward. Based on the OLS of (17.2) and the sample means \bar{x}_1 and \bar{x}_0 of the covariates, we have point estimators

$$\begin{aligned} \hat{\mathcal{E}} &= \hat{\beta}_2^T (\bar{x}_1 - \bar{x}_0), \\ \hat{\mathcal{C}} &= \hat{\beta}_3^T \bar{x}_0 + \hat{\beta}_1, \\ \hat{\mathcal{I}} &= \hat{\beta}_3^T (\bar{x}_1 - \bar{x}_0). \end{aligned}$$

Given the covariates, they are linear transformations of the OLS coefficients.

17.2.4 Chow test

Chow (1960) proposed to test whether the two regressions (17.3) and (17.4) are identical. Under the null hypothesis that $\gamma_0 = \gamma_1$ and $\theta_0 = \theta_1$, he proposed an F test assuming homoskedasticity, which is called the *Chow test* in econometrics. In fact, this is just a special case of the standard F test for the null hypothesis that $\beta_1 = 0$ and $\beta_3 = 0$ in (17.2). Moreover, based on the OLS in (17.2), we can also derive the robust test based on the EHW covariance estimator.

Chow (1960) discussed a subtle case in which one group has a small size rendering the OLS fit underdetermined. I relegate the details to Problem 17.3. Note that under the null hypothesis, $\mathcal{C} = \mathcal{I} = 0$, so the difference in the outcome means is purely due to the difference in the covariate means.

17.3 Difficulties of interaction

Practitioners also interpret the coefficient of the product term of two continuous variables as an interaction. However, this heuristics causes subtle issues. This section reviews some important issues.

17.3.1 Removable interaction

The first issue is about *removable interaction*. I use a simple numerical example to explain the idea.

In the `R` code below, the interaction term is not significant, which is coherent with the true data-generating process that the mean of $\log(y)$ is linear in x_1 and x_2 without interaction.

```
> n = 1000
> x1 = rnorm(n)
> x2 = rnorm(n)
> y = exp(x1 + x2 + rnorm(n))
> ols.fit = lm(log(y) ~ x1*x2)
> summary(ols.fit)
```

Call:

```
lm(formula = log(y) ~ x1 * x2)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.7373	-0.6822	-0.0111	0.7084	3.1039

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.003214	0.031286	0.103	0.918
x1	1.056801	0.030649	34.480	<2e-16 ***
x2	1.009404	0.030778	32.797	<2e-16 ***
x1:x2	-0.017528	0.030526	-0.574	0.566

In the `R` code below, the interaction term is significant because the mean of y is not linear in x_1 and x_2 .

```
> ols.fit = lm(y ~ x1*x2)
> summary(ols.fit)
```

Call:

```
lm(formula = y ~ x1 * x2)
```

Residuals:

Min	1Q	Median	3Q	Max
-35.95	-5.17	-0.97	2.34	513.35

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.2842	0.6686	7.903	7.17e-15 ***
x1	6.7565	0.6550	10.315	< 2e-16 ***
x2	4.9548	0.6577	7.533	1.11e-13 ***
x1:x2	7.3810	0.6524	11.314	< 2e-16 ***

From the example above, we can see that taking the log of the outcome removes the interaction. Therefore, the interaction in the OLS fit of y on (x_1, x_2, x_1x_2) is *removable*. The lesson here is that the significance of interaction in OLS depends on the scale of the outcome.

17.3.2 Main effect in the presence of interaction

The second issue is that including the interaction term complicates the interpretation of the main effects. I also use a simple example to explain the idea.

In the OLS fit below, we observe significant main effects if we do not include the interaction term.

```
> ## data from "https://stats.idre.ucla.edu/stat/data/hsbdemo.dta"
> hsbdemo = read.table("hsbdemo.txt")
> ols.fit = lm(read ~ math + socst, data = hsbdemo)
> summary(ols.fit)
```

Call:

```
lm(formula = read ~ math + socst, data = hsbdemo)
```

Residuals:

Min	1Q	Median	3Q	Max
-18.8729	-4.8987	-0.6286	5.2380	23.6993

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.14654	3.04066	2.350	0.0197 *
math	0.50384	0.06337	7.951	1.41e-13 ***
socst	0.35414	0.05530	6.404	1.08e-09 ***

Then we add the interaction term into the OLS, and suddenly we have significant interaction but not significant main effects.

```
> ols.fit = lm(read ~ math*socst, data = hsbdemo)
> summary(ols.fit)
```

Call:

```
lm(formula = read ~ math * socst, data = hsbdemo)
```

Residuals:

Min	1Q	Median	3Q	Max
-18.6071	-4.9228	-0.7195	4.5912	21.8592

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	37.842715	14.545210	2.602	0.00998 **
math	-0.110512	0.291634	-0.379	0.70514
socst	-0.220044	0.271754	-0.810	0.41908
math:socst	0.011281	0.005229	2.157	0.03221 *

However, if we center the covariates, the main effects are significant again.

```
> hsbdemo$math.c = hsbdemo$math - mean(hsbdemo$math)
> hsbdemo$socst.c = hsbdemo$socst - mean(hsbdemo$socst)
> ols.fit = lm(read ~ math.c*socst.c, data = hsbdemo)
> summary(ols.fit)

Call:
lm(formula = read ~ math.c * socst.c, data = hsbdemo)

Residuals:
    Min       1Q   Median       3Q      Max
-18.6071  -4.9228  -0.7195   4.5912  21.8592

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.615327   0.568685  90.763 < 2e-16 ***
math.c         0.480654   0.063701   7.545 1.65e-12 ***
socst.c        0.373829   0.055546   6.730 1.82e-10 ***
math.c:socst.c 0.011281   0.005229   2.157  0.0322 *

```

Then how should we interpret the main effects above? Clearly, the main effects depend on whether or not we include the interaction term, and depend on how we center the regressors. The following discussion proposes the notion of *average partial or marginal effect* to measure the main effect.

Based on the linear model with interaction

$$E(y_i | x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2},$$

define the main effects as

$$n^{-1} \sum_{i=1}^n \frac{\partial E(y_i | x_{i1}, x_{i2})}{\partial x_{i1}} = n^{-1} \sum_{i=1}^n (\beta_1 + \beta_{12} x_{i2}) = \beta_1 + \beta_{12} \bar{x}_2$$

and

$$n^{-1} \sum_{i=1}^n \frac{\partial E(y_i | x_{i1}, x_{i2})}{\partial x_{i2}} = n^{-1} \sum_{i=1}^n (\beta_2 + \beta_{12} x_{i1}) = \beta_2 + \beta_{12} \bar{x}_1,$$

which are called the *average partial or marginal effects*. So when the covariates are centered at $\bar{x}_1 = \bar{x}_2 = 0$, we can interpret β_1 and β_2 as the main effects. In contrast, the interpretation of the interaction term does not depend on the centering of the covariates because

$$\frac{\partial^2 E(y_i | x_{i1}, x_{i2})}{\partial x_{i1} \partial x_{i2}} = \beta_{12}.$$

17.3.3 Power

The second issue is that usually, statistical tests for interaction do not have enough power. Proposition 17.1 provides a simple explanation. The variance of the interaction equals

$$\text{var}(\hat{\beta}_{12}) = \frac{\sigma_{11}^2}{n_{11}} + \frac{\sigma_{10}^2}{n_{10}} + \frac{\sigma_{01}^2}{n_{01}} + \frac{\sigma_{00}^2}{n_{00}},$$

where $\sigma_{f_1 f_2}^2 = \text{var}(y_i | x_{i1} = f_1, x_{i2} = f_2)$. Therefore, its variance is driven by the smallest value of $n_{11}, n_{10}, n_{01}, n_{00}$. Even when the total sample size is large, one of the subgroup sample sizes can be small, resulting in a large variance of the estimator of the interaction.

17.4 Homework problems

17.1 Interaction and difference-in-differences

Prove Proposition 17.1. Moreover, simplify the HC0 and HC2 versions of the EHW standard errors of the coefficients in terms of $n_{f_1 f_2}$ and $\hat{\sigma}_{f_1 f_2}^2$, where $n_{f_1 f_2}$ is the sample size and $\hat{\sigma}_{f_1 f_2}^2$ is the sample variance of the outcomes for units with $x_{i1} = f_1$ and $x_{i2} = f_2$.

Remark: You can prove the proposition by inverting the 4×4 matrix $X^T X$. However, this method is a little too tedious. Moreover, this proof does not generalize to OLS with $K > 2$ binary covariates. So it is better to find alternative proofs. For the EHW standard errors, you can use the results in Problems 6.3 and 6.4.

17.2 Two OLS

Consider the data $(x_i, z_i, y_i)_{i=1}^n$, where x_i denotes the covariates, z_i denotes the binary group indicator, and y_i denotes the outcome. We can fit two separate OLS:

$$\hat{y}_i = \hat{\gamma}_1 + x_i^T \hat{\beta}_1$$

and

$$\hat{y}_i = \hat{\gamma}_0 + x_i^T \hat{\beta}_0$$

with data in group 1 and group 0, respectively. We can also fit a joint OLS using the pooled data:

$$\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_z z_i + x_i^T \hat{\alpha}_x + z_i x_i^T \hat{\alpha}_{zx}.$$

1. Find $(\hat{\alpha}_0, \hat{\alpha}_z, \hat{\alpha}_x, \hat{\alpha}_{zx})$ in terms of $(\hat{\gamma}_1, \hat{\beta}_1, \hat{\gamma}_0, \hat{\beta}_0)$.
2. Prove that the fitted values \hat{y}_i 's are the same from the separate and the pooled OLS for all units $i = 1, \dots, n$.
3. Prove that the leverage scores h_{ii} 's are the same from the separate and the pooled OLS.

17.3 Chow test when one group size is too small

Assume (17.3) and (17.4) with homoskedastic Normal error terms. Let n_1 and n_0 denote the sample sizes of groups with $z_i = 1$ and $z_i = 0$. Consider the case with n_0 larger than the number of covariates but n_1 smaller than the number of covariates. So we can fit OLS and estimate the variance based on (17.3), but we cannot do so based on (17.4). The statistical test discussed in the main paper does not apply. Chow (1960) proposed the following test based on prediction.

Let $\hat{\gamma}_0$ and $\hat{\theta}_0$ be the coefficients, and $\hat{\sigma}_0^2$ be the variance estimate based on OLS with units $z_i = 0$. Under the null hypothesis that $\gamma_0 = \gamma_1$ and $\theta_0 = \theta_1$, predict the outcomes of the units $z_i = 1$:

$$\hat{y}_i = \hat{\gamma}_0 + \hat{\theta}_0^T x_i$$

with the prediction error

$$d_i = y_i - \hat{y}_i$$

following a multivariate Normal distribution. Propose an F test based on the d_i 's for units with $z_i = 1$.

Remark: It is more convenient to use the matrix form of OLS.

17.4 Invariance of the interaction

In Section 17.3.2, the point estimate and standard error of the coefficient of the interaction term remain the same no matter whether we center the covariates or not. This result holds in general. This problem quantifies this phenomenon.

With scalars x_{i1}, x_{i2}, y_i ($i = 1, \dots, n$), we can fit the OLS

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_{12} x_{i1} x_{i2} + \hat{\varepsilon}_i.$$

Under any location transformations of the covariates $x'_{i1} = x_{i1} - c_1, x'_{i2} = x_{i2} - c_2$, we can fit the OLS

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x'_{i1} + \tilde{\beta}_2 x'_{i2} + \tilde{\beta}_{12} x'_{i1} x'_{i2} + \tilde{\varepsilon}_i.$$

1. Express $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_{12}$ in terms of $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_{12}$. Verify that $\hat{\beta}_{12} = \tilde{\beta}_{12}$.
2. Prove that the EHW standard errors for $\hat{\beta}_{12}$ and $\tilde{\beta}_{12}$ are identical.

Remark: Use the results in Problems 3.4 and 6.4.

When x_1 and x_2 are binary covariates, there are three canonical choices of (c_1, c_2) :

- (C1) $(c_1, c_2) = (0, 0)$: Then the interpretation of $\hat{\beta}_1$ is the effect of covariate x_1 holding the other covariate x_2 at 0, whereas the interpretation of $\hat{\beta}_2$ is the effect of covariate x_2 holding the other covariate x_1 at 0.
- (C2) $(c_1, c_2) = (1/2, 1/2)$: Then the interpretation of $\tilde{\beta}_1$ is the average of the effects of covariate x_1 holding the other covariate x_2 at 0 and 1, whereas the interpretation of $\tilde{\beta}_2$ is the average of the effects of covariate x_2 holding the other covariate x_1 at 0 and 1.
- (C3) $(c_1, c_2) = (\bar{x}_1, \bar{x}_2)$: Then the interpretations of $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are the average marginal effects.

The coefficient of the interaction term is invariant to the location transformations.

See Zhao and Ding (2022) for more general discussions in the context of factorial experiments with multiple factors.

Restricted OLS

Assume that in the standard linear model $Y = X\beta + \varepsilon$, the parameter has linear restrictions:

$$C\beta = r, \quad (18.1)$$

where C is an $l \times p$ matrix and r is a l dimensional vector. Assume that C has linearly independent row vectors; otherwise, some restrictions are redundant. We can use the restricted OLS:

$$\hat{\beta}_r = \arg \min_{b \in \mathbb{R}^p} \|Y - Xb\|^2$$

under the restriction

$$Cb = r.$$

I first give some examples of linear models with restricted parameters, then derive the algebraic properties of the restricted OLS estimator $\hat{\beta}_r$, and finally discuss statistical inference with restricted OLS.

This chapter is mainly of theoretical interest. For readers who focus on practical data analysis, you can skip this chapter when you first read this book.

18.1 Examples

Example 18.1 (Short regression) Partition X into X_1 and X_2 with k and l columns, respectively, with $p = k + l$. The short regression of Y on X_1 yields OLS coefficient $\hat{\beta}_1$. So $(\hat{\beta}_1^T, 0_l^T) = \hat{\beta}_r$ with

$$C = (0_{l \times k}, I_{l \times l}), \quad r = 0_l.$$

Example 18.2 (Testing linear hypothesis) Consider testing the linear hypothesis $C\beta = r$ in the linear model. We have discussed in Chapter 5 the Wald test based on the OLS estimator and its estimated covariance matrix under the Normal linear model. An alternative strategy is to test the hypothesis based on comparing the residual sum of squares under the OLS and restricted OLS. Therefore, we need to compute both $\hat{\beta}$ and $\hat{\beta}_r$.

Example 18.3 (One-way analysis of variance) If x_i contains the intercept and Q_1 dummy variables of a discrete regressor of Q_1 levels, $(f_{i1}, \dots, f_{iQ_1})^T$, then we must impose a restriction on the parameter in the linear model

$$y_i = \alpha + \sum_{j=1}^{Q_1} \beta_j f_{ij} + \varepsilon_i.$$

A canonical choice is $\beta_{Q_1} = 0$, which is equivalent to dropping the last dummy variable due to its redundancy. Another canonical choice is $\sum_{j=1}^{Q_1} \beta_j = 0$. This restriction keeps the

symmetry of the regressors in the linear model and changes the interpretation of β_j as the deviation from the “effect” of level j with respect to the average “effect.” Both are special cases of restricted OLS.

Example 18.4 (Two-way analysis of variance) With two factors of levels Q_1 and Q_2 , respectively, the regressor x_i contains the Q_1 dummy variables of the first factor, $(f_{i1}, \dots, f_{iQ_1})^T$, the Q_2 dummies of the second factor, $(g_{i1}, \dots, g_{iQ_2})^T$, and the $Q_1 Q_2$ dummy variables of the interaction terms, $(f_{i1}g_{i1}, \dots, f_{iQ_1}g_{iQ_2})^T$. We must impose restrictions on the parameters in the linear model

$$y_i = \alpha + \sum_{j=1}^{Q_1} \beta_j f_{ij} + \sum_{k=1}^{Q_2} \gamma_k g_{ik} + \sum_{j=1}^{Q_1} \sum_{k=1}^{Q_2} \delta_{jk} f_{ij} g_{ik} + \varepsilon_i.$$

Similar to the discussion in Example 18.3, two canonical choices of restrictions are

$$\beta_{Q_1} = 0, \quad \gamma_{Q_2} = 0, \quad \delta_{Q_1,k} = \delta_{j,Q_2} = 0, \quad (j = 1, \dots, Q_1; k = 1, \dots, Q_2).$$

and

$$\sum_{j=1}^{Q_1} \beta_j = 0, \quad \sum_{k=1}^{Q_2} \gamma_k = 0, \quad \sum_{j=1}^{Q_1} \delta_{jk} = \sum_{k=1}^{Q_2} \delta_{jk} = 0, \quad (j = 1, \dots, Q_1; k = 1, \dots, Q_2).$$

Problem 8.5 already presented a more advanced example.

18.2 Algebraic properties

I first give an explicit formula of the restricted OLS (Theil, 1971; Rao, 1973). For simplicity, Theorem 18.1 below assumes that $X^T X$ is invertible. This condition may not hold in general; see Examples 18.3 and 18.4. Greene and Seaks (1991) discussed the results without this assumption; see Problem 18.8 for more details.

Theorem 18.1 *If $X^T X$ is invertible, then*

$$\hat{\beta}_r = \hat{\beta} - (X^T X)^{-1} C^T \{C(X^T X)^{-1} C^T\}^{-1} (C\hat{\beta} - r),$$

where $\hat{\beta}$ is the unrestricted OLS coefficient.

Proof of Theorem 18.1: The Lagrangian for the restricted optimization problem is

$$(Y - Xb)^T (Y - Xb) - 2\lambda^T (Cb - r).$$

So the first order condition is

$$2X^T(Y - Xb) - 2C^T\lambda = 0$$

which implies

$$X^T X b = X^T Y - C^T \lambda. \tag{18.2}$$

Solve the linear system in (18.2) to obtain

$$b = (X^T X)^{-1} (X^T Y - C^T \lambda).$$

Using the linear restriction $Cb = r$, we have

$$C(X^T X)^{-1}(X^T Y - C^T \lambda) = r$$

which implies that

$$\lambda = \{C(X^T X)^{-1}C^T\}^{-1}(C\hat{\beta} - r).$$

So the restricted OLS coefficient is

$$\begin{aligned}\hat{\beta}_r &= (X^T X)^{-1}(X^T Y - C^T \lambda) \\ &= \hat{\beta} - (X^T X)^{-1}C^T \lambda \\ &= \hat{\beta} - (X^T X)^{-1}C^T \{C(X^T X)^{-1}C^T\}^{-1}(C\hat{\beta} - r).\end{aligned}$$

Since the objective function is convex and the restrictions are linear, the solution from the first-order condition is indeed the minimizer. \square

In the special case with $r = 0$, Theorem 18.1 has a simpler form.

Corollary 18.1 *Under the restriction (18.1) with $r = 0$, we have*

$$\hat{\beta}_r = M_r \hat{\beta},$$

where

$$M_r = I_p - (X^T X)^{-1}C^T \{C(X^T X)^{-1}C^T\}^{-1}C.$$

Moreover, M_r satisfies the following properties

$$\begin{aligned}M_r(X^T X)^{-1}C^T &= 0, \\ CM_r &= 0, \\ \{I_p - C^T(CC^T)^{-1}C\}M_r &= M_r.\end{aligned}$$

The M_r matrix plays central roles below.

The following result is also an immediate corollary of Theorem 18.1.

Corollary 18.2 *Under the restriction (18.1), we have*

$$\hat{\beta}_r - \beta = M_r(\hat{\beta} - \beta).$$

I leave the proofs of Corollaries 18.1 and 18.2 to Problem 18.1.

18.3 Statistical inference

I first focus on the Gauss–Markov model with the restriction (18.1). As direct consequences of Corollary 18.2, we can show that the restricted OLS estimator is unbiased for β , and obtain its covariance matrix below.

Corollary 18.3 *Assume the Gauss–Markov model under Assumption 4.1 and the restriction (18.1). We have*

$$\begin{aligned}E(\hat{\beta}_r) &= \beta, \\ \text{cov}(\hat{\beta}_r) &= \sigma^2 M_r(X^T X)^{-1}M_r^T.\end{aligned}$$

Moreover, under the Normal linear model with the restriction (18.1), we can derive the exact distribution of the restricted OLS estimator and propose an unbiased estimator for σ^2 .

Theorem 18.2 *Assume the Normal linear model under Assumption 5.1 and the restriction (18.1). We have*

$$\hat{\beta}_r \sim N(\beta, \sigma^2 M_r (X^T X)^{-1} M_r^T).$$

An unbiased estimator for σ^2 is

$$\hat{\sigma}_r^2 = \hat{\varepsilon}_r^T \hat{\varepsilon}_r / (n - p + l),$$

where $\hat{\varepsilon}_r = Y - X \hat{\beta}_r$. Moreover,

$$\hat{\beta}_r \perp \hat{\sigma}_r^2.$$

Corollary 18.3 and Theorem 18.2 extend the results for the OLS estimator. I leave their proofs as Problem 18.3. Based on the results in Theorem 18.2, we can derive the t and F statistics for finite-sample inference of β based on the estimator $\hat{\beta}_r$ and the estimated covariance matrix

$$\hat{\sigma}_r^2 M_r (X^T X)^{-1} M_r^T.$$

I then discuss statistical inference under the heteroskedastic linear model under Assumption 6.1 and the restriction (18.1). Corollary 18.2 implies that

$$\text{cov}(\hat{\beta}_r) = M_r (X^T X)^{-1} X^T \text{diag}\{\sigma_1^2, \dots, \sigma_n^2\} X (X^T X)^{-1} M_r^T.$$

Therefore, the EHW-type estimated covariance matrix is

$$\hat{V}_{\text{EHW},r} = M_r (X^T X)^{-1} X^T \text{diag}\{\hat{\varepsilon}_{i,r}^2, \dots, \hat{\varepsilon}_{n,r}^2\} X (X^T X)^{-1} M_r^T.$$

where the $\hat{\varepsilon}_{i,r}$'s are the residuals from the restricted OLS.

18.4 Final remarks

This chapter follows Theil (1971) and Rao (1973). Bock et al. (1973), Judge et al. (1974), and Tarpey (2000) contained additional results on restricted OLS. Zhao and Ding (2023) used restricted OLS to analyze factorial experiments with covariates.

18.5 Homework problems

18.1 Algebraic details of restricted OLS

Prove Corollaries 18.1 and 18.2.

18.2 Invariance of restricted OLS

Consider an $N \times 1$ vector Y and two $N \times p$ matrices, X and X' , that satisfy $X' = X\Gamma$ for some nonsingular $p \times p$ matrix Γ . The restricted OLS fits of

$$\begin{aligned} Y &= X\hat{\beta}_r + \hat{\epsilon}_r && \text{subject to } C\hat{\beta}_r = r, \\ Y &= \tilde{X}\tilde{\beta}_r + \tilde{\epsilon}_r && \text{subject to } \tilde{C}\tilde{\beta}_r = r, \end{aligned}$$

with $\tilde{X} = X\Gamma$ and $\tilde{C} = C\Gamma$ yield $(\hat{\beta}_r, \hat{\epsilon}_r, \hat{V}_{\text{EHW},r})$ and $(\tilde{\beta}_r, \tilde{\epsilon}_r, \tilde{V}_{\text{EHW},r})$ as the coefficient vectors, residuals, and robust covariances.

Prove that

$$\hat{\beta}_r = \Gamma\tilde{\beta}_r, \quad \hat{\epsilon}_r = \tilde{\epsilon}_r, \quad \hat{V}_{\text{EHW},r} = \Gamma\tilde{V}_{\text{EHW},r}\Gamma^T.$$

18.3 Moments and distribution of restricted OLS

Prove Corollary 18.3 and Theorem 18.2.

18.4 Gauss–Markov theorem for restricted OLS

The Gauss–Markov theorem for $\hat{\beta}_r$ holds, as an extension of Theorem 4.4 for $\hat{\beta}$. Prove Theorem 18.3 below.

Theorem 18.3 (Gauss–Markov theorem for restricted OLS) *Under the Gauss–Markov model with the restrictions (18.1), $\hat{\beta}_r$ is the best linear unbiased estimator in the sense that $\text{cov}(\tilde{\beta}_r) - \text{cov}(\hat{\beta}_r) \succeq 0$ for any linear estimator $\tilde{\beta}_r = \tilde{c} + \tilde{A}_r Y$, with $\tilde{c} \in \mathbb{R}^p$ and $\tilde{A}_r \in \mathbb{R}^{p \times n}$, that satisfies $E(\tilde{\beta}_r) = \beta$ for all β under constraint (18.1).*

Remark: As a corollary of Theorem 18.3, we have

$$(X^T X)^{-1} \succeq M_r (X^T X)^{-1} M_r^T$$

because the restricted OLS estimator is BLUE whereas the unrestricted OLS is not, under the Gauss–Markov theorem with the restriction (18.1).

18.5 Short regression as restricted OLS

The short regression is a special case of the formula of $\hat{\beta}_r$. Show that

$$\hat{\beta}_r = \begin{pmatrix} (X_1^T X_1)^{-1} X_1^T Y \\ 0_l \end{pmatrix}$$

with

$$C = (0_{l \times k}, I_{l \times l}), \quad r = 0_l.$$

In this special case, $p = k + l$.

From the short regression, we can obtain the EHW estimated covariance matrix $\hat{V}_{\text{EHW},1}$. We can also obtain the EHW estimated covariance matrix $\hat{V}_{\text{EHW},r}$ from the restricted OLS.

Prove that

$$\hat{V}_{\text{EHW},r} = \begin{pmatrix} \hat{V}_{\text{EHW},1} & 0 \\ 0 & 0 \end{pmatrix}.$$

18.6 Reducing restricted OLS to OLS

Consider the restricted OLS fit

$$Y = X\hat{\beta}_r + \hat{\varepsilon}_r \quad \text{subject to } C\hat{\beta}_r = 0, \quad (18.3)$$

where $X \in \mathbb{R}^{n \times p}$ and $C \in \mathbb{R}^{l \times p}$.

Let $C_\perp \in \mathbb{R}^{(p-l) \times p}$ be an orthogonal complement of C in the sense that (C_\perp^\top, C^\top) is nonsingular with $C_\perp C^\top = 0$. Define

$$X_\perp = XC_\perp^\top (C_\perp C_\perp^\top)^{-1}.$$

Consider the corresponding unrestricted OLS fit

$$Y = X_\perp \hat{\beta}_\perp + \hat{\varepsilon}_\perp, \quad (18.4)$$

First, prove that the coefficient and residual vectors must satisfy

$$\hat{\beta}_\perp = C_\perp \hat{\beta}_r, \quad \hat{\varepsilon}_\perp = \hat{\varepsilon}_r.$$

Second, prove

$$\hat{V}_{\text{EHW}, \perp} = C_\perp \hat{V}_{\text{EHW}, r} C_\perp^\top,$$

where $\hat{V}_{\text{EHW}, \perp}$ is the EHW robust covariance matrix from (18.4) and $\hat{V}_{\text{EHW}, r}$ is the EHW robust covariance matrix from (18.3).

18.7 Minimum normal OLS estimator as restricted OLS

An application of the formula of $\hat{\beta}_r$ is the minimum norm estimator for under-determined linear equations. When X has more columns than rows, $Y = X\beta$ can have infinitely many solutions, but we may only be interested in the solution with the minimum norm. Assume $p \geq n$ and the rows of X are linearly independent.

Prove that the solution to

$$\min_{b \in \mathbb{R}^p} \|b\|^2 \text{ such that } Y = Xb$$

is

$$\hat{\beta}_m = X^\top (XX^\top)^{-1} Y.$$

18.8 Restricted OLS with degenerate design matrix

Greene and Seaks (1991) pointed out that restricted OLS does not require that $X^\top X$ be invertible, although the proof of Theorem 18.1 does. Modify the proof to show that the restricted OLS and the Lagrange multiplier satisfy

$$\begin{pmatrix} \hat{\beta}_r \\ \lambda \end{pmatrix} = W^{-1} \begin{pmatrix} X^\top Y \\ r \end{pmatrix}, \quad (18.5)$$

as long as

$$W = \begin{pmatrix} X^\top X & C^\top \\ C & 0 \end{pmatrix}$$

is invertible.

Derive the statistical results in parallel with Section 18.3.

Remark: If X has full column rank p , then W must be invertible. Even if X does not have full column rank, W can still be invertible. See Problem 18.9 below for more details.

18.9 Restricted OLS with degenerate design matrix: more algebra

This problem provides more algebraic details for Problem 18.8. Prove Lemma 18.1 below.

Lemma 18.1 *Consider*

$$W = \begin{pmatrix} X^T X & C^T \\ C & 0 \end{pmatrix}$$

where $X^T X$ may not be invertible and C has full row rank.

The matrix W is invertible if and only if $\begin{pmatrix} X \\ C \end{pmatrix}$ has full column rank p .

Remark: When X has full column rank p , then $\begin{pmatrix} X \\ C \end{pmatrix}$ must have full column rank p , which ensures that W is invertible by Lemma 18.1. I made the comment in Problem 18.8.

The invertibility of W plays an important role in other applications. See Benzi et al. (2005) and Bai and Bai (2013) for more general results.



Weighted Least Squares

This chapter will discuss the weighted least squares (WLS), a simple modification of OLS. Computationally, WLS minimizes the weighted average of the squared residuals of the individual observations, with the original OLS as a special case with equal weights. WLS is a useful tool in data analysis at least in the following two cases:

- (C1) In the linear model with heteroskedastic errors, WLS can improve efficiency compared with OLS.
- (C2) With survey data under non-uniform sampling probabilities, WLS can recover the targeted parameters by inverse probability weighting.

Moreover, WLS is a powerful building block to understand other advanced statistics methods. This chapter will introduce the local linear regression, a powerful nonparametric regression method, based on WLS. Chapter 20 will use WLS iteratively to compute the maximum likelihood estimate of the logistic regression.

19.1 Generalized least squares

We can extend the Gauss–Markov model to allow for a general covariance structure of the error term. The following generalized Gauss–Markov model is due to Aitkin (1936).

Assumption 19.1 (Generalized Gauss–Markov model) *We have*

$$Y = X\beta + \varepsilon,$$

with

$$E(\varepsilon) = 0, \quad \text{cov}(\varepsilon) = \sigma^2 \Sigma,$$

where X is a fixed matrix with p linearly independent columns. The unknown parameters are β and σ^2 . The Σ is a known positive definite matrix.

Two leading cases of generalized least squares are

$$\Sigma = \text{diag} \{w_1^{-1}, \dots, w_n^{-1}\}, \tag{19.1}$$

which corresponds to a diagonal covariance matrix, and

$$\Sigma = \text{diag} \{\Sigma_1, \dots, \Sigma_K\}, \tag{19.2}$$

which corresponds to a block diagonal covariance matrix where Σ_k is $n_k \times n_k$ and $\sum_{k=1}^K n_k = n$.

Under Assumption 19.1, we can still use the OLS estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$. It is unbiased because

$$E(\hat{\beta}) = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X \beta = \beta$$

relies only on the assumption $E(\varepsilon) = 0$. It has covariance matrix

$$\begin{aligned} \text{cov}(\hat{\beta}) &= \text{cov}\{(X^T X)^{-1} X^T Y\} \\ &= (X^T X)^{-1} X^T \text{cov}(Y) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1} \end{aligned} \quad (19.3)$$

because $\text{cov}(Y) = \Sigma$. The OLS estimator is the BLUE under the Gauss–Markov model, but it is not under the generalized Gauss–Markov model. Then what is the BLUE? We can transform the model under Assumption 19.1 into the Gauss–Markov model by standardizing the error term to have mean 0 and covariance I_n :

$$\Sigma^{-1/2} Y = \Sigma^{-1/2} X \beta + \Sigma^{-1/2} \varepsilon.$$

Define $Y_* = \Sigma^{-1/2} Y$, $X_* = \Sigma^{-1/2} X$ and $\varepsilon_* = \Sigma^{-1/2} \varepsilon$. The model under Assumption 19.1 reduces to

$$Y_* = X_* \beta + \varepsilon_*,$$

with

$$E(\varepsilon_*) = 0, \quad \text{cov}(\varepsilon_*) = \sigma^2 I_n,$$

which is the Gauss–Markov model for the transformed variables Y_* and X_* . The Gauss–Markov theorem ensures that the BLUE is

$$\begin{aligned} \hat{\beta}_\Sigma &= (X_*^T X_*)^{-1} X_*^T Y_* \\ &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y. \end{aligned}$$

It is unbiased because

$$\begin{aligned} E(\hat{\beta}_\Sigma) &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} E(Y) \\ &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} X \beta \\ &= \beta. \end{aligned}$$

It has covariance matrix¹

$$\begin{aligned} \text{cov}(\hat{\beta}_\Sigma) &= \text{cov}\{(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y\} \\ &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \text{cov}(Y) \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} \\ &= \sigma^2 (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \Sigma \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} \\ &= \sigma^2 (X^T \Sigma^{-1} X)^{-1}. \end{aligned} \quad (19.4)$$

In particular, $\text{cov}(\hat{\beta}_\Sigma)$ is smaller than or equal to $\text{cov}(\hat{\beta})$ in the matrix sense. So based on (19.3) and (19.4), we have the following pure linear algebra inequality:

Corollary 19.1 *If X has linearly independent columns and Σ is invertible, then*

$$(X^T \Sigma^{-1} X)^{-1} \preceq (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}.$$

Problem 19.1 gives a more general result.

¹The matrix $X^T \Sigma^{-1} X$ is positive definite and thus invertible, because

- $\Sigma^{-1} \succeq 0$ implies that $\alpha^T X^T \Sigma X \alpha \geq 0$ for all $\alpha \in \mathbb{R}^p$;
- $\alpha^T X^T \Sigma X \alpha = 0$ if and only if $X \alpha = 0$, which is equivalent to $\alpha = 0$ since X has linearly independent columns.

19.2 Weighted least squares

This chapter focuses on the first covariance structure in (19.1) and Chapter 25 will discuss the second in (19.2). The Σ in (19.1) results in the weighted least squares (WLS) estimator

$$\begin{aligned}\hat{\beta}_w = \hat{\beta}_\Sigma &= (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y \\ &= \left(\sum_{i=1}^n w_i x_i x_i^T \right)^{-1} \sum_{i=1}^n w_i x_i y_i.\end{aligned}$$

From the derivation above, we can also write the WLS estimator as

$$\begin{aligned}\hat{\beta}_w &= \arg \min_{b \in \mathbb{R}^p} (Y - Xb)^T \Sigma^{-1} (Y - Xb) \\ &= \arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^n w_i (y_i - x_i^T b)^2 \\ &= \arg \min_{b \in \mathbb{R}^p} (Y_* - X_* b)^T (Y_* - X_* b) \\ &= \arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^n (y_{*i} - x_{*i}^T b)^2,\end{aligned}$$

where

$$Y_* = W^{1/2} Y, \quad X_* = W^{1/2} X, \quad \text{with } W = \text{diag}(w_1, \dots, w_n),$$

and

$$y_{*i} = w_i^{1/2} y_i, \quad x_{*i} = w_i^{1/2} x_i.$$

So the WLS is equivalent to the OLS with transformed variables, with the weights inversely proportional to the variances of the errors. By this equivalence, WLS inherits many properties of OLS. See the problems in Section 19.6 for more details.

Analogous to OLS, we can derive finite-sample exact inference based on the generalized Normal linear model:

Assumption 19.2 (Generalized Normal Linear Model) For $i = 1, \dots, n$, we have

$$y_i = x_i^T \beta + \varepsilon_i,$$

where x_i is fixed, $\varepsilon_i \sim N(0, \sigma^2/w_i)$, and the ε_i 's are independent across units.

The generalized Normal linear model in Assumption 19.2 is equivalent to

$$y_{*i} = x_{*i}^T \beta + \varepsilon_{*i}$$

with IID $\varepsilon_{*i} \sim N(0, \sigma^2)$. That is, (x_{*i}, y_{*i}) 's satisfy the classic Normal linear model. The `lm` function with `weights` reports the standard error, t -statistic, and p -value based on this model.

Assumption 19.2 requires that the weights fully capture the heteroskedasticity, which is unrealistic in many problems. More generally, we can derive asymptotic inference based on the following heteroskedastic linear model:

Assumption 19.3 For $i = 1, \dots, n$, we have

$$y_i = x_i^T \beta + \varepsilon_i,$$

where x_i is fixed, and the ε_i 's are independent with mean zero and variances σ_i^2 .

Assumption 19.3 is identical to the heteroskedastic linear model under Assumption 6.1. I repeat it here for symmetric presentation. Under Assumption 19.3, it is possible that $w_i \neq 1/\sigma_i^2$, i.e., the variances used to construct the WLS estimator can be misspecified. Even though there is no guarantee that $\hat{\beta}_w$ is BLUE, it is still unbiased. From the decomposition

$$\begin{aligned}\hat{\beta}_w &= \left(\sum_{i=1}^n w_i x_i x_i^T \right)^{-1} \sum_{i=1}^n w_i x_i y_i \\ &= \left(\sum_{i=1}^n w_i x_i x_i^T \right)^{-1} \sum_{i=1}^n w_i x_i (x_i^T \beta + \varepsilon_i) \\ &= \beta + \left(n^{-1} \sum_{i=1}^n w_i x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n w_i x_i \varepsilon_i \right),\end{aligned}$$

we can apply the law of large numbers to show that $\hat{\beta}_w$ is consistent for β and apply the CLT to show that

$$\hat{\beta}_w \stackrel{a}{\sim} N(\beta, V_w),$$

where

$$V_w = n^{-1} \left(n^{-1} \sum_{i=1}^n w_i x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n w_i^2 \sigma_i^2 x_i x_i^T \right) \left(n^{-1} \sum_{i=1}^n w_i x_i x_i^T \right)^{-1}.$$

The EHW robust covariance generalizes to

$$\hat{V}_{\text{EHW},w} = n^{-1} \left(n^{-1} \sum_{i=1}^n w_i x_i x_i^T \right)^{-1} \left(n^{-1} \sum_{i=1}^n w_i^2 \hat{\varepsilon}_{w,i}^2 x_i x_i^T \right) \left(n^{-1} \sum_{i=1}^n w_i x_i x_i^T \right)^{-1},$$

where $\hat{\varepsilon}_{w,i} = y_i - x_i^T \hat{\beta}_w$ is the residual from the WLS. In the sandwich covariance above, w_i appears in the “bread” matrix, but w_i^2 appears in the “middle” or “meat” matrix. This formula appeared in Magee (1998) and Romano and Wolf (2017). The function `hccm` in the **R** package `car` can compute various EHW covariance estimators based on WLS. To save space in the examples below, I report only the standard errors based on the generalized Normal linear model and leave the calculations of the EHW covariances to Problem 19.18.

19.3 WLS motivated by heteroskedasticity

19.3.1 Feasible generalized least squares

Assume that ε has mean zero and covariance matrix $\text{diag}\{\sigma_1^2, \dots, \sigma_n^2\}$. If the σ_i^2 's are known, we can simply apply the WLS above; if they are unknown, we need to estimate them first. This gives the following feasible generalized least squares estimator (FGLS):

- (S1) Run OLS of y_i on x_i to obtain the residuals $\hat{\varepsilon}_i$. Then obtain the squared residuals $\hat{\varepsilon}_i^2$.
- (S2) Run OLS of $\log(\hat{\varepsilon}_i^2)$ on x_i to obtain the fitted values and exponentiate them to obtain $(\hat{\sigma}_i^2)_{i=1}^n$;

(S3) Run WLS of y_i on x_i with weights $\hat{\sigma}_i^{-2}$ to obtain

$$\hat{\beta}_{\text{FGLS}} = \left(\sum_{i=1}^n \hat{\sigma}_i^{-2} x_i x_i^T \right)^{-1} \sum_{i=1}^n \hat{\sigma}_i^{-2} x_i y_i.$$

In (S2), we can change the model based on our understanding of heteroskedasticity. The above FGLS estimator is close to Wooldridge (2012, Chapter 8). Romano and Wolf (2017) propose to regress $\log(\max(\delta^2, \hat{\varepsilon}_i^2))$ on $\log|x_{i1}|, \dots, \log|x_{ip}|$ to estimate the individual variances. Their modification has two features: first, they truncate the small residuals by a pre-specified positive number δ^2 ; second, their regressors are the logs of the absolute values of the original covariates. Rose (1978) and Carroll (1982) proposed to estimate σ_i^2 by using the kernel regression estimator of the residual on covariates. Robinson (1987) discussed the analog using the nearest neighbor method.

Here I use the Boston housing data to compare the OLS and FGLS.

```
> library(mlbench)
> data(BostonHousing)
> ols.fit = lm(medv ~ ., data = BostonHousing)
> dat.res = BostonHousing
> ## log transformation of the squared residuals
> dat.res$medv = log((ols.fit$residuals)^2)
> t.res.ols = lm(medv ~ ., data = dat.res)
> w.fgls = exp(- t.res.ols$fitted.values)
> fgls.fit = lm(medv ~ ., weights = w.fgls, data = BostonHousing)
> round(summary(ols.fit)$coef, 3)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.459	5.103	7.144	0.000
crim	-0.108	0.033	-3.287	0.001
zn	0.046	0.014	3.382	0.001
indus	0.021	0.061	0.334	0.738
chas1	2.687	0.862	3.118	0.002
nox	-17.767	3.820	-4.651	0.000
rm	3.810	0.418	9.116	0.000
age	0.001	0.013	0.052	0.958
dis	-1.476	0.199	-7.398	0.000
rad	0.306	0.066	4.613	0.000
tax	-0.012	0.004	-3.280	0.001
ptratio	-0.953	0.131	-7.283	0.000
b	0.009	0.003	3.467	0.001
lstat	-0.525	0.051	-10.347	0.000

```
> round(summary(fgls.fit)$coef, 3)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.499	4.064	2.338	0.020
crim	-0.081	0.044	-1.825	0.069
zn	0.030	0.011	2.673	0.008
indus	-0.035	0.038	-0.922	0.357
chas1	1.462	1.119	1.306	0.192
nox	-7.161	2.784	-2.572	0.010
rm	5.675	0.364	15.588	0.000
age	-0.044	0.008	-5.501	0.000
dis	-0.927	0.139	-6.683	0.000
rad	0.170	0.051	3.312	0.001
tax	-0.010	0.002	-4.142	0.000
ptratio	-0.700	0.094	-7.447	0.000
b	0.014	0.002	6.545	0.000
lstat	-0.158	0.036	-4.380	0.000

Unfortunately, the coefficients, including the point estimates and standard errors, from OLS and FGLS are quite different for several covariates. This suggests that the linear model may be misspecified. Otherwise, both estimators are consistent for the same true coefficient, and they should not be so different even in the presence of randomness.

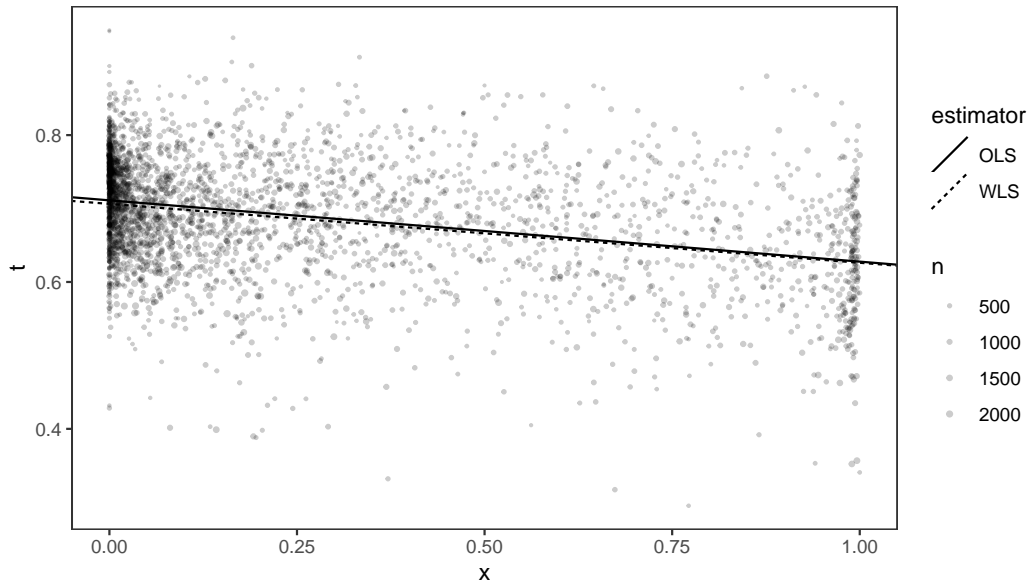


FIGURE 19.1: Fulton data

Romano and Wolf (2017) highlighted the efficiency gain from the FGLS compared with OLS in the presence of heteroskedasticity. DiCiccio et al. (2019) proposed some improved versions of the FGLS estimator even if the variance function is misspecified. However, it is unusual for practitioners to use FGLS even though it can be more efficient than OLS. There are several reasons. First, the EHW standard errors are convenient for correcting the standard error of OLS under heteroskedasticity. Second, the efficiency gain is usually small, and it is even possible that the FGLS is less efficient than OLS when the variance function is misspecified. Third, the linear model is very likely to be misspecified, and if so, OLS and FGLS estimate different parameters. The OLS has the interpretations as the best linear predictor and the best linear approximation of the conditional mean, but the FGLS has more complicated interpretations when the linear model is wrong. Based on these reasons, we need to carefully justify the choice of FGLS over OLS in practical data analyses.

19.3.2 Aggregate data and ecological regression

In some case, (y_i, x_i) come from aggregate data. For example, y_i can be the average test score and x_i can be the average parents' income of students within classroom i . If we believe that the student-level test score and parents' income follow a homoskedastic linear model, then the model based on the classroom average must be heteroskedastic, with the variance inversely proportional to the classroom size. In this case, a natural choice of weight is $w_i = n_i$, the classroom size.

Below I use the `lavoteall` dataset from the R package `ei`. It contains the fraction of black registered voters x , the fraction of voter turnout t , and the total number of people n in each Louisiana precinct. Figure 19.1 is the scatterplot. In this example, OLS and WLS give similar results although n varies a lot across precincts.

```
> lavoteall = read.csv("lavoteall.csv")
> ols.fit = lm(t ~ x, data = lavoteall)
> wls.fit = lm(t ~ x, weights = n, data = lavoteall)
> round(summary(ols.fit)$coef, 3)
```

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.711      0.002 408.211      0
x             -0.083      0.004 -19.953      0
> round(summary(wls.fit)$coef, 3)
      Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.706      0.002 421.662      0
x             -0.080      0.004 -19.938      0

```

In the above, we can interpret the coefficient of \mathbf{x} as the precinct-level relationship between the fraction of black registered voters and the fraction voting. Political scientists are interested in using aggregated data to infer individual voting behavior. Hypothetically, the precinct i has individual data $\{x_{ij}, y_{ij} : j = 1, \dots, n_i\}$, where x_{ij} and y_{ij} are the binary racial and voting status of individual (i, j) ($i = 1, \dots, n; j = 1, \dots, n_i$). However, we only observe the aggregated data $\{\bar{x}_i, \bar{y}_i, n_i : i = 1, \dots, n\}$, where

$$\bar{x}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}, \quad \bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$$

are the fraction of black registered voters and the fraction voting, respectively. Can we infer the individual voting behavior based on the aggregated data? In general, this is almost impossible. Under some assumptions, we can make progress. Goodman's ecological regression below is one possibility.

Assume that for precinct $i = 1, \dots, n$, we have

$$y_{ij} \mid x_{ij} = 1 \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p_{i1}), \quad y_{ij} \mid x_{ij} = 0 \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p_{i0}), \quad (j = 1, \dots, n_i).$$

This is the individual-level model, where the p_{i1} 's and p_{i0} 's measure the association between race and voting. We further assume that they are random and independent of the x_{ij} 's, with means

$$E(p_{i1}) = p_1, \quad E(p_{i0}) = p_0. \quad (19.5)$$

Then we can decompose the aggregated outcome variable as

$$\begin{aligned}
 \bar{y}_i &= n_i^{-1} \sum_{j=1}^{n_i} y_{ij} \\
 &= n_i^{-1} \sum_{j=1}^{n_i} \{x_{ij}y_{ij} + (1 - x_{ij})y_{ij}\} \\
 &= n_i^{-1} \sum_{j=1}^{n_i} \{x_{ij}p_1 + (1 - x_{ij})p_0\} + \varepsilon_i \\
 &= p_1\bar{x}_i + p_0(1 - \bar{x}_i) + \varepsilon_i,
 \end{aligned}$$

where

$$\varepsilon_i = n_i^{-1} \sum_{j=1}^{n_i} \{x_{ij}(y_{ij} - p_1) + (1 - x_{ij})(y_{ij} - p_0)\}.$$

So we have a linear relationship between the aggregated outcome and covariate

$$\bar{y}_i = p_1\bar{x}_i + p_0(1 - \bar{x}_i) + \varepsilon_i,$$

where

$$E(\varepsilon_i \mid \bar{x}_i) = 0.$$

Goodman (1953) suggested to use the OLS of \bar{y}_i on $\{\bar{x}_i, (1 - \bar{x}_i)\}$ to estimate (p_1, p_0) , and Goodman (1959) suggested to use the corresponding WLS with weight n_i since the variance of ε_i has the magnitude n_i^{-1} . Moreover, the variance of ε_i has a rather complicated form of heteroskedasticity, so we should use the EHW standard error for inference. This is called *Goodman's regression* or *ecological regression*. The following R code implements ecological regression based on the `lavoteall` data.

```
> ols.fit = lm(t ~ 0 + x + I(1-x), data = lavoteall)
> wls.fit = lm(t ~ 0 + x + I(1-x), weights = n, data = lavoteall)
> round(summary(ols.fit)$coef, 3)
              Estimate Std. Error t value Pr(>|t|)
x              0.628      0.003 188.292      0
I(1 - x)       0.711      0.002 408.211      0
> round(summary(wls.fit)$coef, 3)
              Estimate Std. Error t value Pr(>|t|)
x              0.626      0.003 194.493      0
I(1 - x)       0.706      0.002 421.662      0
```

The assumption in (19.5) is crucial, which can be too strong when the precinct level p_{i1} 's and p_{i0} 's vary in systematic but unobserved ways. When the assumption is violated, it is possible that the ecological regression yields the opposite result compared to the individual regression. This is called the *ecological fallacy*.

Another obvious problem of ecological regression is that the estimated coefficients may lie outside of the interval $[0, 1]$. Problem 19.19 gives an example.

Gelman et al. (2001) gave an alternative set of assumptions justifying the ecological regression. King (1997) proposed some extensions. Robinson (1950) warned that the ecological correlation might not inform individual correlation. Freedman et al. (1991) warned that the assumptions underlying the ecological regression might not be plausible in practice.

19.4 WLS Motivated by Survey Weights

WLS can be used in other settings unrelated to heteroskedasticity. A leading case is survey sampling. Most discussions in this book are based on IID samples, or, at least, the sample represents the population of interest. Sometimes, researchers over sample some units and under sample some other units from a population of interest.

If we have the large population with size N , then the ideal OLS estimator of the y_i 's on the x_i 's is

$$\hat{\beta}_{\text{ideal}} = \left(\sum_{i=1}^N x_i x_i^T \right)^{-1} \sum_{i=1}^N x_i y_i.$$

However, we do not have all the data points in the large population, but sample each data point independently with probability

$$\pi_i = \text{pr}(I_i = 1 \mid x_i, y_i),$$

where I_i is a binary indicator for being included in the sample. Conditioning on $X_N = (x_i)_{i=1}^N$ and $Y_N = (y_i)_{i=1}^N$, $\hat{\beta}_{\text{ideal}}$ is a fixed number, and an estimator is the following WLS

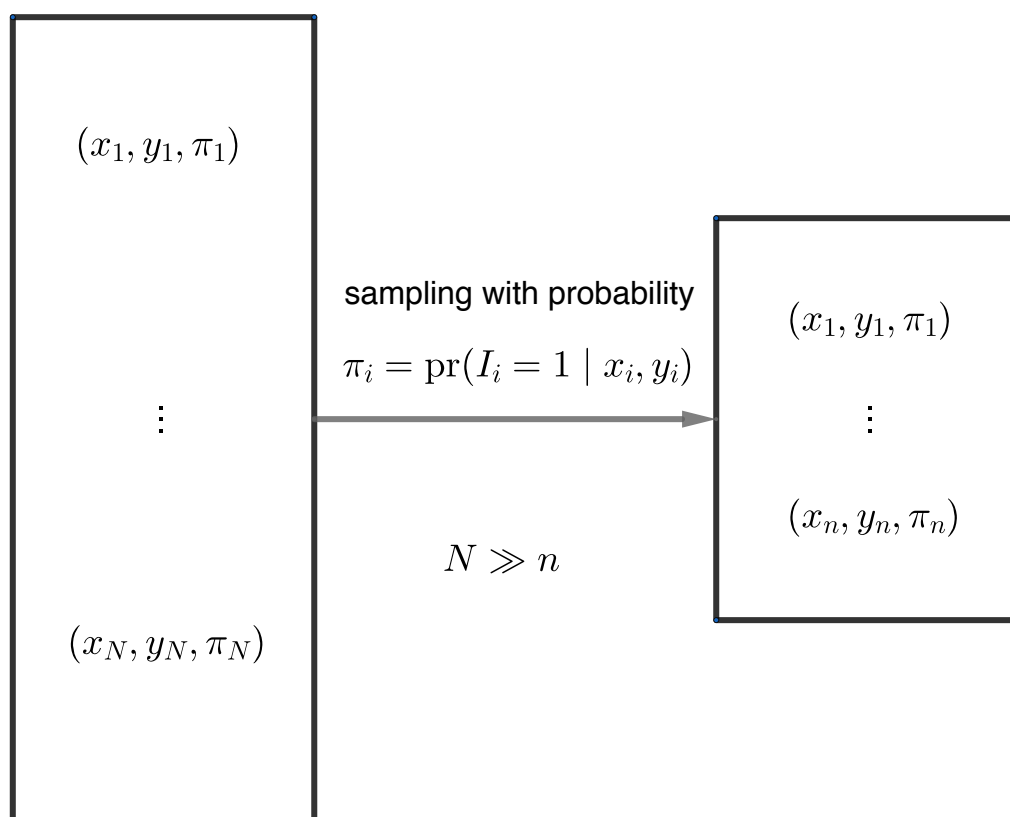


FIGURE 19.2: Survey sampling

estimator²

$$\begin{aligned}\hat{\beta}_{1/\pi} &= \left(\sum_{i=1}^N \frac{I_i}{\pi_i} x_i x_i^T \right)^{-1} \sum_{i=1}^N \frac{I_i}{\pi_i} x_i y_i \\ &= \left(\sum_{i=1}^n \pi_i^{-1} x_i x_i^T \right)^{-1} \sum_{i=1}^n \pi_i^{-1} x_i y_i,\end{aligned}$$

with weights inversely proportional to the sampling probability. This inverse probability weighting estimator is reasonable because

$$\begin{aligned}E \left(\sum_{i=1}^N \frac{I_i}{\pi_i} x_i x_i^T \mid X_N, Y_N \right) &= \sum_{i=1}^N x_i x_i^T, \\ E \left(\sum_{i=1}^N \frac{I_i}{\pi_i} x_i y_i \mid X_N, Y_N \right) &= \sum_{i=1}^N x_i y_i.\end{aligned}$$

The inverse probability weighting estimators are called the Horvitz–Thompson estimators (Horvitz and Thompson, 1952), which are the cornerstones of survey sampling.

Below I illustrate the use of sampling weight using the data from Angrist et al. (2006). Below, `perwt` represents sampling weight.

```
> library(foreign)
> census00 = read.dta("census00.dta")
> ols.fit = lm(logwk ~ age + educ + exper + exper2 + black,
+             data = census00)
> wls.fit = lm(logwk ~ age + educ + exper + exper2 + black,
+             weights = perwt, data = census00)
> round(summary(ols.fit)$coef, 3)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.167      0.128   40.308   0.000
age            -0.015      0.007   -2.201   0.028
educ             0.130      0.007   19.712   0.000
exper2           0.000      0.000    2.243   0.025
black          -0.247      0.008  -29.173   0.000
> round(summary(wls.fit)$coef, 3)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.074      0.127   40.030   0.000
age            -0.008      0.007   -1.263   0.207
educ             0.123      0.007   18.807   0.000
exper2           0.000      0.000    1.294   0.196
black          -0.257      0.008  -32.013   0.000
```

19.5 WLS as a Building Block for Local linear regression

Calculus tells us that locally, we can approximate any smooth function $f(x)$ by a linear function even though the original function can be highly nonlinear:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0),$$

²The notation $\sum_{i=1}^N$ and $\sum_{i=1}^n$ can be confusing here. The summation $\sum_{i=1}^N$ is over all units in the population, whereas the summation $\sum_{i=1}^n$ is over all units in the observed data.

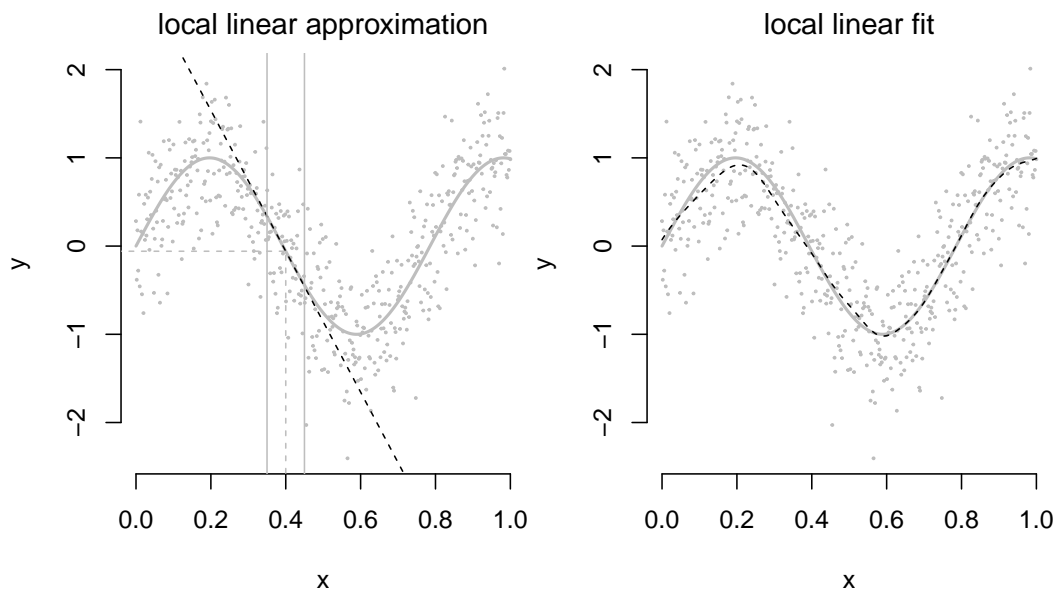


FIGURE 19.3: Local linear regression

when x is near x_0 . The left panel of Figure 19.3 shows that in the neighborhood of $x_0 = 0.4$, even a sine function can be well approximated by a line. Based on data $(x_i, y_i)_{i=1}^n$, if we want to predict the mean value of y given $x = x_0$, then we can predict based on a line with the local data points close to x_0 . It is also reasonable to down weight the points that are far from x_0 , which motivates the following WLS:

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{a, b} \sum_{i=1}^n w_i \{y_i - a - b(x_i - x_0)\}^2$$

with $w_i = K\{(x_i - x_0)/h\}$, where $K(\cdot)$ is called the kernel function and h is called the bandwidth parameter. With the fitted line $\hat{y}(x) = \hat{\alpha} + \hat{\beta}(x - x_0)$, the predicted value at $x = x_0$ is the intercept $\hat{\alpha}$.

Technically, $K(\cdot)$ can be any density function, and two canonical choices are the standard Normal density $K(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$ and the Epanechnikov kernel $K(t) = 0.75(1-t^2)1(|t| \leq 1)$. The choice of the kernel does not matter that much. The choice of the bandwidth matters much more. With a larger bandwidth, we have a poorer linear approximation, leading to bias; with a smaller bandwidth, we have fewer data points, leading to larger variance. In practice, we face a bias-variance trade-off. In practice, we can either use cross-validation or other criterion to select h .

In general, we can approximate a smooth function by a polynomial:

$$f(x) \approx \sum_{k=0}^K \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k,$$

when x is near x_0 . So we can even fit a polynomial function locally, which is called the *local polynomial regression* (Fan and Gijbels, 1996). In the `R` package `kernsmooth`, the function `locpoly` fits local polynomial regression, and the function `dpi11` selects h based on Ruppert et al. (1995). The default specification of `locpoly` is the local linear regression. The right panel of Figure 19.3 shows the local linear fit of the data.

```

> library("KernSmooth")
KernSmooth 2.23 loaded
Copyright M. P. Wand 1997-2009
> n = 500
> x = seq(0, 1, length.out = n)
> fx = sin(8*x)
> y = fx + rnorm(n, 0, 0.5)
> library("KernSmooth")
> n = 500
> x = seq(0, 1, length.out = n)
> fx = sin(8*x)
> y = fx + rnorm(n, 0, 0.5)
> par(mfrow = c(1, 2), mar = c(4, 4, 2, 0.05))
> plot(y ~ x, pch = 19, cex = 0.2, col = "grey", bty = "n",
+      main = "local linear approximation", font.main = 1)
> lines(fx ~ x, lwd = 2, col = "grey")
>
> x0 = 0.4
> y0 = sin(8*x0)
> segments(x0, -3, x0, y0, lty = 2, col = "grey")
> segments(-4, y0, x0, y0, lty = 2, col = "grey")
>
> ylinear = sin(8*x0) + 8*cos(8*x0)*(x - x0)
> lines(ylinear ~ x, lty = 2)
> abline(v = x0 - 0.05, col = "grey")
> abline(v = x0 + 0.05, col = "grey")
>
>
> plot(y ~ x, pch = 19, cex = 0.2, col = "grey", bty = "n",
+      main = "local linear fit", font.main = 1)
> lines(fx ~ x, lwd = 2, col = "grey")
> h = dpill(x, y)
> locp.fit = locpoly(x, y, bandwidth = h)
> lines(locp.fit, lty = 2)

```

19.6 Homework problems

19.1 A linear algebra fact related to WLS

Theorem 19.1 below extends Corollary 19.1. Prove Theorem 19.1.

Theorem 19.1 *We have*

$$(X^T \Sigma^{-1} X)^{-1} \preceq (X^T \Omega X)^{-1} X^T \Omega \Sigma \Omega X (X^T \Omega X)^{-1},$$

as long as the inverse matrices exist. Moreover, the equality holds when $\Omega = \Sigma^{-1}$.

Remark: With $\Omega = I_n$, Theorem 19.1 reduces to Corollary 19.1. To prove Theorem 19.1, we can compare the covariance matrices of $\hat{\beta}_\Sigma$ and $\hat{\beta}_\Omega$ under the generalized Gauss–Markov model.

19.2 Generalized least squares with a block diagonal covariance

Partition X and Y into

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_K \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_K \end{pmatrix}$$

corresponding to Σ in (19.2) such that $X_k \in \mathbb{R}^{n_k \times p}$ and $Y_k \in \mathbb{R}^{n_k}$.

Prove that the generalized least squares estimator is

$$\hat{\beta}_\Sigma = \left(\sum_{k=1}^K X_k^\top \Sigma_k^{-1} X_k \right)^{-1} \left(\sum_{k=1}^K X_k^\top \Sigma_k^{-1} Y_k \right).$$

19.3 Univariate WLS

Prove the following Galtonian formula for the univariate WLS:

$$\min_{a,b} \sum_{i=1}^n w_i (y_i - a - bx_i)^2$$

has the minimizer

$$\begin{aligned} \hat{\beta}_w &= \frac{\sum_{i=1}^n w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}, \\ \hat{\alpha}_w &= \bar{y}_w - \hat{\beta}_w \bar{x}_w, \end{aligned}$$

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 means of the covariate and outcome.

19.4 Difference-in-means with weights

This problem extends Problem 19.3.

With a binary covariate x_i , show that the coefficient of x_i in the WLS of y_i on $(1, x_i)$ with weights w_i ($i = 1, \dots, n$) equals $\bar{y}_{w,1} - \bar{y}_{w,0}$, where

$$\begin{aligned} \bar{y}_{w,1} &= \frac{\sum_{i=1}^n w_i x_i y_i}{\sum_{i=1}^n w_i x_i}, \\ \bar{y}_{w,0} &= \frac{\sum_{i=1}^n w_i (1 - x_i) y_i}{\sum_{i=1}^n w_i (1 - x_i)} \end{aligned}$$

are the weighted averages of the outcome under treatment and control, respectively.

19.5 Asymptotic Normality of WLS and robust covariance estimator

Under the heteroskedastic linear model, prove that $\hat{\beta}_w$ is consistent and asymptotically Normal, and that $n\hat{V}_w$ is consistent for the asymptotic covariance of $\sqrt{n}(\hat{\beta}_w - \beta)$. Specify the regularity conditions.

19.6 WLS in ANOVA

This problem extends Problems 5.6 and 6.3.

For units $i = 1, \dots, n$, assume y_i denotes the outcome, x_i denotes the p -vector with entries as the dummy variables for a discrete covariate with p levels, $w_i > 0$ denotes a weight, and $\pi_i > 0$ denotes another weight that is a function of x_i only (for example, $\pi_i = n_j/n$ if $x_i = e_j$). Run the following regressions:

- (R1) WLS of y_i on x_i with weight w_i for $i = 1, \dots, n$ to obtain the coefficient vector $\hat{\beta}$ and EHW covariance matrix \hat{V} ;
- (R2) WLS of y_i on x_i with weight $w_i \pi_i$ for $i = 1, \dots, n$ to obtain the coefficient vector $\hat{\beta}'$ and EHW covariance matrix \hat{V}' .

Prove that $\hat{\beta} = \hat{\beta}'$ with the j th entry

$$\hat{\beta}_j = \hat{\beta}'_j = \frac{\sum_{i:x_i=e_j} w_i y_i}{\sum_{i:x_i=e_j} w_i},$$

and moreover, $\hat{V} = \hat{V}'$ are diagonal with the (j, j) th entry

$$\hat{V}_{jj} = \hat{V}'_{jj} = \frac{\sum_{i:x_i=e_j} w_i^2 (y_i - \hat{\beta}_j)^2}{(\sum_{i:x_i=e_j} w_i)^2}.$$

19.7 WLS with aggregate data

Due to privacy considerations, we may aggregate individual data $(x_i, y_i)_{i=1}^n$ into group means $(\bar{x}_j, \bar{y}_j)_{j=1}^m$ where

$$\bar{x}_j = \frac{1}{|\mathcal{G}_j|} \sum_{i \in \mathcal{G}_j} x_i, \quad \bar{y}_j = \frac{1}{|\mathcal{G}_j|} \sum_{i \in \mathcal{G}_j} y_i$$

with units $\{1, \dots, n\}$ partitioned into disjoint sets $\mathcal{G}_1, \dots, \mathcal{G}_m$.

Assume that the individual data satisfy the Gauss–Markov model

$$Y = X\beta + \varepsilon$$

with regression coefficient β and homoskedastic variance σ^2 . Based on the individual data, we can obtain the OLS estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$, which is BLUE.

Derive the BLUE for β based on the aggregate data, denoted by $\tilde{\beta}$. Prove that $\text{cov}(\tilde{\beta}) \geq \text{cov}(\hat{\beta})$.

Remark: Prais and Aitchison (1954) provided a formal discussion of regression with aggregate data. Lancaster (1968) discussed the possibility of $\text{cov}(\tilde{\beta}) \leq \text{cov}(\hat{\beta})$ when the error terms in the individual model have different variances.

19.8 An infeasible generalized least squares estimator

Can we skip Step S2 in Section 19.3.1 and directly apply the following WLS estimator:

$$\hat{\beta}_{\text{IGLS}} = \left(\sum_{i=1}^n \hat{\varepsilon}_i^{-2} x_i x_i^T \right)^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^{-2} x_i y_i,$$

with $\hat{\varepsilon}_i = y_i - x_i^T \hat{\beta}$ is the residual from the OLS. If so, give a theoretical justification; if not, give a counterexample. Evaluate the finite-sample properties of $\hat{\beta}_{\text{IGLS}}$ using simulated data.

19.9 FWL theorem in WLS

Theorem 19.2 below extends Theorem 7.1. Prove Theorem 19.2.

Theorem 19.2 Consider the WLS with an $n \times 1$ vector Y , an $n \times k$ matrix X_1 , an $n \times l$ matrix X_2 , and weights w_i 's. Consider the coefficient $\hat{\beta}_{w,2}$ in the long WLS fit

$$Y = X_1 \hat{\beta}_{w,1} + X_2 \hat{\beta}_{w,2} + \hat{\varepsilon}_w.$$

It equals the coefficient of $\tilde{X}_{w,2}$ in the WLS fit of Y on $\tilde{X}_{w,2}$, where $\tilde{X}_{w,2}$ are the residual vectors from the column-wise WLS of X_2 on X_1 . It also equals the coefficient of $\tilde{X}_{w,2}$ in the WLS fit of \tilde{Y}_w on $\tilde{X}_{w,2}$, where \tilde{Y}_w is the residual vector from the WLS of Y on X_1 .

19.10 Cochran's formula in WLS

Theorem 19.3 extends Theorem 9.1. Prove Theorem 19.3.

Theorem 19.3 Consider the WLS with an $n \times 1$ vector Y , an $n \times k$ matrix X_1 , an $n \times l$ matrix X_2 , and weights w_i 's. We can fit the following WLS:

$$\begin{aligned} Y &= X_1 \hat{\beta}_{w,1} + X_2 \hat{\beta}_{w,2} + \hat{\varepsilon}_w, \\ Y &= X_2 \tilde{\beta}_{w,2} + \tilde{\varepsilon}_w, \\ X_1 &= X_2 \hat{\delta}_w + \hat{U}_w, \end{aligned}$$

where $\hat{\varepsilon}_w, \tilde{\varepsilon}_w, \hat{U}_w$ are the residuals. The last WLS fit means the WLS fit of each column of X_1 on X_2 .

We have

$$\tilde{\beta}_{w,2} = \hat{\beta}_{w,2} + \hat{\delta}_w \hat{\beta}_{w,1}.$$

19.11 EHW robust covariance estimator in WLS

We have proved in Section 19.1 that the coefficients from WLS are identical to those from OLS with transformed variables. Further, prove that the corresponding HC0 version of EHW covariance estimators are also identical.

19.12 Invariance of covariance estimators in WLS

Problem 6.4 states the invariance of covariance estimators in OLS. Show that the same result holds for covariance estimators in WLS.

19.13 Ridge regression with weights

Define the ridge regression with weights w_i 's, and derive the formula for the ridge coefficient.

19.14 Coordinate descent algorithm in lasso with weights

Define the lasso with weights w_i 's, and give the coordinate descent algorithm for solving the weighted lasso problem.

19.15 General leave-one-out formula via WLS

With data (X, Y) , we can define $\hat{\beta}_{[-i]}(w)$ as the WLS estimator of Y on X with weights $w_{i'} = 1(i' \neq i) + w1(i' = i)$ for $i' = 1, \dots, n$, where $0 \leq w \leq 1$. It reduces to the OLS estimator $\hat{\beta}$ when $w = 1$ and the leave-one-out OLS estimator $\hat{\beta}_{[-i]}$ when $w = 0$.

Prove the general formula

$$\hat{\beta}_{[-i]}(w) = \hat{\beta} - \frac{1-w}{1-(1-w)h_{ii}}(X^T X)^{-1}x_i \hat{\varepsilon}_i$$

recalling that h_{ii} is the leverage score and $\hat{\varepsilon}_i$ is the residual of observation i .

Remark: Based on the above formula, we can compute the derivative of $\hat{\beta}_{[-i]}(w)$ with respect to w :

$$\frac{\partial \hat{\beta}_{[-i]}(w)}{\partial w} = \frac{1}{\{1 - (1-w)h_{ii}\}^2} (X^T X)^{-1} x_i \hat{\varepsilon}_i,$$

which reduces to

$$\frac{\partial \hat{\beta}_{[-i]}(0)}{\partial w} = \frac{1}{(1-h_{ii})^2} (X^T X)^{-1} x_i \hat{\varepsilon}_i$$

at $w = 0$ and

$$\frac{\partial \hat{\beta}_{[-i]}(1)}{\partial w} = (X^T X)^{-1} x_i \hat{\varepsilon}_i$$

at $w = 1$. Pregibon (1981) reviewed related formulas for OLS. Broderick et al. (2020) discussed related formulas for general statistical models.

19.16 Hat matrix and leverage score in WLS

Based on the WLS estimator $\hat{\beta}_w = (X^T W X)^{-1} X^T W Y$ with $W = \text{diag}(w_1, \dots, w_n)$, we have the predicted vector

$$\hat{Y}_w = X \hat{\beta}_w = X (X^T W X)^{-1} X^T W Y.$$

This motivates the definition of the hat matrix

$$H_w = X (X^T W X)^{-1} X^T W$$

such that $\hat{Y}_w = H_w Y$.

First, prove the following basic properties of the hat matrix:

$$\begin{aligned} W H_w &= H_w^T W, \\ X^T W (I_n - H_w) &= 0. \end{aligned}$$

(Note that H_w is not a projection matrix in general.)

Second, prove an extended version of Theorem 11.1: with $x_i = (1, x_{i2}^T)^T$, the (i, i) th diagonal element of H_w satisfies

$$h_{w,ii} = \frac{w_i}{\sum_{i'=1}^n w_{i'}} (1 + D_{w,i}^2)$$

where

$$D_{w,i}^2 = (x_{i2} - \bar{x}_{w,2})^T S_w^{-1} (x_{i2} - \bar{x}_{w,2})$$

with $\bar{x}_{w,2} = \sum_{i=1}^n w_i x_{i2} / \sum_{i=1}^n w_i$ being the weighted average of x_{i2} 's and $S_w = \sum_{i=1}^n w_i (x_{i2} - \bar{x}_{w,2})(x_{i2} - \bar{x}_{w,2})^T / \sum_{i=1}^n w_i$ being the corresponding sample covariance matrix.

Remark: Li and Valliant (2009) presented the basic properties of H_w for WLS in the context of survey data.

19.17 Leave-one-out formula for WLS

Use the notation in Problem 19.16. Let $\hat{\beta}_w$ be the WLS estimator of Y on X with weights w_i 's. Let $\hat{\beta}_{w[-i]}$ be the WLS estimator without using the i th observation.

Prove that

$$\hat{\beta}_{w[-i]} = \hat{\beta}_w - \frac{w_i}{1 - h_{w,ii}} (X^T W X)^{-1} x_i \hat{\varepsilon}_{w,i}.$$

19.18 EHW standard errors in WLS

Report the EHW standard errors in the examples in Sections 19.3.1, 19.3.2, and 19.4.

19.19 Another example of ecological inference

The `fultongen` dataset in the `ri` package contains aggregated data from 289 precincts in Fulton County, Georgia. The variable `t` represents the fraction voting in 1994 and `x` the fraction in 1992. The variable `n` represents the total number of people. Run ecological regression similar to Section 19.3.2.

Part VII

Generalized Linear Models



Logistic Regression for Binary Outcomes

Many applications have binary outcomes $y_i \in \{0, 1\}$. This chapter discusses statistical models of binary outcomes, focusing on the logistic regression, also called the logit regression for simplicity.

20.1 Regression with binary outcomes

20.1.1 Linear probability model

For simplicity, we can still use the linear model for a binary outcome. It is also called the *linear probability model*:

$$y_i = x_i^T \beta + \varepsilon_i \quad \text{with} \quad E(\varepsilon_i | x_i) = 0$$

because the conditional probability of $y_i = 1$ given x_i is a linear function of x_i :

$$\text{pr}(y_i = 1 | x_i) = E(y_i | x_i) = x_i^T \beta.$$

An advantage of this linear model is that the interpretation of the coefficient remains the same as linear models for general outcomes:

$$\frac{\partial \text{pr}(y_i = 1 | x_i)}{\partial x_{ij}} = \beta_j,$$

that is, β_j measures the partial impact of x_{ij} on the probability of y_i .

A minor technical issue is that the linear probability model implies heteroskedasticity because

$$\begin{aligned} \text{var}(y_i | x_i) &= \text{pr}(y_i = 1 | x_i)(1 - \text{pr}(y_i = 1 | x_i)) \\ &= x_i^T \beta (1 - x_i^T \beta). \end{aligned}$$

Therefore, we must use the EHW covariance based on OLS. We can also use the feasible generalized least squares (FGLS) in Chapter 19.3.1 to improve efficiency over OLS.

A more severe problem with the linear probability model is its plausibility in general. We may not believe that a linear model is the correct model for a binary outcome because the probability $\text{pr}(y_i = 1 | x_i)$ on the left-hand side is bounded between 0 and 1, but the linear combination $x_i^T \beta$ on the right-hand side can be unbounded for general covariates and coefficients. Nevertheless, the OLS decomposition $y_i = x_i^T \beta + \varepsilon_i$ works for any $y_i \in \mathbb{R}$, so it is applicable for binary y_i .

Sometimes, practitioners feel that the linear model is not natural for binary outcomes because the predicted value can be outside the range of $[0, 1]$. Therefore, it is more reasonable to build a model that automatically accommodates the binary feature of the outcome.

20.1.2 General link functions

A linear combination of general covariates may be outside the range of $[0, 1]$, but we can find a monotone transformation to force it to lie within the interval $[0, 1]$. This motivates us to consider the following model:

$$\text{pr}(y_i = 1 \mid x_i) = g(x_i^\top \beta),$$

where $g(\cdot) : \mathbb{R} \rightarrow [0, 1]$ is a monotone function, and its inverse is often called the link function. Mathematically, the distribution function of any continuous random variable is a monotone function that maps from \mathbb{R} to $[0, 1]$. So we have infinitely many choices for $g(\cdot)$. Four canonical choices “logit”, “probit”, “cauchit”, and “cloglog” are below which are the standard options in `R`:

name	functional form
logit	$g(z) = \frac{e^z}{1+e^z}$
probit	$g(z) = \Phi(z)$
cauchit	$g(z) = \frac{1}{\pi} \arctan(z) + \frac{1}{2}$
cloglog	$g(z) = 1 - \exp(-e^z)$

The above $g(z)$ ’s correspond to different distribution functions. The $g(z)$ for the logit model¹ is the distribution function of the standard logistic distribution with density

$$g'(z) = \frac{e^z}{(1+e^z)^2} = g(z) \{1 - g(z)\}. \quad (20.1)$$

The $g(z)$ for the probit model² is the distribution function of a standard Normal distribution. The $g(z)$ for the cauchit model is the distribution function of the standard Cauchy distribution with density

$$g'(z) = \frac{1}{\pi(1+z^2)}.$$

The $g(z)$ for the cloglog model is the distribution function of the standard log-Weibull distribution with density

$$g'(z) = \exp(z - e^z).$$

I will give more motivations for the first three link functions in Section 20.7.1 and for the fourth link function in Problem 22.6.

Figure 20.1 shows the distributions and densities of the corresponding link functions. The distribution functions are quite similar for all links, but the density for cloglog is asymmetric although all other three densities are symmetric.

This chapter will focus on the logit model, and extensions to other models are conceptually straightforward. We can also write the logit model as

$$\text{pr}(y_i = 1 \mid x_i) \equiv \pi(x_i, \beta) = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}}, \quad (20.2)$$

for the conditional probability of y_i given x_i , or, equivalently,

$$\text{logit} \{\text{pr}(y_i = 1 \mid x_i)\} \equiv \log \frac{\text{pr}(y_i = 1 \mid x_i)}{1 - \text{pr}(y_i = 1 \mid x_i)} = x_i^\top \beta,$$

for the log of the odds of y_i given x_i , with the logit function

$$\text{logit}(\pi) = \log \frac{\pi}{1 - \pi}.$$

¹Berkson (1944) was an early use of the logit model.

²Bliss (1934) was an early use of the probit model.

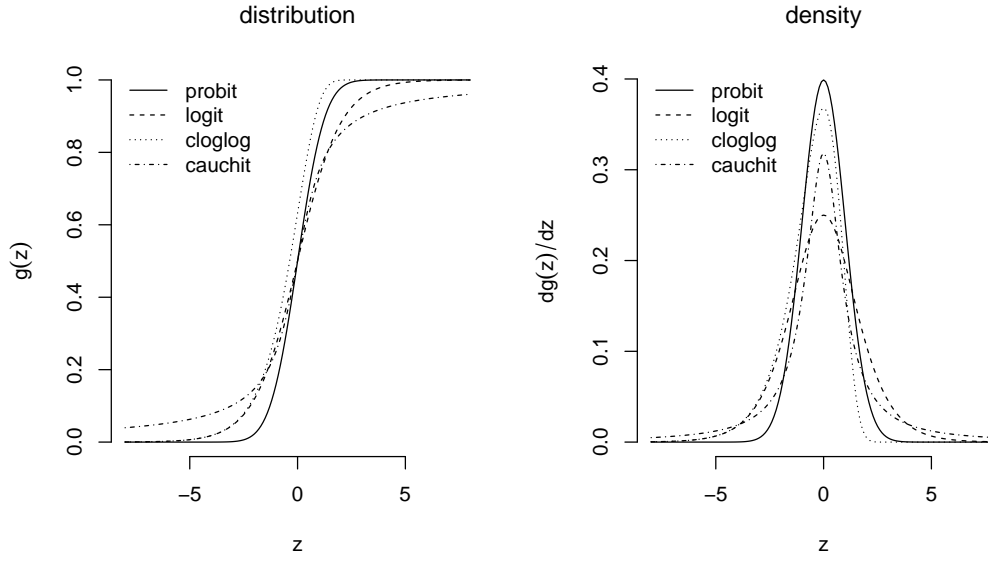


FIGURE 20.1: Distributions and densities corresponding to the link functions

Because y_i is a binary random variable, its probability completely determines its distribution. So we can also write the logit model in the following form:

Assumption 20.1 (binary logistic regression model) *We have*

$$y_i \mid x_i \sim \text{Bernoulli} \left(\frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}} \right).$$

The observations are independent across units. The β is the unknown parameter.

Each coefficient β_j measures the impact of x_{ij} on the log odds of the outcome:

$$\frac{\partial}{\partial x_{ij}} \text{logit}\{\text{pr}(y_i = 1 \mid x_i)\} = \beta_j.$$

Epidemiologists also call β_j the conditional log odds ratio³ because

$$\begin{aligned} \beta_j &= \text{logit}\{\text{pr}(y_i = 1 \mid \dots, x_{ij} + 1, \dots)\} - \text{logit}\{\text{pr}(y_i = 1 \mid \dots, x_{ij}, \dots)\} \\ &= \log \frac{\text{pr}(y_i = 1 \mid \dots, x_{ij} + 1)}{1 - \text{pr}(y_i = 1 \mid \dots, x_{ij} + 1)} - \log \frac{\text{pr}(y_i = 1 \mid \dots, x_{ij}, \dots)}{1 - \text{pr}(y_i = 1 \mid \dots, x_{ij}, \dots)} \\ &= \log \left\{ \frac{\text{pr}(y_i = 1 \mid \dots, x_{ij} + 1, \dots)}{1 - \text{pr}(y_i = 1 \mid \dots, x_{ij} + 1, \dots)} \bigg/ \frac{\text{pr}(y_i = 1 \mid \dots, x_{ij}, \dots)}{1 - \text{pr}(y_i = 1 \mid \dots, x_{ij}, \dots)} \right\}, \end{aligned}$$

that is, the change of the log odds of y_i if we increase x_{ij} by a unit holding other covariates unchanged. Qualitatively, if $\beta_j > 0$, then larger values of x_{ij} lead to larger probabilities of $y_i = 1$; if $\beta_j < 0$, then larger values of x_{ij} lead to smaller probabilities of $y_i = 1$.

³In probability theory, if p is the probability, then $p/(1 - p)$ is called the odds. It is a terminology from gambling.

20.2 Maximum likelihood estimator of the logistic model

Because we have specified a fully parametric model for y_i given x_i , we can estimate β using the maximum likelihood. With independent observations, the likelihood function for general binary outcomes is⁴

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n f(y_i | x_i) \\ &= \prod_{i=1}^n \{\pi(x_i, \beta) \text{ if } y_i = 1 \text{ or } 1 - \pi(x_i, \beta) \text{ if } y_i = 0\} \\ &= \prod_{i=1}^n \{\pi(x_i, \beta)\}^{y_i} \{1 - \pi(x_i, \beta)\}^{1-y_i}. \end{aligned}$$

Under the logit form (20.2), the likelihood function simplifies to

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n \left\{ \frac{\pi(x_i, \beta)}{1 - \pi(x_i, \beta)} \right\}^{y_i} \{1 - \pi(x_i, \beta)\} \\ &= \prod_{i=1}^n \left(e^{x_i^\top \beta} \right)^{y_i} \frac{1}{1 + e^{x_i^\top \beta}} \\ &= \prod_{i=1}^n \frac{e^{y_i x_i^\top \beta}}{1 + e^{x_i^\top \beta}}. \end{aligned}$$

The log-likelihood function is

$$\log L(\beta) = \sum_{i=1}^n \left\{ y_i x_i^\top \beta - \log(1 + e^{x_i^\top \beta}) \right\},$$

the score function is

$$\begin{aligned} \frac{\partial \log L(\beta)}{\partial \beta} &= \sum_{i=1}^n \left(x_i y_i - \frac{x_i e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}} \right) \\ &= \sum_{i=1}^n x_i \left(y_i - \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}} \right) \\ &= \sum_{i=1}^n x_i \{y_i - g(x_i^\top \beta)\} \\ &= \sum_{i=1}^n x_i \{y_i - \pi(x_i, \beta)\}, \end{aligned}$$

⁴The notation can be confusing because β denotes both the true parameter and the dummy variable for the likelihood function. It is somewhat standard in statistics, when we do not want to introduce additional notation. In previous chapters of this book, I use β to denote the true parameter in the linear model and b to denote the parameter in the OLS objective function.

and the Hessian matrix

$$\begin{aligned}
 \frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^T} &= \left(\frac{\partial^2 \log L(\beta)}{\partial \beta_j \partial \beta_{j'}} \right)_{1 \leq j, j' \leq p} \\
 &= - \sum_{i=1}^n x_i \frac{\partial g(x_i^T \beta)}{\partial \beta^T} \\
 &\stackrel{(20.1)}{=} - \sum_{i=1}^n x_i x_i^T g(x_i^T \beta) \{1 - g(x_i^T \beta)\} \\
 &= - \sum_{i=1}^n \pi(x_i, \beta) \{1 - \pi(x_i, \beta)\} x_i x_i^T.
 \end{aligned}$$

For any $\alpha \in \mathbb{R}^p$, we have

$$\alpha^T \frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^T} \alpha = - \sum_{i=1}^n \pi(x_i, \beta) \{1 - \pi(x_i, \beta)\} (\alpha^T x_i)^2 \leq 0$$

so the Hessian matrix is negative semi-definite. If it is negative definite, then the likelihood function has a unique maximizer.

The maximum likelihood estimate (MLE) must satisfy the following score or Normal equation:

$$\sum_{i=1}^n x_i \left\{ y_i - \pi(x_i, \hat{\beta}) \right\} = \sum_{i=1}^n x_i \left(y_i - \frac{e^{x_i^T \hat{\beta}}}{1 + e^{x_i^T \hat{\beta}}} \right) = 0.$$

If we view $\pi(x_i, \hat{\beta})$ as the fitted probability for y_i , then $y_i - \pi(x_i, \hat{\beta})$ is the residual, and the score equation is similar to that of OLS. Moreover, if x_i contains 1, then

$$\sum_{i=1}^n \left\{ y_i - \pi(x_i, \hat{\beta}) \right\} = 0,$$

which implies

$$n^{-1} \sum_{i=1}^n y_i = n^{-1} \sum_{i=1}^n \pi(x_i, \hat{\beta}).$$

That is, the average of the outcomes equals the average of their fitted values.

However, the score equation is nonlinear, and in general, there is no explicit formula for the MLE. We usually use Newton's method to solve for the MLE based on the linearization of the score equation. Starting from the old value β^{old} , we can approximate the score equation by a linear equation:

$$0 = \frac{\partial \log L(\beta)}{\partial \beta} \cong \frac{\partial \log L(\beta^{\text{old}})}{\partial \beta} + \frac{\partial^2 \log L(\beta^{\text{old}})}{\partial \beta \partial \beta^T} (\beta - \beta^{\text{old}}),$$

and then update

$$\beta^{\text{new}} = \beta^{\text{old}} - \left\{ \frac{\partial^2 \log L(\beta^{\text{old}})}{\partial \beta \partial \beta^T} \right\}^{-1} \frac{\partial \log L(\beta^{\text{old}})}{\partial \beta}.$$

Using the matrix form, we can gain more insight from Newton's method. Recall that

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix},$$

and define

$$\Pi^{\text{old}} = \begin{pmatrix} \pi(x_1, \beta^{\text{old}}) \\ \vdots \\ \pi(x_n, \beta^{\text{old}}) \end{pmatrix}, \quad W^{\text{old}} = \text{diag} [\pi(x_i, \beta^{\text{old}}) \{1 - \pi(x_i, \beta^{\text{old}})\}]_{i=1}^n.$$

Then

$$\begin{aligned} \frac{\partial \log L(\beta^{\text{old}})}{\partial \beta} &= X^T(Y - \Pi^{\text{old}}), \\ \frac{\partial^2 \log L(\beta^{\text{old}})}{\partial \beta \partial \beta^T} &= -X^T W^{\text{old}} X, \end{aligned}$$

and Newton's method simplifies to

$$\begin{aligned} \beta^{\text{new}} &= \beta^{\text{old}} + (X^T W^{\text{old}} X)^{-1} X^T (Y - \Pi^{\text{old}}) \\ &= (X^T W^{\text{old}} X)^{-1} \{X^T W^{\text{old}} X \beta^{\text{old}} + X^T (Y - \Pi^{\text{old}})\} \\ &= (X^T W^{\text{old}} X)^{-1} X^T W^{\text{old}} Z^{\text{old}}, \end{aligned}$$

where

$$Z^{\text{old}} = X \beta^{\text{old}} + (W^{\text{old}})^{-1} (Y - \Pi^{\text{old}}).$$

So we can obtain β^{new} based on the WLS fit of Z^{old} on X with weights W^{old} , the diagonal elements of which are the conditional variances of the y_i 's given the x_i 's at β^{old} . The `glm` function in `R` uses the Fisher scoring algorithm, which is identical to Newton's method for the logit model.⁵ Sometimes, it is also called the iteratively reweighted least squares algorithm.

20.3 Statistics with the logit model

20.3.1 Inference

Based on the general theory of MLE, $\hat{\beta}$ is consistent for β and is asymptotically Normal. Approximately, we can conduct statistical inference based on

$$\hat{\beta} \stackrel{\text{a}}{\sim} N \left\{ \beta, \left(-\frac{\partial^2 \log L(\hat{\beta})}{\partial \beta \partial \beta^T} \right)^{-1} \right\} = N \left\{ \beta, (X^T \hat{W} X)^{-1} \right\},$$

where

$$\hat{W} = \text{diag} [\pi(x_i, \hat{\beta}) \{1 - \pi(x_i, \hat{\beta})\}]_{i=1}^n.$$

Based on this, the `glm` function reports the point estimate, standard error, z -value, and p -value for each coordinate of β . It is almost identical to the output of the `lm` function, except that the interpretation of the coefficient becomes the conditional log odds ratio.

⁵The Fisher scoring algorithm uses a slightly different approximation:

$$0 = \frac{\partial \log L(\beta)}{\partial \beta} \cong \frac{\partial \log L(\beta^{\text{old}})}{\partial \beta} + E \left\{ \frac{\partial^2 \log L(\beta^{\text{old}})}{\partial \beta \partial \beta^T} \mid X \right\} (\beta - \beta^{\text{old}}),$$

with the expected Fisher information instead of the observed Fisher information. For other link functions, the Fisher scoring algorithm is different from Newton's method.

I use the data from Hirano et al. (2000) to illustrate logistic regression, where the main interest is the effect of the encouragement of receiving the flu shot via email on the binary indicator of flu-related hospitalization. We can fit a logistic regression using the `glm` function in R with `family = binomial(link = logit)`.

```
> flu = read.table("fludata.txt", header = TRUE)
> flu = within(flu, rm(receive))
> assign.logit = glm(outcome ~ .,
+                    family = binomial(link = logit),
+                    data = flu)
> summary(assign.logit)

Call:
glm(formula = outcome ~ ., family = binomial(link = logit), data = flu)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1957  -0.4566  -0.3821  -0.3048   2.6450

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.199815   0.408684  -5.383 7.34e-08 ***
assign       -0.197528   0.136235  -1.450 0.14709
age          -0.007986   0.005569  -1.434 0.15154
copd         0.337037   0.153939   2.189 0.02857 *
dm           0.454342   0.143593   3.164 0.00156 **
heartd       0.676190   0.153384   4.408 1.04e-05 ***
race        -0.242949   0.143013  -1.699 0.08936 .
renal        1.519505   0.365973   4.152 3.30e-05 ***
sex          -0.212095   0.144477  -1.468 0.14210
liverd       0.098957   1.084644   0.091 0.92731

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1667.9  on 2860  degrees of freedom
Residual deviance: 1598.4  on 2851  degrees of freedom
AIC: 1618.4

Number of Fisher Scoring iterations: 5
```

Three subtle issues arise in the above code. First, `flu = within(flu, rm(receive))` drops `receive`, which is the indicator of whether a patient received the flu shot or not. The reason is that `assign` is randomly assigned but `receive` is subject to selection bias, that is, patients receiving the flu shot can be quite different from patients not receiving the flu shot.

Second, the `Null deviance` and `Residual deviance` are defined as $-2\log L(\tilde{\beta})$ and $-2\log L(\hat{\beta})$, respectively, where $\tilde{\beta}$ is the MLE assuming that all coefficients except the intercept are zero, and $\hat{\beta}$ is the MLE without any restrictions. They are not of independent interest, but their difference is: Wilks' theorem ensures that

$$\{-2\log L(\tilde{\beta})\} - \{-2\log L(\hat{\beta})\} = 2\log \frac{L(\hat{\beta})}{L(\tilde{\beta})} \stackrel{a}{\sim} \chi_{p-1}^2.$$

So we can test whether the coefficients of the covariates are all zero, which is analogous to the joint F test in linear models.

```
> pchisq(assign.logit$null.deviance - assign.logit$deviance,
+        df = assign.logit$df.null - assign.logit$df.residual,
+        lower.tail = FALSE)
[1] 1.912952e-11
```

Third, the AIC is defined as $-2 \log L(\hat{\beta}) + 2p$, where p is the number of parameters in the logit model. This is also the general formula of AIC for other parametric models; recall its form under Normal linear model in Chapter 13.4.2.

20.3.2 Prediction

The logit model is often used for prediction or classification since the outcome is binary. With the MLE $\hat{\beta}$, we can predict the probability of being one as $\hat{\pi}_{n+1} = g(x_{n+1}^T \hat{\beta})$ for a unit with covariate value x_{n+1} , and we can easily dichotomize the fitted probability to predict the outcome itself by $\hat{y}_{n+1} = 1(\hat{\pi}_{n+1} \geq c)$, for example, with $c = 0.5$.

We can even quantify the uncertainty in the fitted probability based on a linear approximation (i.e., the delta method in Proposition C.11). Based on

$$\begin{aligned}\hat{\pi}_{n+1} &= g(x_{n+1}^T \hat{\beta}) \\ &\cong g(x_{n+1}^T \beta) + g'(x_{n+1}^T \beta) x_{n+1}^T (\hat{\beta} - \beta) \\ &= g(x_{n+1}^T \beta) + g(x_{n+1}^T \beta) \{1 - g(x_{n+1}^T \beta)\} x_{n+1}^T (\hat{\beta} - \beta),\end{aligned}$$

we can approximate the asymptotic variance of $\hat{\pi}_{n+1}$ by

$$[g(x_{n+1}^T \beta) \{1 - g(x_{n+1}^T \beta)\}]^2 x_{n+1}^T (X^T \hat{W} X)^{-1} x_{n+1}.$$

We can use the `predict` function in `R` to calculate the predicted values based on a `glm` object in the same way as the linear model. If we specify `type="response"`, then we obtain the fitted probabilities; if we specify `se.fit = TRUE`, then we also obtain the standard errors of the fitted probabilities. In the following, I predict the probabilities of flu-related hospitalization if a patient receives the email encouragement or not, fixing other covariates at their empirical means.

```
> emp.mean = apply(flu, 2, mean)
> data.ave = rbind(emp.mean, emp.mean)
> data.ave[1, 1] = 1
> data.ave[2, 1] = 0
> data.ave = data.frame(data.ave)
> data.ave
```

	assign	outcome	age	copd	dm	heartd	race
emp.mean	1	0.08528487	65.26949	0.2820692	0.2785739	0.5735757	0.6550157
emp.mean.1	0	0.08528487	65.26949	0.2820692	0.2785739	0.5735757	0.6550157

```

      renal      sex      liverd
emp.mean  0.01328207 0.6682978 0.003145753
emp.mean.1 0.01328207 0.6682978 0.003145753
> predict(assign.logit, newdata = data.ave,
+         type = "response", se.fit = TRUE)
$fit
      emp.mean emp.mean.1
0.06981828 0.08378818

$se.fit
      emp.mean emp.mean.1
0.006689665 0.007526307

$residual.scale
[1] 1
```

20.4 More on the interpretations of the coefficients

Many practitioners find the coefficients in the logit model difficult to interpret. Another measure of the impact of the covariate on the outcome is the *average marginal effect or average partial effect*.⁶ For a continuous covariate x_{ij} , the average marginal effect is defined as

$$\begin{aligned}\text{AME}_j &= n^{-1} \sum_{i=1}^n \frac{\partial \text{pr}(y_i = 1 \mid x_i)}{\partial x_{ij}} \\ &= n^{-1} \sum_{i=1}^n g'(x_i^\top \beta) \beta_j,\end{aligned}$$

which reduces to the following form for the logit model

$$\text{AME}_j = \beta_j \times n^{-1} \sum_{i=1}^n \pi(x_i, \beta) \{1 - \pi(x_i, \beta)\}.$$

For a binary covariate x_{ij} , the average marginal effect is defined as

$$\text{AME}_j = n^{-1} \sum_{i=1}^n \{\text{pr}(y_i = 1 \mid \dots, x_{ij} = 1, \dots) - \text{pr}(y_i = 1 \mid \dots, x_{ij} = 0, \dots)\}.$$

The `margins` function in the `margins` package can compute the average marginal effects and the corresponding standard errors. In particular, the average marginal effect of `assign` is not significant as shown below.

```
> library("margins")
> ape = margins(assign.logit)
> summary(ape)
      factor      AME      SE      z      p    lower    upper
age      -0.0006 0.0004 -1.4322 0.1521 -0.0014 0.0002
assign   -0.0150 0.0103 -1.4480 0.1476 -0.0352 0.0053
copd      0.0255 0.0117  2.1830 0.0290  0.0026 0.0485
dm        0.0344 0.0109  3.1465 0.0017  0.0130 0.0559
heartd    0.0512 0.0118  4.3441 0.0000  0.0281 0.0743
liverd    0.0075 0.0822  0.0912 0.9273 -0.1536 0.1686
race     -0.0184 0.0109 -1.6958 0.0899 -0.0397 0.0029
renal     0.1151 0.0278  4.1461 0.0000  0.0607 0.1696
sex      -0.0161 0.0110 -1.4660 0.1426 -0.0376 0.0054
```

The interaction term is much more complicated. Contradictory suggestions are given across fields. Consider the following model

$$\text{pr}(y_i = 1 \mid x_{i1}, x_{i2}) = g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}).$$

If the link is logit, then epidemiologists interpret $e^{\beta_{12}}$ as the interaction between x_{i1} and x_{i2} on the odds ratio scale. Consider a simple case with binary x_{i1} and x_{i2} . Given $x_{i2} = 1$, the odds ratio of x_{i1} on y_i equals $e^{\beta_1 + \beta_{12}}$; given $x_{i2} = 0$, the odds ratio of x_{i1} on y_i equals e^{β_1} . Therefore, the ratio of the two odds ratio equals $e^{\beta_{12}}$. When we measure effects on the odds ratio scale, the logistic model is a natural choice. The interaction term in the logistic model indeed measures the interaction of x_{i1} and x_{i2} .

⁶Recall its definition in Chapter 17.3.2.

Ai and Norton (2003) gave a different suggestion. We have two ways to define the interaction effect: first,

$$n^{-1} \sum_{i=1}^n \frac{\partial \text{pr}(y_i = 1 \mid x_{i1}, x_{i2})}{\partial (x_{i1} x_{i2})} = n^{-1} \sum_{i=1}^n g'(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}) \beta_{12};$$

second,

$$\begin{aligned} & n^{-1} \sum_{i=1}^n \frac{\partial^2 \text{pr}(y_i = 1 \mid x_{i1}, x_{i2})}{\partial x_{i1} \partial x_{i2}} \\ &= n^{-1} \sum_{i=1}^n \frac{\partial}{\partial x_{i2}} \left\{ \frac{\partial \text{pr}(y_i = 1 \mid x_{i1}, x_{i2})}{\partial x_{i1}} \right\} \\ &= n^{-1} \sum_{i=1}^n \frac{\partial}{\partial x_{i2}} \{g'(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2})(\beta_1 + \beta_{12} x_{i2})\} \\ &= n^{-1} \sum_{i=1}^n \{g''(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2})(\beta_2 + \beta_{12} x_{i1})(\beta_1 + \beta_{12} x_{i2}) + g'(z_i) \beta_{12}\}. \end{aligned}$$

Although the first one is more straightforward based on the definition of the average partial effect, the second one is more reasonable based on the natural definition of interaction based on the mixed derivative. Note that even if $\beta_{12} = 0$, the second definition of interaction does not necessarily equal 0 since

$$n^{-1} \sum_{i=1}^n \frac{\partial^2 \text{pr}(y_i = 1 \mid x_{i1}, x_{i2})}{\partial x_{i1} \partial x_{i2}} = n^{-1} \sum_{i=1}^n g''(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) \beta_1 \beta_2.$$

This is due to the nonlinearity of the link function. The second definition quantifies interaction based on the probability itself, whereas the parameter β_{12} in the logistic model measures the interaction on the odds ratio scale.

20.5 Does the link function matter?

First, I generate data from a simple one-dimensional logistic model.

```
> n = 100
> x = rnorm(n, 0, 3)
> prob = 1/(1 + exp(-1 + x))
> y = rbinom(n, 1, prob)
```

Then I fit the data with the linear probability model and binary models with four link functions.

```
> lpmfit = lm(y ~ x)
> probitfit = glm(y ~ x, family = binomial(link = "probit"))
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> logitfit = glm(y ~ x, family = binomial(link = "logit"))
> cloglogfit = glm(y ~ x, family = binomial(link = "cloglog"))
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> cauchitfit = glm(y ~ x, family = binomial(link = "cauchit"))
```

The coefficients are quite different because the coefficients measure the association between x and y on difference scales. These parameters are not directly comparable. Nevertheless, the signs of the coefficients are all negative.

```
> betacoeff = c(lpmfit$coef[2],
+               probitfit$coef[2],
+               logitfit$coef[2],
+               cloglogfit$coef[2],
+               cauchitfit$coef[2])
> names(betacoeff) = c("lpm", "probit", "logit", "cloglog", "cauchit")
> round(betacoeff, 2)
      lpm  probit  logit cloglog cauchit
-0.10  -0.83  -1.47  -1.07  -2.09
```

However, if we care only about the prediction, then these five models give very similar results.

```
> table(y, lpmfit$fitted.values>0.5)

y    FALSE  TRUE
0      31     9
1       5    55
> table(y, probitfit$fitted.values>0.5)

y    FALSE  TRUE
0      31     9
1       5    55
> table(y, logitfit$fitted.values>0.5)

y    FALSE  TRUE
0      31     9
1       5    55
> table(y, cloglogfit$fitted.values>0.5)

y    FALSE  TRUE
0      31     9
1       5    55
> table(y, cauchitfit$fitted.values>0.5)

y    FALSE  TRUE
0      34     6
1       7    53
> table(y, cauchitfit$fitted.values>0.5)

y    FALSE  TRUE
0      34     6
1       7    53
```

Figure 20.2 shows the fitted probabilities versus the true probabilities $\text{pr}(y_i = 1 | x_i)$. The patterns are quite similar although the linear probability model can give fitted probabilities outside $[0, 1]$. When we use the cutoff point 0.5 to predict the binary outcome, the problem of the linear probability model becomes rather minor.

An interesting fact is that the coefficients from the logit model approximately equal those from the probit model multiplied by 1.7, a constant that minimizes $\max_y |g_{\text{logit}}(by) - g_{\text{probit}}(y)|$. We can easily compute this constant numerically:

```
> d.logit.probit = function(b){
+   x = seq(-20, 20, 0.00001)
+   max(abs(plogis(b*x) - pnorm(x)))
+ }
> optimize(d.logit.probit, c(-10, 10))
$minimum
[1] 1.701743

$objective
[1] 0.009457425

>
```

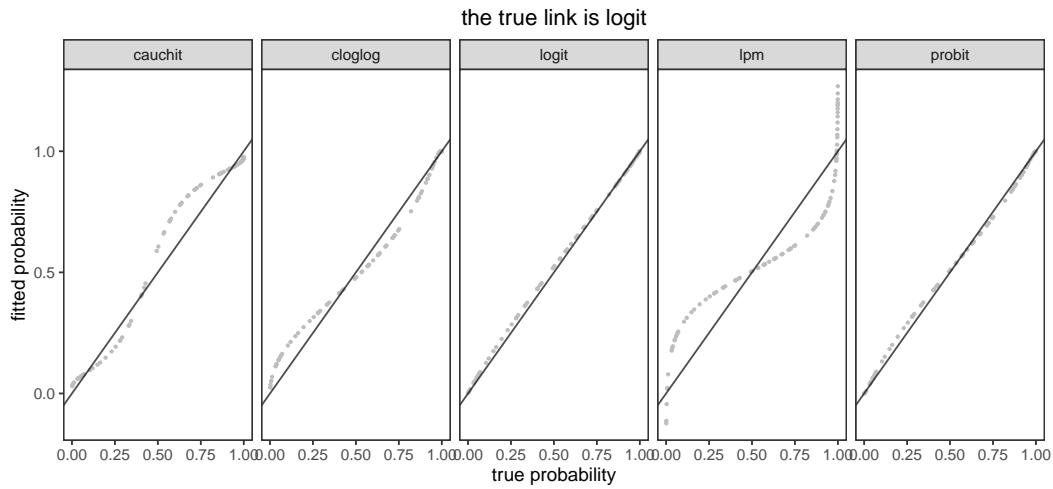


FIGURE 20.2: Comparing the fitted probabilities from different link functions

```
> d.logit.cauchit = function(b){
+   x = seq(-20, 20, 0.00001)
+   max(abs(plogis(b*x) - pcauchy(x)))
+ }
> optimize(d.logit.cauchit, c(-10, 10))
$minimum
[1] 0.8590545

$objective
[1] 0.04945328

>
> f.cloglog = function(z){
+   1 - exp(-exp(z))
+ }
> d.logit.cloglog = function(b){
+   x = seq(-20, 20, 0.00001)
+   max(abs(plogis(b*x) - f.cloglog(x)))
+ }
> optimize(d.logit.cloglog, c(-10, 10))
$minimum
[1] 1.47175

$objective
[1] 0.1321229
```

Based on the above calculation, the maximum difference is approximately 0.009. Therefore, the logit and probit link functions are extremely close up to the scaling factor 1.7.⁷ However, $\min_b \max_y |g_{\text{logit}}(by) - g_*(y)|$ is much larger for the link functions of cauchit and cloglog.

⁷See Problem 20.1 for a related heuristic argument.

20.6 Extensions of the logistic regression

20.6.1 Penalized logistic regression

Similar to the high dimensional linear model, we can also extend the logit model to a penalized version. Since the objective function for the original logit model is the log-likelihood, we can minimize the following penalized log-likelihood function:

$$\arg \min_{\beta_0, \beta_1, \dots, \beta_p} -\frac{1}{n} \sum_{i=1}^n \ell_i(\beta) + \lambda \sum_{j=1}^p \{\alpha \beta_j^2 + (1 - \alpha) |\beta_j|\},$$

where

$$\ell_i(\beta) = y_i(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) - \log(1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}})$$

is the log-likelihood function based on the i th observation. When $\alpha = 1$, it gives the ridge analog of the logistic regression; when $\alpha = 0$, it gives the lasso analog; when $\alpha \in (0, 1)$, it gives the elastic net analog. The `R` package `glmnet` uses the coordinate descent algorithm based on a quadratic approximation of the log-likelihood function. We can select the tuning parameter λ based on cross-validation.

20.6.2 Case-control study

A nice property of the logit model is that it works not only for the cohort study with data from conditional distribution $y_i | x_i$ but also for the case-control study with data from the conditional distribution $x_i | y_i$.⁸ The former is a prospective study while the latter is a retrospective study. Below, I will explain the basic idea in Prentice and Pyke (1979).

Assume that $(x_i, y_i, s_i)_{i=1}^n$ are IID with

$$\text{pr}(y_i = 1 | x_i) = \frac{e^{\beta_0 + x_i^T \beta}}{1 + e^{\beta_0 + x_i^T \beta}} \quad (20.3)$$

and

$$\text{pr}(s_i = 1 | x_i, y_i) = \text{pr}(s_i = 1 | y_i) = \begin{cases} p_1, & \text{if } y_i = 1, \\ p_0, & \text{if } y_i = 0. \end{cases} \quad (20.4)$$

But we only have data with $s_i = 1$ with p_1 and p_0 often unknown. Fortunately, conditioning on $s_i = 1$, we have the following result.

Theorem 20.1 Under (20.3) and (20.4), we have

$$\text{pr}(y_i = 1 | x_i, s_i = 1) = \frac{e^{\delta + \beta_0 + x_i^T \beta}}{1 + e^{\delta + \beta_0 + x_i^T \beta}},$$

where $\delta = \log(p_1/p_0)$.

Proof of Theorem 20.1: We have

$$\begin{aligned} & \text{pr}(y_i = 1 | x_i, s_i = 1) \\ &= \frac{\text{pr}(y_i = 1 | x_i) \text{pr}(s_i = 1 | x_i, y_i = 1)}{\text{pr}(y_i = 1 | x_i) \text{pr}(s_i = 1 | x_i, y_i = 1) + \text{pr}(y_i = 0 | x_i) \text{pr}(s_i = 1 | x_i, y_i = 0)} \end{aligned}$$

⁸Breslow (1996) provided a scholarly review of the statistics of the case-control study in epidemiology.

by Bayes' formula. Under the logit model, we have

$$\begin{aligned}
 \text{pr}(y_i = 1 \mid x_i, s_i = 1) &= \frac{\frac{e^{\beta_0 + x_i^T \beta}}{1 + e^{\beta_0 + x_i^T \beta}} p_1}{\frac{e^{\beta_0 + x_i^T \beta}}{1 + e^{\beta_0 + x_i^T \beta}} p_1 + \frac{1}{1 + e^{\beta_0 + x_i^T \beta}} p_0} \\
 &= \frac{e^{\beta_0 + x_i^T \beta} p_1}{e^{\beta_0 + x_i^T \beta} p_1 + p_0} \\
 &= \frac{e^{\beta_0 + x_i^T \beta} p_1 / p_0}{e^{\beta_0 + x_i^T \beta} p_1 / p_0 + 1} \\
 &= \frac{e^{\delta + \beta_0 + x_i^T \beta}}{1 + e^{\delta + \beta_0 + x_i^T \beta}}.
 \end{aligned}$$

□

Theorem 20.1 ensures that conditioning on $s_i = 1$, the model of y_i given x_i is still logit with the intercept changing from β_0 to $\beta_0 + \log(p_1/p_0)$. Although we cannot consistently estimate the intercept without knowing (p_1, p_0) , we can still estimate all the slopes consistently. Kagan (2001) showed that the logistic link is the only one that enjoys this property.

I will end this subsection with a case study. Samarani et al. (2019) hypothesized that variation in the inherited activating Killer-cell Immunoglobulin-like Receptor genes in humans is associated with their innate susceptibility/resistance to developing Crohn's disease. They used a case-control study from three cities (Manitoba, Montreal, and Ottawa) in Canada to investigate the potential association. The following logistic regression uses all the data. Problem 20.8 requires separate analyses based on `center`.

```

> dat = read.csv("samarani.csv")
> pool.glm = glm(case_comb ~ ds1 + ds2 + ds3 + ds4_a +
+               ds4_b + ds5 + ds1_3 + center,
+               family = binomial(link = logit),
+               data = dat)
> summary(pool.glm)

Call:
glm(formula = case_comb ~ ds1 + ds2 + ds3 + ds4_a + ds4_b + ds5 +
     ds1_3 + center, family = binomial(link = logit), data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9982  -0.9274  -0.5291   1.0113   2.2289

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.39681    0.21768  -11.011  < 2e-16 ***
ds1             0.55945    0.14437   3.875  0.000107 ***
ds2             0.42531    0.14758   2.882  0.003954 **
ds3             0.81377    0.14503   5.611  2.01e-08 ***
ds4_a          0.30270    0.30679   0.987  0.323802
ds4_b          0.29199    0.17726   1.647  0.099511 .
ds5            0.92049    0.14852   6.198  5.72e-10 ***
ds1_3          0.49982    0.14706   3.399  0.000677 ***
centerMontreal -0.05816    0.15889  -0.366  0.714316
centerOttawa   0.14164    0.20251   0.699  0.484292

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1403.7  on 1020  degrees of freedom
Residual deviance: 1192.0  on 1011  degrees of freedom

```

AIC: 1212

Number of Fisher Scoring iterations: 3

20.7 Other model formulations

20.7.1 Latent linear model

Let $y_i = 1(y_i^* \geq 0)$ where

$$y_i^* = x_i^T \beta + \varepsilon_i$$

and $-\varepsilon_i$ has distribution function $g(\cdot)$ and is independent of x_i . From this latent linear model, we can verify that

$$\begin{aligned} \text{pr}(y_i = 1 \mid x_i) &= \text{pr}(y_i^* \geq 0 \mid x_i) \\ &= \text{pr}(x_i^T \beta + \varepsilon_i \geq 0 \mid x_i) \\ &= \text{pr}(-\varepsilon_i \leq x_i^T \beta \mid x_i) \\ &= g(x_i^T \beta). \end{aligned}$$

So the $g(\cdot)$ function can be interpreted as the distribution function of the error term in the latent linear model.

This latent variable formulation provides another way to interpret the coefficients in the models for binary data. It is a powerful way to generate models for more complex data. We will see another example in the next chapter.

20.7.2 Inverse model

Assume that

$$y_i \sim \text{Bernoulli}(q), \quad (20.5)$$

and

$$x_i \mid y_i = 1 \sim N(\mu_1, \Sigma), \quad x_i \mid y_i = 0 \sim N(\mu_0, \Sigma), \quad (20.6)$$

where x_i does not contain 1. This is called the linear discriminant model. We can verify that $y_i \mid x_i$ follows a logit model as shown in Theorem 20.2 below (Cornfield et al., 1961; Cornfield, 1962).

Theorem 20.2 *Under (20.5) and (20.6), we have*

$$\text{logit}\{\text{pr}(y_i = 1 \mid x_i)\} = \alpha + x_i^T \beta,$$

where

$$\alpha = \log \frac{q}{1-q} - \frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0), \quad (20.7)$$

$$\beta = \Sigma^{-1} (\mu_1 - \mu_0). \quad (20.8)$$

Proof of Theorem 20.2: Using Bayes' formula, we have

$$\begin{aligned}\text{pr}(y_i = 1 \mid x_i) &= \frac{\text{pr}(y_i = 1, x_i)}{\text{pr}(x_i)} \\ &= \frac{\text{pr}(y_i = 1)\text{pr}(x_i \mid y_i = 1)}{\text{pr}(y_i = 1)\text{pr}(x_i \mid y_i = 1) + \text{pr}(y_i = 0)\text{pr}(x_i \mid y_i = 0)} \\ &= \frac{e^\Delta}{1 + e^\Delta},\end{aligned}$$

where

$$\begin{aligned}\Delta &= \log \frac{\text{pr}(y_i = 1)\text{pr}(x_i \mid y_i = 1)}{\text{pr}(y_i = 0)\text{pr}(x_i \mid y_i = 0)} \\ &= \log \frac{q \{(2\pi)^p \det(\Sigma)\}^{-1/2} \exp \{-(x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1)/2\}}{(1-q) \{(2\pi)^p \det(\Sigma)\}^{-1/2} \exp \{-(x_i - \mu_0)^T \Sigma^{-1} (x_i - \mu_0)/2\}} \\ &= \log \frac{q \exp \{-(x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1)/2\}}{(1-q) \exp \{-(x_i - \mu_0)^T \Sigma^{-1} (x_i - \mu_0)/2\}} \\ &= \log \frac{q \exp \{-(2x_i^T \Sigma^{-1} \mu_1 + \mu_1^T \Sigma^{-1} \mu_1)/2\}}{(1-q) \exp \{-(2x_i^T \Sigma^{-1} \mu_0 + \mu_0^T \Sigma^{-1} \mu_0)/2\}} \\ &= \log \frac{q}{1-q} - \frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0) + x_i^T \Sigma^{-1} (\mu_1 - \mu_0).\end{aligned}$$

So $y_i \mid x_i$ follows a logistic model with α and β given in (21.12). \square

We can easily obtain the moment estimators for the unknown parameters under (20.5) and (20.6). Let $n_1 = \sum_{i=1}^n y_i$ and $n_0 = n - n_1$. The moment estimator for q is $\hat{q} = n_1/n$, the sample mean of the y_i 's. The moment estimators for μ_1 and μ_0 are

$$\begin{aligned}\hat{\mu}_1 &= n_1^{-1} \sum_{i=1}^n y_i x_i, \\ \hat{\mu}_0 &= n_0^{-1} \sum_{i=1}^n (1 - y_i) x_i,\end{aligned}$$

the sample means of the x_i 's for units with $y_i = 1$ and $y_i = 0$, respectively. The moment estimator for Σ is

$$\hat{\Sigma} = \left[\sum_{i=1}^n y_i (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^T + \sum_{i=1}^n (1 - y_i) (x_i - \hat{\mu}_0)(x_i - \hat{\mu}_0)^T \right] / (n - 2),$$

the pooled covariance matrix, after centering the x_i 's by the y -specific means. Based on Theorem 20.2, we can obtain estimates $\hat{\alpha}$ and $\hat{\beta}$ by replacing the true parameters with their moment estimators. This gives us another way to fit the logistic model, which does not require iteration.

Efron (1975) compared the above moment estimator and the MLE derived from the logistic model. Since the linear discriminant model imposes stronger assumptions, the estimator based on Theorem 20.2 is more efficient if the model is correct. In contrast, the MLE derived from the logistic model is more robust because it does not impose the Normality assumption on x_i .

20.8 Homework problems

20.1 Moments of the logistic distribution

Theorem 20.3 states the first two moments of the logistic distribution. Prove Theorem 20.3.

Theorem 20.3 Assume ε is the logistic distribution with CDF

$$\text{pr}(\varepsilon \leq z) = \frac{e^z}{1 + e^z}.$$

Then ε has mean $E(\varepsilon) = 0$ and variance

$$\text{var}(\varepsilon) = \frac{\pi^2}{3}.$$

Remark: Theorem 20.3 appeared in deCanis and Stine (1986). The variance calculation involves non-trivial integrals. You may use Euler's formula $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. For most readers, you can ignore this problem unless you want to review calculus.

An interesting implication of the variance is that with the same data, the logistic regression coefficients are about $\sqrt{\pi^2/3} \approx 1.8$ times the probit regression coefficients. Based on the latent linear model representations of the logistic and probit models, the coefficients are only comparable when the corresponding ε_i has the same variance. Of course, the logistic distribution and standard Normal distribution differ in many other ways. The calculation based on the square root of the variance ratio is only an approximation. Based on another heuristic argument, Chapter 20.5 derived 1.7 as the approximated ratio.

20.2 Invariance of logistic regression

This problem extends Problem 3.4.

Assume $\tilde{x}_i^T = x_i^T \Gamma$ with an invertible Γ . Run logistic regression of y_i 's on x_i 's to obtain the coefficient $\hat{\beta}$ and fitted probabilities $\hat{\pi}_i$'s. Run another logistic regression of y_i 's on \tilde{x}_i 's to obtain the coefficient $\tilde{\beta}$ and fitted probabilities $\tilde{\pi}_i$'s.

Prove that $\tilde{\beta} = \Gamma \hat{\beta}$ and $\tilde{\pi}_i = \hat{\pi}_i$ for all i 's.

20.3 Two logistic regressions

This is an extension of Problem 17.2.

Given data $(x_i, z_i, y_i)_{i=1}^n$ where x_i denotes the covariates, $z_i \in \{1, 0\}$ denotes the binary group indicator, and y_i denotes the binary outcome. We can fit two separate logistic regressions:

$$\text{logit}\{\text{pr}(y_i = 1 \mid z_i = 1, x_i)\} = \gamma_1 + x_i^T \beta_1$$

and

$$\text{logit}\{\text{pr}(y_i = 1 \mid z_i = 0, x_i)\} = \gamma_0 + x_i^T \beta_0$$

with the treated and control data, respectively. We can also fit a joint logistic regression using the pooled data:

$$\text{logit}\{\text{pr}(y_i = 1 \mid z_i, x_i)\} = \alpha_0 + \alpha_z z_i + x_i^T \alpha_x + z_i x_i^T \alpha_{zx}.$$

Let the parameters with hats denote the MLEs, for example, $\hat{\gamma}_1$ is the MLE for γ_1 . Find $(\hat{\alpha}_0, \hat{\alpha}_z, \hat{\alpha}_x, \hat{\alpha}_{zx})$ in terms of $(\hat{\gamma}_1, \hat{\beta}_1, \hat{\gamma}_0, \hat{\beta}_0)$.

20.4 Likelihood for probit model

Write down the likelihood function for the probit model, and derive the steps for Newton's method and Fisher scoring for computing the MLE. How do we estimate the asymptotic covariance matrix of the MLE?

20.5 Logit and general exponential family

Efron (1975) pointed out an extension of Theorem 20.2. Prove Theorem 20.4 below and find the formulas of α and β in terms of $(\theta_1, \theta_0, \eta)$.

Theorem 20.4 Under (20.5) and

$$f(x_i | y_i = y) = g(\theta_y, \eta) h(x_i, \eta) \exp(x_i^T \theta_y), \quad (y = 0, 1)$$

with parameters $(\theta_1, \theta_0, \eta)$, we have

$$\text{logit}\{\text{pr}(y_i = 1 | x_i)\} = \alpha + x_i^T \beta$$

for some α and β .

Remark: As a sanity check, you can compare this problem with Theorem 20.2.

20.6 Empirical comparison of logistic regression and linear discriminant analysis

This problem is related to Chapter 20.7.2.

Compare the performance of logistic regression and linear discriminant analysis in terms of prediction accuracy. You should simulate at least three cases:

- (C1) the model for linear discriminant analysis is correct;
- (C2) the model for linear discriminant analysis is incorrect but the model for logistic regression is correct;
- (C3) the model for logistic regression is incorrect.

20.7 Quadratic discriminant analysis

Theorem 20.5 below extends Theorem 20.2. Prove Theorem 20.5.

Theorem 20.5 Assume that

$$y_i \sim \text{Bernoulli}(q),$$

and

$$x_i | y_i = 1 \sim N(\mu_1, \Sigma_1), \quad x_i | y_i = 0 \sim N(\mu_0, \Sigma_0),$$

where $x_i \in \mathbb{R}^p$ does not contain 1.

We have that

$$\text{logit}\{\text{pr}(y_i = 1 | x_i)\} = \alpha + x_i^T \beta + x_i^T \Lambda x_i,$$

where

$$\begin{aligned} \alpha &= \log \frac{q}{1-q} - \frac{1}{2} \log \frac{\det(\Sigma_1)}{\det(\Sigma_0)} - \frac{1}{2} (\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_0^T \Sigma_0^{-1} \mu_0), \\ \beta &= \Sigma_1^{-1} \mu_1 - \Sigma_0^{-1} \mu_0, \\ \Lambda &= -\frac{1}{2} (\Sigma_1^{-1} - \Sigma_0^{-1}). \end{aligned}$$

Remark: Theorem 20.5 extends the linear discriminant model in Section 20.7.2 to the quadratic discriminant model by allowing for heteroskedasticity in the conditional Normality of x given y (Cornfield, 1962). It implies the logistic model with the linear, quadratic, and interaction terms of the basic covariates.

20.8 Data analysis

Reanalyze the data in Chapter 20.6.2, stratifying the analysis based on `center`. Do the results vary significantly across centers?

20.9 R^2 in logistic regression

Recall that R^2 in the linear model measures the linear dependence of the outcome on the covariates. However, the definition of R^2 is not obvious in the logistic model. The `glm` function in `R` does not return any R^2 for the logistic regression.

Recall the following equivalent definitions of R^2 in the linear model

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\rho}_{y\hat{y}}^2 = \frac{(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}))^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}. \end{aligned}$$

The fitted values are $\hat{\pi}_i = \pi(x_i, \hat{\beta})$ in the logistic model, which have mean \bar{y} with the intercept included in the model. Analogously, we can define R^2 in the logistic model as

$$\begin{aligned} R_{\text{model}}^2 &= \frac{\text{SS}_M}{\text{SS}_T}, \\ R_{\text{residual}}^2 &= 1 - \frac{\text{SS}_R}{\text{SS}_T}, \\ R_{\text{correlation}}^2 &= \hat{\rho}_{y\hat{\pi}}^2 = \frac{C_{y\hat{\pi}}^2}{\text{SS}_T \text{SS}_M}, \end{aligned}$$

where

$$\begin{aligned} \text{SS}_T &= \sum_{i=1}^n (y_i - \bar{y})^2, \\ \text{SS}_M &= \sum_{i=1}^n (\hat{\pi}_i - \bar{y})^2, \\ \text{SS}_R &= \sum_{i=1}^n (y_i - \hat{\pi}_i)^2, \\ C_{y\hat{\pi}} &= \sum_{i=1}^n (y_i - \bar{y})(\hat{\pi}_i - \bar{y}). \end{aligned}$$

These three definitions are not equivalent in general. In particular, R_{model}^2 differs from R_{residual}^2 since

$$\text{SS}_T = \text{SS}_M + \text{SS}_R + 2C_{\hat{\pi}\hat{\pi}}$$

where

$$C_{\hat{\pi}\hat{\pi}} = \sum_{i=1}^n (y_i - \hat{\pi}_i)(\hat{\pi}_i - \bar{y}).$$

1. Prove that $R_{\text{model}}^2 \geq 0$, $R_{\text{correlation}}^2 \geq 0$ with equality holding if $\hat{\pi}_i = \bar{y}$ for all i . Prove that $R_{\text{model}}^2 \leq 1$, $R_{\text{residual}}^2 \leq 1$, $R_{\text{correlation}}^2 \leq 1$ with equality holding if $y_i = \hat{\pi}_i$ for all i .

Note that R_{residual}^2 may be negative. Give an example.

2. Define

$$\bar{\pi}_1 = \frac{\sum_{i=1}^n y_i \hat{\pi}_i}{\sum_{i=1}^n y_i}, \quad \bar{\pi}_0 = \frac{\sum_{i=1}^n (1 - y_i) \hat{\pi}_i}{\sum_{i=1}^n (1 - y_i)}$$

as the average of the fitted values for units with $y_i = 1$ and $y_i = 0$, respectively. Define

$$D = \bar{\pi}_1 - \bar{\pi}_0.$$

Prove that

$$D = (R_{\text{model}}^2 + R_{\text{residual}}^2)/2 = \sqrt{R_{\text{model}}^2 R_{\text{correlation}}^2}.$$

Further prove that $D \geq 0$ with equality holding if $\hat{\pi}_i = \bar{y}$ for all i , and $D \leq 1$ with equality holding if $y_i = \hat{\pi}_i$ for all i .

3. McFadden (1974) defined the following R^2 :

$$R_{\text{mcfadden}}^2 = 1 - \frac{\log L(\hat{\beta})}{\log L(\tilde{\beta})}$$

recalling that $\tilde{\beta}$ is the MLE assuming that all coefficients except the intercept are zero, and $\hat{\beta}$ is the MLE without any restrictions. This R^2 must lie within $[0, 1]$.

Verify that under the Normal linear model, the above formula does not reduce to the usual R^2 in OLS.

4. Cox and Snell (1989) defined the following R^2 :

$$R_{\text{CS}}^2 = 1 - \left(\frac{L(\tilde{\beta})}{L(\hat{\beta})} \right)^{2/n}.$$

Verify that under the Normal linear model, the above formula reduces to the usual R^2 .

Remark: Tjur (2009) gave an excellent discussion of R_{model}^2 , R_{residual}^2 , $R_{\text{correlation}}^2$ and D . Nagelkerke (1991) pointed out that the upper bound of this R_{CS}^2 is $1 - (L(\tilde{\beta}))^{2/n} < 1$ and proposed to modify it as

$$R_{\text{nagelkerke}}^2 = \frac{1 - \left(\frac{L(\tilde{\beta})}{L(\hat{\beta})} \right)^{2/n}}{1 - (L(\tilde{\beta}))^{2/n}}$$

to ensure that its upper bound is 1. However, this modification seems purely ad hoc. Although D is an appealing definition of R^2 for the logistic model, it does not generalize to other models. McKelvey and Zavoina (1975) suggested to define R^2 based on the latent linear representation of the logistic regression in Chapter 20.7.1:

$$R^2 = \frac{\hat{\beta}^T S_{xx} \hat{\beta}}{\hat{\beta}^T S_{xx} \hat{\beta} + \pi^2/3},$$

where $\hat{\beta}$ is the MLE of the coefficient, S_{xx} is the sample covariance matrix of the x_i 's, and $\pi^2/3$ is the variance of the standard logistic distribution by Theorem 20.3. Hu et al. (2006) studied the asymptotic properties of some of the R^2 s above.

21

Logistic Regressions for Categorical Outcomes

Categorical outcomes are common in empirical research. There are two types of categorical outcomes:

- (T1) Nominal. For example, the outcome denotes the preference for fruits (apple, orange, and pear) or transportation services (Uber, Lyft, or BART if you live in the San Francisco Bay Area).
- (T2) Ordinal. For example, the outcome denotes the course evaluation at Berkeley (1, 2, ..., 7) or Amazon review (1 to 5 stars).

This chapter discusses statistical modeling strategies for categorical outcomes, including two classes of models corresponding to the nominal and ordinal outcomes, respectively.

21.1 Multinomial distribution

A categorical random variable y taking values in $\{1, \dots, K\}$ with probabilities $\text{pr}(y = k) = \pi_k$ ($k = 1, \dots, K$) is often called a multinomial distribution, denoted by

$$y \sim \text{Multinomial}\{1; (\pi_1, \dots, \pi_K)\}, \quad (21.1)$$

where $\sum_{k=1}^K \pi_k = 1$. We can calculate the mean and covariance matrix of y :

Proposition 21.1 *If y is the Multinomial random variable in (21.1), then $(1(y = 1), \dots, 1(y = K - 1))$ has mean $(\pi_1, \dots, \pi_{K-1})$ and covariance matrix*

$$\begin{pmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_{K-1} \\ -\pi_1\pi_2 & \pi_2(1 - \pi_2) & \cdots & -\pi_2\pi_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_1\pi_{K-1} & -\pi_2\pi_{K-1} & \cdots & \pi_{K-1}(1 - \pi_{K-1}) \end{pmatrix}. \quad (21.2)$$

As a byproduct, the matrix in (21.2) must be positive semi-definite.

Proof of Proposition 21.1: Without loss of generality, I will calculate the (1,1)th and the (1,2)th element of the matrix. First, $1(y = 1)$ is Bernoulli with probability π_1 , so the (1,1)th element equals $\text{var}(1(y = 1)) = \pi_1(1 - \pi_1)$. Similarly, the (2,2)th element equals $\text{var}(1(y = 2)) = \pi_2(1 - \pi_2)$.

Second, $1(y = 1) + 1(y = 2)$ is Bernoulli with probability $\pi_1 + \pi_2$, so $\text{var}(1(y = 1) + 1(y = 2)) = (\pi_1 + \pi_2)(1 - \pi_1 - \pi_2)$. Therefore, the $(1, 2)$ -th element equals

$$\begin{aligned} & \text{cov}(1(y = 1), 1(y = 2)) \\ &= \{\text{var}(1(y = 1) + 1(y = 2)) - \text{var}(1(y = 1)) - \text{var}(1(y = 2))\} / 2 \\ &= \{(\pi_1 + \pi_2)(1 - \pi_1 - \pi_2) - \pi_1(1 - \pi_1) - \pi_2(1 - \pi_2)\} / 2 \\ &= -\pi_1\pi_2. \end{aligned}$$

□

With independent samples of $(x_i, y_i)_{i=1}^n$, we want to model y_i based on covariates x_i ¹:

$$y_i \mid x_i \sim \text{Multinomial}[1; \{\pi_1(x_i), \dots, \pi_K(x_i)\}],$$

where $\sum_{k=1}^K \pi_k(x_i) = 1$ for all x_i . We can write the probability mass function of $\text{pr}(y_i \mid x_i)$ as

$$\begin{aligned} \pi_{y_i}(x_i) &= \prod_{k=1}^K \{\pi_k(x_i) \text{ if } y_i = k\} \\ &= \prod_{k=1}^K \{\pi_k(x_i)\}^{1(y_i=k)}. \end{aligned}$$

Here $\pi_k(x_i)$ is a general function of x_i . The remaining parts of this chapter will discuss the canonical choices of $\pi_k(x_i)$ for nominal and ordinal outcomes.

21.2 Multinomial logistic model for nominal outcomes

21.2.1 Modeling

Viewing category K as the reference level, we can model the ratio of the probabilities of categories k and K as

$$\log \frac{\pi_k(x_i)}{\pi_K(x_i)} = x_i^\top \beta_k \quad (k = 1, \dots, K-1)$$

which implies that

$$\pi_k(x_i) = \pi_K(x_i) e^{x_i^\top \beta_k} \quad (k = 1, \dots, K-1).$$

Due to the normalization

$$1 = \sum_{k=1}^K \pi_k(x_i) = \sum_{k=1}^K \pi_K(x_i) e^{x_i^\top \beta_k} = \pi_K(x_i) \sum_{k=1}^K e^{x_i^\top \beta_k},$$

we have

$$\pi_K(x_i) = 1 / \sum_{k=1}^K e^{x_i^\top \beta_k}.$$

¹An alternative strategy is to model $1(y_i = k) \mid x_i$ for each k . The advantage of this strategy is that it reduces to binary logistic models. The disadvantage of this strategy is that it does not model the whole distribution of y_i and can lose efficiency in estimation.

Therefore,

$$\pi_k(x_i) = \frac{e^{x_i^\top \beta_k}}{\sum_{l=1}^K e^{x_i^\top \beta_l}}, \quad (k = 1, \dots, K-1).$$

A more compact form is the following definition of the multinomial logistic model.

Definition 21.1 (multinomial logistic regression model) *We have*

$$y_i \mid x_i \sim \text{Multinomial}[1; \{\pi_1(x_i), \dots, \pi_K(x_i)\}],$$

with

$$\pi_k(x_i) = \pi_k(x_i, \beta) = \frac{e^{x_i^\top \beta_k}}{\sum_{l=1}^K e^{x_i^\top \beta_l}}, \quad (k = 1, \dots, K). \quad (21.3)$$

The observations are independent across units. The $\beta = (\beta_1, \dots, \beta_{K-1})$ denotes the unknown parameter, with $\beta_K = 0$ for the reference category.

From the ratio form of (21.3), we can only identify $\beta_k - \beta_K$ for all $k = 1, \dots, K$. So for convenience, we impose the restriction $\beta_K = 0$ in Definition 21.1.

Similar to the binary logistic regression model, we can interpret the coefficients as the conditional log odds ratio compared to the reference level:

$$\begin{aligned} \beta_{k,j} &= \log \frac{\pi_k(\dots, x_{ij} + 1, \dots)}{\pi_K(\dots, x_{ij} + 1, \dots)} - \log \frac{\pi_k(\dots, x_{ij}, \dots)}{\pi_K(\dots, x_{ij}, \dots)} \\ &= \log \left\{ \frac{\pi_k(\dots, x_{ij} + 1, \dots)}{\pi_K(\dots, x_{ij} + 1, \dots)} \bigg/ \frac{\pi_k(\dots, x_{ij}, \dots)}{\pi_K(\dots, x_{ij}, \dots)} \right\}. \end{aligned}$$

21.2.2 MLE

The likelihood function for the multinomial logistic model is

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n \prod_{k=1}^K \{\pi_k(x_i)\}^{1(y_i=k)} \\ &= \prod_{i=1}^n \prod_{k=1}^K \left\{ \frac{e^{x_i^\top \beta_k}}{\sum_{l=1}^K e^{x_i^\top \beta_l}} \right\}^{1(y_i=k)} \\ &= \prod_{i=1}^n \left[\left\{ \prod_{k=1}^K e^{1(y_i=k) x_i^\top \beta_k} \right\} / \sum_{k=1}^K e^{x_i^\top \beta_k} \right], \end{aligned}$$

and the log-likelihood function is

$$\log L(\beta) = \sum_{i=1}^n \left[\sum_{k=1}^K 1(y_i = k) x_i^\top \beta_k - \log \left\{ \sum_{k=1}^K e^{x_i^\top \beta_k} \right\} \right].$$

The score function is

$$\frac{\partial \log L(\beta)}{\partial \beta} = \begin{pmatrix} \frac{\partial \log L(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial \log L(\beta)}{\partial \beta_{K-1}} \end{pmatrix} \in \mathbb{R}^{p(K-1)}$$

with

$$\begin{aligned}
\frac{\partial \log L(\beta)}{\partial \beta_k} &= \sum_{i=1}^n \left\{ x_i 1(y_i = k) - \frac{x_i e^{x_i^T \beta_k}}{\sum_{l=1}^K e^{x_i^T \beta_l}} \right\} \\
&= \sum_{i=1}^n x_i \left\{ 1(y_i = k) - \frac{e^{x_i^T \beta_k}}{\sum_{l=1}^K e^{x_i^T \beta_l}} \right\} \\
&= \sum_{i=1}^n x_i \{1(y_i = k) - \pi_k(x_i, \beta)\} \in \mathbb{R}^p, \quad (k = 1, \dots, K-1).
\end{aligned}$$

The Hessian matrix is

$$\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^T} = \begin{pmatrix} \frac{\partial^2 \log L(\beta)}{\partial \beta_1 \partial \beta_1^T} & \frac{\partial^2 \log L(\beta)}{\partial \beta_1 \partial \beta_2^T} & \cdots & \frac{\partial^2 \log L(\beta)}{\partial \beta_1 \partial \beta_{K-1}^T} \\ \frac{\partial^2 \log L(\beta)}{\partial \beta_2 \partial \beta_1^T} & \frac{\partial^2 \log L(\beta)}{\partial \beta_2 \partial \beta_2^T} & \cdots & \frac{\partial^2 \log L(\beta)}{\partial \beta_2 \partial \beta_{K-1}^T} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L(\beta)}{\partial \beta_{K-1} \partial \beta_1^T} & \frac{\partial^2 \log L(\beta)}{\partial \beta_{K-1} \partial \beta_2^T} & \cdots & \frac{\partial^2 \log L(\beta)}{\partial \beta_{K-1} \partial \beta_{K-1}^T} \end{pmatrix} \in \mathbb{R}^{p(K-1) \times p(K-1)} \quad (21.4)$$

with the (k, k) th block

$$\begin{aligned}
\frac{\partial^2 \log L(\beta)}{\partial \beta_k \partial \beta_k^T} &= - \sum_{i=1}^n x_i \frac{\partial}{\partial \beta_k^T} \left(\frac{e^{x_i^T \beta_k}}{\sum_{l=1}^K e^{x_i^T \beta_l}} \right) \\
&= - \sum_{i=1}^n x_i x_i^T \frac{e^{x_i^T \beta_k} \sum_{l=1}^K e^{x_i^T \beta_l} - e^{x_i^T \beta_k} e^{x_i^T \beta_k}}{(\sum_{l=1}^K e^{x_i^T \beta_l})^2} \\
&= - \sum_{i=1}^n \pi_k(x_i, \beta) \{1 - \pi_k(x_i, \beta)\} x_i x_i^T \in \mathbb{R}^{p \times p} \quad (k = 1, \dots, K-1)
\end{aligned}$$

and the (k, l) th block

$$\begin{aligned}
\frac{\partial^2 \log L(\beta)}{\partial \beta_k \partial \beta_l^T} &= - \sum_{i=1}^n x_i \frac{\partial}{\partial \beta_l^T} \left(\frac{e^{x_i^T \beta_k}}{\sum_{l=1}^K e^{x_i^T \beta_l}} \right) \\
&= - \sum_{i=1}^n x_i x_i^T \frac{-e^{x_i^T \beta_k} e^{x_i^T \beta_l}}{(\sum_{l=1}^K e^{x_i^T \beta_l})^2} \\
&= \sum_{i=1}^n \pi_k(x_i, \beta) \pi_l(x_i, \beta) x_i x_i^T \in \mathbb{R}^{p \times p} \quad (k \neq l : k, l = 1, \dots, K-1).
\end{aligned}$$

We can verify that the Hessian matrix is negative semi-definite based on Proposition 21.1, which is left as Problem 21.2.

In **R**, the function `multinom` in the `nnnet` package uses Newton's method to fit the multinomial logistic model. We can make inference about the parameters based on the asymptotic Normality of the MLE. Based on a new observation with covariate x_{n+1} , we can make prediction based on the fitted probabilities $\pi_k(x_{n+1}, \hat{\beta})$, and furthermore classify it into K categories based on

$$\hat{y}_{n+1} = \arg \max_{1 \leq k \leq K} \pi_k(x_{n+1}, \hat{\beta}).$$

21.3 A latent variable representation for the multinomial logistic regression

We can view the multinomial logistic regression (21.3) as an extension of the binary logistic regression model. The binary logistic regression has a latent variable representation as shown in Section 20.7.1. The multinomial logistic regression model in Definition 21.1 also has a latent variable representation below.

Assume

$$\begin{cases} U_{i1} &= x_i^T \beta_1 + \varepsilon_{i1}, \\ \vdots & \\ U_{iK} &= x_i^T \beta_K + \varepsilon_{iK}, \end{cases}$$

where $\varepsilon_{i1}, \dots, \varepsilon_{iK}$ are IID standard Gumbel random variables.² Using the language of economics, (U_{i1}, \dots, U_{iK}) are the utilities associated with the choices $(1, \dots, K)$. So unit i chooses k if k has the highest utility:

$$y_i = k \quad \text{if } U_{ik} > U_{il} \text{ for all } l \neq k.$$

We can show that this latent variable model implies (21.3). This follows from the lemma below, which is due to McFadden (1974).³ When $K = 2$, it also gives another latent variable representation for the binary logistic regression, which is different from the one in Section 20.7.1.

Lemma 21.1 *Assume*

$$\begin{cases} U_1 &= V_1 + \varepsilon_1, \\ \vdots & \\ U_K &= V_K + \varepsilon_K, \end{cases}$$

where $\varepsilon_1, \dots, \varepsilon_K$ are IID standard Gumbel. Define

$$y = \arg \max_{1 \leq l \leq K} U_l$$

as the index corresponding to the maximum of the U_k 's. We have

$$\text{pr}(y = k) = \frac{e^{V_k}}{\sum_{l=1}^K e^{V_l}}.$$

Proof of Lemma 21.1: Recall that the standard Gumbel random variable has CDF $F(z) = \exp(-e^{-z})$ and density $f(z) = \exp(-e^{-z})e^{-z}$.

The event " $y = k$ " is equivalent to the event " $U_k > U_l$ for all $l \neq k$ ", so

$$\begin{aligned} \text{pr}(y = k) &= \text{pr}(U_k > U_l, l = 1, \dots, k-1, k+1, \dots, K) \\ &= \text{pr}(V_k + \varepsilon_k > V_l + \varepsilon_l, l = 1, \dots, k-1, k+1, \dots, K) \\ &= \int_{-\infty}^{\infty} \text{pr}(V_k + z > V_l + \varepsilon_l, l = 1, \dots, k-1, k+1, \dots, K) f(z) dz \end{aligned}$$

²See Appendix B.1.3 for a review of the Gumbel distribution.

³Daniel McFadden shared the 2000 Nobel Memorial Prize in Economic Sciences with James Heckman.

where the last line follows from conditioning on ε_k . By the independence of the ε 's, we have

$$\begin{aligned}\text{pr}(y = k) &= \int_{-\infty}^{\infty} \prod_{l \neq k} \text{pr}(\varepsilon_l < V_k - V_l + z) f(z) dz \\ &= \int_{-\infty}^{\infty} \prod_{l \neq k} \exp(-e^{-V_k + V_l - z}) \exp(-e^{-z}) e^{-z} dz \\ &= \int_{-\infty}^{\infty} \exp\left(-\sum_{l \neq k} e^{-V_k + V_l} e^{-z}\right) \exp(-e^{-z}) e^{-z} dz.\end{aligned}$$

Changing of variables $t = e^{-z}$ with $dz = -1/t dt$, we obtain

$$\begin{aligned}\text{pr}(y = k) &= \int_0^{\infty} \exp\left(-t \sum_{l \neq k} e^{-V_k + V_l}\right) \exp(-t) dt \\ &= \int_0^{\infty} \exp(-t C_k) dt,\end{aligned}$$

where

$$C_k = 1 + \sum_{l \neq k} e^{-V_k + V_l}.$$

The integral simplifies to $1/C_k$ due to the density of the exponential distribution. Therefore,

$$\begin{aligned}\text{pr}(y = k) &= \frac{1}{1 + \sum_{l \neq k} e^{-V_k + V_l}} \\ &= \frac{e^{V_k}}{e^{V_k} + \sum_{l \neq k} e^{V_l}} \\ &= \frac{e^{V_k}}{\sum_{l=1}^K e^{V_l}}.\end{aligned}$$

□

Lemma 21.1 is remarkable and elegant. I will use it again in Section 21.6.

21.4 Proportional odds model for ordinal outcomes

For ordinal outcomes, we can still use the multinomial logistic model, but by doing this, we discard the ordering information in the outcome. Now I will introduce the proportional odds model for ordinal outcomes, which is more parsimonious and more interpretable than the multinomial logistic model.

Motivated by the latent linear representation of the binary logistic model in Chapter 20.7.1, we imagine that the ordinal outcome arises from discretizing a continuous latent variable, as shown in the model below.

Definition 21.2 (proportional odds model for an ordinal outcome) Assume that the latent variable y_i^* satisfies the following linear model:

$$y_i^* = x_i^T \beta + \varepsilon_i, \quad (21.5)$$

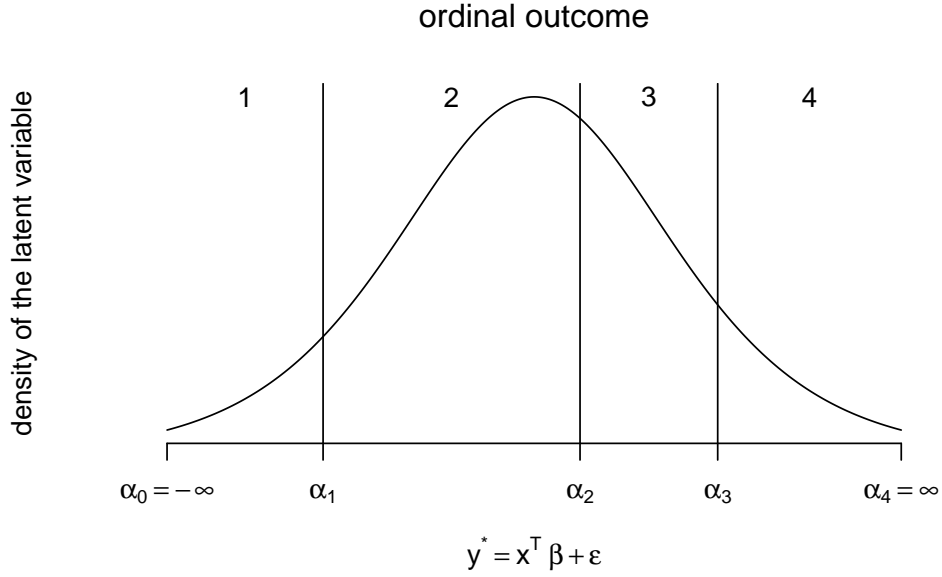


FIGURE 21.1: Latent variable representation of the ordinal outcome

with $\varepsilon_i \perp\!\!\!\perp x_i$, where

$$\text{pr}(\varepsilon_i \leq z \mid x_i) = g(z)$$

with $g(z) = e^z / (1 + e^z)$ is the CDF of the standard logistic distribution. Discretize y_i^* to obtain the observed y_i :

$$y_i = k, \quad \text{if } \alpha_{k-1} < y_i^* \leq \alpha_k, \quad (k = 1, \dots, K) \quad (21.6)$$

where

$$-\infty = \alpha_0 < \alpha_1 < \dots < \alpha_{K-1} < \alpha_K = \infty.$$

The observations are independent across units. The unknown parameters are $(\beta, \alpha_1, \dots, \alpha_{K-1})$.

In Definition 21.2, the distribution of the latent error term $g(\cdot)$ is known and specified as the standard logistic distribution. In general, $g(\cdot)$ can be the CDF of other random variables. For instance, when $g(\cdot)$ is the CDF of the standard Normal distribution, the model is called the “ordered Probit model.” I will focus on the proportional odds logistic model in the main text and defer the details for the ordered Probit model to Problem 21.5. Figure 21.1 illustrates the data generating process of Definition 21.2 with $K = 4$.

Based on the latent linear model, we can compute

$$\begin{aligned} \text{pr}(y_i \leq k \mid x_i) &= \text{pr}(y_i^* \leq \alpha_k \mid x_i) \\ &= \text{pr}(x_i^\top \beta + \varepsilon_i \leq \alpha_k \mid x_i) \\ &= \text{pr}(\varepsilon_i \leq \alpha_k - x_i^\top \beta \mid x_i) \\ &= g(\alpha_k - x_i^\top \beta). \end{aligned}$$

With the proportional odds model, we have

$$\text{pr}(y_i \leq k \mid x_i) = \frac{e^{\alpha_k - x_i^\top \beta}}{1 + e^{\alpha_k - x_i^\top \beta}}$$

or

$$\text{logit}\{\text{pr}(y_i \leq k \mid x_i)\} = \log \frac{\text{pr}(y_i \leq k \mid x_i)}{\text{pr}(y_i > k \mid x_i)} = \alpha_k - x_i^T \beta. \quad (21.7)$$

The model has the “proportional odds” property because

$$\frac{\text{pr}(y_i \leq k \mid \dots, x_{ij} + 1, \dots)}{\text{pr}(y_i > k \mid \dots, x_{ij} + 1, \dots)} \bigg/ \frac{\text{pr}(y_i \leq k \mid \dots, x_{ij}, \dots)}{\text{pr}(y_i > k \mid \dots, x_{ij}, \dots)} = e^{-\beta_j}$$

which is a positive constant not depending on k .

The sign of $x_i^T \beta$ is negative due to the latent variable representation. Some textbooks and software packages use a positive sign, but the function `polr` in package `MASS` of `R` uses (21.7). This book follows `polr`.

The proportional odds model implies a quite complicated form of the probability for each category:

$$\text{pr}(y_i = k \mid x_i) = \frac{e^{\alpha_k - x_i^T \beta}}{1 + e^{\alpha_k - x_i^T \beta}} - \frac{e^{\alpha_{k-1} - x_i^T \beta}}{1 + e^{\alpha_{k-1} - x_i^T \beta}}, \quad (k = 1, \dots, K).$$

So the likelihood function is

$$\begin{aligned} L(\beta, \alpha_1, \dots, \alpha_{K-1}) &= \prod_{i=1}^n \prod_{k=1}^K \{\text{pr}(y_i = k \mid x_i)\}^{1(y_i=k)} \\ &= \prod_{i=1}^n \prod_{k=1}^K \left(\frac{e^{\alpha_k - x_i^T \beta}}{1 + e^{\alpha_k - x_i^T \beta}} - \frac{e^{\alpha_{k-1} - x_i^T \beta}}{1 + e^{\alpha_{k-1} - x_i^T \beta}} \right)^{1(y_i=k)}. \end{aligned}$$

The log-likelihood function is concave (Pratt, 1981; Burridge, 1981), and it is strictly concave in most cases. The function `polr` in the `R` package `MASS` computes the MLE of the proportional odds model using the BFGS algorithm (which is similar to Newton’s method but does not involve calculating the Hessian). It uses the explicit formulas of the gradient of the log-likelihood function, and computes the Hessian matrix numerically. I relegate the formulas of the gradient to Problem 21.4. For more details of the Hessian matrix, see Agresti (2010), which is a textbook discussion on modeling ordinal data.

21.5 A case study

I use an observational dataset from the Karolinska Institute in Stockholm, Sweden to illustrate the application of logistic regressions. Rubin (2008) used this dataset to investigate whether it is better for cardia cancer patients to be treated in a large-volume hospital compared with a small-volume hospital, where volume is defined by the number of patients with cardia cancer treated in recent years. I use the following variables:

- `highdiag` indicating whether the patient was diagnosed at a high-volume hospital,
- `hightreat` indicating whether the patient was treated at a high-volume hospital,
- `age` representing the age of the patient,
- `rural` indicating whether the patient was from a rural area,

- survival representing the years of survival after diagnosis with three categories (“1”, “2-4”, “5+”).

```
karolinska = read.table("karolinska.txt", header = TRUE)
karolinska = karolinska[, c("highdiag", "hightreat",
                           "age", "rural",
                           "male", "survival")]
```

21.5.1 Binary logistic for the treatment

We have two choices of the treatment: `highdiag` and `hightreat`. The logistic fit of `highdiag` on covariates is shown below.

```
> diagglm = glm(highdiag ~ age + rural + male,
+               data = karolinska,
+               family = binomial(link = "logit"))
> summary(diagglm)
```

Call:

```
glm(formula = highdiag ~ age + rural + male, family = binomial(link = "logit"),
    data = karolinska)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.06147	-0.98645	-0.05759	1.01391	1.75696

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	3.46604	1.14545	3.026	0.002479	**
age	-0.03124	0.01481	-2.110	0.034854	*
rural	-1.26322	0.34530	-3.658	0.000254	***
male	-0.97524	0.41303	-2.361	0.018216	*

The logistic fit of `hightreat` is shown below.

```
> treatglm = glm(hightreat ~ age + rural + male,
+                data = karolinska,
+                family = binomial(link = "logit"))
> summary(treatglm)
```

Call:

```
glm(formula = hightreat ~ age + rural + male, family = binomial(link = "logit"),
    data = karolinska)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2912	-0.9978	0.5387	0.8408	1.4810

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	6.44683	1.49544	4.311	1.63e-05	***
age	-0.06297	0.01890	-3.332	0.000862	***
rural	-1.28777	0.39572	-3.254	0.001137	**
male	-0.74856	0.45285	-1.653	0.098329	.

Both treatments are associated with the covariates. `hightreat` is more strongly associated with age. Rubin (2008) argued that `highdiag` is more random than `hightreat`, and may have weaker association with other hidden covariates. For each model below, I fit the data twice corresponding to two choices of treatment. Overall, we should trust the results with `highdiag` more based on Rubin (2008)’s argument.

21.5.2 Binary logistic for the outcome

I first fit binary logistic models for the dichotomized outcome indicating whether the patient survived longer than a year after diagnosis.

```
> karolinska$loneyear = (karolinska$survival != "1")
> loneyearglm = glm(loneyear ~ highdiag + age + rural + male,
+                   data = karolinska,
+                   family = binomial(link = "logit"))
> summary(loneyearglm)
```

Call:

```
glm(formula = loneyear ~ highdiag + age + rural + male, family = binomial(link = "logit"),
    data = karolinska)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.1755	-0.9936	-0.7739	1.3024	1.8557

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.22919	1.15545	-1.064	0.2874
highdiag	0.13684	0.36586	0.374	0.7084
age	-0.00389	0.01411	-0.276	0.7829
rural	0.33360	0.35798	0.932	0.3514
male	0.86706	0.44034	1.969	0.0489 *

The regressor `highdiag` is not significant in the above regression.

```
> loneyearglm = glm(loneyear ~ hightreat + age + rural + male,
+                   data = karolinska,
+                   family = binomial(link = "logit"))
> summary(loneyearglm)
```

Call:

```
glm(formula = loneyear ~ hightreat + age + rural + male, family = binomial(link = "logit"),
    data = karolinska)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.3767	-0.9683	-0.6784	1.0813	2.0833

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.353977	1.317942	-2.545	0.01093 *
hightreat	1.417458	0.455603	3.111	0.00186 **
age	0.008725	0.014840	0.588	0.55655
rural	0.633278	0.368525	1.718	0.08572 .
male	1.079973	0.452191	2.388	0.01693 *

The regressor `hightreat` is significant in the above regression.

21.5.3 Multinomial logistic for the outcome

I then fit multinomial logistic models for the outcome with three categories.

```
> library(nnet)
> yearmultinom = multinom(survival ~ highdiag + age + rural + male,
+                          data = karolinska)
# weights: 18 (10 variable)
initial value 173.580742
iter 10 value 134.331992
final value 134.130815
converged
```

```
> summary(yearmultinom)
Call:
multinom(formula = survival ~ highdiag + age + rural + male,
  data = karolinska)

Coefficients:
  (Intercept)    highdiag      age      rural      male
2-4    -1.075818 -0.06973187 -0.004624030 0.1744256 0.5028786
5+     -4.180416  0.64036289 -0.001846453 0.7365111 2.1628717

Std. Errors:
  (Intercept)    highdiag      age      rural      male
2-4     1.286987 0.4113006 0.01596377 0.4014718 0.4716831
5+      2.003581 0.5816365 0.02148936 0.5741017 1.0741239

Residual Deviance: 268.2616
AIC: 288.2616
> predict(yearmultinom, type = "probs")[1:5, ]
      1      2-4      5+
1 0.5950631 0.2647047 0.14023222
2 0.5941802 0.2655369 0.14028293
3 0.8081376 0.1718963 0.01996613
4 0.5950631 0.2647047 0.14023222
5 0.6366929 0.2260086 0.13729849
```

The regressor `highdiag` is not significant above. The `predict` function gives the fitted probabilities for all categories of the outcome.

```
> yearmultinom = multinom(survival ~ hightreat + age + rural + male,
+                           data = karolinska)
# weights:  18 (10 variable)
initial value 173.580742
iter  10 value 129.548642
final value 129.283739
converged
> summary(yearmultinom)
Call:
multinom(formula = survival ~ hightreat + age + rural + male,
  data = karolinska)

Coefficients:
  (Intercept) hightreat      age      rural      male
2-4    -3.312433  1.326354 0.008527561 0.5186654 0.7514451
5+     -5.935172  1.627711 0.008978103 0.9063831 2.2780877

Std. Errors:
  (Intercept) hightreat      age      rural      male
2-4     1.463258 0.5141127 0.01660648 0.4085976 0.4806953
5+      2.190305 0.7320788 0.02244867 0.5645595 1.0739669

Residual Deviance: 258.5675
AIC: 278.5675
```

The regressor `hightreat` is significant above.

21.5.4 Proportional odds logistic for the outcome

The multinomial logisitic model does not reflect the ordering information of the outcome. I will fit the proportional odds models below.

```
> library(MASS)
> yearpo = polr(factor(survival) ~ highdiag + age + rural + male,
+               Hess = TRUE,
+               data = karolinska)
```

```
> summary(yearpo)
Call:
polr(formula = factor(survival) ~ highdiag + age + rural + male,
      data = karolinska, Hess = TRUE)
```

Coefficients:

	Value	Std. Error	t value
highdiag	0.216755	0.35892	0.6039
age	-0.002881	0.01378	-0.2091
rural	0.371898	0.35313	1.0532
male	0.943955	0.43588	2.1656

Intercepts:

	Value	Std. Error	t value
1 2-4	1.4079	1.1309	1.2450
2-4 5+	2.9284	1.1514	2.5434

Residual Deviance: 271.0778

AIC: 283.0778

```
> predict(yearpo, type = "probs")[1:5, ]
      1      2-4      5+
1 0.5862465 0.2800892 0.13366427
2 0.5855475 0.2804542 0.13399823
3 0.8087341 0.1421065 0.04915948
4 0.5862465 0.2800892 0.13366427
5 0.6205983 0.2615112 0.11789050
```

The regressor `highdiag` is not significant above. The `predict` function gives the fitted probabilities of three categories.

```
> yearpo = polr(factor(survival) ~ hightreat + age + rural + male,
+               Hess = TRUE,
+               data = karolinska)
> summary(yearpo)
Call:
polr(formula = factor(survival) ~ hightreat + age + rural + male,
      data = karolinska, Hess = TRUE)
```

Coefficients:

	Value	Std. Error	t value
hightreat	1.399538	0.44518	3.1438
age	0.008032	0.01438	0.5584
rural	0.638862	0.35450	1.8022
male	1.122698	0.44377	2.5299

Intercepts:

	Value	Std. Error	t value
1 2-4	3.3273	1.2752	2.6092
2-4 5+	4.9258	1.3106	3.7583

Residual Deviance: 260.2831

AIC: 272.2831

```
> predict(yearpo, type = "probs")[1:5, ]
      1      2-4      5+
1 0.7007473 0.2197669 0.07948581
2 0.7024290 0.2186709 0.07890008
3 0.7736340 0.1705068 0.05585924
4 0.7007473 0.2197669 0.07948581
5 0.5305401 0.3176927 0.15176720
```

The regressor `hightreat` is significant above.

21.6 Discrete choice models

21.6.1 Model

The covariates in model (21.3) depend only on individuals. McFadden (1974) extends it to allow for choice-specific covariates z_{ik} . His formulation is based on the latent utility representation:

$$\begin{cases} U_{i1} &= z_{i1}^T \theta + \varepsilon_{i1}, \\ \vdots & \\ U_{iK} &= z_{iK}^T \theta + \varepsilon_{iK}, \end{cases}$$

where $\varepsilon_{i1}, \dots, \varepsilon_{iK}$ are IID standard Gumbel. Unit i chooses k if k has the highest utility. Lemma 21.1 implies that

$$\pi_k(z_i) = \pi_k(z_i, \theta) = \frac{e^{z_{ik}^T \theta}}{\sum_{l=1}^K e^{z_{il}^T \theta}}, \quad (k = 1, \dots, K). \quad (21.8)$$

Model (21.8) seems rather similar to model (21.3). However, there are many subtle differences. First, a component of z_{ik} may vary only with choice k , for example, it can represent the price of choice k . Partition z_{ik} into three types of covariates: x_i that only vary across individuals, c_k that only vary across choices, and w_{ik} that vary across both individuals and choices. Model (21.8) reduces to

$$\pi_k(z_i) = \frac{e^{x_i^T \theta_x + c_k^T \theta_c + w_{ik}^T \theta_w}}{\sum_{l=1}^K e^{x_i^T \theta_x + c_l^T \theta_c + w_{il}^T \theta_w}} = \frac{e^{c_k^T \theta_c + w_{ik}^T \theta_w}}{\sum_{l=1}^K e^{c_l^T \theta_c + w_{il}^T \theta_w}},$$

that is, the individual-specific covariates drop out. Therefore, z_{ik} in model (21.8) does not contain covariates that vary only with individuals. In particular, z_{ik} in model (21.8) does not contain the constant, but in contrast, the x_i in model (21.3) usually contains the intercept by default.

Second, if we want to use individual-specific covariates in the model, they must have choice-specific coefficients. So a more general model unifying (21.3) and (21.8) is

$$\pi_k(x_i, w_{ik}, \theta, \beta) = \frac{e^{w_{ik}^T \theta + x_i^T \beta_k}}{\sum_{l=1}^K e^{w_{il}^T \theta + x_i^T \beta_l}}, \quad (k = 1, \dots, K). \quad (21.9)$$

Equivalently, we can create pseudo covariates z_{ik} as the original w_{ik} together with interaction of x_i and the dummy for choice k . For example, if $K = 3$ and x_i contain the intercept and a scalar individual-specific covariate x'_i , then (z_{i1}, z_{i2}, z_{i3}) are

$$\begin{pmatrix} z_{i1} \\ z_{i2} \\ z_{i3} \end{pmatrix} = \begin{pmatrix} w_{i1} & 1 & 0 & x'_i & 0 \\ w_{i2} & 0 & 1 & 0 & x'_i \\ w_{i3} & 0 & 0 & 0 & 0 \end{pmatrix},$$

where $K = 3$ is the reference level. So with augmented covariates, the discrete choice model (21.8) is strictly more general than the multinomial logistic model (21.3). In the special case with $K = 2$, model (21.8) reduces to

$$\text{pr}(y_i = 1 \mid x_i) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

where $x_i = z_{i1} - z_{i2}$.

21.6.2 MLE

Based on the model specification (21.8), the log likelihood function is

$$\log L(\theta) = \sum_{i=1}^n \sum_{k=1}^K 1(y_i = k) \left(z_{ik}^T \theta - \log \sum_{l=1}^K e^{z_{il}^T \theta} \right).$$

So the score function is

$$\begin{aligned} \frac{\partial}{\partial \theta} \log L(\theta) &= \sum_{i=1}^n \sum_{k=1}^K 1(y_i = k) \left(z_{ik} - \frac{\sum_{l=1}^K e^{z_{il}^T \theta} z_{il}}{\sum_{l=1}^K e^{z_{il}^T \theta}} \right) \\ &= \sum_{i=1}^n \sum_{k=1}^K 1(y_i = k) \{z_{ik} - E(z_{ik}; \theta)\}, \end{aligned}$$

where $E(\cdot; \theta)$ is the average value of $\{z_{i1}, \dots, z_{iK}\}$ over the probability mass function

$$p_k(\theta) = e^{z_{ik}^T \theta} / \sum_{l=1}^K e^{z_{il}^T \theta}.$$

The Hessian matrix is

$$\begin{aligned} &\frac{\partial^2}{\partial \theta \partial \theta^T} \log L(\theta) \\ &= - \sum_{i=1}^n \sum_{k=1}^K 1(y_i = k) \frac{\sum_{l=1}^K e^{z_{il}^T \theta} z_{il} z_{il}^T \sum_{l=1}^K e^{z_{il}^T \theta} - \sum_{l=1}^K e^{z_{il}^T \theta} z_{il} \sum_{l=1}^K e^{z_{il}^T \theta} z_{il}^T}{(\sum_{l=1}^K e^{z_{il}^T \theta})^2} \\ &= - \sum_{i=1}^n \sum_{k=1}^K 1(y_i = k) \text{cov}(z_{ik}; \theta), \end{aligned}$$

where $\text{cov}(\cdot; \theta)$ is the covariance matrix of $\{z_{i1}, \dots, z_{iK}\}$ over the probability mass function defined above. From these formulas, we can compute the MLE using Newton's method and obtain its asymptotic distribution based on the inverse of the Fisher information matrix.

21.6.3 Example

The R package `mlogit` provides a function `mlogit` to fit the general discrete logistic model (Croissant, 2020). Here I use an example from this package to illustrate the model fitting of `mlogit`.

```
> library("nnet")
> library("mlogit")
> data("Fishing")
> head(Fishing)
```

	mode	price.beach	price.pier	price.boat	price.charter
1	charter	157.930	157.930	157.930	182.930
2	charter	15.114	15.114	10.534	34.534
3	boat	161.874	161.874	24.334	59.334
4	pier	15.134	15.134	55.930	84.930
5	boat	106.930	106.930	41.514	71.014
6	charter	192.474	192.474	28.934	63.934

	catch.beach	catch.pier	catch.boat	catch.charter	income
1	0.0678	0.0503	0.2601	0.5391	7083.332
2	0.1049	0.0451	0.1574	0.4671	1250.000
3	0.5333	0.4522	0.2413	1.0266	3750.000

4	0.0678	0.0789	0.1643	0.5391	2083.333
5	0.0678	0.0503	0.1082	0.3240	4583.332
6	0.5333	0.4522	0.1665	0.3975	4583.332

The dataset `Fishing` is in the “wide” format, where `mode` denotes the choice of four modes of fishing (beach, pier, boat and charter), `price` and `catch` denote the price and catching rates which are choice-specific, `income` is individual-specific. We need to first transform the dataset into “long” format.

```
> Fish = dfidx(Fishing,
+             varying = 2:9,
+             shape = "wide",
+             choice = "mode")
> head(Fish)
~~~~~
first 10 observations out of 4728
~~~~~
   mode  income  price  catch  idx
1 FALSE 7083.332 157.930 0.0678 1:each
2 FALSE 7083.332 157.930 0.2601 1:boat
3 TRUE  7083.332 182.930 0.5391 1:rter
4 FALSE 7083.332 157.930 0.0503 1:pier
5 FALSE 1250.000  15.114 0.1049 2:each
6 FALSE 1250.000  10.534 0.1574 2:boat
7 TRUE  1250.000  34.534 0.4671 2:rter
8 FALSE 1250.000  15.114 0.0451 2:pier
9 FALSE 3750.000 161.874 0.5333 3:each
10 TRUE 3750.000  24.334 0.2413 3:boat
```

Using only choice-specific covariates, we have the following fitted model:

```
> summary(mlogit(mode ~ 0 + price + catch, data = Fish))

Call:
mlogit(formula = mode ~ 0 + price + catch, data = Fish, method = "nr")

Frequencies of alternatives:choice
  beach  boat charter  pier
0.11337 0.35364 0.38240 0.15059

nr method
6 iterations, 0h:0m:0s
g'(-H)^-1g = 0.000179
successive function values within tolerance limits

Coefficients :
      Estimate Std. Error z-value Pr(>|z|)
price -0.0204765  0.0012231 -16.742 < 2.2e-16 ***
catch  0.9530982  0.0894134  10.659 < 2.2e-16 ***
```

If we do not enforce `0 + price`, we allow for intercepts that vary across choices:

```
> summary(mlogit(mode ~ price + catch, data = Fish))

Call:
mlogit(formula = mode ~ price + catch, data = Fish, method = "nr")

Frequencies of alternatives:choice
  beach  boat charter  pier
0.11337 0.35364 0.38240 0.15059

nr method
7 iterations, 0h:0m:0s
g'(-H)^-1g = 6.22E-06
successive function values within tolerance limits
```

```

Coefficients :
              Estimate Std. Error z-value Pr(>|z|)
(Intercept):boat    0.8713749   0.1140428   7.6408 2.154e-14 ***
(Intercept):charter  1.4988884   0.1329328  11.2755 < 2.2e-16 ***
(Intercept):pier    0.3070552   0.1145738   2.6800 0.0073627 **
price              -0.0247896   0.0017044 -14.5444 < 2.2e-16 ***
catch               0.3771689   0.1099707   3.4297 0.0006042 ***

```

Using only individual-specific covariates, we have the following fitted model:

```

> summary(mlogit(mode ~ 0 | income, data = Fish))

Call:
mlogit(formula = mode ~ 0 | income, data = Fish, method = "nr")

Frequencies of alternatives:choice
  beach    boat charter    pier
0.11337 0.35364 0.38240 0.15059

nr method
4 iterations, 0h:0m:0s
g'(-H)^-1g = 8.32E-07
gradient close to zero

```

```

Coefficients :
              Estimate Std. Error z-value Pr(>|z|)
(Intercept):boat    7.3892e-01   1.9673e-01   3.7560 0.0001727 ***
(Intercept):charter  1.3413e+00   1.9452e-01   6.8955 5.367e-12 ***
(Intercept):pier    8.1415e-01   2.2863e-01   3.5610 0.0003695 ***
income:boat         9.1906e-05   4.0664e-05   2.2602 0.0238116 *
income:charter      -3.1640e-05   4.1846e-05  -0.7561 0.4495908
income:pier        -1.4340e-04   5.3288e-05  -2.6911 0.0071223 **

```

It is equivalent to fitting the multinomial logistic model using the original data.

```

> summary(multinom(mode ~ income, data = Fishing))
# weights: 12 (6 variable)
initial value 1638.599935
iter 10 value 1477.150646
final value 1477.150569
converged
Call:
multinom(formula = mode ~ income, data = Fishing)

```

```

Coefficients:
      (Intercept)      income
pier    0.8141506 -1.434028e-04
boat    0.7389178  9.190824e-05
charter  1.3412901 -3.163844e-05

```

```

Std. Errors:
      (Intercept)      income
pier  5.816490e-09  2.668383e-05
boat  3.209473e-09  2.057825e-05
charter 3.921689e-09 2.116425e-05

```

```

Residual Deviance: 2954.301
AIC: 2966.301

```

The most general model includes all covariates.

```

> summary(mlogit(mode ~ price + catch | income, data = Fish))

Call:
mlogit(formula = mode ~ price + catch | income, data = Fish,

```

```

method = "nr")

Frequencies of alternatives:choice
  beach    boat charter    pier
0.11337 0.35364 0.38240 0.15059

nr method
7 iterations, 0h:0m:0s
g'(-H)^-1g = 1.37E-05
successive function values within tolerance limits

Coefficients :
              Estimate Std. Error z-value Pr(>|z|)
(Intercept):boat    5.2728e-01  2.2279e-01   2.3667 0.0179485 *
(Intercept):charter  1.6944e+00  2.2405e-01   7.5624 3.952e-14 ***
(Intercept):pier    7.7796e-01  2.2049e-01   3.5283 0.0004183 ***
price              -2.5117e-02  1.7317e-03 -14.5042 < 2.2e-16 ***
catch              3.5778e-01  1.0977e-01   3.2593 0.0011170 **
income:boat         8.9440e-05  5.0067e-05   1.7864 0.0740345 .
income:charter      -3.3292e-05  5.0341e-05  -0.6613 0.5084031
income:pier         -1.2758e-04  5.0640e-05  -2.5193 0.0117582 *

```

21.6.4 More comments

The assumption of Gumbel error terms is very strong. However, relaxing this assumption leads to much more complicated forms of the conditional probabilities of the outcome. The model (21.8) implies that

$$\frac{\pi_k(z_i)}{\pi_l(z_i)} = \exp\{(z_{ik} - z_{il})^T \theta\},$$

so the choice between k and l does not depend on the existence of other choices. This is called the independence of irrelevant alternatives (IIA) assumption. This is often a plausible assumption. However, it may be violated. For example, with the apple and orange, someone chooses the apple; but with the apple, orange, and banana, they may choose the orange.

The model (21.8) is the basic form of the discrete choice model. Train (2009) is a monograph on this topic, which provides many extensions.

21.7 Homework problems

21.1 Inverse model for the multinomial logit model

Theorem 21.1 below extends Theorem 20.2. Prove Theorem 21.1.

Assume

$$y_i \sim \text{Multinomial}(1; q_1, \dots, q_K), \quad (21.10)$$

and

$$x_i \mid y_i = k \sim N(\mu_k, \Sigma), \quad (21.11)$$

where x_i does not contain 1. We can verify that $y_i \mid x_i$ follows a multinomial logit model as shown in the theorem below.

Theorem 21.1 Under (21.10) and (21.11), we have

$$\text{pr}(y_i = k \mid x_i) = \frac{e^{\alpha_k + x_i^\top \beta_k}}{\sum_{l=1}^K e^{\alpha_l + x_i^\top \beta_l}},$$

where

$$\alpha_k = \log q_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k, \quad \beta_k = \Sigma^{-1} \mu_k.$$

21.2 Hessian matrix in the multinomial logit model

Prove that the Hessian matrix (21.4) of the log-likelihood function of the multinomial logit model is negative semi-definite.

Remark: Use Proposition 21.1.

21.3 Iteratively reweighted least squares algorithm for the multinomial logit model

Similar to the binary logistic model, Newton's method for computing the MLE for the multinomial logit model can be written as iteratively reweighted least squares. Give the details.

21.4 Score function of the proportional odds model

Derive the explicit formulas of the score function of the proportional odds model.

21.5 Ordered Probit regression

If we choose $\varepsilon_i \mid x_i \sim N(0, 1)$ in (21.5), then the corresponding model is called the ordered Probit regression. Write down the likelihood function and derive the score function for this model.

Remark: You can use the function `polr` in `R` to fit this model with the specification `method = "probit"`.

21.6 Case-control study and multinomial logistic model

Theorem 21.2 below extends Theorem 20.1. Prove Theorem 21.2.

Theorem 21.2 Assume

$$\text{pr}(y_i = k \mid x_i) = \frac{e^{\alpha_k + x_i^\top \beta_k}}{\sum_{l=1}^K e^{\alpha_l + x_i^\top \beta_l}}$$

and

$$\text{pr}(s_i = 1 \mid y_i = k, x_i) = \text{pr}(s_i = 1 \mid y_i = k) = p_k$$

for $k = 1, \dots, K$. Then we have

$$\text{pr}(y_i = k \mid x_i, s_i = 1) = \frac{e^{\tilde{\alpha}_k + x_i^\top \beta_k}}{\sum_{l=1}^K e^{\tilde{\alpha}_l + x_i^\top \beta_l}}$$

with $\tilde{\alpha}_k = \alpha_k + \log p_k$ for $k = 1, \dots, K$.

Regression Models for Count Outcomes

A random variable for counts can take values in $\{0, 1, 2, \dots\}$. This type of variable is common in applied statistics. For example, it can represent

- (E1) how many times you visit the gym every week,
- (E2) how many lectures you have missed in the “Linear Model” course,
- (E3) how many traffic accidents happened in certain areas during certain periods.

This chapter focuses on statistical modeling of those outcomes given covariates. Hilbe (2014) is a textbook focusing on count outcome regressions.

22.1 Some random variables for counts

I first review four canonical choices of random variables for modeling count data.

22.1.1 Poisson

A random variable y is Poisson(λ), if its probability mass function is

$$\text{pr}(y = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad (k = 0, 1, 2, \dots)$$

which sums to 1 by the Taylor expansion formula $e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$. Propositions 22.1 and 22.2 below review the basic properties of Poisson(λ) random variable. I relegate their proofs to Problem 22.1.

Proposition 22.1 *If $y \sim \text{Poisson}(\lambda)$, then*

$$E(y) = \text{var}(y) = \lambda.$$

Proposition 22.2 *If y_1, \dots, y_K are mutually independent with $y_k \sim \text{Poisson}(\lambda_k)$ for $k = 1, \dots, K$, then*

$$y_1 + \dots + y_K \sim \text{Poisson}(\lambda),$$

and

$$(y_1, \dots, y_K) \mid y_1 + \dots + y_K = n \sim \text{Multinomial}(n, (\lambda_1/\lambda, \dots, \lambda_K/\lambda)),$$

where $\lambda = \lambda_1 + \dots + \lambda_K$.

Conversely, if $S \sim \text{Poisson}(\lambda)$ with $\lambda = \lambda_1 + \dots + \lambda_K$, and $(y_1, \dots, y_K) \mid S = n \sim \text{Multinomial}(n, (\lambda_1/\lambda, \dots, \lambda_K/\lambda))$, then y_1, \dots, y_K are mutually independent with $y_k \sim \text{Poisson}(\lambda_k)$ for $k = 1, \dots, K$.

Where does the Poisson random variable come from? One way to generate Poisson is through independent Bernoulli random variables. I will review Le Cam (1960)'s theorem below. Its proof is beyond the scope of this book.

Theorem 22.1 *Suppose X_i 's are independent Bernoulli random variables with probabilities p_i 's ($i = 1, \dots, n$). Define $\lambda_n = \sum_{i=1}^n p_i$ and $S_n = \sum_{i=1}^n X_i$. Then*

$$\sum_{k=0}^{\infty} \left| \text{pr}(S_n = k) - e^{-\lambda_n} \frac{\lambda_n^k}{k!} \right| \leq 2 \sum_{i=1}^n p_i^2.$$

As a special case, if $p_i = \lambda/n$, then Theorem 22.1 implies

$$\sum_{k=0}^{\infty} \left| \text{pr}(S_n = k) - e^{-\lambda} \frac{\lambda^k}{k!} \right| \leq 2 \sum_{i=1}^n (\lambda/n)^2 = \lambda^2/n \rightarrow 0.$$

So the sum of IID Bernoulli random variables is approximately Poisson, if the probability has order $1/n$. This is called the law of rare events, or Poisson limit theorem, or Le Cam's theorem. By Theorem 22.1, we can use Poisson as a model for the sum of many rare events.

22.1.2 Negative-Binomial

The Poisson distribution restricts that the mean must be the same as the variance. It cannot capture the feature of overdispersed data with the variance larger than the mean.¹ The Negative-Binomial is an extension of the Poisson that allows for overdispersion. The definition below, due to Fisher et al. (1943), is different from its standard definition, but it is more natural as an extension of the Poisson.² Define y as the Negative-Binomial random variable, denoted by $\text{NB}(\mu, \theta)$ with $\mu > 0$ and $\theta > 0$, if

$$\begin{cases} y \mid \lambda & \sim \text{Poisson}(\lambda), \\ \lambda & \sim \text{Gamma}(\theta, \theta/\mu). \end{cases} \quad (22.1)$$

So the Negative-Binomial is the Poisson with a random Gamma intensity, that is, the Negative-Binomial is a scale mixture of the Poisson. If $\theta \rightarrow \infty$, then λ is a point mass at μ and the Negative-Binomial reduces to $\text{Poisson}(\mu)$. Proposition 22.3 below gives the probability mass function and moments of the Negative-Binomial. I relegate its proof to Problem 22.2.

Proposition 22.3 *The Negative-Binomial random variable defined in (22.1) has the probability mass function*

$$\text{pr}(y = k) = \frac{\Gamma(k + \theta)}{\Gamma(k + 1)\Gamma(\theta)} \left(\frac{\theta}{\mu + \theta} \right)^{\theta} \left(\frac{\mu}{\mu + \theta} \right)^k, \quad (k = 0, 1, 2, \dots),$$

¹This book focused on the issue of overdispersion. It is also possible to have underdispersed data with the variance smaller than the mean. See Puig et al. (2023) for the probabilistic mechanism that leads to underdispersion.

²With IID Bernoulli(p) trials, the Negative-Binomial distribution, denoted by $y \sim \text{NB}'(r, p)$, is the number of success before the r th failure. Its probability mass function is

$$\text{pr}(y = k) = \binom{k + r - 1}{k} (1 - p)^r p^k, \quad (k = 0, 1, 2, \dots)$$

If $p = \mu/(\mu + \theta)$ and $r = \theta$ then these two definitions coincide. This definition is more restrictive because r must be an integer.

and moments

$$\begin{aligned} E(y) &= \mu, \\ \text{var}(y) &= \mu + \frac{\mu^2}{\theta}. \end{aligned}$$

By Proposition 22.3, the variance of Negative-Binomial is always larger than its mean, with a finite θ . The dispersion parameter θ controls the variance of the Negative-Binomial. With the same mean, the Negative-Binomial has a larger variance than Poisson. Figure 22.1 further compares the log probability mass functions of the Negative-Binomial and Poisson. It shows that the Negative-Binomial has a slightly higher probability at zero but much heavier tails than the Poisson.

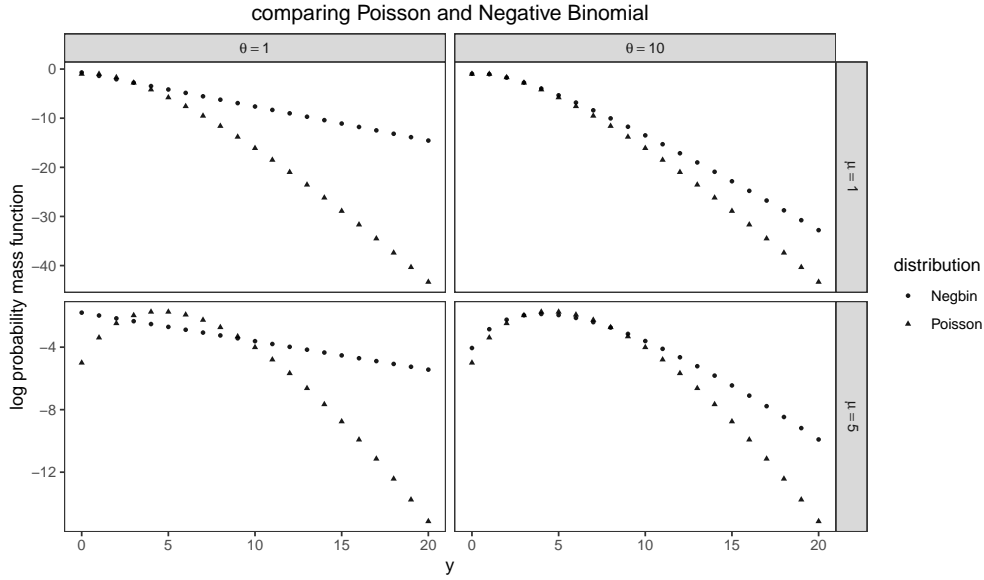


FIGURE 22.1: Comparing the log probabilities of the Poisson and Negative-Binomial with the same mean

22.1.3 Zero-inflated count distributions

Many count distributions have larger masses at zero compared to Poisson and Negative-Binomial. Therefore, it is also important to have more general distributions capturing this feature of empirical data. We can simply add an additional zero component to the Poisson or the Negative-Binomial.

A zero-inflated Poisson random variable y is a mixture of two components: a point mass at 0 and a $\text{Poisson}(\lambda)$ random variable, with probabilities p and $1 - p$, respectively. So y has the probability mass function

$$\text{pr}(y = k) = \begin{cases} p + (1 - p)e^{-\lambda}, & \text{if } k = 0, \\ (1 - p)e^{-\lambda} \frac{\lambda^k}{k!}, & \text{if } k = 1, 2, \dots \end{cases}$$

It has the first two moments below:

Proposition 22.4 *The zero-inflated Poisson random variable has the first two moments:*

$$E(y) = (1 - p)\lambda, \quad \text{var}(y) = (1 - p)\lambda(1 + p\lambda).$$

A zero-inflated Negative-Binomial random variable y is a mixture of two components: a point mass at zero and a $\text{NB}(\mu, \theta)$ random variable, with probabilities p and $1 - p$, respectively. So y has probability mass function

$$\text{pr}(y = k) = \begin{cases} p + (1 - p) \left(\frac{\theta}{\mu + \theta} \right)^\theta, & \text{if } k = 0, \\ (1 - p) \frac{\Gamma(k + \theta)}{\Gamma(k + 1)\Gamma(\theta)} \left(\frac{\theta}{\mu + \theta} \right)^\theta \left(\frac{\mu}{\mu + \theta} \right)^k, & \text{if } k = 1, 2, \dots \end{cases}$$

It has the first two moments below:

Proposition 22.5 *The zero-inflated Negative-Binomial random variable has the first two moments:*

$$E(y) = (1 - p)\mu, \quad \text{var}(y) = (1 - p)\mu(1 + \mu/\theta + p\mu).$$

I leave the proofs of Propositions 22.4 and 22.5 to Problem 22.4.

22.2 Regression models for counts

To model a count outcome y_i given x_i , we can still use OLS. However, a problem with OLS is that the predicted value can be negative. This can be easily fixed by running OLS of $\log(y_i + 1)$ given x_i . However, this still does not reflect the fact that y_i is a count outcome. For example, these two OLS fits cannot easily make a prediction for the probabilities $\text{pr}(y_i \geq 1 \mid x_i)$ or $\text{pr}(y_i > 3 \mid x_i)$. A more direct approach is to model the conditional distribution of y_i given x_i using the distributions reviewed in Section 22.1.

22.2.1 Poisson regression

I first discuss the Poisson regression model.

Assumption 22.1 (Poisson regression model) *We have*

$$y_i \mid x_i \sim \text{Poisson}(\lambda_i)$$

with

$$\lambda_i = \lambda(x_i, \beta) = e^{x_i^\top \beta}.$$

The observations are independent across units. The β is the unknown parameter.

Under Assumption 22.1, the mean and variance of $y_i \mid x_i$ are

$$E(y_i \mid x_i) = \text{var}(y_i \mid x_i) = e^{x_i^\top \beta}.$$

Because

$$\log E(y_i \mid x_i) = x_i^\top \beta,$$

this model is sometimes called the log-linear model, with the coefficient β_j interpreted as the conditional log mean ratio:

$$\log \frac{E(y_i \mid \dots, x_{ij} + 1, \dots)}{E(y_i \mid \dots, x_{ij}, \dots)} = \beta_j.$$

The likelihood function for independent Poisson random variables is³

$$L(\beta) = \prod_{i=1}^n e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!} \propto \prod_{i=1}^n e^{-\lambda_i} \lambda_i^{y_i},$$

and omitting the constants, we can write the log-likelihood function as

$$\log L(\beta) = \sum_{i=1}^n (-\lambda_i + y_i \log \lambda_i) = \sum_{i=1}^n \left(-e^{x_i^T \beta} + y_i x_i^T \beta \right).$$

The score function is

$$\begin{aligned} \frac{\partial \log L(\beta)}{\partial \beta} &= \sum_{i=1}^n \left(-x_i e^{x_i^T \beta} + x_i y_i \right) \\ &= \sum_{i=1}^n x_i \left(y_i - e^{x_i^T \beta} \right) \\ &= \sum_{i=1}^n x_i \{ y_i - \lambda(x_i, \beta) \}, \end{aligned}$$

and the Hessian matrix is

$$\begin{aligned} \frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^T} &= - \sum_{i=1}^n x_i \frac{\partial}{\partial \beta^T} \left(e^{x_i^T \beta} \right) \\ &= - \sum_{i=1}^n e^{x_i^T \beta} x_i x_i^T, \end{aligned}$$

which is negative semi-definite. When the Hessian is negative definite, the MLE is unique. The MLE must satisfy that

$$\sum_{i=1}^n x_i \left(y_i - e^{x_i^T \hat{\beta}} \right) = \sum_{i=1}^n x_i \{ y_i - \lambda(x_i, \hat{\beta}) \} = 0.$$

We can solve this nonlinear equation using Newton's method:

$$\begin{aligned} \beta^{\text{new}} &= \beta^{\text{old}} - \left\{ \frac{\partial^2 \log L(\beta^{\text{old}})}{\partial \beta \partial \beta^T} \right\}^{-1} \frac{\partial \log L(\beta^{\text{old}})}{\partial \beta} \\ &= \beta^{\text{old}} - (X^T W^{\text{old}} X)^{-1} X^T (Y - \Lambda^{\text{old}}), \end{aligned}$$

where

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and

$$\Lambda^{\text{old}} = \begin{pmatrix} \exp(x_1^T \beta^{\text{old}}) \\ \vdots \\ \exp(x_n^T \beta^{\text{old}}) \end{pmatrix}, \quad W^{\text{old}} = \text{diag} \{ \exp(x_i^T \beta^{\text{old}}) \}_{i=1}^n.$$

³The notation \propto means “proportional to.” Dropping those constants does not change the MLE.

Similar to the derivation for the logit model, we can simplify Newton's method to

$$\beta^{\text{new}} = (X^T W^{\text{old}} X)^{-1} X^T W^{\text{old}} Z^{\text{old}},$$

where

$$Z^{\text{old}} = X\beta^{\text{old}} + (W^{\text{old}})^{-1}(Y - \Lambda^{\text{old}}).$$

So we have an iterative reweighted least squares algorithm. In **R**, we can use the `glm` function with “family = poisson(link = “log”)” to fit the Poisson regression, which uses Newton's method.

Statistical inference under Poisson regression relies on the Normal approximation to the MLE:

$$\hat{\beta} \stackrel{\text{a}}{\sim} N \left\{ \beta, \left(-\frac{\partial^2 \log L(\hat{\beta})}{\partial \beta \partial \beta^T} \right)^{-1} \right\} = N \left\{ \beta, (X^T \hat{W} X)^{-1} \right\},$$

where $\hat{W} = \text{diag}\{\exp(x_i^T \hat{\beta})\}_{i=1}^n$.

After obtaining the MLE, we can predict the mean $E(y_i | x_i)$ by $\hat{\lambda}_i = e^{x_i^T \hat{\beta}}$. Because Poisson regression is a fully parametrized model, we can also predict any other probability quantities involving $y_i | x_i$. For example, we can predict $\text{pr}(y_i = 0 | x_i)$ by $e^{-\hat{\lambda}_i}$, and $\text{pr}(y_i \geq 3 | x_i)$ by $1 - e^{-\hat{\lambda}_i}(1 + \hat{\lambda}_i + \hat{\lambda}_i^2/2)$. To account for the uncertainty, we can use the delta method⁴ to approximate the standard errors of the predictors.

22.2.2 Negative-Binomial regression

I then discuss the Negative-Binomial regression.

Assumption 22.2 (Negative-Binomial regression model) *We have*

$$y_i | x_i \sim \text{NB}(\mu_i, \theta)$$

with

$$\mu_i = e^{x_i^T \beta}.$$

The observations are independent across units. The (β, θ) are the unknown parameters.

Under Assumption 22.2, the mean and variance of $y_i | x_i$ are

$$E(y_i | x_i) = e^{x_i^T \beta},$$

and

$$\text{var}(y_i | x_i) = e^{x_i^T \beta} (1 + e^{x_i^T \beta} / \theta).$$

It is also a log-linear model.

The log-likelihood function for Negative-Binomial regression is $\log L(\beta, \theta) = \sum_{i=1}^n l_i(\beta, \theta)$ with

$$\begin{aligned} l_i(\beta, \theta) &= \log \Gamma(y_i + \theta) - \log \Gamma(y_i + 1) - \log \Gamma(\theta) \\ &\quad + \theta \log \left(\frac{\theta}{\mu_i + \theta} \right) + y_i \log \left(\frac{\mu_i}{\mu_i + \theta} \right), \end{aligned}$$

where $\mu_i = e^{x_i^T \beta}$ has partial derivative $\partial \mu_i / \partial \beta = e^{x_i^T \beta} x_i = \mu_i x_i$. We can use Newton's algorithm or Fisher scoring algorithm to compute the MLE $(\hat{\beta}, \hat{\theta})$ which requires deriving

⁴Review Appendix C.3 if you are unfamiliar with it.

the first and second derivatives of $\log L(\beta, \theta)$ with respect to (β, θ) . I will derive some important components and relegate other details to Problem 22.3. First,

$$\frac{\partial \log L(\beta, \theta)}{\partial \beta} = \sum_{i=1}^n (1 + \mu_i/\theta)^{-1} (y_i - \mu_i) x_i.$$

The corresponding first-order condition can be viewed as the estimating equation of Poisson regression with weights $(1 + \mu_i/\theta)^{-1}$. Second,

$$\frac{\partial^2 \log L(\beta, \theta)}{\partial \beta \partial \theta} = \sum_{i=1}^n \frac{\mu_i}{(\mu_i + \theta)^2} (y_i - \mu_i) x_i.$$

We can verify

$$E \left\{ \frac{\partial^2 \log L(\beta, \theta)}{\partial \beta \partial \theta} \mid X \right\} = 0$$

since each term inside the summation has conditional expectation zero. This implies that the Fisher information matrix is diagonal, so $\hat{\beta}$ and $\hat{\theta}$ are asymptotically independent.

The `glm.nb` in the `MASS` package iterate between β and θ : given θ , update β based on Fisher scoring; given β , update θ based on Newton's algorithm. It reports standard errors based on the inverse of the Fisher information matrix.⁵

22.2.3 Zero-inflated regressions

I finally discuss the zero-inflated analogues of the Poisson and Negative-Binomial regressions.

Assumption 22.3 (zero-inflated Poisson regression model) *We have*

$$y_i \mid x_i \sim \begin{cases} 0, & \text{with probability } p_i, \\ \text{Poisson}(\lambda_i), & \text{with probability } 1 - p_i, \end{cases}$$

where

$$p_i = \frac{e^{x_i^\top \gamma}}{1 + e^{x_i^\top \gamma}}, \quad \lambda_i = e^{x_i^\top \beta}.$$

The observations are independent across units. The (γ, β) are the unknown parameters.

Assumption 22.4 (zero-inflated Negative-Binomial regression model) *We have*

$$y_i \mid x_i \sim \begin{cases} 0, & \text{with probability } p_i, \\ \text{NB}(\mu_i, \theta), & \text{with probability } 1 - p_i, \end{cases}$$

where

$$p_i = \frac{e^{x_i^\top \gamma}}{1 + e^{x_i^\top \gamma}}, \quad \mu_i = e^{x_i^\top \beta}.$$

The observations are independent across units. The (γ, β, θ) are the unknown parameters.

To avoid over-parametrization, we can also restrict some coefficients to be zero. The `zeroinfl` function in the `R` package `pscl` can fit the zero-inflated Poisson and Negative-Binomial regressions.

⁵The command `rnbreg` in `Stata` uses the BHHH algorithm by default, which may give slightly different numbers compared with `R`. The BHHH algorithm is similar to Newton's algorithm but avoids calculating the Hessian matrix.

22.3 A case study

I will use the dataset from Royer et al. (2015) to illustrate the regressions for count outcomes. From the regression formula below, I will estimate the effects of two treatments `incentive_commit` and `incentive` on the number of visits to the gym, controlling for two pre-treatment covariates `target` and `member_gym_pre`.

```
library("ggplot2")
library("gridExtra")
library("foreign")
library("MASS")
gym1 = read.dta("gym_treatment_exp_weekly.dta")
f.reg = weekly_visit ~ incentive_commit + incentive +
  target + member_gym_pre
```

22.3.1 Linear, Poisson, and Negative-Binomial regressions

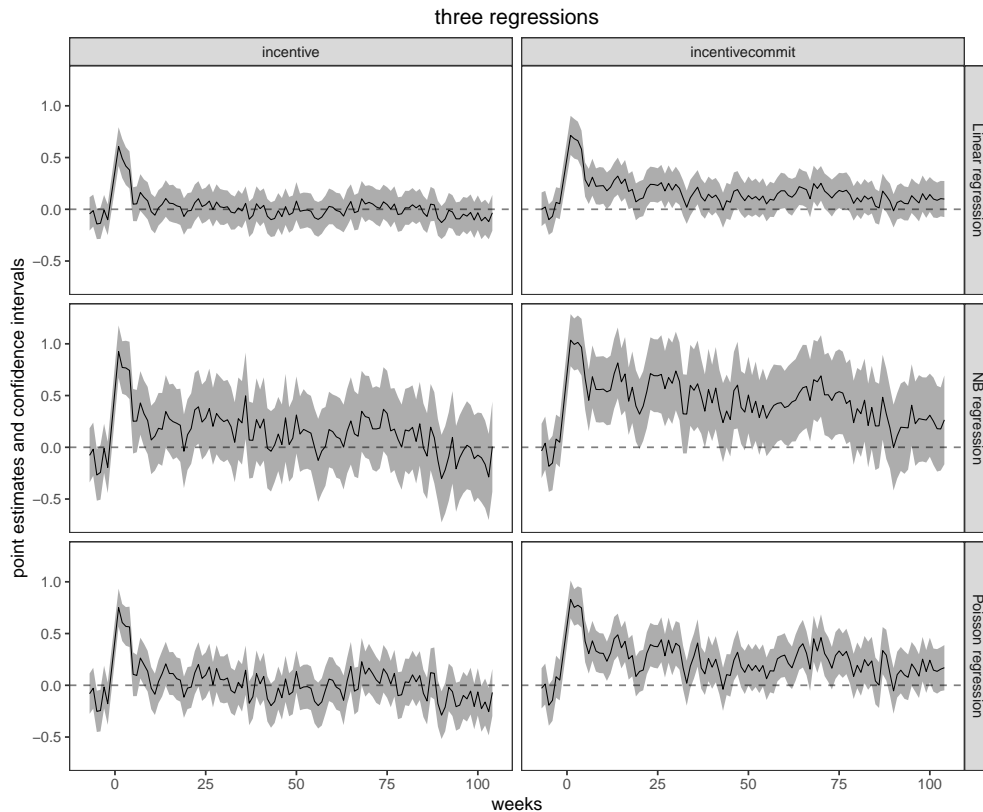


FIGURE 22.2: Linear, Poisson, and Negative-Binomial regressions

Each worker was observed over time. Therefore, I run regressions with the outcome data observed in each week. In the following, I compute the linear regression coefficients, standard errors, and AICs.

```
> weekids = sort(unique(gym1$incentive_week))
```

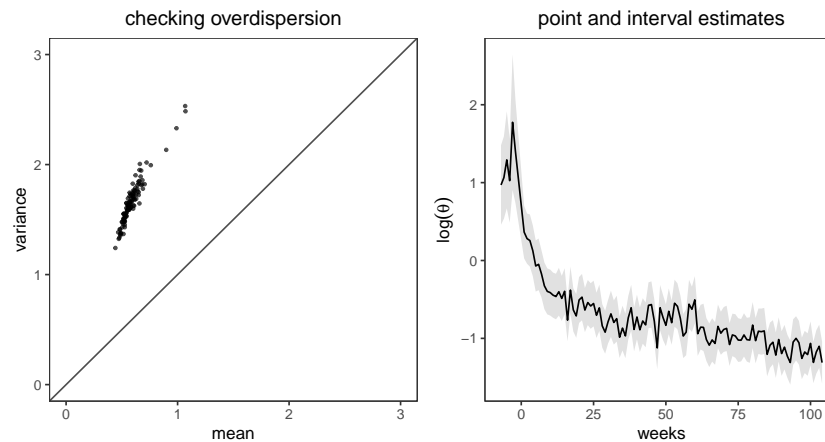


FIGURE 22.3: Overdispersion of the data

```

> lweekkids          = length(weekkids)
> coefincentivecommit = 1:lweekkids
> coefincentive       = 1:lweekkids
> seincentivecommit   = 1:lweekkids
> seincentive         = 1:lweekkids
> AIClm               = 1:lweekkids
> for(i in 1:lweekkids)
+ {
+   gymweek = gym1[which(gym1$incentive_week == weekkids[i]), ]
+   regweek = lm(f.reg, data = gymweek)
+   regweekcoef = summary(regweek)$coef
+
+   coefincentivecommit[i] = regweekcoef[2, 1]
+   coefincentive[i]       = regweekcoef[3, 1]
+   seincentivecommit[i]   = regweekcoef[2, 2]
+   seincentive[i]         = regweekcoef[3, 2]
+
+   AIClm[i]               = AIC(regweek)
+ }

```

By changing the line with `lm` to

```
regweek = glm(f.reg, family = poisson(link = "log"), data = gymweek)
```

and

```
regweek = glm.nb(f.reg, data = gymweek)
```

I obtain the corresponding results from Poisson and Negative-Binomial regressions, respectively. Figure 22.2 compares the regression coefficients with the associated confidence intervals over time. Three regressions give very similar patterns: `incentive_commit` has both short-term and long-term effects, but `incentive` only has short-term effects.

The left panel of Figure 22.3 shows that variances are larger than the means for outcomes from all weeks, and the right panel of Figure 22.3 shows the point estimates and confidence intervals of θ from Negative-Binomial regressions. Overall, overdispersion seems an important feature of the data.

22.3.2 Zero-inflated regressions

Figure 22.4 plots the histograms of the outcomes from four weeks before and four weeks after the experiment. Eight histograms all show severe zero inflation because most workers

just did not go to the gym regardless of the treatments. Therefore, it seems crucial to accommodate the zeros in the models.

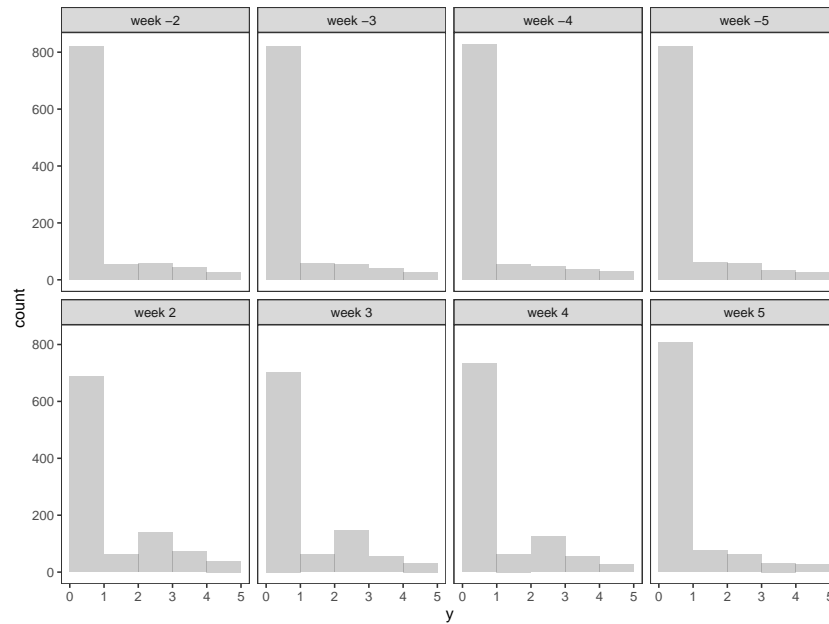


FIGURE 22.4: Zero-inflation of the data

I now fit zero-inflated Poisson regressions. The model has parameters for the zero component and parameters for the Poisson components.

```
> library("pscl")
> coefincentivecommit0 = coefincentivecommit
> coefincentive0       = coefincentive
> seincentivecommit0   = seincentivecommit
> seincentive0         = seincentive
> AIC0poisson          = AICnb
> for(i in 1:1weekkids)
+ {
+   gymweek = gym1[which(gym1$incentive_week == weekkids[i]), ]
+   regweek = zeroinfl(f.reg, dist = "poisson", data = gymweek)
+   regweekcoef = summary(regweek)$coef
+
+   coefincentivecommit[i] = regweekcoef$count[2, 1]
+   coefincentive[i]       = regweekcoef$count[3, 1]
+   seincentivecommit[i]   = regweekcoef$count[2, 2]
+   seincentive[i]         = regweekcoef$count[3, 2]
+
+   coefincentivecommit0[i] = regweekcoef$zero[2, 1]
+   coefincentive0[i]       = regweekcoef$zero[3, 1]
+   seincentivecommit0[i]   = regweekcoef$zero[2, 2]
+   seincentive0[i]         = regweekcoef$zero[3, 2]
+
+   AIC0poisson[i]         = AIC(regweek)
+ }
```

Replacing the line with `zeroinfl` by

```
regweek = zeroinfl(f.reg, dist = "negbin", data = gymweek)
```

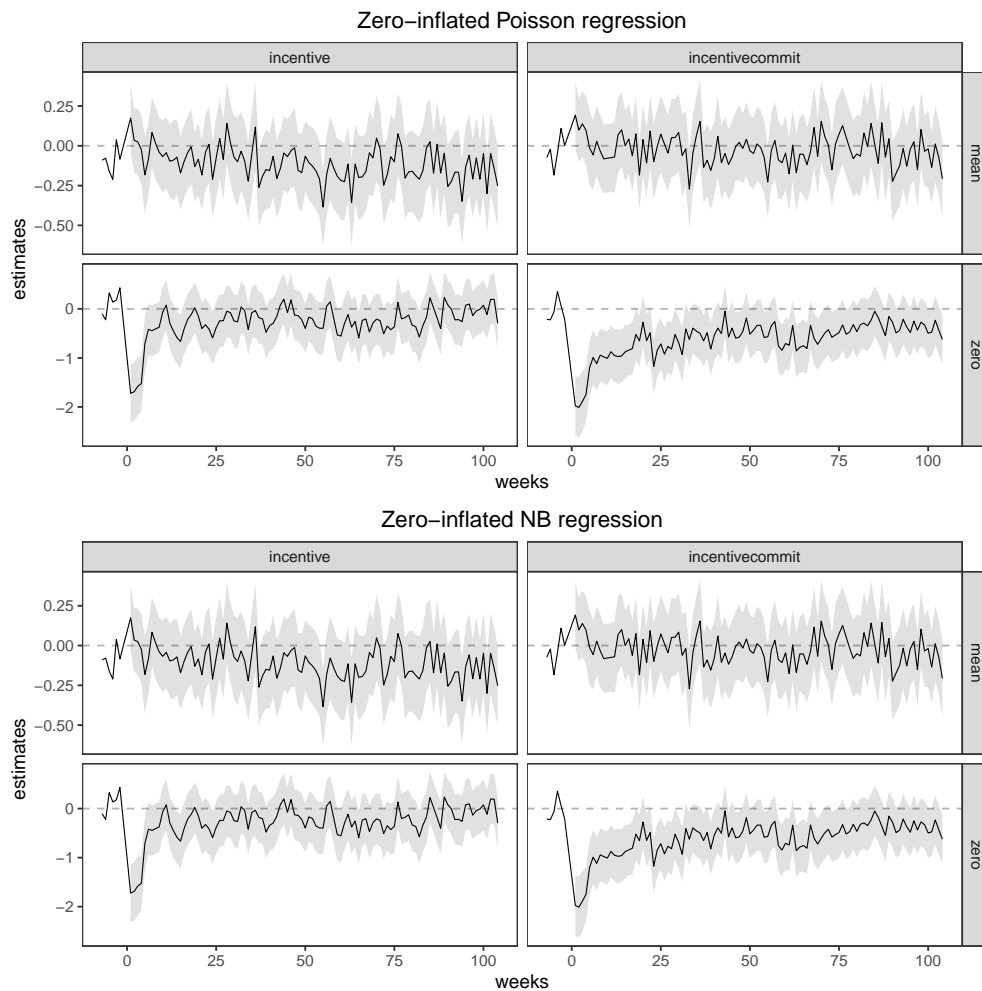


FIGURE 22.5: Zero-inflated regressions

we can fit the corresponding zero-inflated Negative-Binomial regressions. Figure 22.5 plots the point estimates and the confidence intervals of the coefficients of the treatment. It shows that the treatments do not have effects on the Poisson or Negative-Binomial components, but have effects on the zero components. This suggests that the treatments affect the outcome mainly by changing the workers' behavior of whether to go to the gym.

Another interesting result is the large $\hat{\theta}$'s from the zero-inflated Negative-Binomial regression:

```
> quantile(gymtheta, probs = c(0.01, 0.25, 0.5, 0.75, 0.99))
 1% 25% 50% 75% 99%
12.3 13.1 13.7 14.4 15.7
```

Once the zero-inflated feature has been modeled, it is not crucial to account for the overdispersion. It is reasonable because the maximum outcome is five, ruling out heavy-tailedness. This is further corroborated by the following comparison of the AICs from five regression models.

```
> diff.aic = AIC0nb - AIC0poisson
> quantile(diff.aic, probs = c(0.01, 0.25, 0.5, 0.75, 0.99))
```

1% 25% 50% 75% 99%
 2.000000 2.000013 2.000024 2.000031 2.002898

Figure 22.6 shows that zero-inflated Poisson regressions have the smallest AICs, beating the zero-inflated Negative-Binomial regressions, which are more flexible but have more parameters to estimate.

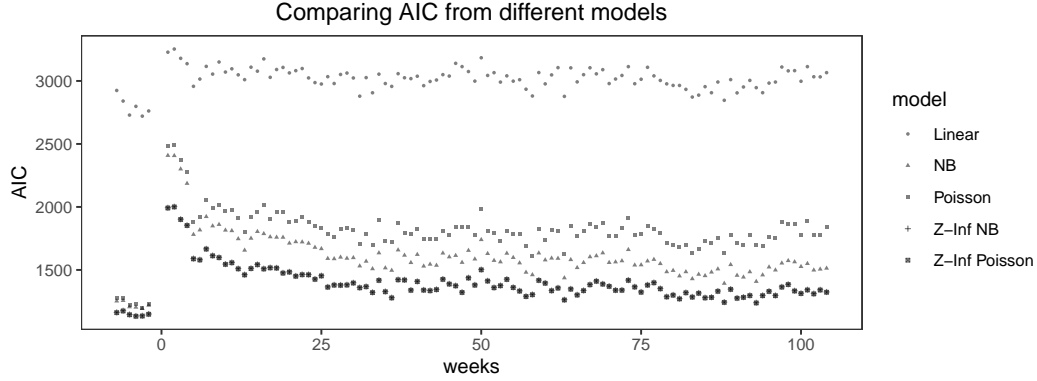


FIGURE 22.6: Comparing AICs from five regression models

22.4 Homework problems

22.1 Basic properties of Poisson

Prove Propositions 22.1 and 22.2.

22.2 Basic properties of Negative Binomial

Prove Proposition 22.3.

22.3 Newton's method for Negative-Binomial regression

Calculate the score function and Hessian matrix based on the log-likelihood function of the Negative-Binomial regression. What is the joint asymptotic distribution of the MLE $(\hat{\beta}, \hat{\theta})$?

22.4 Moments of Zero-inflated Poisson and Negative-Binomial

Prove Propositions 22.4 and 22.5.

22.5 Overdispersion and zero-inflation

Prove that for a zero-inflated Poisson, if $p \leq 1/2$ then $E(y) < \text{var}(y)$ always holds. What is the condition for $E(y) < \text{var}(y)$ when $p > 1/2$?

22.6 Poisson latent variable and the binary regression model with the cloglog link

Assume that $y_i^* | x_i \sim \text{Poisson}(e^{x_i^T \beta})$, and define $y_i = 1(y_i^* > 0)$ as the indicator that y_i^* is not zero.