Peng Ding

Linear Model and Extensions

Lecture notes for my students in Stat 230A at University of California, Berkeley

Contents

Acronyms

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

CLT	central limit theorem
CV	cross-validation
EHW	Eicker–Huber–White (robust standard error)
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
MLE	maximum likelihood estimate
OLS	ordinary least squares
RSS	residual sum of squares
WLS	weighted least squares

Symbols

β	regression coefficient
ε	error term
H	hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$
I_n	identity matrix of dimension $n \times n$
x_i	covariate vector for unit i
X	covariate matrix
Y	outcome vector
y_i	outcome for unit i
$\perp\!\!\!\perp$	independence and conditional independence

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	lalonde	LaLonde data
nnet	multinom	Multinomial logistic regression
quantreg	rq	quantile regression
survival	coxph	Cox proportional hazards regression
	survdiff	log rank test
	survfit	Kaplan–Meier curve

Part I

Introduction

1

Motivations for Statistical Models

1.1 Data and statistical models

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

Unit	outcome/response	design/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}

We often use

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

to denote the n-dimensional outcome or response 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 $n \times p$ design/covariate/feature/predictor matrix. In most cases, the first column of X contains constants 1s.

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

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

2. Prediction of Y^* based on new data X^*. 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 to predict whether an email is spam or not based on the frequencies of the most commonly occuring words and punctuation marks in the email? How to predict cancer patients survival time based on some clinical measures?

3. Causal effect of some components in X on Y. What if we change some components of X? How do we measure the impact of 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, FDA approves drugs based on randomized controlled trials because these trials 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 wage.

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

1. X and Y are univariate and continuous. In Francis Galton's[1] classic example, X is parents' average height and Y is children's average height (Galton, 1886). Galton derived the following formula:

$$y = \bar{y} + \hat{\rho}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}(x - \bar{x}) \Longleftrightarrow \frac{y - \bar{y}}{\hat{\sigma}_y} = r\frac{x - \bar{x}}{\hat{\sigma}_x},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the standard deviations, and $\hat{\rho}$ is the sample Pearson correlation coefficient:

$$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

This is the famous formula of "regression towards mediocrity" or "regression towards the mean". Galton first used the name "regression", and it is widely used in statistics now.

2. Y univariate and continuous, and X multivariate of mixed types. In the `R` package `ElemStatLearn`, the dataset `prostate` has an outcome of interest `lpsa` and some potential predictors are below

covariate name	meaning
lcavol	log cancer volume
lweight	log prostate weight
age	age

[1] Who 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.

3. Y binary or indicator of two classes, and X 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 the woman was in the labor force in 1975, and some useful covariates are

covariate name	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

4. Y categorical without ordering
5. Y categorical and ordered
6. Y counts, for example, the number of times one go to gym last week
7. Y time-to-event outcome. For example, survival time of patients and time to find the next job. the former is called survival analysis and the latter is called duration analysis.
8. Y multivariate and correlated.

1.2 Why linear models?

Why linear models, not nonlinear models?

1. Linear models are simple but non-trivial starting point for learning.
2. Linear models often provide more insights because we can derive explicit formulas based on elegant algebra and geometry.
3. Linear models can handle nonlinearity by incorporating nonlinear terms, for example, X can contain the polynomials or nonlinear transformations of the original covariates.
4. Linear models can be good approximations to nonlinear data generating processes.
5. Linear models are simpler than nonlinear models, but they do not necessarily perform worse than more complicated models. We have finite data, and simpler models do not suffer from overfitting.

If you are interested in nonlinear models, you can take another machine learning course.

2

Ordinary Least Squares with a Univariate Covariate

2.1 Univariate ordinary least squares

Figure 2.1 shows the scatterplot of Galton's data.

FIGURE 2.1: Galton's data

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

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

What do we mean by the "best" fit? Gauss proposed to use the following

criterion, called the ordinary least squares (OLS):

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

The OLS criterion is based on the squared terms of the "misfits" $y_i - \alpha - \beta x_i$. 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). We will discuss LAD later in this course.

How to solve the OLS minimization problem? The objective function is quadratic, and as α and β diverge, it diverges to infinity. So it must has a unique minimizer $(\hat{\alpha}, \hat{\beta})$ which satisfies the first order condition:

$$\begin{cases} -\frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) & = 0, \\ -\frac{2}{n} \sum_{i=1}^{n} x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) & = 0. \end{cases}$$

These two equations are called the Normal Equations of OLS. The first equation implies

$$\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}, \tag{2.1}$$

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

$$\overline{xy} = \hat{\alpha} \bar{x} + \hat{\beta} \overline{x^2}, \tag{2.2}$$

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 (2.1)$\times \bar{x}$ from (2.2), we have

$$\hat{\text{cov}}(x, y) = \hat{\beta} \hat{\text{var}}(x) \Longrightarrow \hat{\beta} = \frac{\hat{\text{cov}}(x, y)}{\hat{\text{var}}(x)},$$

so the OLS coefficient of x equals the sample covariance between x and y divided by the sample variance of x. From (2.1), we obtain that

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

Finally, the fitted line is

$$\begin{aligned} & y = \hat{\alpha} + \hat{\beta} x = \bar{y} - \hat{\beta} \bar{x} + \hat{\beta} x \\ & \Longrightarrow y - \bar{y} = \hat{\beta} (x - \bar{x}) \\ & \Longrightarrow y - \bar{y} = \frac{\hat{\text{cov}}(x, y)}{\hat{\text{var}}(x)} (x - \bar{x}) = \frac{\hat{\rho} \hat{\sigma}_x \hat{\sigma}_y}{\hat{\sigma}_x^2} (x - \bar{x}) \\ & \Longrightarrow \frac{y - \bar{y}}{\hat{\sigma}_y} = \hat{\rho} \frac{x - \bar{x}}{\hat{\sigma}_x}, \end{aligned}$$

which is the Galtonian formula mentioned in the last lecture.

2.2 Galton's data

```
> 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
```

2.3 Homework problems

2.1 Univariate OLS without intercept

Find

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

2.2 Pairwise slopes

Given $(x_i, y_i)_{i=1}^n$, show that Galton's slope equals

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

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

$$b_{ij} = (y_i - y_j)/(x_i - x_j)$$

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

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

is the weight proportional to the squared distance between x_i and x_j. In the above formulas, we define $b_{ij} = 0$ if $x_i = x_j$. Wu (1986) and Gelman and Park (2009) used this formula.

Part II

Ordinary least squares and statistical inference

3

Ordinary Least Squares with Multiple Covariates

3.1 The ordinary least squares formula

Recall that we have outcome vector

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

and covariate matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix} = (X_1, \ldots, X_p)$$

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

We want to find the best linear fit of the data

$$(x_{i,}\hat{y}_i = x_i^{\mathrm{T}}\hat{\beta})_{i=1}^n$$

in the sense that

$$\hat{\beta} = \arg\min_b n^{-1} \sum_{i=1}^n (y_i - x_i^{\mathrm{T}} b)^2 = \arg\min_b n^{-1} \|Y - Xb\|^2,$$

where $\hat{\beta}$ is called the ordinary least squares (OLS) coefficient, the $\hat{y}_i$'s are the fitted values, and the $y_i - \hat{y}_i$'s are called the residuals.

The objective function is quadratic in b which diverges to infinity when b diverges to infinity. So it has a unique minimizer $\hat{\beta}$ satisfying the first order condition

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

which simplifies to

$$\sum_{i=1}^{n} x_i(y_i - x_i^{\mathrm{T}}\hat{\beta}) = 0 \Longleftrightarrow X^{\mathrm{T}}(Y - X\hat{\beta}) = 0. \tag{3.1}$$

The above equation (3.1) is called the Normal equation of the OLS, which implies the main theorem:

Theorem 3.1 *The OLS coefficient equals*

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

if $X^{\mathrm{T}}X$ is non-degenerate.

The non-degeneracy of $X^{\mathrm{T}}X$ in Theorem 3.1 requires that for any non-zero vector $\alpha \in \mathbb{R}^p$,

$$\alpha^{\mathrm{T}}X^{\mathrm{T}}X\alpha = \|X\alpha\|^2 \neq 0 \Longleftrightarrow X\alpha \neq 0,$$

i.e., the columns of X are *linearly independent*. This effectively rules of redundant columns in the design matrix X. If X_1 can be represented by other columns $X_1 = c_2X_2 + \cdots + c_pX_p$ for some $(c_2, \ldots, c_p)$, then $X^{\mathrm{T}}X$ is degenerate.

3.2 The geometry of OLS

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: First, we have

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

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 (3.1) of the OLS:

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

FIGURE 3.1: The geometry of OLS

□

The formula of the decomposition looks like the Pythagorean Theorem, and indeed it is a general form of it with the geometry illustrated by Figure 3.1. For any $b = (b_1, \ldots, b_p)^{\mathrm{T}} \in \mathbb{R}^p$ and $X = (X_1, \ldots, X_p) \in \mathbb{R}^{n \times p}$, $Xb = b_1 X_1 + \cdots + b_p X_p$ 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 combination of the column vectors of X constitutes the column space of X, denoted by $\mathcal{C}(X)$. 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 $X_1, \ldots, X_p$. This geometric intuition in turn implies that

$$X_1^{\mathrm{T}}\hat{\varepsilon} = 0, \ldots, X_p^{\mathrm{T}}\hat{\varepsilon} = 0 \Longleftrightarrow X^{\mathrm{T}}\hat{\varepsilon} = 0$$

which is essentially the Normal equation (3.1):

$$X_1^{\mathrm{T}}(Y - X\hat{\beta}) = 0, \ldots, X_p^{\mathrm{T}}(Y - X\hat{\beta}) = 0 \Longleftrightarrow X^{\mathrm{T}}(Y - X\hat{\beta}) = 0.$$

The above argument gives an geometric derivation of the OLS formula in Theorem 3.1.

If X contains a column of intercepts $1_n = (1, \ldots, 1)^{\mathrm{T}}$, then

$$1_n^{\mathrm{T}} \hat{\varepsilon} = 0 \Longrightarrow n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i = 0,$$

so the residuals are automatically centered.

3.3 The projection matrix from OLS

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

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

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

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

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 matrix because

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

The rank equals the trace of a projection matrix, so the rank of H equals

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

The projection matrix H has the following geometric interpretations.

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

(G1) $Hv = v \Longleftrightarrow v \in \mathcal{C}(X)$;

(G2) $Hw = 0 \Longleftrightarrow w \perp \mathcal{C}(X)$.

Proof of Proposition 3.1: I first prove (1). If $v \in \mathcal{C}(X)$, then $v = Xb$ for some b, which implies that $Hv = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Xb = Xb = v$. Conversely, if $v = Hv$, then $v = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}v = Xu$ with $u = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}v$, which ensures that $v \in \mathcal{C}(X)$.

I then prove (2). If $w \perp \mathcal{C}(X)$, then w is orthogonal to all column vectors of X implying that

$$X_j^{\mathrm{T}}w = 0 \quad (j = 1, \ldots, p) \Longrightarrow X^{\mathrm{T}}w = 0 \Longrightarrow Hw = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}w = 0.$$

Conversely, if $Hw = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}w = 0$, then $w^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}w = 0$. Because $(X^{\mathrm{T}}X)^{-1}$ is positive definite, we have $X^{\mathrm{T}}w = 0$ ensuring that $w \perp \mathcal{C}(X)$. □

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

$$\hat{y}_i = \sum_{i'=1}^{n} h_{ii'}y_{i'} = h_{ii}y_i + \sum_{i'\neq i} h_{ii'}y_{i'}.$$

It shows that the predicted value $\hat{y}_i$ is a linear combination of all the responses. Moreover, if X contains a column of intercepts $1_n = (1, \ldots, 1)^{\mathrm{T}}$, then

$$H1_n = 1_n \Longrightarrow \sum_{i'=1}^{n} h_{ii'} = 1 \quad (i = 1, \ldots, n),$$

which implies that $\hat{y}_i$ is a weighted average of all the responses with the possibly negative weights sum to be one.

3.4 Homework problems

3.1 Univariate and multivariate OLS

Derive the univariate OLS based on the multivariate OLS formula.

3.2 OLS via vector and matrix calculus

Using vector and matrix calculus, show that the OLS estimator minimizes $(Y - Xb)^{\mathrm{T}}(Y - Xb)$.

3.3 Invariance of OLS

Assume that $X^{\mathrm{T}}X$ is non-degenerate and Γ is a $p\times p$ orthogonal matrix. Define $\tilde{X} = X\Gamma$. Give the formulas of the coefficient of $\tilde{X}$ and the fitted values in the OLS fit of Y on $\tilde{X}$. How do they depend on Γ?

3.4 Computation of OLS

Most software packages, for example, `R`, do not calculate the inverse of $X^{\mathrm{T}}X$ directly. Instead, they first find the reduced QR decomposition of X: $X = QR$ where Q is an $n \times p$ matrix with orthogonal columns and R is an $p \times p$ upper triangular matrix. Show that the OLS coefficient satisfies

$$R\hat{\beta} = Q^{\mathrm{T}}Y,$$

which can be easily solved by back substitution because R is upper triangular. Show that $H = QQ^{\mathrm{T}}$, so h_{ii} equals the squared length of the i-th row of Q.

3.5 OLS with multiple responses

For each unit $i = 1, \ldots, n$, we have multiple responses $y_i = (y_{i1}, \ldots, y_{iq})^{\mathrm{T}}$ and multiple covariates $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. Define

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix} = \begin{pmatrix} y_1^{\mathrm{T}} \\ \vdots \\ y_n^{\mathrm{T}} \end{pmatrix} = (Y_1, \ldots, Y_q)$$

and

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix} = (X_1, \ldots, X_p)$$

as the $n \times q$ response matrix and $n \times p$ covariate matrix, respectively. Define the multiple OLS matrix as

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

Show that $\hat{B} = (\hat{B}_1, \ldots, \hat{B}_q)$, where

$$\hat{B}_1 = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y_1, \ldots, \hat{B}_q = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y_q.$$

This result tells us that if the multiple OLS for a vector outcomes reduces to multiple independent OLS fits.

3.6 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 $X_k \in \mathbb{R}^{n_k \times p}$ and $Y_k \in \mathbb{R}^{n_k}$ are the covariate matrix and outcome vector in subsample k. Let $\hat{\beta}$ be the OLS coefficient based on the full sample, and $\hat{\beta}_k$ be the OLS coefficient based on subsample k. Show that $\hat{\beta} = \sum_{k=1}^{K} W_k \hat{\beta}_k$, where the weight matrix equals

$$W_k = \left(\sum_{k'=1}^{K} X_{k'}^{\mathrm{T}} X_{k'} \right)^{-1} X_k^{\mathrm{T}} X_k.$$

4

The Gauss–Markov Model and Theorem

4.1 Gauss–Markov model

Without any stochastic assumptions, the OLS in the last lecture is purely algebraic. From now on, we want to discuss the statistical properties of $\hat{\beta}$ and associated quantities, so we need to invoke some modeling assumptions. A simple starting point is the following Gauss–Markov model with a fixed design matrix X and unknown parameters (β, σ^2):

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

with the regularity condition that $X^{\mathrm{T}}X$ is non-degenerate and stochastic assumptions

$$E(\varepsilon) = 0, \quad \mathrm{cov}(\varepsilon) = \sigma^2 I_n.$$

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

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i, \qquad (i = 1, \ldots, n)$$

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

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

4.2 Properties of the OLS estimator

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

Theorem 4.1 *Under the Gauss–Markov model,*

$$E(\hat{\beta}) = \beta$$

and

$$\mathrm{cov}(\hat{\beta}) = \sigma^2 (X^{\mathrm{T}} X)^{-1}.$$

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

$$E(\hat{\beta}) = E\left\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\right\} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}E(Y) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\beta = \beta.$$

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

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

□

We can decompose the response vector as

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

where the fitted value vector is $\hat{Y} = X\hat{\beta} = HY$ and the residual vector is $\hat{\varepsilon} = Y - \hat{Y} = (I_n - H)Y$. 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. In particular, $HX = X$, $(I_n - H)X = 0$, and they are orthogonal:*

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

These follows from simple linear algebra, and I leave the proof of Lemma 4.1 as a homework problem. It states that H and $I_n - H$ are orthogonal 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 that

$$\hat{Y}^{\mathrm{T}}\hat{\varepsilon} = Y^{\mathrm{T}}H^{\mathrm{T}}(I_n - H)Y = Y^{\mathrm{T}}H(I_n - H)Y = 0.$$

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

Theorem 4.2 *Under the Gauss–Markov model,*

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

and

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

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

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

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

is a linear transformation of Y. It has mean

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

and covariance matrix

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

where the last step follows from Lemma 4.1. □

Although the original responses and error terms are uncorrelated between units with $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$, the fitted values and the residuals are correlated with $\operatorname{cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -h_{ij}$ 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 it 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 $\hat{\varepsilon}_i = y_i - x_i^{\mathrm{T}}\hat{\beta}$. It seems intuitive to estimate σ^2 by

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

where

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

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 original

ε_i. So RSS has mean

$$\sum_{i=1}^{n} \sigma^2(1 - h_{ii}) = \sigma^2 \{n - \text{trace}(H)\} = \sigma^2(n - p),$$

which implies the following theorem.

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 the Gauss–Markov model.*

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 large samples 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 and are there any 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. It is for this reason that I call the corresponding model the Gauss–Markov model.

Theorem 4.4 *Under the Gauss–Markov model, the OLS estimator* $\hat{\beta}$ *for* β *is the best linear unbiased estimator (BLUE) in the sense that* $\text{cov}(\hat{\beta}) \preceq \text{cov}(\tilde{\beta})$ *for any estimator* $\tilde{\beta}$ *satisfying*

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

(C2) $\tilde{\beta}$ *is unbiased for* β*.*

Before proving the theorem, we need to understand its meaning and immediate implications. We do not compare the OLS with any arbitrary estimators. In fact, we restrict to the estimators that are linear and unbiased. 1 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. 2 requires that $\tilde{\beta}$ is an unbiased estimator for β. More precisely, $E(\tilde{\beta}) = \beta$ holds for any value of β.

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 many modern applications with many covariates, some biased estimators can perform better than unbiased estimators if they have smaller variance. We will discuss these estimators later.

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

$$c^{\mathrm{T}}\mathrm{cov}(\hat{\beta})c \preceq c^{\mathrm{T}}\mathrm{cov}(\tilde{\beta})c \Longleftrightarrow \mathrm{var}(c^{\mathrm{T}}\hat{\beta}) \leq \mathrm{var}(c^{\mathrm{T}}\tilde{\beta}),$$

that is, any linear transformation of the OLS estimator has variance smaller than or equal to the same linear transformation of any other estimators. In particular, if $c = (0, \ldots, 1, \ldots, 0)^{\mathrm{T}}$ with only the jth coordinate being 1, then the above inequality implies that

$$\mathrm{var}(\hat{\beta}_j) \leq \mathrm{var}(\tilde{\beta}_j), \quad (j = 1, \ldots, p).$$

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

Now we prove the theorem.

Proof of Theorem 4.4: We must verify that the OLS estimator itself satisfies 1 and 2. We have $\hat{\beta} = \hat{A}Y$ with $\hat{A} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, and it is unbiased based on Theorem 4.1.

First, the unbiasedness requirement implies that

$$E(\tilde{\beta}) = \beta \Longrightarrow E(AY) = AE(Y) = AX\beta = \beta \Longrightarrow AX\beta = \beta$$

for any value of β. So $AX = I_p$ must hold. In particular, the OLS estimator satisfies $\hat{A}X = I_p$.

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

$$\begin{aligned} & \mathrm{cov}(\tilde{\beta}) \\ = \; & \mathrm{cov}(AY) \\ = \; & \mathrm{cov}(AY - \hat{A}Y + \hat{A}Y) \\ = \; & \mathrm{cov}\left\{(A - \hat{A})Y\right\} + \mathrm{cov}(\hat{A}Y) \\ & + \mathrm{cov}\left\{(A - \hat{A})Y, \hat{A}Y\right\} + \mathrm{cov}\left\{\hat{A}Y, (A - \hat{A})Y\right\}. \end{aligned}$$

The last two terms are in fact zero. By symmetry, we only need to show that

the last term is zero:

$$\begin{aligned}
& \operatorname{cov}\left\{\hat{A}Y, (A-\hat{A})Y\right\} \\
= \;& \hat{A}\operatorname{cov}(Y)(A-\hat{A})^{\mathrm{T}} \\
= \;& \sigma^2\hat{A}(A-\hat{A})^{\mathrm{T}} \\
= \;& \sigma^2(\hat{A}A^{\mathrm{T}} - \hat{A}\hat{A}^{\mathrm{T}}) \\
= \;& \sigma^2\left\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}A^{\mathrm{T}} - (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}\right\} \\
= \;& \sigma^2\left\{(X^{\mathrm{T}}X)^{-1}I_p - (X^{\mathrm{T}}X)^{-1}\right\} \qquad (\text{because } AX = I_p) \\
= \;& 0.
\end{aligned}$$

The above covariance decomposition simplifies to

$$\operatorname{cov}(\tilde{\beta}) = \operatorname{cov}\left\{(A-\hat{A})Y\right\} + \operatorname{cov}(\hat{A}Y) = \operatorname{cov}\left\{(A-\hat{A})Y\right\} + \operatorname{cov}(\hat{\beta}).$$

So we have

$$\operatorname{cov}(\tilde{\beta}) - \operatorname{cov}(\hat{\beta}) = \operatorname{cov}\left\{(A-\hat{A})Y\right\} = \operatorname{cov}(\tilde{\beta}-\hat{\beta}) \succeq 0,$$

which implies that $\operatorname{cov}(\hat{\beta}) \preceq \operatorname{cov}(\tilde{\beta})$. □

Theorem 4.4 is elegant but abstract. It says that in some sense, we can just 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, and have infinitely many linear unbiased estimators in general. A class of linear unbiased estimators that will be discussed more thoroughly later are the weighted least squares estimator

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

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

$$E(\tilde{\beta}) = E\left\{(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}Y\right\} = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}X\beta = \beta.$$

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.

I will give an extension and application of the Gauss–Markov Theorem as homework problems.

4.5 Homework problems

4.1 Projection matrices

Prove Lemma 4.1.

4.2 BLUE estimator for the mean

Assume that y_i has mean μ and σ^2 with y_i $(i = 1, \ldots, n)$ being uncorrelated. A linear estimator of the mean μ is $\hat{\mu} = \sum_{i=1}^{n} a_i y_i$, which is not unique. Find the best linear unbiased estimator for μ.

4.3 Gauss–Markov Theorem for prediction

Under the Gauss–Markov model, the OLS predictor $\hat{Y} = X\hat{\beta}$ for the mean $X\beta$ is the best linear unbiased predictor in the sense that $\mathrm{cov}(\hat{Y}) \preceq \mathrm{cov}(\tilde{Y})$ 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) $\tilde{Y}$ is unbiased for $X\beta$.

Prove this theorem.

4.4 Consequence of useless regressors

Partition the covariate matrix and parameter as

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

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 $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ and $\tilde{\beta}_1 = (X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}Y$ be the coefficient based on a partial OLS of Y on X_1 only. Show that $\mathrm{cov}(\hat{\beta}_1) \succeq \mathrm{cov}(\tilde{\beta}_1)$.

5

Gaussian/Normal Linear Model: Inference and Prediction

Under the Gauss–Markov model, we have shown that

$$E(\hat{\beta}) = \beta, \quad \text{cov}(\hat{\beta}) = \sigma^2 (X^{\mathrm{T}} X)^{-1}, \quad E(\hat{\sigma}^2) = \sigma^2.$$

Although these characterize the first two moments 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 a model with stronger assumptions. It will focus on the Gaussian/Normal linear model:

$$Y \sim \mathrm{N}(X\beta, \sigma^2 I_n) \iff y_i \overset{\text{IND}}{\sim} \mathrm{N}(x_i^{\mathrm{T}} \beta, \sigma^2), \qquad (i = 1, \ldots, n),$$

where X is fixed such that $X^{\mathrm{T}} X$ is non-degenerate, and (β, σ^2) are fixed but unknown parameters. We can equivalently write the model as a linear function of covariates with error terms:

$$Y = X\beta + \varepsilon \iff y_i = x_i^{\mathrm{T}} \beta + \varepsilon_i, \qquad (i = 1, \ldots, n),$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ or $\varepsilon_i \overset{\text{IID}}{\sim} \mathrm{N}(0, \sigma^2)$. The modeling assumption is extremely strong, but it is canonical in statistics. It allows us to derive elegant formulas, and also justifies the output of the linear regression function in many statistical packages.

5.1 Joint distribution of $(\hat{\beta}, \hat{\sigma}^2)$

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 the Gaussian linear model,*

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

so $\hat{\beta} \perp\!\!\!\perp \hat{\varepsilon}$; $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$, *and* $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$.

Proof of Theorem 5.1: First,

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

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

$$\mathrm{cov}(\hat{\beta}, \hat{\varepsilon}) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{cov}(Y)(I_n - H)^{\mathrm{T}} = \sigma^2(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(I_n - H^{\mathrm{T}}) = 0$$

which holds because $(I_n - H)X = 0$.

Second, because $\hat{\sigma}^2 = \mathrm{RSS}/(n-p) = \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}/(n-p)$ 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, which follows from the Normality of $\hat{\varepsilon}/\sigma$ with the projection matrix $I_n - H$ as its covariance matrix. □

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

Theorem 5.2 *Under the Gaussian linear model,*

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

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

Recall that we have shown that $Y = \hat{Y} + \hat{\varepsilon}$ with $\hat{Y} \perp \hat{\varepsilon}$ by the OLS properties. Now Theorem 5.2 further ensures that $\hat{Y} \perp\!\!\!\perp \hat{\varepsilon}$. The former says that $\hat{Y}$ and $\hat{\varepsilon}$ are orthogonal, which is a pure linear algebra fact without assumptions. The latter says that $\hat{Y}$ and $\hat{\varepsilon}$ are independent which is a statistical property under the Gaussian linear model.

5.2 Pivotal quantities and statistical inference

5.2.1 Scalar parameters

We first consider statistical inference for $c^{\mathrm{T}}\beta$, a one-dimensional linear function of β where $c \in \mathbb{R}^p$. For example, if $c = e_j \equiv (0, \ldots, 1, \ldots, 0)^{\mathrm{T}}$ with only the jth element being one, then $c^{\mathrm{T}}\beta = \beta_j$ is the jth 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 $c = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^{\mathrm{T}} = e_j - e_{j'}$.

Theorem 5.1 implies that

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

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

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

does not follow $\mathrm{N}(0,1)$ anymore. In fact, we can show that it is a t distribution.

Theorem 5.3 *Under the Gaussian linear model, 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

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

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

$$\begin{aligned} T_c = \frac{c^{\mathrm{T}}\hat{\beta} - c^{\mathrm{T}}\beta}{\sqrt{\hat{\sigma}^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}} &= \frac{c^{\mathrm{T}}\hat{\beta} - c^{\mathrm{T}}\beta}{\sqrt{\sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}} \Big/ \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \\ &\sim \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \\ &\sim t_{n-p}. \end{aligned}$$

□

In Theorem 5.1, the left-hand side depends on the observed data and the unknown true parameter, but the right-hand side is a random variable depending on only the dimension (n,p) of X, but neither the data nor the true parameter. 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

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

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 is close to that of $\mathrm{N}(0,1)$. For example, $t_{97.5\%,n-p} \approx 1.96$, the 97.5% quantile of $\mathrm{N}(0,1)$, which is the critical value for the 95% confidence interval.

We often call $\sqrt{\hat{\sigma}^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c} \equiv \hat{\mathrm{se}}_c$ the (estimated) standard error of

$c^{\mathrm{T}}\hat{\beta}$. Using this definition, we can equivalently write the above probability statement as

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

We use

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

as a $1-\alpha$ level confidence interval. By duality of confidence interval and hypothesis testing, we can also construct a level α test for $c^{\mathrm{T}}\beta$.

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

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 $l > 1$, then

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

correspond to the joint value of l parameters $c_1^{\mathrm{T}}\beta, \ldots,$ and $c_l^{\mathrm{T}}\beta$.

Example 5.1 *If*

$$C = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \Longrightarrow C\beta = \begin{pmatrix} \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

then $C\beta$ 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 $(\beta_2, \ldots, , \beta_p)$.

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

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

We will discuss this partitioned regression in more details later.

Now we will focus on the generic problem of inferring $C\beta$. To avoid degeneracy, we assume that C has linearly independent rows.

Theorem 5.1 implies that

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

and therefore the standardized quadratic form has a chi-squared distribution

$$(C\hat{\beta} - C\beta)^{\mathrm{T}} \left\{\sigma^2 C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}\right\}^{-1} (C\hat{\beta} - C\beta) \sim \chi^2_l.$$

The above chi-squared distribution follows from the property of the quadratic form of a Normal, but it requires that $\sigma^2 C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}$ is a positive definite matrix. This is true, but I relegate the technical details as a homework problem. 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 the following theorem.

Theorem 5.4 *Under the Gaussian linear model, for a fixed matrix $C \in \mathbb{R}^{l\times p}$, we have*

$$F_C \equiv \frac{(C\hat{\beta} - C\beta)^{\mathrm{T}} \left\{C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}\right\}^{-1} (C\hat{\beta} - C\beta)}{l\hat{\sigma}^2} \sim F_{l,n-p}.$$

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

$$\begin{aligned} F_C =& \frac{(C\hat{\beta} - C\beta)^{\mathrm{T}} \left\{\sigma^2 C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}\right\}^{-1} (C\hat{\beta} - C\beta)/l}{\hat{\sigma}^2/\sigma^2} \\ \sim& \frac{\chi^2_l/l}{\chi^2_{n-p}/(n-p)} \qquad (\text{where } \chi^2_l \perp\!\!\!\perp \chi^2_{n-p}) \\ \sim& F_{l,n-p}. \end{aligned}$$

□

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

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

where $f_{1-\alpha,l,n-p}$ is the upper α quantile of the $F_{l,n-p}$ distribution. By duality of 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. In 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,

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

with the same (β, σ^2).

First, we can predict the mean of y_{n+1} which is $x_{n+1}^{\mathrm{T}}\beta$. 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 $x_{n+1}^{\mathrm{T}}\hat{\beta}$ with $1-\alpha$ level prediction interval

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

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

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

and therefore

$$\begin{aligned}\frac{y_{n+1} - x_{n+1}^{\mathrm{T}}\hat{\beta}}{\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 x_{n+1}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_{n+1}}} &= \frac{y_{n+1} - x_{n+1}^{\mathrm{T}}\hat{\beta}}{\sqrt{\sigma^2 + \sigma^2 x_{n+1}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_{n+1}}} \Big/ \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \\ &\sim \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \\ &\sim t_{n-p}\end{aligned}$$

is a pivotal quantity. The squared prediction error

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

has two components. The first one has magnitude close to σ^2 which is a constant order. The second one has magnitude decreasing in n if $n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$ 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 $\hat{\mathrm{pe}}_{x_{n+1}}$, we can construct the following $1-\alpha$ level prediction interval:

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

5.4 Examples and R implementation

5.4.1 Univariate regression

OLS fit

```
> library("HistData")
> galton_fit = lm(childHeight ~ midparentHeight,
+                 data = GaltonFamilies)
> round(summary(galton_fit)$coef, 3)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      22.636      4.265   5.307        0
midparentHeight   0.637      0.062  10.345        0
```

predictions

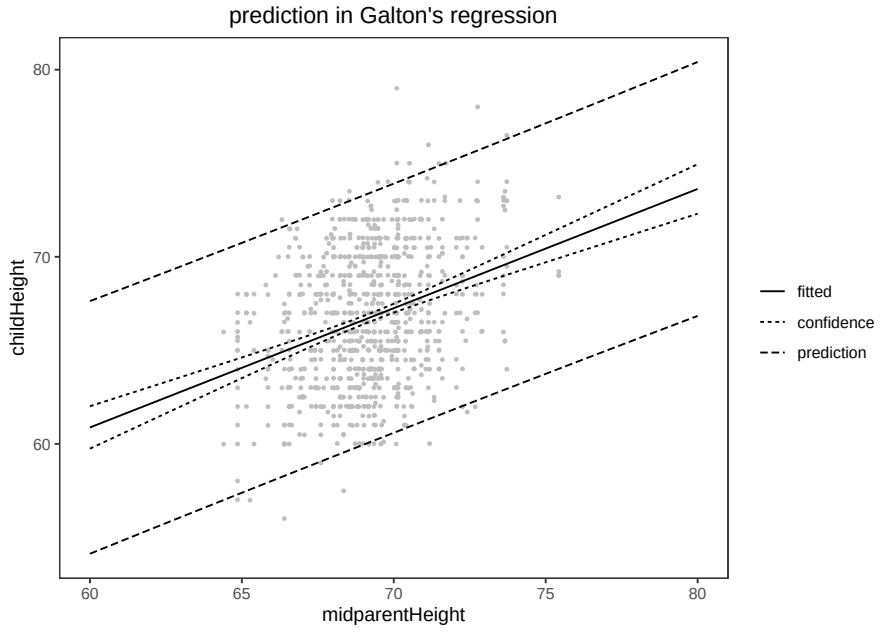
```
> new_mph  = seq(60, 80, by = 0.5)
> new_data = data.frame(midparentHeight = new_mph)
> new_ci   = predict(galton_fit, new_data,
+                    interval = "confidence")
> new_pi   = predict(galton_fit, new_data,
+                    interval = "prediction")
> round(head(cbind(new_ci, new_pi)), 3)
     fit    lwr    upr    fit    lwr    upr
1 60.878 59.744 62.012 60.878 54.126 67.630
2 61.197 60.122 62.272 61.197 54.454 67.939
3 61.515 60.499 62.531 61.515 54.782 68.249
4 61.834 60.877 62.791 61.834 55.109 68.559
5 62.153 61.254 63.051 62.153 55.436 68.869
6 62.471 61.632 63.311 62.471 55.762 69.180
```

5.4.2 Multivariate regression

OLS fit

```
> library("Matching")
> data(lalonde)
> ## OLS fit
> lalonde_fit = lm(re78 ~ ., data = lalonde)
> round(summary(lalonde_fit)$coef, 3)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   256.670   3521.682   0.073    0.942
age            53.571     45.806   1.170    0.243
educ          400.770    228.822   1.751    0.081
black       -2037.333   1173.716  -1.736    0.083
hisp          425.819   1564.553   0.272    0.786
married      -146.329    882.257  -0.166    0.868
nodegr        -15.179   1005.885  -0.015    0.988
re74            0.123      0.088   1.405    0.161
re75            0.020      0.150   0.131    0.896
u74          1380.285   1187.917   1.162    0.246
u75         -1071.215   1024.878  -1.045    0.297
treat        1670.709    641.132   2.606    0.009
```

Joint testing




```
> 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 *
---
Signif. codes:
```

```
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

counterfactual prediction

```
> ## predictions: confidence and prediction intervals
> 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
```

5.5 Homework problems

5.1 MLE

Under the Gaussian linear model, show that the maximum likelihood estimator (MLE) for β is the OLS estimator, but the MLE 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.

5.2 MLE with Laplace errors

Assume that $y_i = x_i^{\mathrm{T}}\beta + \sigma\varepsilon_i$ where the ε_i's are i.i.d. Laplace distribution with density $f(\varepsilon) = 2^{-1}e^{-|\varepsilon|}$ $(i = 1, \ldots, n)$. Find the MLEs for (β, σ^2).

5.3 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 under which the rows of X_{n+1} are not linearly independent.

5.4 Analysis of Variance (ANOVA) with multi-level treatment

Let x_i be the indicator vector for J treatment levels in a completely randomized experiment, for example, $x_i = e_j = (0, \ldots, 1, \ldots, 0)^{\mathrm{T}}$ with the jth element being one if unit i receives treatment level j $(j = 1, \ldots, J)$. Let y_i be the outcome of unit i $(i = 1, \ldots, n)$. 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

$$H_0 : \beta_1 = \cdots = \beta_J$$

in the Gaussian linear model $Y = X\beta + \varepsilon$ assuming $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$. This is a special case of testing $C\beta = 0$. Find C and show that the 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 the proof.

5.5 An application

The `R` package `sampleSelection` 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 is `lnmeddol` which denotes the log of medical expenses. Use linear regression to investigate the relationship between the outcome and various important covariates. The solution of this problem is not unique, but do justify your choice of covariates and model, and interpret the results.

5.6 A technical detail

Verify that $C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}$ is positive definite if the columns of X are linearly independent and the rows of C are linearly independent.

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

Show that when C containing only one row c^{T}, then $t_c^2 = F_C$.

6

The Frisch–Waugh–Lovell Theorem

6.1 Long and short regressions

If we partition X and β into

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

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

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

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 $\hat{\varepsilon}$ and $\tilde{\varepsilon}$ are the residual vectors from the long and short regressions. These two regressions are of great interest in practice. For example, we can at least ask the following questions:

(Q1) if the true β_1 is zero, then what is the consequence of including it in the long regression?

(Q2) if the true β_1 is not zero, then what is the consequence of omitting it 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 gives us any information about β_1?

Many problems in statistics are related to the long and short regressions. We will discuss application in next chapter.

6.2 The main theorem

The following theorem helps to answer these questions.

Theorem 6.1 *The OLS estimator in the short regression is* $\tilde{\beta}_2 = (X_2^{\mathrm{T}} X_2)^{-1} X_2^{\mathrm{T}} Y$*, and the OLS estimator for* β_2 *in the long regression has the following equivalent forms*

$$\hat{\beta}_2 = \left[(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\right]_{\text{last } l \text{ elements}} \tag{6.1}$$

$$= \left\{X_2^{\mathrm{T}}(I_n - H_1)X_2\right\}^{-1} X_2^{\mathrm{T}}(I_n - H_1)Y \quad \textit{where } H_1 = X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1 \tag{6.2}$$

$$= (\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}}Y \quad \textit{where } \tilde{X}_2 = (I_n - H_1)X_2 \tag{6.3}$$

$$= (\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}}\tilde{Y} \quad \textit{where } \tilde{Y} = (I_n - H_1)Y. \tag{6.4}$$

This result 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 (Yule, 1907; Fisher, 1925; Cochran, 1938). A similar form is called Cochran's formula, which will appear as a homework problem.

Before proving Theorem 6.1, I will first discuss its meanings and interpretations. Equation (6.1) follows from the definition of the OLS coefficient. The matrix $I_n - H_1$ in equation (6.2) is the projection matrix onto the space orthogonal to the column space of X_1. Equation (6.3) 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 of X_2 on X_1. 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 (6.4) is a slight modification of equation (6.3), 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 of Y on X_1. From (6.3) and (6.4), it is not crucial to residualize Y or not, but it is crucial to residualize X_2.

The forms (6.3) and (6.4) suggest the interpretation of $\hat{\beta}_2$ as the "impact" of X_2 on Y holding X_1 constant, or in econometric terms, the "impact" of X_2 on Y *ceteris paribus*. The English meaning of this Latin phase is "with other conditions remaining the same." However, taking this interpretation too serious is problematic because Theorem 6.1 is a pure algebraic result without any distributional assumptions. We cannot hold X_1 constant using pure algebra. Sometimes, we can manipulate the value of X_1 in an experimental setting, but this relies on the data collecting process.

There are many ways to prove Theorem 6.1. Below I take a detour to give a unnecessarily complicated proof because some intermediate steps will be useful for later parts of the book. I will give some simpler proofs in the homework problems. The following proof relies on a lemma.

Lemma 6.1 *The inverse of* $X^{\mathrm{T}}X$ *is*

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

where

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

I leave the proof of Lemma 6.1 as a homework problem. With Lemma 6.1, we can easily prove Theorem 6.1.

Proof of Theorem 6.1: The OLS coefficient is

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

Then using Lemma 6.1, we can simplify $\hat{\beta}_2$ as

$$
\begin{aligned}
\hat{\beta}_2 &= S_{21}X_1^{\mathrm{T}}Y + S_{22}X_2^{\mathrm{T}}Y \\
&= -(\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}X_2^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}Y + (\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}X_2^{\mathrm{T}}Y \\
&= -(\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}X_2^{\mathrm{T}}H_1Y + (\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}X_2^{\mathrm{T}}Y \\
&= (\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}X_2^{\mathrm{T}}(I_n - H_1)Y \qquad (6.5)\\
&= (\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}}Y. \qquad (6.6)
\end{aligned}
$$

Equation (6.5) is the form (6.2), and equation (6.6) is the form (6.3). Because we also have $X_2^{\mathrm{T}}(I_n - H_1)Y = X_2^{\mathrm{T}}(I_n - H_1)^2Y = \tilde{X}_2^{\mathrm{T}}\tilde{Y}$, we can write $\hat{\beta}_2$ as $\hat{\beta}_2 = (\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}}\tilde{Y}$, giving the form (6.4). □

Corollary 6.1 *If $X_1^{\mathrm{T}}X_2 = 0$, 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 6.1: Corollary 6.1 follows directly from

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

and Theorem 6.1. □

When the columns of X_1 and X_2 are orthogonal, adding X_1 or not in the regression does not change the value of the OLS coefficient of X_2. So the long regression and short regression are equivalent with orthogonal covariates.

6.3 A numerical example

We first generate data and compute the coefficients from the long regression.

```
> n   = 100
> X1  = rnorm(n)
> X2  = 0.5*X1 + rnorm(n)
> Y   = 1 + X1 + X2 + rnorm(n)
> lm(Y ~ X1 + X2)$coef
(Intercept)          X1          X2
  1.0097883   0.8813489   1.0133137
```

We can see that the coefficient of X_2 equals those from two partial regressions below:

```
> reg2.1 = lm(X2 ~ X1)
> regY.1 = lm(Y ~ X1)
> lm(Y ~ 0 + I(reg2.1$residual))$coef
I(reg2.1$residual)
          1.013314
> lm(I(regY.1$residual) ~ 0 + I(reg2.1$residual))$coef
I(reg2.1$residual)
          1.013314
```

Moreover, it is does matter whether we include the intercept or not in the partial regressions because the residuals are centered.

```
> lm(Y ~ I(reg2.1$residual))$coef
       (Intercept) I(reg2.1$residual)
         0.8217285          1.0133137
> lm(I(regY.1$residual) ~ I(reg2.1$residual))$coef
       (Intercept) I(reg2.1$residual)
      5.536334e-17       1.013314e+00
```

6.4 Homework problems

6.1 Inverse of a block matrix

Prove Lemma 6.1 and the following alternative form:

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

where $H_2 = X_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)^{-1}X_2$, $\tilde{X}_1 = (I_n - H_2)X_1$, and

$$\begin{aligned} Q_{11} &= (\tilde{X}_1^{\mathrm{T}}\tilde{X}_1)^{-1}, \\ Q_{12} &= -(\tilde{X}_1^{\mathrm{T}}\tilde{X}_1)^{-1}X_1^{\mathrm{T}}X_2(X_2^{\mathrm{T}}X_2)^{-1}, \\ Q_{21} &= Q_{12}^{\mathrm{T}}, \\ Q_{22} &= (X_2^{\mathrm{T}}X_2)^{-1} + (X_2^{\mathrm{T}}X_2)^{-1}X_2^{\mathrm{T}}X_1(\tilde{X}_1^{\mathrm{T}}\tilde{X}_1)^{-1}X_1^{\mathrm{T}}X_2(X_2^{\mathrm{T}}X_2)^{-1}. \end{aligned}$$

6.2 Another (simpler) proof of the FWL Theorem

From the OLS decomposition of the long regression $\hat{Y} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}$, first show that

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

then show that

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

The FWL Theorem follows immediately. The FWL Theorem implicitly uses the fact that $X_2^{\mathrm{T}}(I_n - H_1)X_2$ is invertible. Why is it true?

6.3 Residuals in the FWL Theorem

Based on the FWL Theorem, we can obtain $\hat{\beta}_2$ from (6.3) or (6.4). Do the residuals from the partial regressions (6.3) and (6.4) equal to $\hat{\varepsilon}$?

6.4 Multivariate regression via univariate regressions

The FWL Theorem states that the OLS coefficient in the long regression can be obtained from several short regressions. Consider the most extreme case, if you only know how to compute univariate regressions, how can you compute $\hat{\beta}_j$, the j-th coordinate in the long regression?

Hint: The coefficient in the OLS fit of a vector a on a vector b equals $a^{\mathrm{T}}b/b^{\mathrm{T}}b$.

6.5 The sample version of 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 assume any randomness. We can fit the following OLS:

$$\begin{aligned} Y &= X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}, \\ Y &= X_2\tilde{\beta}_2 + \tilde{\varepsilon}, \\ X_1 &= X_2\hat{\delta} + \hat{U}, \end{aligned}$$

where $\hat{\varepsilon}, \tilde{\varepsilon}, \hat{U}$ are the residuals. The last OLS fit means the OLS fit of each column of X_1 on X_2, and therefore the corresponding residual $\hat{U}$ is an $n \times k$ matrix. Show that

$$\tilde{\beta}_2 = \hat{\beta}_2 + \hat{\delta}\hat{\beta}_1.$$

Note that this is a pure linear algebra fact similar to the FWL Theorem.

7

Applications of the Frisch–Waugh–Lovell Theorem

The FWL theorem has many applications, and I will highlight some in this chapter.

7.1 Centering regressors

As a special case, partition $X = (X_1, X_2)$ with $X_1 = 1_n$. The projection matrix

$$H_1 = 1_n(1_n^{\mathrm{T}}1_n)^{-1}1_n = n^{-1}1_n1_n^{\mathrm{T}} = \begin{pmatrix} n^{-1} & \cdots & n^{-1} \\ \vdots & & \vdots \\ n^{-1} & \cdots & n^{-1} \end{pmatrix} \equiv J_n$$

contains n^{-1}'s as its elements, and $C_n = I_n - n^{-1}1_n1_n^{\mathrm{T}}$ is the projection matrix orthogonal to 1_n. Multiplying any vector by J_n is equivalent to obtaining its mean, and multiplying any vector by C_n is equivalent to centering that vector, for example,

$$J_nY = \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} = \bar{y}1_n, \quad C_nY = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}.$$

More generally, multiplying any matrix by J_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,

$$J_nX_2 = \begin{pmatrix} \bar{x}_2^{\mathrm{T}} \\ \vdots \\ \bar{x}_2^{\mathrm{T}} \end{pmatrix} = 1_n\bar{x}_2^{\mathrm{T}}, \quad C_nX_2 = \begin{pmatrix} x_{12}^{\mathrm{T}} - \bar{x}_2^{\mathrm{T}} \\ \vdots \\ x_{n2}^{\mathrm{T}} - \bar{x}_2^{\mathrm{T}} \end{pmatrix},$$

where X_2 contains row vectors $x_{12}, \ldots, x_{n2}$ 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_nX_2 in the OLS fit of C_nY on C_nX_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: its quadratic form equals the sample variance multiplied by $n-1$, for example,

$$\begin{aligned} Y^{\mathrm{T}} C_n Y &= Y^{\mathrm{T}}(I_n - n^{-1}1_n 1_n^{\mathrm{T}})Y \\ &= (y_1 - \bar{y}, \ldots, y_n - \bar{y}) \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)\hat{\sigma}_y^2, \end{aligned}$$

where $\hat{\sigma}_y^2$ is the sample variance of Y. For an $n \times p$ matrix X,

$$\begin{aligned} X^{\mathrm{T}} C_n X &= \begin{pmatrix} X_1^{\mathrm{T}} \\ \vdots \\ X_p^{\mathrm{T}} \end{pmatrix} C_n \begin{pmatrix} X_1 & \cdots & X_p \end{pmatrix} \\ &= \begin{pmatrix} X_1^{\mathrm{T}} C_n X_1 & \cdots & X_1^{\mathrm{T}} C_n X_p \\ \vdots & & \vdots \\ X_p^{\mathrm{T}} C_n X_1 & \cdots & X_p^{\mathrm{T}} C_n X_p \end{pmatrix} \\ &= (n-1) \begin{pmatrix} \hat{\sigma}_{11} & \cdots & \hat{\sigma}_{1p} \\ \vdots & & \vdots \\ \hat{\sigma}_{p1} & \cdots & \hat{\sigma}_{pp} \end{pmatrix}, \end{aligned}$$

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 $(n-1)^{-1} X^{\mathrm{T}} C_n X$ equals the sample covariance matrix of X. For these reason, I choose the notation C_n with "C" for both "centering" and "covariance."

In another important special case, X_1 contains the dummies for a discrete variable, for example, the indicators for different treatment levels or groups.

With k groups, X_1 can take the following two forms:

$$X_1 = \begin{pmatrix} 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & & 1 \\ 1 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}_{n \times k} \quad \text{or} \quad X_1 = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{n \times k}, \quad (7.1)$$

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. From the forms of X_1, the observations are sorted according to the group indicators. If we regress Y on X_1, the residual vector is

$$Y - \begin{pmatrix} \bar{y}_{[1]} \\ \vdots \\ \bar{y}_{[1]} \\ \vdots \\ \bar{y}_{[k]} \\ \vdots \\ \bar{y}_{[k]} \end{pmatrix}, \quad (7.2)$$

where $\bar{y}_{[1]}, \ldots, \bar{y}_{[k]}$ are the averages of the outcome within groups $1, \ldots, k$. 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 is 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.

7.2 Partial correlation coefficient and Simpson's paradox

The 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 Pearson correlation coefficient based on the residuals from 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 partial correlation coefficient between x and y given w as

$$\hat{\rho}_{yx|w} = \frac{\sum_{i=1}^n \hat{\varepsilon}_{x,i}\hat{\varepsilon}_{y,i}}{\sqrt{\sum_{i=1}^n \hat{\varepsilon}_{x,i}^2}\sqrt{\sum_{i=1}^n \hat{\varepsilon}_{y,i}^2}}.$$

In the above definition, we do not center the residuals because they have zero sample means due to the inclusions of the intercept in the OLS fits 1 and 2. The partial correlation coefficient determines the OLS coefficient of $\hat{\varepsilon}_y$ on $\hat{\varepsilon}_x$:

$$\hat{\beta}_{yx|w} = \frac{\sum_{i=1}^n \hat{\varepsilon}_{x,i}\hat{\varepsilon}_{y,i}}{\sum_{i=1}^n \hat{\varepsilon}_{x,i}^2} = \hat{\rho}_{yx|w}\sqrt{\frac{\sum_{i=1}^n \hat{\varepsilon}_{y,i}^2}{\sum_{i=1}^n \hat{\varepsilon}_{x,i}^2}} = \hat{\rho}_{yx|w}\frac{\hat{\sigma}_{y|w}}{\hat{\sigma}_{x|w}}, \tag{7.3}$$

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 1 and 2. Equation (7.3) is the Galtonian formula for multiple regression. Based on the FWL Theorem, $\hat{\beta}_{yx|w}$ is also the OLS coefficient of X in the long regression of Y on $(1, X, W)$.

To investigate the relationship between y and x, different researchers may run different regressions. One may run OLS of Y on X and $(1, W)$, and the other may run OLS of Y on X and $(1, W')$, 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*. 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 an empty set. Simpson's paradox is then equivalent to

$$\hat{\rho}_{yx|w}\hat{\rho}_{yx} < 0.$$

The following theorem gives an expression linking $\hat{\rho}_{yx|w}$ and $\hat{\rho}_{yx}$.

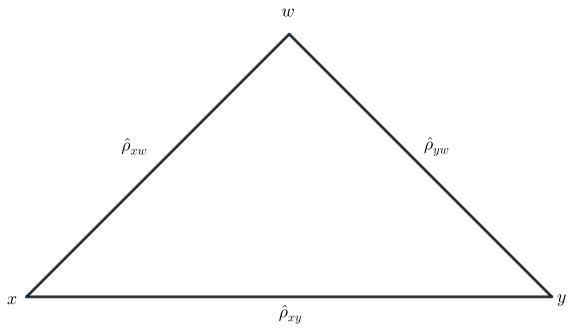


FIGURE 7.1: Correlations among three variables

Theorem 7.1 *For $Y, X, W \in \mathbb{R}^n$,*

$$\hat{\rho}_{yx|w} = \frac{\hat{\rho}_{yx} - \hat{\rho}_{yw}\hat{\rho}_{xw}}{\sqrt{1-\hat{\rho}_{yw}^2}\sqrt{1-\hat{\rho}_{xw}^2}}.$$

Its proof is purely algebraic, so I leave it as a homework problem.

Based on data $(y_i, x_i, w_i)_{i=1}^n$, we can compute the sample correlation matrix

$$\hat{R} = \begin{pmatrix} 1 & \hat{\rho}_{yx} & \hat{\rho}_{yw} \\ \hat{\rho}_{xy} & 1 & \hat{\rho}_{xw} \\ \hat{\rho}_{wy} & \hat{\rho}_{wx} & 1 \end{pmatrix},$$

which is symmetric and positive semi-definite. Since $\hat{R}$ is 3×3 and the $\hat{\rho}$'s are smaller than or equal to one, its positive semi-definiteness is equivalent to $\det(\hat{R}) \geq 0$. With this only constraint, Simpson's paradox happens if and only if

$$\hat{\rho}_{yx}(\hat{\rho}_{yx} - \hat{\rho}_{yw}\hat{\rho}_{xw}) < 0 \Longleftrightarrow \hat{\rho}_{yx}^2 < \hat{\rho}_{yx}\hat{\rho}_{yw}\hat{\rho}_{xw}.$$

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 7.2. 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.

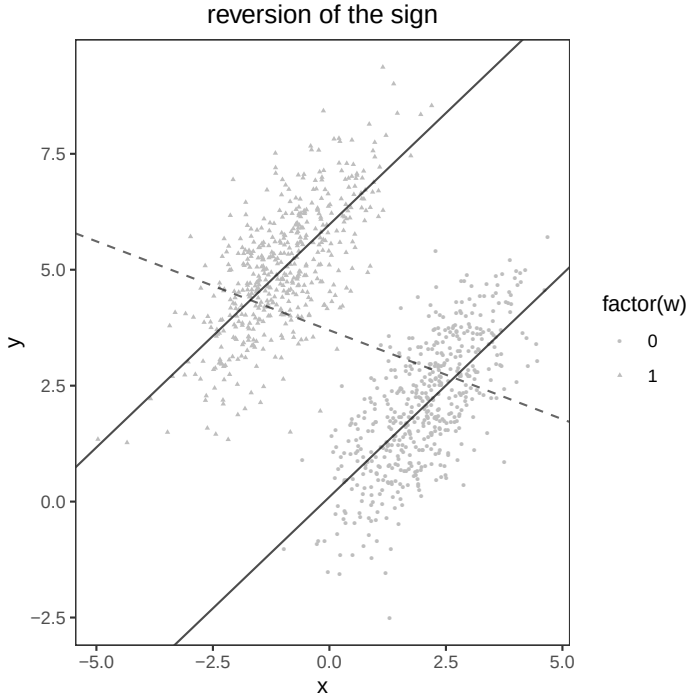


FIGURE 7.2: An Example of Simpson's Paradox

7.3 Hypothesis testing and analysis of variance

Partition X and β into

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

where $X_1 \in \mathbb{R}^{n\times k}, X_2 \in \mathbb{R}^{n\times l}, \beta_1 \in \mathbb{R}^k$ and $\mathbb{R}^l$. We are often interested in testing

$$H_0 : \beta_2 = 0$$

in the long regression

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon, \tag{7.4}$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$. If H_0 holds, then a short regression suffices:

$$Y = X_1\beta + \varepsilon \tag{7.5}$$

because X_2 is redundant. This is special case of testing $C\beta = 0$ with

$$C = \begin{pmatrix} 0_{l\times k} & I_{l\times l} \end{pmatrix}.$$

As discussed before, we can use

$$\hat{\beta}_2 \sim \mathrm{N}(0, \sigma^2 S^{22})$$

with S^{22} being the $(2,2)$th block of $(X^{\mathrm{T}}X)^{-1}$, to construct the Wald-type statistic for hypothesis testing:

$$F_{\text{Wald}} = \frac{\hat{\beta}_2^{\mathrm{T}}(S^{22})^{-1}\hat{\beta}_2}{l\hat{\sigma}^2} \sim F_{l,n-p}.$$

In this section, I will test H_0 from an alternative perspective based on comparing the residual sum of squares in the long regression (7.4) and the short regression (7.5). This technique is called the analysis of variance (ANOVA), first proposed by R. A. Fisher in design of experiments. Intuitively, if $\beta_2 = 0$, then the residual vectors from the long regression (7.4) and the short regression (7.5) 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

$$\text{RSS}_{\text{long}} = Y^{\mathrm{T}}(I_n - H)Y$$

and

$$\text{RSS}_{\text{short}} = Y^{\mathrm{T}}(I_n - H_1)Y$$

as the residual sum of squares from the long and short regressions, respectively. By the definition of OLS, it must be true that

$$\text{RSS}_{\text{long}} \leq \text{RSS}_{\text{short}} \Longrightarrow \text{RSS}_{\text{short}} - \text{RSS}_{\text{long}} = Y^{\mathrm{T}}(H - H_1)Y \geq 0. \tag{7.6}$$

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

$$F_{\text{ANOVA}} = \frac{\text{RSS}_{\text{short}} - \text{RSS}_{\text{long}}}{l\hat{\sigma}^2}. \tag{7.7}$$

Theorem 7.2 *Under the Gaussian linear model, if $\beta_2 = 0$, then $F_{\text{ANOVA}} \sim F_{l,n-p}$. In fact, $F_{\text{ANOVA}} = F_{Wald}$ which is a numerical result without assuming the Gaussian linear model.*

Proof of Theorem 7.2: The proof has two steps.

Step 1.

I will prove the distribution of F_{ANOVA} directly. First, I will use the following basic facts repeatedly:

$$HX = X \Longrightarrow H \begin{pmatrix} X_1 & X_2 \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \Longrightarrow HX_1 = X_1, \quad (7.8)$$

$$HH_1 = H_1, \qquad H_1 H = H_1, \quad (7.9)$$

$$(H - H_1)X_1 = 0; \quad (7.10)$$

$H - H_1$ is a projection matrix of rank $p - k = l$, $I_n - H$ is a projection matrix of rank $n - p$, and they are orthogonal:

$$(H - H_1)(I_n - H) = 0. \quad (7.11)$$

I leave it as a homework problem to verify these.

Second, the residual vector from the long regression is $\hat{\varepsilon} = (I_n - H)Y = (I_n - H)(X\beta + \varepsilon) = (I_n - H)\varepsilon$, so the residual sum of squares is

$$\text{RSS}_{\text{long}} = \hat{\varepsilon}^{\text{T}}\hat{\varepsilon} = \varepsilon^{\text{T}}(I_n - H)\varepsilon;$$

since $\beta_2 = 0$, the residual vector from the short regression is $\tilde{\varepsilon} = (I_n - H_1)Y = (I_n - H_1)(X_1\beta_1 + \varepsilon) = (I_n - H_1)\varepsilon$, so the residual sum of squares is

$$\text{RSS}_{\text{short}} = \tilde{\varepsilon}^{\text{T}}\tilde{\varepsilon} = \varepsilon^{\text{T}}(I_n - H_1)\varepsilon.$$

Let $\varepsilon_0 = \varepsilon/\sigma \sim \text{N}(0, I_n)$ be a standard Normal random vector, then we can write F_{ANOVA} as

$$\begin{aligned} F_{\text{ANOVA}} &= \frac{\varepsilon^{\text{T}}(H - H_1)\varepsilon/l}{\varepsilon^{\text{T}}(I_n - H)\varepsilon/(n-p)} \\ &= \frac{\varepsilon_0^{\text{T}}(H - H_1)\varepsilon_0/l}{\varepsilon_0^{\text{T}}(I_n - H)\varepsilon_0/(n-p)} \\ &= \frac{\|(H - H_1)\varepsilon_0\|^2/l}{\|(I_n - H)\varepsilon_0\|^2/(n-p)}. \end{aligned} \quad (7.12)$$

Third, we have the following joint Normality using the basic fact (7.11):

$$\begin{pmatrix} (H - H_1)\varepsilon_0 \\ (I_n - H)\varepsilon_0 \end{pmatrix} = \begin{pmatrix} H - H_1 \\ I_n - H \end{pmatrix} \varepsilon_0 \sim \text{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} H - H_1 & 0 \\ 0 & I_n - H \end{pmatrix} \right\}.$$

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 implies that their squared lengths are chi-squared:

$$\|(H - H_1)\varepsilon_0\|^2 \sim \chi^2_l, \qquad \|(I_n - H)\varepsilon_0\|^2 \sim \chi^2_{n-p},$$

and they are independent. These facts, coupled with (7.12), imply that $F_{\text{ANOVA}} \sim F_{l,n-p}$.

Step 2.

I will show that $F_{\text{ANOVA}} = F_{\text{Wald}}$ which gives an indirect proof for the distribution of F_{ANOVA}. Using the FWL Theorem that $\hat{\beta}_2 = (\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}}Y$ and the fact that $S^{22} = (\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}$, we can rewrite F_{Wald} as

$$\begin{aligned} F_{\text{Wald}} &= \frac{Y^{\mathrm{T}}\tilde{X}_2(\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}}\tilde{X}_2(\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}}Y}{l\hat{\sigma}^2} \\ &= \frac{Y^{\mathrm{T}}\tilde{X}_2(\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}}Y}{l\hat{\sigma}^2} \\ &= \frac{Y^{\mathrm{T}}\tilde{H}_2 Y}{l\hat{\sigma}^2}, \end{aligned} \tag{7.13}$$

where $\tilde{H}_2 = \tilde{X}_2(\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}}$ is the projection matrix onto the column space of $\tilde{X}_2$. Comparing (7.13) with (7.6) and (7.7), we only need to show that $H - H_1 = \tilde{H}_2$, which is a linear algebra fact following from the formula of the inverse of the block matrix $X^{\mathrm{T}}X$. I relegate the proof as a homework problem. □

We can use the `anova` function in `R` to compute the F statistic and the p-value. Below we revisit the `lalonde` data, obtaining identical result as before.

```
> lalonde_fit1 = lm(re78 ~ ., data = lalonde)
> lalonde_fit2 = lm(re78 ~ treat, data = lalonde)
> anova(lalonde_fit2, lalonde_fit1)
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 *
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

In fact, we can conduct 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.

```
> lalonde_fit3 = lm(re78 ~ 1, data = lalonde)
> anova(lalonde_fit3, lalonde_fit2, lalonde_fit1)
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 *
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

7.4 Homework problems

7.1 General centering

What is the projection matrix onto the column space of X_1 defined in (7.1)? Verify (7.2).

7.2 Formula of the partial correlation coefficient

Prove Theorem 7.1.

7.3 Examples of Simpson's Paradox

Given three numerical examples of (Y, X, W) which causes Simpson's Paradox. Report the mean and covariance matrix for each example.

7.4 Simpson's Paradox in reality

Find a real-life dataset with Simpson's Paradox.

7.5 Basic properties of projection matrices

Show (7.9) and (7.10). Verify that the rank of $H - H_1$ is l and the rank of $I_n - H$ is $n - p$.

7.6 Decomposition of the projection matrix

Show that $H - H_1 = \tilde{H}_2$.

7.7 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 $\hat{\rho}_{x_1 x_2} > 0$. Show that the corresponding OLS coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are negatively correlated under the Gauss–Markov Model.

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. If the partial correlation coefficient between X_1 and X_2 given X_3 is positive, then show that the corresponding OLS coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are negatively correlated under the Gauss–Markov Model.

8

Asymptotic Inference in OLS with Possibly Non-Normal and Heteroskedastic Errors

8.1 Motivation

Standard software packages, for example, R, report point estimator, standard error, and p-value for each coordinate of β based on the Gaussian linear model:

$$Y = X\beta + \varepsilon \sim \mathrm{N}(X\beta, \sigma^2 I_n).$$

Statistical inference based on this model is finite-sample exact. However, the assumptions of this model is extremely strong: the functional form is linear, the error terms are additive with distributions not dependent on X, and the error terms are IID Normal with the same variance. If we do not believe these assumptions, can we still trust the associated statistical inference?

8.1.1 Numerical examples

We start with some simple numerical examples. The first one is the ideal Gaussian 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, we generate outcomes from a simple linear model $y_i = x_i + \varepsilon_i$ with $\varepsilon_i \overset{\text{IID}}{\sim} \mathrm{N}(0, \sigma^2 = 1)$. Over 5000 replications of the data, we compute the OLS coefficient $\hat{\beta}$ of x_i and reported two standard errors. One standard error is the one under Gaussian linear model which is also the default choice of the `lm` function of R; the other standard error, computed by the `hccm` function in the R package `car`, will be introduced later, which is a main topic of this chapter. The $(1, 1)$ the panel of Figure 8.1 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 the estimated standard errors are close to the true one.

To investigate the impact of Normality, we 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 8.1 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 Normality does not seem a crucial assumption for the validity of the inference procedure under the Gaussian linear model.

We 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 8.1 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.

Finally, we generate heteroskadastic 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 8.1 corresponds to this setting, which has a similar pattern as the $(2, 1)$ panel. So Normality is not crucial, but the homoskedasticity is.

8.1.2 Goal of this chapter

In this chapter, we will still impose the linearity assumption, but relax the distributional assumption on the error terms. We assume the following heteroskedastic linear model:

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i,$$

where the x_i's are fixed, and the ε_i's are independent with mean zero and variance σ_i^2 $(i = 1, \ldots, n)$. Because the error terms can have different variances, they are not IID in general. Their variances can be functions of the

x_i's, and the variances σ_i^2 are n free numbers. 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. The tool we use in this chapter is the asymptotic analysis, assuming that the sample size n is large so that certain limiting theorems hold.

The asymptotic analysis later will show that if the error terms are homoskedastic, i.e., $\sigma_i^2 = \sigma^2$ for all $i = 1, \ldots, n$, we can still trust the statistical inference based on the Gaussian linear model as long the central limit theorem (CLT) for the OLS estimator holds with $n \to \infty$. If the error terms are heteroskedastic, i.e., their variances are different, we need to adjust the standard error with the so-called Eicker–Huber–White robust standard error. I will give technical details below.

8.2 Consistency of OLS

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 useful lemma.

Lemma 8.1 *The OLS estimator has the following representation:*

$$\hat{\beta} - \beta = \left(n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \right)^{-1} n^{-1} \sum_{i=1}^{n} x_i \varepsilon_i.$$

Proof 1 *Since $y_i = x_i^{\mathrm{T}} \beta + \varepsilon_i$, we have*

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

Using Lemma 8.1, we can show that $E(\hat{\beta}) = \beta$ and $\text{cov}(\hat{\beta}) = n^{-1}B_n^{-1}M_nB_n^{-1}$ where

$$B_n = n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}, \qquad M_n = n^{-1}\sum_{i=1}^{n} \sigma_i^2 x_i x_i^{\mathrm{T}}.$$

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 concentrates near its mean β. This is the idea of consistency, formally stated below.

Assumption 8.1 *$B_n \to B$ and $M_n \to M$ where B and M are finite with B invertible.*

Theorem 8.1 *Under Assumption 8.1, $\hat{\beta} \to \beta$ in probability.*

Proof 2 *We only need to show that $n^{-1}\sum_{i=1}^{n} x_i\varepsilon_i \to 0$ in probability. It has mean zero and covariance matrix M_n/n, so it converges to zero in probability using a proposition in the appendix.*

8.3 Asymptotic Normality of OLS

Intuitively, $n^{-1}\sum_{i=1}^{n} x_i\varepsilon_i$ is the sample average of some independent terms, and therefore, the classic Lindberg–Feller Theorem guarantees that it enjoys a CLT under some regularity conditions. Consequently, $\hat{\beta}$ also enjoys a CLT with mean β and covariance matrix $n^{-1}B_n^{-1}M_nB_n^{-1}$. 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 8.2 *Under Assumption 8.1, if there exist a $\delta > 0$ such that $d_{2+\delta,n} \to d_{2+\delta} < \infty$, then $\sqrt{n}(\hat{\beta} - \beta) \to \text{N}(0, B^{-1}MB^{-1})$ in distribution.*

Proof 3 *The key is to show the CLT for $n^{-1/2}\sum_{i=1}^{n} x_i\varepsilon_i$, and the CLT for $\hat{\beta}$ holds due to the Slutsky's Theorem. Define*

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

with mean zero and finite covariance, and we need to verify the two conditions required by the Lindeberg–Feller CLT. First, the Lyapunov condition holds

because

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

Second,

$$\sum_{i=1}^{n}\operatorname{cov}(z_{n,i}) = n^{-1}\sum_{i=1}^{n}\sigma_i^2 x_i x_i^{\mathrm{T}} = M_n \to M.$$

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

The asymptotic covariance $B^{-1}MB^{-1}$ has a sandwich form, justifying the choice of notation B for the "bread" and M for the "meat."

8.4 Eicker–Huber–White standard error

8.4.1 Sandwich variance estimator

The CLT in Theorem 8.2 shows that

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

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

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

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:

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

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 8.3 *Under Assumption 8.1, if*

$$n^{-1}\sum_{i=1}^{n}\mathrm{var}(\varepsilon_i^2)x_{ij_1}^2x_{ij_2}^2 \to c_{j_1j_2} < \infty, \tag{8.1}$$

$$n^{-1}\sum_{i=1}^{n}x_{ij_1}x_{ij_2}x_{ij_3}x_{ij_4} \to c_{j_1j_2j_3j_4} < \infty, \tag{8.2}$$

$$n^{-1}\sum_{i=1}^{n}\sigma_i^2x_{ij_1}^2x_{ij_2}^2x_{ij_3}^2 \to c_{j_1j_2j_3} < \infty, \tag{8.3}$$

for any $j_1, j_2, j_3, j_4 = 1, \ldots, p$, then $\hat{M}_n \to M$ in probability.

Proof of Theorem 8.3: Assumption 8.1 ensures that $\hat{\beta} \to \beta$ in probability by Theorem 8.1. Condition (8.1) and Markov's inequality ensures that $\tilde{M}_n - M_n \to 0$ in probability. So we only need to show that $\hat{M}_n - \tilde{M}_n \to 0$ in probability. The $(j_1, j_2)$the element of their difference is

$$\begin{aligned}
(\hat{M}_n - \tilde{M}_n)_{j_1,j_2} &= n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2x_{i,j_1}x_{i,j_2} - n^{-1}\sum_{i=1}^{n}\varepsilon_i^2x_{i,j_1}x_{i,j_2} \\
&= n^{-1}\sum_{i=1}^{n}\left[\left(\varepsilon_i + x_i^{\mathrm{T}}\beta - x_i^{\mathrm{T}}\hat{\beta}\right)^2 - \varepsilon_i^2\right]x_{i,j_1}x_{i,j_2} \\
&= n^{-1}\sum_{i=1}^{n}\left[\left(x_i^{\mathrm{T}}\beta - x_i^{\mathrm{T}}\hat{\beta}\right)^2 + 2\varepsilon_i\left(x_i^{\mathrm{T}}\beta - x_i^{\mathrm{T}}\hat{\beta}\right)\right]x_{i,j_1}x_{i,j_2} \\
&= (\beta - \hat{\beta})^{\mathrm{T}}n^{-1}\sum_{i=1}^{n}x_ix_i^{\mathrm{T}}x_{i,j_1}x_{i,j_2}(\beta - \hat{\beta}) \\
&\quad + 2(\beta - \hat{\beta})^{\mathrm{T}}n^{-1}\sum_{i=1}^{n}x_ix_{i,j_1}x_{i,j_2}\varepsilon_i,
\end{aligned}$$

which converges to zero in probability because the first term converges to zero due to condition (8.2) and the second term converges to zero in probability due to Markov's inequality and condition (8.3). □

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

$$\hat{V} = n^{-1}\left(n^{-1}\sum_{i=1}^{n}x_ix_i^{\mathrm{T}}\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2x_ix_i^{\mathrm{T}}\right)\left(n^{-1}\sum_{i=1}^{n}x_ix_i^{\mathrm{T}}\right)^{-1},$$

which is called the Eicker–Huber–White robust covariance matrix. In matrix form, we can rewrite it as

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

where $\hat{\Omega} = \mathrm{diag}\left\{\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\right\}$. Eicker (1967) first proposed to use $\hat{V}$, which was

later popularized by White (1980a) in economics. Huber (1967) and Fuller (1975) discussed related problems with variance estimators similar to $\hat{V}$, and Miller (1974) and Hinkley (1977) motivated the use of $\hat{V}$ based on an idea called *jackknife*. The square root of the diagonal terms of $\hat{V}$, denoted by $\hat{\mathrm{se}}_j$ $(j = 1, \ldots, p)$, 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 hypothesis based on

$$\hat{\beta} \overset{\text{a}}{\sim} \mathrm{N}(\beta, \hat{V}),$$

and in particular, we can infer each element of the coefficient based on $\hat{\beta}_j \overset{\text{a}}{\sim} \mathrm{N}(\beta_j, \hat{\mathrm{se}}_j^2)$.

8.4.2 Other "HC" standard errors

Since White (1980a) published his paper, several modifications of $\hat{V}$ appeared aiming for better finite-sample properties. I summarize them below: recall that h_{ii}'s are the diagonal elements of the projection matrix H, and define

$$\tilde{V} = n^{-1} \left(n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} \tilde{\varepsilon}_i^2 x_i x_i^{\mathrm{T}} \right) \left(n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \right)^{-1},$$

where $\tilde{\varepsilon}_i$ can be

$$\tilde{\varepsilon}_i = \begin{cases} \hat{\varepsilon}_i, & \text{(HC0)} \\ \hat{\varepsilon}_i \sqrt{\frac{n}{n-p}}, & \text{(HC1)} \\ \hat{\varepsilon}_i / \sqrt{1 - h_{ii}}, & \text{(HC2)} \\ \hat{\varepsilon}_i / (1 - h_{ii}), & \text{(HC3)} \\ \hat{\varepsilon}_i / (1 - h_{ii})^{\min\{2, n h_{ii}/(2p)\}}, & \text{(HC4).} \end{cases}$$

See MacKinnon and White (1985) and Long and Ervin (2000) for reviews. Using simulation studies, Long and Ervin (2000) recommended HC3.

8.4.3 Special case with homoskedasticity

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

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

which simplifies the covariance of $\hat{\beta}$ to $\mathrm{cov}(\hat{\beta}) = \sigma^2 B_n^{-1}$, and the asymptotic Normality to $\sqrt{n}(\hat{\beta} - \beta) \to \mathrm{N}(0, \sigma^2 B)$ 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 8.1, justifying the use of

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

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

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

It is slightly different from the inference based on t and F distributions. But with large n, the difference is very small. To end this chapter, I give a formal result on the consistency of $\hat{\sigma}^2$.

Theorem 8.4 *Under Assumption 8.1, if $\sigma_i^2 = \sigma^2 < \infty$ for all $i = 1, \ldots, n$, then $\hat{\sigma}^2 \to \sigma^2$ in probability.*

Proof of Theorem 8.4: By the law of large numbers, $n^{-1}\sum_{i=1}^n \varepsilon_i^2 \to \sigma^2$ [check, errors are not iid in general] in probability, and $n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i^2$ has the same probability limit as $\hat{\sigma}^2$. So we only need to show that $n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i^2 - n^{-1}\sum_{i=1}^n \varepsilon_i^2 \to 0$ in probability. Their difference is

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

where the last step follows from Lemma 8.1. So the difference converges to zero in probability because $\hat{\beta} - \beta \to 0$ in probability by Theorem 8.1 and $B_n \to B$ by Assumption 8.1. □

8.5 Examples

8.5.1 LaLonde experimental data

In the `lalonde` data, different standard errors give similar t-values.

```
> library("Matching")
> data(lalonde)
> lalonde_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)  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
```

8.5.2 Data from King and Roberts (2015)

In the following example from King and Roberts (2015), the robust standard errors for some coefficients are quite different:

```
> library(foreign)
> library(car)
> dat = read.dta("Article␣for␣ISQ␣(aid).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
```

But if we use the log transformation on the outcome, then all standard errors give similar t-values.

```
> 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 heteroskedasticity or the poor approximation of the linear model. We will discuss the model misspecification issue later.

8.5.3 Boston housing data

```
> library("mlbench")
> library("car")
> 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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

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 help.

```
> 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
```

8.6 Homework problems

8.1 Testing linear hypotheses under heteroskedasticity

Under the heteroskedastic linear model, how to test the hypotheses

$$H_0 : c^{\mathrm{T}}\beta = 0, \quad c \in \mathbb{R}^p$$

and

$$H_0 : C\beta = 0, \quad C \in \mathbb{R}^{l \times p}?$$

8.2 Two-sample problem

1. Assume that $z_1, \ldots, z_m \overset{\text{IID}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$ and $w_1, \ldots, w_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$, and test $\mu_1 = \mu_2$. Show that the t statistic with pooled variance estimator have the following distribution:

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

where the sample means are

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

and the sample variances are

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

2. Assume that $z_1, \ldots, z_m$ are IID with mean μ_1 and variance σ_1^2, and $w_1, \ldots, w_n$ are IID with mean μ_2 and variance σ_2^2, and test $\mu_1 = \mu_2$. Show that under the null hypothesis, the following t statistic has an asymptotically Normal distribution:

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

in distribution. The names "`equal`" and "`unequal`" are motivated by the "`var.equal`" parameter of the `R` function `t.test`.

3. We can write the above problems as hypothesis testing in linear regression $Y = X\beta + \varepsilon$ with

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

and hypothesis $H_0 : \beta_1 = 0$. Based on homoskedasticity or heteroskedasticity of the error terms, we can compute two t statistics. Show that the former is identical to t_{equal} and the latter is identical to t_{unequal} with HC2.

8.3 ANOVA with heteroskedasticity

continue lecture 5.

If $y_i \mid i \in \mathcal{T}_j$ has mean β_j and variance σ_j^2, find the ols and ehw se and compare them.

HC0 and HC2?

8.4 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).

8.5 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 8.5.2. Replicate another one using OLS (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/`).

8.6 Unbiased sandwich variance estimator under the Gauss–Markov model

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

8.7 FWL theorem for the EHW standard error

Based on the OLS fit $\hat{Y} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}$, the second component $\hat{\beta}_2$ has estimated covariance $\hat{V}$ assuming homoskedasticity or $\hat{V}_{\text{EHW}}$ allowing for heteroskedasticity.

Based on the FLW theorem, we can also obtain $\hat{\beta}_2$ from the following partial regression. First, from the OLS fit of Y on X_1 we obtain the residual vector $\tilde{Y}$. Then, from the column-wise OLS fit of X_2 on X_1 we obtain the residual matrix $\tilde{X}_2$. Finally, from the OLS fit of $\tilde{Y}$ on $\tilde{X}_2$ we obtain the coefficient $\tilde{\beta}_2$. Then based on the final regression, $\tilde{\beta}_2$ has estimated covariance $\tilde{V}$ assuming homoskedasticity or $\tilde{V}_{\text{EHW}}$ allowing for heteroskedasticity.

Show that $(n-K-L)\hat{V} = (n-L)\tilde{V}$ and $\hat{V}_{\text{EHW}} = \tilde{V}_{\text{EHW}}$.

8.8 EHW in long and short regressions

Problem 6.5 gives Cochran's formula related to the coefficients from three OLS fits. There are at least two ways to estimate the covariance of $\tilde{\beta}_2$. First, from the second OLS fit, the EHW covariance estimator is

$$\tilde{V} = (X_2^{\mathrm{T}}X_2)^{-1}X_2^{\mathrm{T}}\mathrm{diag}(\tilde{\varepsilon}^2)X_2(X_2^{\mathrm{T}}X_2)^{-1}.$$

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

$$\hat{V} = (\hat{\delta}, I_l)(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{diag}(\hat{\varepsilon}^2)X(X^{\mathrm{T}}X)^{-1}\begin{pmatrix}\hat{\delta}^{\mathrm{T}}\\ I_l\end{pmatrix}.$$

Show that

$$\hat{V} = (X_2^{\mathrm{T}}X_2)^{-1}X_2^{\mathrm{T}}\mathrm{diag}(\hat{\varepsilon}^2)X_2(X_2^{\mathrm{T}}X_2)^{-1}.$$

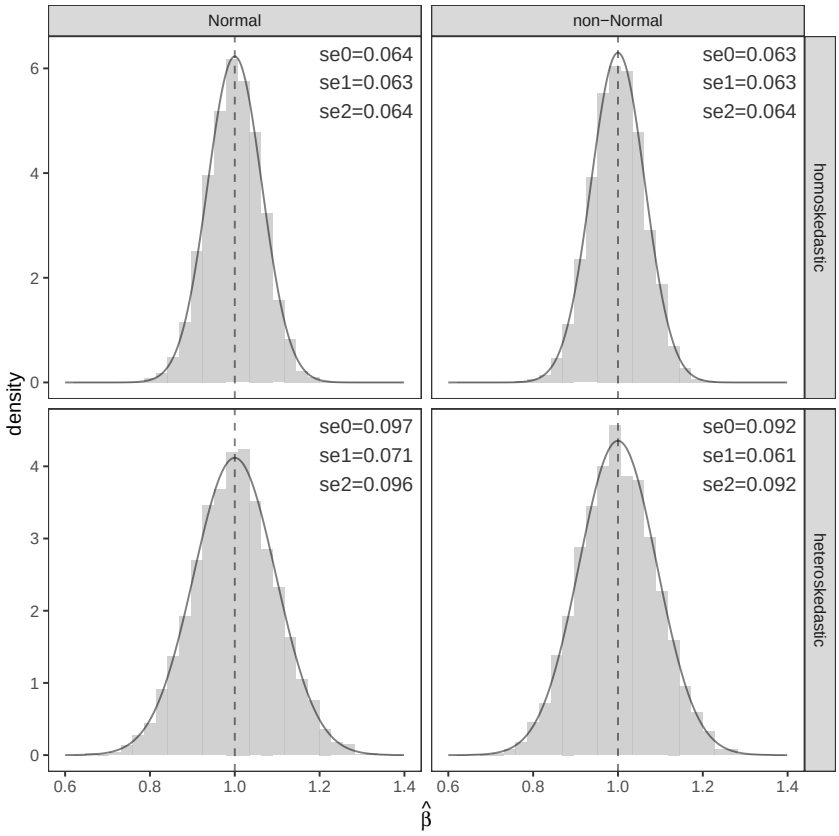


FIGURE 8.1: Simulation with 5000 replications: "se0" denotes the standard error of $\hat{\beta}$, "se1" denotes the estimated standard error based on the homoskedasticity assumption, and "se2" denotes the Eicker–Huber–White standard error allowing for heteroskedasticity. The density curves are Normal with mean 1 and standard deviation se0.

Part III

Model fitting and checking

9

Multiple Correlation Coefficient

In this chapter, I will introduce the multiple correlated coefficient, usually called the R^2. It can achieve two goals: first, it extends the Pearson correlation coefficient between two scalars to a measure of correlation between a scalar and a vector; second, it measures how well multiple covariates can linearly represent an outcome.

9.1 Equivalent definitions of R^2

I start with the standard definition of R^2. Based on the OLS of Y on X with X including 1_n, we define

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

We have discussed before that including 1_n in the OLS ensures that

$$n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i = 0 \Longrightarrow n^{-1}\sum_{i=1}^n y_i = n^{-1}\sum_{i=1}^n \hat{y}_i \Longrightarrow \bar{y} = \bar{\hat{y}},$$

i.e., the average of the fitted values equals the average of the original observed outcomes. So I use $\bar{y}$ for both the means of outcomes and the fitted values. 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 9.1 *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 this lemma as a homework problem. Lemma 9.1 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 $\sum_{i=1}^n (y_i - \hat{y}_i)^2$. From Lemma 9.1, R^2 must be lie within the interval $[0, 1]$ which measure the proportion of the regression sum of squares in the total sum of squares.

We can also verify that R^2 is the squared Pearson correlation coefficient between Y and $\hat{Y}$.

Theorem 9.1 *We have* $R^2 = \hat{\rho}^2_{y\hat{y}}$ *where*

$$\hat{\rho}_{y\hat{y}} = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}}. \tag{9.1}$$

I leave the proof of this theorem as a homework problem. Although the 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 vectors $Y - \bar{y}1_n$ and $\hat{Y} - \bar{y}1_n$.

In terms of long and short regressions, we can partition the design matrix into $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$ with $X_1 = 1_n$, then the OLS fit of the long regression is

$$Y = 1_n\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}, \tag{9.2}$$

and the OLS fit of the short regression is

$$Y = 1_n\tilde{\beta}_1 + \tilde{\varepsilon}, \qquad \text{with } \tilde{\beta} = \bar{y} \text{ and } H_1 = I_n - n^{-1}1_n1_n^{\mathrm{T}}. \tag{9.3}$$

The total sum of squares is the residual sum of squares from the short regression $\sum_{i=1}^n (y_i - \bar{y})^2 = Y^{\mathrm{T}}(I_n - H_1)Y$, so we can also write R^2 as

$$R^2 = \frac{Y^{\mathrm{T}}(I_n - H_1)Y - Y^{\mathrm{T}}(I_n - H)Y}{Y^{\mathrm{T}}(I_n - H_1)Y} = \frac{Y^{\mathrm{T}}(H - H_1)Y}{Y^{\mathrm{T}}(I_n - H_1)Y}.$$

9.2 R^2 and the F statistic

Under the Gaussian linear model

$$Y = 1_n\beta_1 + X_2\beta_2 + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2 I_n), \tag{9.4}$$

we can use the F statistic to test whether $\beta_2 = 0$. This F statistic is a monotone function of R^2. Most standard software packages report both.

Theorem 9.2 *We have*

$$F = \frac{n-p}{p-1} \times \frac{R^2}{1-R^2}.$$

Proof of Theorem 9.2: Based on the long regression (9.2) and the short regression (9.3), we have

$$F = \frac{(\mathrm{RSS}_{\mathrm{short}} - \mathrm{RSS}_{\mathrm{long}})/(p-1)}{\mathrm{RSS}_{\mathrm{long}}/(n-p)}$$

and

$$R^2 = \frac{\text{RSS}_{\text{short}} - \text{RSS}_{\text{long}}}{\text{RSS}_{\text{short}}}.$$

So the conclusion follows. □

Corollary 9.1 *Under the Gaussian linear model (9.4), if $\beta_2 = 0$, then*

$$R^2 \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right).$$

Proof of Corollary 9.1: The F statistic can be represented as

$$F = \frac{\chi^2_{p-1}/(p-1)}{\chi^2_{n-p}/(n-p)}$$

where $\chi^2_{p-1} \perp\!\!\!\perp \chi^2_{n-p}$. Using Theorem 9.2, we have

$$\frac{R^2}{1-R^2} = F \times \frac{p-1}{n-p} = \frac{\chi^2_{p-1}}{\chi^2_{n-p}} \Longrightarrow R^2 = \frac{\chi^2_{p-1}}{\chi^2_{p-1} + \chi^2_{n-p}}.$$

Because $\chi^2_{p-1} \sim \text{Gamma}\left(\frac{p-1}{2}, \frac{1}{2}\right)$ and $\chi^2_{n-p} \sim \text{Gamma}\left(\frac{n-p}{2}, \frac{1}{2}\right)$, we have

$$R^2 = \frac{\text{Gamma}\left(\frac{p-1}{2}, \frac{1}{2}\right)}{\text{Gamma}\left(\frac{p-1}{2}, \frac{1}{2}\right) + \text{Gamma}\left(\frac{n-p}{2}, \frac{1}{2}\right)} \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right),$$

which follows from the Beta–Gamma duality in Theorem A2.1. □

9.3 Numerical examples

LaLonde data

```
> library("Matching")
> data(lalonde)
> ols.fit = lm(re78 ~ ., y = TRUE, data = lalonde)
> ols.summary = summary(ols.fit)
> r2 = ols.summary$r.squared
> r2 - (cor(ols.fit$y, ols.fit$fitted.values))^2
[1] -1.665335e-16
>
> fstat = ols.summary$fstatistic
> fstat
    value     numdf     dendf
  2.43349  11.00000 433.00000
> fstat[1] - fstat[3]/fstat[2]*r2/(1-r2)
value
    0
```

Neumeyer data

```
> library(foreign)
> dat = read.dta("Article␣for␣ISQ␣(aid).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$r.squared
> r2 - (cor(ols.fit$y, ols.fit$fitted.values))^2
[1] -1.110223e-16
>
> fstat = ols.summary$fstatistic
> fstat
    value     numdf     dendf
 40.69519  12.00000 468.00000
> fstat[1] - fstat[3]/fstat[2]*r2/(1-r2)
value
    0
```

9.4 Homework problems

9.1 ANOVA

Prove Lemma 9.1.

9.2 R^2 and Pearson Correlation Coefficient

Prove Theorem 9.1.

9.3 Exact distribution of $\hat{\rho}$

Under the Gaussian linear model with univariate x_i and its coefficient being 0, find the exact distribution of $\hat{\rho}_{xy}$.

10

Leverage Scores and Leave-One-Out Formulas

10.1 Leverage scores

We have seen the use of the hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ in previous chapters. Because

$$H = \begin{pmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix} (X^{\mathrm{T}}X)^{-1} \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix},$$

its (i,j)th element equals $h_{ij} = x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_j$. In this chapter, we will pay special attention to its diagonal elements

$$h_{ii} = x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i$$

often called the *leverage scores*, which play important roles in many discussions later.

First, because H is a rank p projection matrix, we have

$$\sum_{i=1}^{n} h_{ii} = \mathrm{trace}(H) = \mathrm{rank}(H) = p \Longrightarrow n^{-1}\sum_{i=1}^{n} h_{ii} = p/n.$$

So the average of the leverage scores equals p/n and the maximum of the leverage scores must be larger than or equal to p/n, which is close to zero when p is small relative to n.

Second, because $H = H^2$ and $H = H^{\mathrm{T}}$, we have

$$h_{ii} = \sum_{j=1}^{n} h_{ij}h_{ji} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2 \geq h_{ii}^2 \Longrightarrow h_{ii} \in [0,1].$$

So each leverage score is bounded between zero and one.

Third, because $\hat{Y} = HY$, we have

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j \Longrightarrow \frac{\partial \hat{y}_i}{\partial y_i} = h_{ii}.$$

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 $\hat{y}_i = x_i^{\mathrm{T}}\hat{\beta}$ under the Gauss–Markov model:

$$\mathrm{var}(\hat{y}_i) = x_i^{\mathrm{T}}\mathrm{cov}(\hat{\beta})x_i = \sigma^2 x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i = \sigma^2 h_{ii}.$$

So the variance of $\hat{y}_i$ increases with h_{ii}.

The final property of h_{ii} is less obvious: it measures whether observation i is an outlier based on its covariate value, that is, how far x_i is from the center of the data. Partition the design matrix as $X = \begin{pmatrix} 1_n & X_2 \end{pmatrix}$ with $H_1 = n^{-1}1_n 1_n^{\mathrm{T}}$. The covariates X_2 has center $\bar{x}_2 = n^{-1}\sum_{i=1}^n x_{i2}$ and sample covariance

$$S = (n-1)^{-1}\sum_{i=1}^n (x_{i2} - \bar{x}_2)(x_{i2} - \bar{x}_2)^{\mathrm{T}} = (n-1)^{-1}X_2^{\mathrm{T}}(I_n - H_1)X_2.$$

The sample Mahalanobis distance between x_{i2} and the center $\bar{x}_2$ is

$$D_i^2 = (x_{i2} - \bar{x}_2)^{\mathrm{T}} S^{-1}(x_{i2} - \bar{x}_2).$$

The following theorem shows that h_{ii} is a monotone function of D_i^2:

Theorem 10.1 *We have*

$$h_{ii} = \frac{1}{n} + \frac{D_i^2}{n-1}, \tag{10.1}$$

so $h_{ii} \geq 1/n$.

Proof of Theorem 10.1: The definition of D_i^2 implies that it is the (i,i)th element of the following matrix:

$$\begin{aligned}
&\begin{pmatrix} x_{12} - \bar{x}_2 \\ \vdots \\ x_{n2} - \bar{x}_2 \end{pmatrix}^{\mathrm{T}} S^{-1} \begin{pmatrix} x_{12} - \bar{x}_2 & \cdots & x_{n2} - \bar{x}_2 \end{pmatrix} \\
=&(I_n - H_1)X_2 \left\{(n-1)^{-1}X_2^{\mathrm{T}}(I_n - H_1)X_2\right\}^{-1} X_2^{\mathrm{T}}(I_n - H_1) \\
=&(n-1)\tilde{X}_2(\tilde{X}_2^{\mathrm{T}}\tilde{X}_2)^{-1}\tilde{X}_2^{\mathrm{T}} \\
=&(n-1)\tilde{H}_2 \\
=&(n-1)(H - H_1).
\end{aligned}$$

Therefore,

$$D_i^2 = (n-1)(h_{ii} - 1/n)$$

which implies (10.1). □

These are the basic properties of the leverage scores. We will see their roles frequently in later parts of the chapter.

Huber's asymptotics?

10.2 Leave-one-out formulas

To measure the impact of the ith observation on the final OLS estimator, a natural approach is to delete the ith row from the full data

$$X = \begin{pmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

and check how much the OLS estimator changes. Let $X_{[-i]}$ and $Y_{[-i]}$ denote the leave-i-out data, and define

$$\hat{\beta}_{[-i]} = (X_{[-i]}^{\mathrm{T}} X_{[-i]})^{-1} X_{[-i]}^{\mathrm{T}} Y_{[-i]} \tag{10.2}$$

as the corresponding OLS estimator. We can fit n OLS by deleting the ith row ($i = 1, \ldots, n$). However, this is computationally intensive especially when n is large. The following theorem shows that we need only to fit OLS once.

Theorem 10.2 *Recalling that $\hat{\beta}$ is the full data OLS, $\hat{\varepsilon}_i$ is the residual and h_{ii} is the leverage score for the ith observation, we have*

$$\hat{\beta}_{[-i]} = \hat{\beta} - (1 - h_{ii})^{-1} (X^{\mathrm{T}} X)^{-1} x_i \hat{\varepsilon}_i.$$

Proof of Theorem 10.2: From (10.2), we need to invert

$$X_{[-i]}^{\mathrm{T}} X_{[-i]} = \sum_{i' \neq i} x_{i'} x_{i'}^{\mathrm{T}} = X^{\mathrm{T}} X - x_i x_i^{\mathrm{T}}$$

and calculate

$$X_{[-i]}^{\mathrm{T}} Y_{[-i]} = \sum_{i' \neq i} x_i y_i = X^{\mathrm{T}} Y - x_i y_i,$$

which are the original $X^{\mathrm{T}} X$ and $X^{\mathrm{T}} Y$ with slight modifications. Using the following Sherman–Morrison formula:

$$(A + uv^{\mathrm{T}})^{-1} = A^{-1} - (1 + v^{\mathrm{T}} A^{-1} u)^{-1} A^{-1} u v^{\mathrm{T}} A^{-1}$$

with $A = X^{\mathrm{T}} X, u = x_i$, and $v = -x_i$ we can invert $X_{[-i]}^{\mathrm{T}} X_{[-i]}$ as

$$\begin{aligned} (X_{[-i]}^{\mathrm{T}} X_{[-i]})^{-1} &= (X^{\mathrm{T}} X)^{-1} + \left\{1 - x_i^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} x_i\right\}^{-1} (X^{\mathrm{T}} X)^{-1} x_i x_i^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} \\ &= (X^{\mathrm{T}} X)^{-1} + (1 - h_{ii})^{-1} (X^{\mathrm{T}} X)^{-1} x_i x_i^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1}. \end{aligned}$$

Therefore,

$$\begin{aligned}
\hat{\beta}_{[-i]} &= (X_{[-i]}^{\mathrm{T}} X_{[-i]})^{-1} X_{[-i]}^{\mathrm{T}} Y_{[-i]} \\
&= \left\{ (X^{\mathrm{T}}X)^{-1} + (1-h_{ii})^{-1} (X^{\mathrm{T}}X)^{-1} x_i x_i^{\mathrm{T}} (X^{\mathrm{T}}X)^{-1} \right\} (X^{\mathrm{T}}Y - x_i y_i) \\
&= (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}}Y - (X^{\mathrm{T}}X)^{-1} x_i y_i \\
&\quad + (1-h_{ii})^{-1} (X^{\mathrm{T}}X)^{-1} x_i x_i^{\mathrm{T}} (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}}Y - (1-h_{ii})^{-1} (X^{\mathrm{T}}X)^{-1} x_i x_i^{\mathrm{T}} (X^{\mathrm{T}}X)^{-1} x_i y_i \\
&= \hat{\beta} - (X^{\mathrm{T}}X)^{-1} x_i y_i + (1-h_{ii})^{-1} (X^{\mathrm{T}}X)^{-1} x_i x_i^{\mathrm{T}} \hat{\beta} - h_{ii} (1-h_{ii})^{-1} (X^{\mathrm{T}}X)^{-1} x_i y_i \\
&= \hat{\beta} - (1-h_{ii})^{-1} (X^{\mathrm{T}}X)^{-1} x_i y_i + (1-h_{ii})^{-1} (X^{\mathrm{T}}X)^{-1} x_i \hat{y}_i \\
&= \hat{\beta} - (1-h_{ii})^{-1} (X^{\mathrm{T}}X)^{-1} x_i \hat{\varepsilon}_i.
\end{aligned}$$

□

With the leave-i-out OLS estimator $\hat{\beta}_{[-i]}$, we can define the predicted residual

$$\hat{\varepsilon}_{[-i]} = y_i - x_i^{\mathrm{T}} \hat{\beta}_{[-i]},$$

which is different from the residual $\hat{\varepsilon}_i$. The predicted residual based on leave-one-out can often measure the performance of the prediction better because it mimics the real problem of predicting a future observation. In contrast, the original residual based on full data often gives overly optimistic measure of prediction. This is related to the overfitting issue discussed later. Under the Gauss–Markov model, it is straightforward to verify that the original residual has mean zero and variance

$$\mathrm{var}(\hat{\varepsilon}_i) = \sigma^2 (1 - h_{ii}), \tag{10.3}$$

and the predicted residual has mean zero and variance

$$\mathrm{var}(\hat{\varepsilon}_{[-i]}) = \mathrm{var}(y_i - x_i^{\mathrm{T}} \hat{\beta}_{[-i]}) = \sigma^2 + \sigma^2 x_i^{\mathrm{T}} (X_{[-i]}^{\mathrm{T}} X_{[-i]})^{-1} x_i. \tag{10.4}$$

The following theorem further simplifies the predicted residual and its variance.

Theorem 10.3 *We have $\hat{\varepsilon}_{[-i]} = \hat{\varepsilon}_i/(1-h_{ii})$, and under the Gauss–Markov model,*

$$\mathrm{var}(\hat{\varepsilon}_{[-i]}) = \sigma^2/(1-h_{ii}). \tag{10.5}$$

Proof of Theorem 10.3: By definition, we have

$$\begin{aligned}
\hat{\varepsilon}_{[-i]} &= y_i - x_i^{\mathrm{T}} \hat{\beta}_{[-i]} \\
&= y_i - x_i^{\mathrm{T}} \left\{ \hat{\beta} - (1-h_{ii})^{-1} (X^{\mathrm{T}}X)^{-1} x_i \hat{\varepsilon}_i \right\} \\
&= y_i - x_i^{\mathrm{T}} \hat{\beta} + (1-h_{ii})^{-1} x_i^{\mathrm{T}} (X^{\mathrm{T}}X)^{-1} x_i \hat{\varepsilon}_i \\
&= \hat{\varepsilon}_i + h_{ii} (1-h_{ii})^{-1} \hat{\varepsilon}_i \\
&= \hat{\varepsilon}_i/(1-h_{ii}).
\end{aligned} \tag{10.6}$$

Combing (10.3) and (10.6) , we can derive its variance formula. □

Comparing formulas (10.4) and (10.5), we obtain that

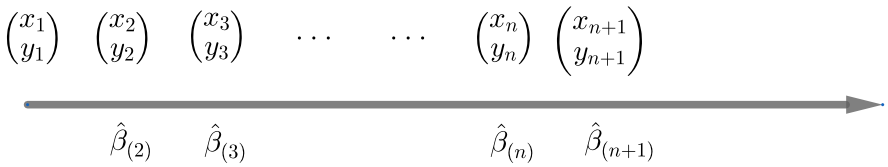
$$1 + x_i^{\mathrm{T}}(X_{[-i]}^{\mathrm{T}}X_{[-i]})^{-1}x_i = (1-h_{ii})^{-1} = \left\{1 - x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i\right\}^{-1}.$$

This is not an obvious linear algebra identity, but it follows from the two ways to calculate the variance of the predicted residual.

10.3 Applications of the leave-one-out formulas

10.3.1 Gauss updating formula

Consider an online setting in which the data points come sequentially as illustrated in 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 the following theorem.

Theorem 10.4 *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 $\gamma_{(n+1)} = (X_{(n+1)}^{\mathrm{T}}X_{(n+1)})^{-1}x_{n+1}$ and $\hat{\varepsilon}_{[n+1]} = y_{n+1} - x_{n+1}^{\mathrm{T}}\hat{\beta}_{(n)}$ is the predicted residual of the $(n+1)$the outcome based on the OLS of the first n observations.

Proof of Theorem 10.4: This is the reverse form of the leave-one-out formula. We can view the first $n+1$ data points as the full data, and $\hat{\beta}_{(n)}$ as the the OLS estimator leaving the $(n+1)$the observation out. Applying Theorem

10.2, we have

$$\begin{aligned}\hat{\beta}_{(n)} &= \hat{\beta}_{(n+1)} - (X_{(n+1)}^{\mathrm{T}} X_{(n+1)})^{-1} x_{n+1} \frac{\hat{\varepsilon}_{n+1}}{1 - h_{n+1,n+1}} \\ &= \hat{\beta}_{(n+1)} - \gamma_{(n+1)} \hat{\varepsilon}_{[n+1]},\end{aligned}$$

where $\hat{\varepsilon}_{n+1}$ is the $(n+1)$the residual based on the full data OLS, and the $(n+1)$the predicted residual equals $\hat{\varepsilon}_{[n+1]} = \hat{\varepsilon}_{n+1}/(1 - h_{n+1,n+1})$ based on Theorem 10.3. □

Theorem 10.4 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 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 10.4 gives an algorithm for sequentially computing the OLS estimators. Using the Sherman–Morrison formula for updating the inverse of $X_{(n+1)}^{\mathrm{T}} X_{(n+1)}$ based on the inverse of $X_{(n)}^{\mathrm{T}} X_{(n)}$, we have an even simpler algorithm below:

(G1) Start with $V_{(n)} = (X_{(n)}^{\mathrm{T}} X_{(n)})^{-1}$ and $\hat{\beta}_{(n)}$.

(G2) Update

$$V_{(n+1)} = V_{(n)} - \left\{1 + x_{n+1}^{\mathrm{T}} V_{(n)} x_{n+1}\right\}^{-1} V_{(n)} x_{n+1} x_{n+1}^{\mathrm{T}} V_{(n)}.$$

(G3) Calculate $\gamma_{(n+1)} = V_{(n+1)} x_{n+1}$ and $\hat{\varepsilon}_{[n+1]} = y_{n+1} - x_{n+1}^{\mathrm{T}} \hat{\beta}_{(n)}$.

(G4) Update $\hat{\beta}_{(n+1)} = \hat{\beta}_{(n)} + \gamma_{(n+1)} \hat{\varepsilon}_{[n+1]}$.

10.3.2 Outlier detection based on residuals

Under the Gaussian linear model $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$, we know some basic probabilistic properties of the residual vector:

$$E(\hat{\varepsilon}) = 0, \qquad \mathrm{var}(\hat{\varepsilon}) = \sigma^2 (I_n - H).$$

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 Gaussian linear model.

The first quantity is the standardized residual

$$\mathrm{standr}_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}}.$$

We may hope that it has mean zero and variance one. However, because of the estimated variance, it is not easy to quantify the exact distribution of standr_i.

The second quantity is the studentized residual based on the predicted residual:

$$\text{studr}_i = \frac{\hat{\varepsilon}_{[-i]}}{\sqrt{\hat{\sigma}^2_{[-i]}/(1-h_{ii})}} = \frac{y_i - x_i^{\mathrm{T}}\hat{\beta}_{[-i]}}{\sqrt{\hat{\sigma}^2_{[-i]}/(1-h_{ii})}},$$

where $\hat{\beta}_{[-i]}$ and $\hat{\sigma}^2_{[-i]}$ are the estimates of the coefficient and variance based on the leave-i-out OLS. Because $(y_i, \hat{\beta}_{[-i]}, \hat{\sigma}^2_{[-i]})$ are mutually independent under the Gaussian linear model, we can show that

$$\text{studr}_i \sim t_{n-p-1}. \tag{10.7}$$

Because we know its distribution, we can compare it to the quantiles of the t distribution.

The third quantity is Cook's distance (Cook, 1977):

$$\begin{aligned}\text{cook}_i &= (\hat{\beta}_{[-i]} - \hat{\beta})^{\mathrm{T}} X^{\mathrm{T}} X (\hat{\beta}_{[-i]} - \hat{\beta})/(p\hat{\sigma}^2) \\ &= (X\hat{\beta}_{[-i]} - X\hat{\beta})^{\mathrm{T}} (X\hat{\beta}_{[-i]} - X\hat{\beta})/(p\hat{\sigma}^2),\end{aligned}$$

where the first form measures the change of the OLS estimator and the second form measures the change of the predicted values based on leaving-one-out. It has a slightly different motivation, but eventually it is related to the previous two due to the leave-on-out formulas.

Theorem 10.5 *Cook's distance is related to the standardized residual via:*

$$\text{cook}_i = \text{standr}_i^2 \times \frac{h_{ii}}{p(1-h_{ii})}.$$

First, I generate data from a univariate Gaussian linear model without outliers. In `R`, we can apply `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 10.1.

```
> n = 100
> x = seq(0, 1, length = n)
> y = 1 + 3*x + rnorm(n)
```

If I add 8 to the outcome of the last observation, the plots change to the second column of Figure 10.1. If I add 8 to the 50th observation, the plots change to the last column of Figure 10.1. 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 Gaussian linear model.

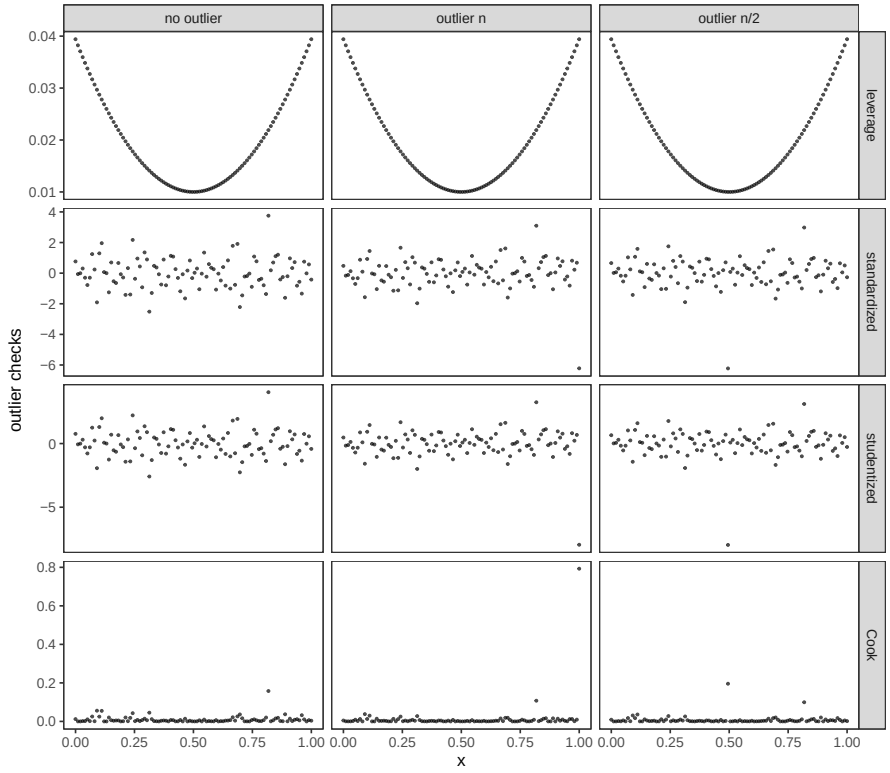


FIGURE 10.1: Outlier detections

10.3.3 Jackknife

Jackknife is a general strategy for bias reduction and variance estimation proposed by Quenouille (1949, 1956) and Tukey (1958). Based on independent data Z_i's, how to estimate the variance of a general estimator $\hat{\theta}(Z_1, \ldots, Z_n)$ 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 $\hat{\theta}_{\text{J}} = n^{-1}\sum_{i=1}^{n}\tilde{\theta}_i$, and the jackknife variance estimator is

$$\hat{V}_{\text{J}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\theta}_{[-i]} - \hat{\theta}_{\text{J}})(\hat{\theta}_{[-i]} - \hat{\theta}_{\text{J}})^{\text{T}}.$$

We have already shown that 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

$$\begin{aligned}\tilde{\beta}_i &= n\hat{\beta} - (n-1)\hat{\beta}_{[-i]} \\ &= n\hat{\beta} - (n-1)\left\{\hat{\beta} - (1-h_{ii})^{-1}(X^{\mathrm{T}}X)^{-1}x_i\hat{\varepsilon}_i\right\} \\ &= \hat{\beta} + (n-1)(1-h_{ii})^{-1}(X^{\mathrm{T}}X)^{-1}x_i\hat{\varepsilon}_i.\end{aligned}$$

The jackknife point estimator is

$$\hat{\beta}_{\mathrm{J}} = \hat{\beta} + \frac{n-1}{n}\left(n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} x_i \frac{\hat{\varepsilon}_i}{1-h_{ii}}\right).$$

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. But their difference is quite small. I omit their difference in the derivation, and leave the exact form of the jackknife point and variance estimators as a homework problem. Assuming that $\hat{\beta}_{\mathrm{J}} \cong \hat{\beta}$, we can continue to calculate the approximate jackknife variance estimator:

$$\begin{aligned}\hat{V}_{\mathrm{J}} &\cong \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\beta}_i - \hat{\beta})(\tilde{\beta}_i - \hat{\beta})^{\mathrm{T}} \\ &= \frac{n-1}{n}(X^{\mathrm{T}}X)^{-1}\sum_{i=1}^{n}\left(\frac{\hat{\varepsilon}_i}{1-h_{ii}}\right)^2 x_i x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1},\end{aligned}$$

which is almost identical to the HC3 form of the EHW standard error. Miller (1974) first analyzed the jackknife in OLS. Hinkley (1977) modified the original jackknife and proposed a version that is identical to HC1, and Wu (1986) proposed some further modification 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.

10.4 Homework problems

10.1 Invariance of the hat matrix and leverage scores

Show 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.

10.2 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 2 can recover the OLS coefficient, the leverages scores from 2 are not the same as those from the long regression. Show that the summation of the corresponding leverage scores from 1 and 2 equal the leverage scores from the long regression.

10.3 Leverage scores in experiments

In general, the leverage scores have complex forms, but when the covariates are dummy variables, the leverage scores have more explicit forms. Compute the leverage scores in the following examples.

(1) In a treatment-control experiment with n_1 treated and n_0 control units, the covariate matrix X contains 1 and a dummy variable for the treatment.

(2) In an experiment with n_j units receiving treatment level j ($j = 1, \ldots, J$), the covariate matrix X contains J dummy variables for the treatment levels without 1.

10.4 Implementing the Gauss updating formula

Implement the algorithm in 1–4, and try it on simulated data.

10.5 The distribution of the studentized residual

Show (10.7).

10.6 Leave-one-out coefficient

Show that

$$\hat{\beta} = \sum_{i=1}^{n} w_i \hat{\beta}_{[-i]},$$

and find the weights w_i's. Show they are positive and sum to one. Does $\hat{\beta} = n^{-1} \sum_{i=1}^{n} \hat{\beta}_{[-i]}$ hold in general?

10.7 Cook's distance and the standardized residual

Prove Theorem 10.5.

10.8 The relationship between the standardized and studentized residual

Show that

1. $(n - p - 1)\hat{\sigma}^2_{[-i]} = (n - p)\hat{\sigma}^2 - \hat{\varepsilon}_i^2/(1 - h_{ii})$,

2. there is a monotone relationship between the standardized and studentized residual:
$$\text{studr}_i = \text{standr}_i \sqrt{\frac{n-p-1}{n-p-\text{standr}_i^2}}.$$

10.9 Bias and variance of the jackknife point estimator

Show that under the Gauss–Markov model, $\hat{\beta}_{\text{J}}$ is unbiased but $\text{cov}(\hat{\beta}_{\text{J}}) \succeq \text{cov}(\hat{\beta})$.

10.10 Exact form of the jackknife variance estimate

Without any approximations, derive the exact form of $\hat{V}_{\text{J}}$ for the OLS coefficient.

11

Population Ordinary Least Squares and Inference with a Misspecified Linear Model

11.1 Conditional expectation and its best linear approximation

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. The following theorem states that the conditional expectation $E(y \mid x)$ is the best predictor in terms of the mean squared error.

Theorem 11.1 *For any function $m(x)$, we have the decomposition*

$$E\left[\{y - m(x)\}^2\right] = E\left[\{E(y \mid x) - m(x)\}^2\right] + E\{\text{var}(y \mid x)\}, \tag{11.1}$$

provided the existence of the moments in (11.1)*. This implies*

$$E(y \mid x) = \arg\min_{m(\cdot)} E\left[\{y - m(x)\}^2\right]$$

with the minimum value equaling $E\{\text{var}(y \mid x)\}$, the expectation of the conditional variance of y given x.

Theorem 11.1 is well-known in probability theory. I relegate its proof as a homework problem. We have finite data points, but the function $E(y \mid x)$ lies in an infinite dimensional space. Nonparametric estimation of $E(y \mid x)$ is generally a hard problem especially with a multidimensional x. As a starting point, we often use a linear function of x to approximate $E(y \mid x)$ and define the population OLS coefficient as

$$\beta = \arg\min_{b \in \mathbb{R}^p} \mathcal{L}(b),$$

where

$$\begin{aligned}
\mathcal{L}(b) &= E\left\{(y - x^{\mathrm{T}} b)^2\right\} \\
&= E\left\{y^2 + b^{\mathrm{T}} x x^{\mathrm{T}} b - 2 y x^{\mathrm{T}} b\right\} \\
&= E(y^2) + b^{\mathrm{T}} E\left(x x^{\mathrm{T}}\right) b - 2E\left(y x^{\mathrm{T}}\right) b
\end{aligned}$$

is a quadratic function of b. From the first order condition, we have

$$\frac{\partial \mathcal{L}(b)}{\partial b} \mid_{b=\beta} = 2E\left(xx^{\mathrm{T}}\right)\beta - 2E\left(xy\right) = 0 \Longrightarrow \beta = \left\{E\left(xx^{\mathrm{T}}\right)\right\}^{-1} E\left(xy\right);$$

from the second order condition, we have

$$\frac{\partial^2 \mathcal{L}(b)}{\partial b \partial b^{\mathrm{T}}} = 2E\left(xx^{\mathrm{T}}\right) \geq 0.$$

So β is the unique minimizer of $\mathcal{L}(b)$. The above derivation shows that $x^{\mathrm{T}}\beta$ is the best linear predictor, and the following theorem states precisely that $x^{\mathrm{T}}\beta$ is the best linear approximation to the possibly nonlinear conditional mean $E(y \mid x)$.

Theorem 11.2 *We have*

$$\beta = \arg\min_{b \in \mathbb{R}^p} E\left[\left\{E(y \mid x) - x^{\mathrm{T}}b\right\}^2\right]$$

and also

$$\beta = \left\{E\left(xx^{\mathrm{T}}\right)\right\}^{-1} E\left\{xE(y \mid x)\right\}.$$

As a special case, the covariate contains 1 and a scalar x, the OLS coefficient has the following form.

Corollary 11.1 *For scalar x and y,*

$$(\alpha, \beta) = \arg\min_{(a,b)} E(y - a - bx)^2,$$

where $\alpha = E(y) - E(x)\beta$ and

$$\beta = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} = \rho_{xy}\sqrt{\frac{\mathrm{var}(y)}{\mathrm{var}(x)}}$$

I leave the proofs of the theorem and corollary as homework problems.

11.2 Population OLS decomposition and the FWL Theorem

With β, we can define

$$\varepsilon = y - x^{\mathrm{T}}\beta \tag{11.2}$$

as the population residual. Because we do not have an upper-case letter for ε in Greek, this notation may cause confusion with previous discussion on OLS

where ε denotes the vector of residuals. Here ε is a scalar. By the definition of β, we can verify

$$E(x\varepsilon) = E\left\{x(y - x^{\mathrm{T}}\beta)\right\} = E(xy) - E(xx^{\mathrm{T}})\beta = 0. \tag{11.3}$$

If we include 1 as a component of x, then $E(\varepsilon) = 0$, which, coupled with (11.3), implies that $\mathrm{cov}(x, \varepsilon) = 0$. 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 (11.2) as

$$y = x^{\mathrm{T}}\beta + \varepsilon, \tag{11.4}$$

which holds by the definition of the population OLS coefficient and residual without any modeling assumption. We call (11.4) the population OLS decomposition.

To aid interpretation of the population OLS coefficient, we have the population FWL Theorem.

Theorem 11.3 *In the population OLS decomposition*

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1} + \varepsilon, \tag{11.5}$$

the coefficient of x_k has two equivalent forms

$$\beta_k = \frac{\mathrm{cov}(\tilde{x}_k, y)}{\mathrm{var}(\tilde{x}_k)} = \frac{\mathrm{cov}(\tilde{x}_k, \tilde{y})}{\mathrm{var}(\tilde{x}_k)},$$

where $\tilde{x}_k$ is the residual from the population OLS decomposition

$$x_k = \gamma_0 + \gamma_1 x_1 + \cdots + \ no\ x_k + \gamma_{p-1}x_{p-1} + \tilde{x}_k, \tag{11.6}$$

and $\tilde{y}$ is the residual from the population OLS decomposition

$$y = \delta_0 + \delta_1 x_1 + \cdots + \ no\ x_k + \delta_{p-1}x_{p-1} + \tilde{y}.$$

Proof of Theorem 11.3: We use (11.5) to simplify

$$\begin{aligned}\mathrm{cov}(\tilde{x}_k, y) &= \mathrm{cov}\left(\tilde{x}_k, \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1} + \varepsilon\right) \\ &= \beta_k \underbrace{\mathrm{cov}(\tilde{x}_k, x_k)}_{C_1} + \sum_{l \neq k} \beta_l \underbrace{\mathrm{cov}(\tilde{x}_k, x_l)}_{C_2} + \underbrace{\mathrm{cov}(\tilde{x}_k, \varepsilon)}_{C_3}.\end{aligned}$$

From (11.6), C_1 simplifies to

$$\begin{aligned}C_1 &= \mathrm{cov}(\tilde{x}_k, x_k) \\ &= \mathrm{cov}(\tilde{x}_k, \gamma_0 + \gamma_1 x_1 + \cdots + \ \mathrm{no}\ x_k + \gamma_{p-1}x_{p-1} + \tilde{x}_k) \\ &= \mathrm{cov}(\tilde{x}_k, \tilde{x}_k) \\ &= \mathrm{var}(\tilde{x}_k),\end{aligned}$$

and C_2 simplifies to

$$C_2 = \text{cov}(\tilde{x}_k, x_l) = 0, \qquad (l \neq k).$$

Because $\tilde{x}_k$ is the linear combination of x, the OLS decomposition (11.5) implies that $C_3 = 0$. Therefore,

$$\text{cov}(\tilde{x}_k, y) = \beta_k \text{var}(\tilde{x}_k) \Longrightarrow \beta_k = \frac{\text{cov}(\tilde{x}_k, y)}{\text{var}(\tilde{x}_k)},$$

proving the first identity.

To prove the second identify, we only need to show that $\text{cov}(\tilde{x}_k, y) = \text{cov}(\tilde{x}_k, \tilde{y})$, or, equivalently,

$$\text{cov}(\tilde{x}_k, \delta_0 + \delta_1 x_1 + \cdots + \text{ no } x_k + \delta_{p-1} x_{p-1}) = 0,$$

which holds because of the OLS decomposition (11.6). □

11.3 Population R^2 and partial correlation coefficient

Assume that covariates and outcome are center with mean zero and covariance matrix

$$\text{cov} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \sigma_y^2 \end{pmatrix}.$$

We define the population R^2 as

$$R^2 = \frac{\Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}}{\sigma_y^2},$$

and give several equivalent definitions below. Let β be the population OLS coefficient, and $\hat{y} = E(y) + x^{\text{T}}\beta$ be the best linear predictor.

Theorem 11.4 *R^2 equals the ratio of the variance of the best linear predictor of y and the variance of y itself:*

$$R^2 = \frac{\text{var}(\hat{y})}{\text{var}(y)}.$$

Proof of Theorem 11.4: Because of the centering of x, we can verify that

$$\begin{aligned} \text{var}(\hat{y}) &= \beta^{\text{T}} \text{cov}(x) \beta \\ &= \text{cov}(y, x) \text{cov}(x)^{-1} \text{cov}(x) \text{cov}(x)^{-1} \text{cov}(x, y) \\ &= \text{cov}(y, x) \text{cov}(x)^{-1} \text{cov}(x, y). \end{aligned}$$

□

Theorem 11.5 *R^2 equals the maximum value of the squared Pearson correlation coefficient between y and a linear combination of x:*

$$R^2 = \max_b \rho^2(y, x^{\mathrm{T}}b) = \rho^2(y, \hat{y}).$$

Proof of Theorem 11.5: We have

$$\rho^2(y, x^{\mathrm{T}}b) = \frac{\mathrm{cov}^2(y, x^{\mathrm{T}}b)}{\mathrm{var}(y)\mathrm{var}(x^{\mathrm{T}}b)} = \frac{b^{\mathrm{T}}\Sigma_{xy}\Sigma_{yx}b}{\sigma_y^2 \times b^{\mathrm{T}}\Sigma_{xx}b}.$$

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

$$\sigma_y^2 \times \rho^2(y, x^{\mathrm{T}}b) = \frac{\gamma^{\mathrm{T}}\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yx}\Sigma_{xx}^{-1/2}\gamma}{\gamma^{\mathrm{T}}\gamma},$$

which has maximum value equaling the maximum eigenvalue of $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yx}\Sigma_{xx}^{-1/2}$. This matrix has rank one, so it at most has one non-zero eigenvalue and it must equal the trace of the matrix. So

$$\begin{aligned}
\max_b \rho^2(y, x^{\mathrm{T}}b) &= \sigma_y^{-2}\mathrm{trace}(\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yx}\Sigma_{xx}^{-1/2}) \\
&= \sigma_y^{-2}\mathrm{trace}(\Sigma_{xy}\Sigma_{yx}\Sigma_{xx}^{-1/2}\Sigma_{xx}^{-1/2}) \\
&= \sigma_y^{-2}\mathrm{trace}(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}) \\
&= \sigma_y^{-2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \\
&= R^2.
\end{aligned}$$

We can easily verify that $R^2 = \rho^2(y, \hat{y})$. □

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

$$y = \hat{y} + \tilde{y}, \qquad x = \hat{x} + \tilde{x},$$

where

$$\tilde{y} = \{y - E(y)\} - \{w - E(w)\}^{\mathrm{T}}\beta_y, \qquad \tilde{x} = \{x - E(x)\} - \{w - E(w)\}^{\mathrm{T}}\beta_x,$$

and 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 on the population level.

With a scalar w, we have more explicit formula below.

Theorem 11.6 *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 leave the proof of Theorem 11.6 as a homework problem.

11.4 Inference and prediction in the population OLS

11.4.1 Inference with the EHW standard errors

Based on the IID data $(x_i, y_i)_{i=1}^n$, we can easily obtain the moment estimator for the population OLS coefficient

$$\hat{\beta} = \left(n^{-1} \sum_{i=1}^n x_i x_i^{\mathrm{T}} \right)^{-1} \left(n^{-1} \sum_{i=1}^n x_i y_i \right),$$

and the residuals $\hat{\varepsilon}_i = y_i - x_i^{\mathrm{T}} \hat{\beta}$. Assuming finite fourth moments of (x, y), we can use the law of large numbers to show that $n^{-1} \sum_{i=1}^n x_i x_i^{\mathrm{T}} \to E(xx^{\mathrm{T}})$ and $n^{-1} \sum_{i=1}^n x_i y_i \to E(xy)$, so $\hat{\beta} \to \beta$ in probability; we can use the CLT to show that $n^{-1/2} \sum_{i=1}^n x_i \varepsilon_i \to \mathrm{N}(0, M)$ in distribution, where $M = E(\varepsilon^2 xx^{\mathrm{T}})$, so

$$\sqrt{n}(\hat{\beta} - \beta) \to \mathrm{N}(0, V = B^{-1} M B^{-1}) \tag{11.7}$$

in distribution, where $B = E(xx^{\mathrm{T}})$. So the moment estimator for the asymptotic variance of $\hat{\beta}$ is again the Eicker–Huber–White robust covariance estimator:

$$\hat{V} = n^{-1} \left(n^{-1} \sum_{i=1}^n x_i x_i^{\mathrm{T}} \right)^{-1} \left(n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i^{\mathrm{T}} \right)^{-1} \left(n^{-1} \sum_{i=1}^n x_i x_i^{\mathrm{T}} \right)^{-1}. \tag{11.8}$$

Following the almost the same proof of Theorem 8.3, we can show that $\hat{V}$ is consistent for the asymptotic covariance V. I summarize the formal results below.

Theorem 11.7 *Assume that* $(x_i, y_i)_{i=1}^n \overset{\text{IID}}{\sim} (x, y)$ *with* $E(\|x\|_4^4) < \infty$ *and* $E(y^4) < \infty$. *We have* (11.7) *and* $n\hat{V} \to V$ *in probability.*

So the EHW standard error is not only robust to heteroskedasticity of the errors but also robust to the misspecification of the linear model (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 $E(y \mid x)$.

Boston housing data

```
> library("mlbench")
> library("car")
> 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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

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
```

Comparison of the standard errors:

```
> 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
```

Log transformation of the outcome does not remove the large discrepancy of the standard errors completely.

11.4.2 Conformal prediction based on exchangeability

Without the Gaussian linear model assumption, the prediction interval based on the t pivotal quantity is no longer valid in general. Fortunately, 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 of the data points

$$(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y_{n+1}).$$

Pretending that we know the value $y_{n+1} = h^*$, we can fit OLS using $n + 1$ data points and obtain residuals

$$\hat{\varepsilon}_i(h^*) = y_i - x_i^{\mathrm{T}} \hat{\beta}(h^*), \quad (i = 1, \ldots, n + 1)$$

where we emphasize the dependence of the OLS coefficient and residuals on the unknown h^*. The absolute values of the residuals $|\hat{\varepsilon}_i(h^*)|$'s are also exchangeable, so the rank of $|\hat{\varepsilon}_{n+1}(h^*)|$, denoted by $\hat{R}_{n+1}(h^*)$, must have a uniform distribution over $\{1, 2, \ldots, n, n + 1\}$, a known distribution not depending on anything else. It is a pivotal quantity satisfying

$$\mathrm{pr}\left\{\hat{R}_{n+1}(h^*) \leq \lceil (1 - \alpha)(n + 1) \rceil\right\} \geq 1 - \alpha.$$

Equivalently, this is a statement linking the unknown quantity h^* and observed data, so it gives a confidence set for h^* at level $1 - \alpha$. In practice, we can use a grid search to solve for the inequality involving h^*.

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)
>
> ols.fit.full = lm(medv ~ ., data = BostonHousing,
+                   x = TRUE, y = TRUE)
> beta     = ols.fit.full$coef
> e.sigma  = summary(ols.fit.full)$sigma
> X        = ols.fit.full$x
> Y        = ols.fit.full$y
> Gram.inv = solve(t(X)%*%X)
```

```
> hatmat   = X%*%Gram.inv%*%t(X)
> 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])
+   pred.i = sum(X[i, ]*beta.i)
+   lower.i = pred.i - cvt*e.sigma/(1 - leverage[i])
+   upper.i = pred.i + cvt*e.sigma/(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)
+
+ }
```

The variable `loo.pred` has five columns corresponding to the point predictors, lower and upper intervals based on the Gaussian 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  -3.182107 16.44913  -3.5 16.7
[2,]  8.806641  -2.245882 19.85916  -2.6 20.1
[3,] 12.044154   2.479117 21.60919   2.2 21.8
[4,] 11.025253   1.427318 20.62319   1.2 21.0
[5,] -5.181154 -15.248387  4.88608 -15.0  4.9
[6,]  8.324114  -1.763129 18.41136  -2.0 18.8
```

Figure 11.1 plots the observed outcomes and the prediction intervals for the 20 observations with the outcomes at the bottom, middle and top. The Gaussian and conformal intervals are almost indistinguishable. For the observations with the highest outcome, the predictions are quite poor. Surprisingly, the overall coverage rates across observations are close to 95% for both methods.

```
> ## coverage rates
> apply(loo.cov==1, 2, mean)
[1] 0.9505929 0.9525692
```

Figure 11.2 compares the lengths of the two prediction intervals. Although

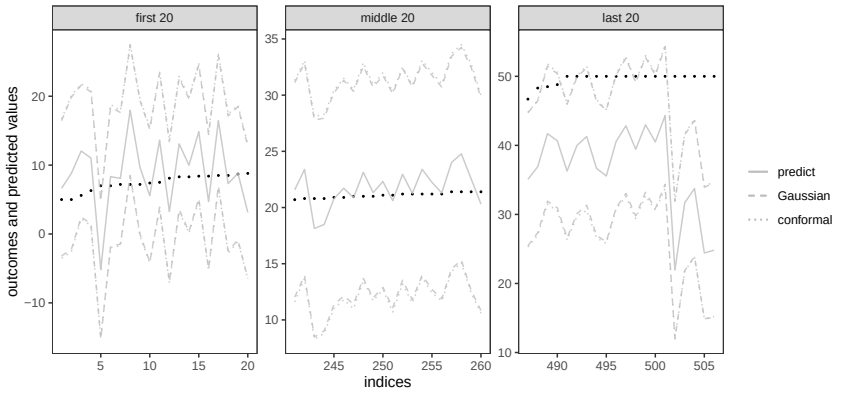


FIGURE 11.1: Leave-one-out prediction intervals with the Boston housing data

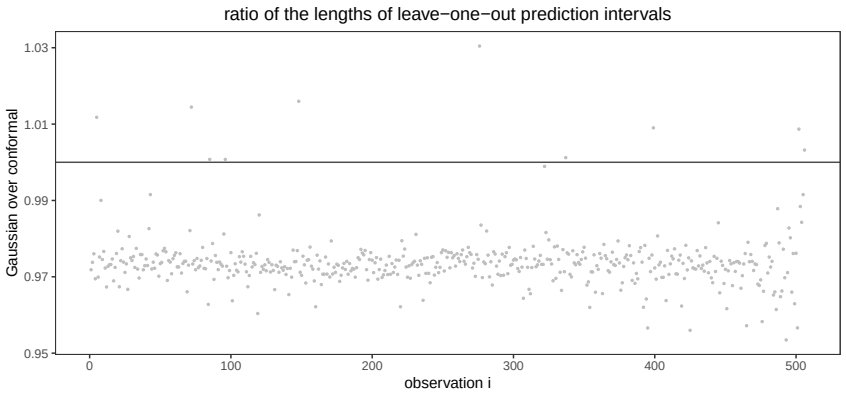


FIGURE 11.2: Boston housing data

the conformal prediction intervals are slightly wider than the Gaussian prediction interval, the differences are rather small, with the ratio of the length above 0.97.

11.5 Population OLS and the restricted mean model

I will give a simple example.

Example 11.1 *Assume that* $x \sim F(x)$, $\varepsilon \sim \mathrm{N}(0,1), x \perp\!\!\!\perp \varepsilon$, *and* $y = x^2 + \varepsilon$.

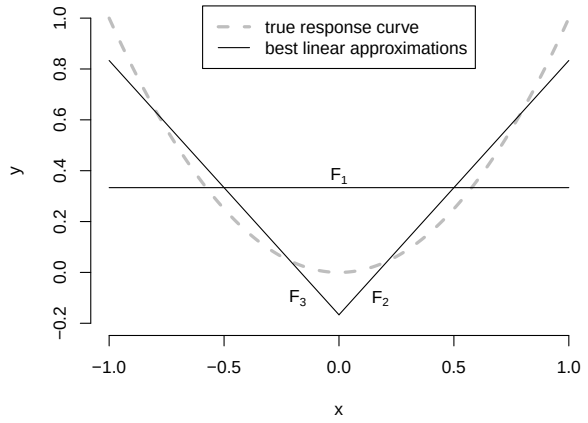


FIGURE 11.3: Best linear approximations correspond to three different distributions of x.

1. *If* $x \sim F_1(x)$ *is uniform within* $[-1, 1]$*, then the best linear approximation is* $1/3 + 0 \cdot x$*.*
2. *If* $x \sim F_2(x)$ *is uniform within* $[0, 1]$*, then the best linear approximation is* $-1/6 + x$*.*
3. *If* $x \sim F_3(x)$ *is uniform within* $[-1, 0]$*, then the best linear approximation is* $-1/6 - x$*.*

Figure 11.3 shows the true data generating process and the best linear approximations.

From the above, we can see that the best linear approximation depends on the distribution of X. This in some sense complicates the interpretation of β from the population OLS decomposition (Buja et al., 2019a). More importantly, this can cause problems if we care about the external validity of statistical inference (Sims, 2010, page 66).

However, if we believe the following restricted mean model

$$E(y \mid x) = x^{\mathrm{T}}\beta \tag{11.9}$$

or, equivalently,

$$y = x^{\mathrm{T}}\beta + \varepsilon, \qquad E(\varepsilon \mid x) = 0,$$

then the population OLS coefficient is the true parameter of interest:

$$\begin{aligned}\left\{E(xx^{\mathrm{T}})\right\}^{-1}E(xy) &= \left\{E(xx^{\mathrm{T}})\right\}^{-1}E\left\{xE(y \mid x)\right\} \\ &= \left\{E(xx^{\mathrm{T}})\right\}^{-1}E(xx^{\mathrm{T}}\beta) \\ &= \beta.\end{aligned}$$

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: the regression model and the correlation model, as shown in Figure 11.4. The left-hand side represents the regression model, or the restricted mean model (11.9). In the regression model, we first generate x and ε under some restrictions, for example, $E(\varepsilon \mid x) = 0$, and then generate the outcome based on $y = x^{\mathrm{T}}\beta + \varepsilon$, 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 $x^{\mathrm{T}}\beta$ and the leftover residual ε. The latter ensures $E(\varepsilon x) = 0$, but the former requires $E(\varepsilon \mid x) = 0$. So the former imposes a stronger assumption (see a homework problem).

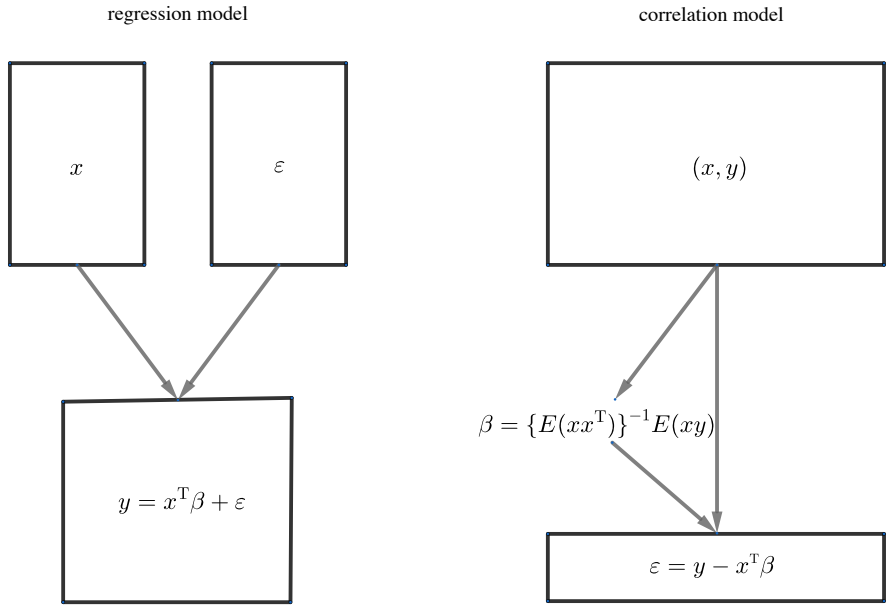


FIGURE 11.4: Freedman's classification of OLS

11.6 Homework problems

11.1 Conditional mean

Prove Theorem 11.1.

11.2 Best linear approximation

Prove Theorem 11.2.

11.3 Univariate population OLS

Prove Corollary 11.1.

11.4 Asymptotics for the population OLS

Give a complete proof for Theorem 11.7.

11.5 Population Cochran's formula

Assume $(y_i, x_{1i}, x_{2i})_{i=1}^n$ are iid, where y_i is a scalar, x_{i1} has dimension k, and x_{i2} has dimension l. We have the following OLS decompositions of random variables

$$\begin{aligned} y_i &= \beta_1^{\mathrm{T}} x_{i1} + \beta_2^{\mathrm{T}} x_{2i} + \varepsilon_i, && (11.10) \\ y_i &= \tilde{\beta}_2^{\mathrm{T}} x_{i2} + \tilde{\varepsilon}_i, && (11.11) \\ x_{i1} &= \delta^{\mathrm{T}} x_{i2} + u_i. && (11.12) \end{aligned}$$

Equation (11.10) is the population long regression, and Equation (11.11) is called the population short regression. In Equation (11.12), δ is a $k \times l$ matrix because it is OLS decomposition of a vector on a vector. You can view (11.12) as OLS decomposition of each component of x_{i1} on x_{i2}.

Show that $\tilde{\beta}_2 = \beta_2 + \delta\beta_1$.

11.6 Canonical correlation analysis

With vector x and y, find the best linear combinations that have the maximum squared Pearson correlation coefficient:

$$\arg\max_{a,b} \rho^2(y^{\mathrm{T}} a, x^{\mathrm{T}} b).$$

We define the maximum value as $\textsc{cc}(X, Y)$, and call it the canonical correlation between X and Y. Show that $\textsc{cc}(X, Y) \geq 0$ and $\textsc{cc}(X, Y) = 0$ if $X \perp\!\!\!\perp Y$.

11.7 Population partial correlation coefficient

Prove Theorem 11.6.

11.8 Independence and correlation

With scalar x and y, show that if $x \perp\!\!\!\perp y$, then $\rho_{yx} = 0$. With another variable w, if $x \perp\!\!\!\perp y \mid w$, does $\rho_{yx|w} = 0$ hold? If so, give a proof; otherwise, give a counterexample.

11.9 Best linear approximation of a quadratic curve

Verify the best linear approximations in Example 11.1.

11.10 Best linear approximation of a cubic curve

Assume that $x \sim \mathrm{N}(0, 1)$, $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, $x \perp\!\!\!\perp \varepsilon$, and $y = x^3 + \varepsilon$. Find the best linear approximation of y based on $(1, x)$.

11.11 Regression versus correlation models

Show that $E(\varepsilon \mid x) = 0$ implies $E(\varepsilon x) = 0$.

11.12 Consistency of the EHW variance estimator

Show that $n\hat{V}$ is consistent for the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta)$, with $\hat{V}$ defined in (11.8). You may need to impose moment conditions on (x, y) for certain laws of large numbers to hold.

11.13 Leave-one-out formula in conformal prediction

11.14 Conformal prediction for multiple outcomes

Assuming exchangeability of

$$(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y_{n+1}), \ldots, (x_{n+k}, y_{n+k}).$$

Construct joint conformal prediction regions for $(y_{n+1}, \ldots, y_{n+k})$ based on (X, Y) and $(x_{n+1}, \ldots, x_{n+k})$.

Part IV

Overfitting and model selection

12

Perils of Overfitting

12.1 David Freedman's simulation

Freedman (1983) simulated data from the following the Gaussian linear model $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ and $\beta = (\mu, 0, \ldots, 0)$. He then computed the sample R^2. Since the covariates does 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 unreasonably large when p has the same magnitude as n.

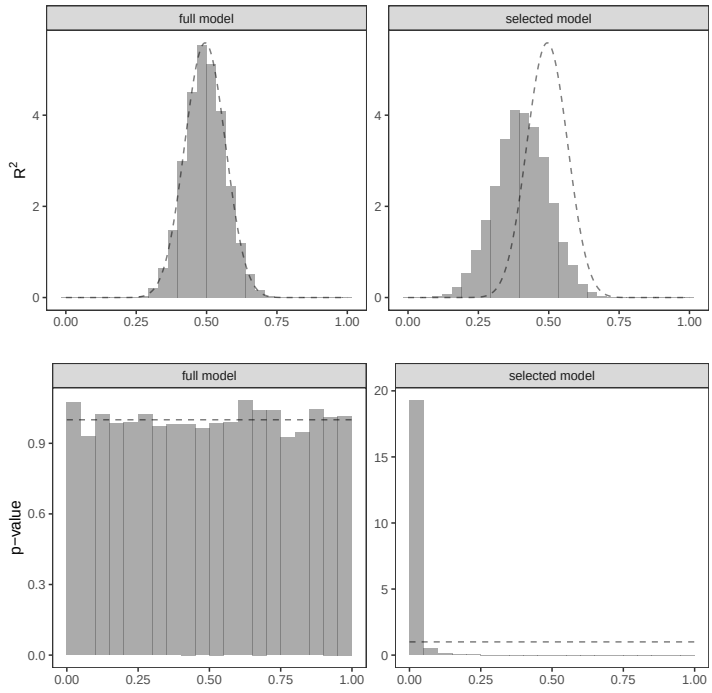


FIGURE 12.1: Freedman's simulation

Figure 12.1 shows the results from the replication of Freedman's simulation

with $n = 100$ and $p = 50$. 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 derived before:

$$R^2 \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right),$$

with the density shown in the $(1,1)$th and $(1,2)$th subfigure of Figure 12.1. 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

$$\text{var}(R^2) = \frac{\frac{p-1}{2} \times \frac{n-p}{2}}{\left(\frac{p-1}{2} + \frac{n-p}{2}\right)^2 \left(\frac{p-1}{2} + \frac{n-p}{2} + 1\right)} = \frac{2(p-1)(n-p)}{(n-1)^2(n+1)}.$$

When $p/n \to 0$, we have

$$E(R^2) \to 0, \quad \text{var}(R^2) \to 0,$$

so Markov's inequality implies that $R^2 \to 0$ in probability. However, when $p/n \to \gamma \in (0,1)$, we have

$$E(R^2) \to \gamma, \quad \text{var}(R^2) \to 0,$$

so Markov's inequality implies that $R^2 \to \gamma$ 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 p-values 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.

The above simulation and calculation gives an important warning: we cannot over interpret the sample R^2 since 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 chase noise rather than 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 in model selection.

12.2 Variance inflation factor

Theorem 12.1 *Consider fixed design matrix X. Let $\hat{\beta}_j$ be the coefficient of x_j of the OLS fit of Y on $1_n, X_1, \ldots, X_s$ with $s \leq p$. Under the model $y_i = f(x_i) + \varepsilon_i$ with an unknown $f(\cdot)$ and the ε_i's uncorrelated with mean zero and variance σ^2, the variance of $\hat{\beta}_j$ is*

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \times \frac{1}{1 - R_j^2},$$

where R_j^2 is the sample R^2 from the linear regression of X_j on 1_n and all other covariates.

Theorem 12.1 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, the coefficient is

$$\tilde{\beta}_j = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j) y_i}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

with variance

$$\text{var}(\tilde{\beta}_j) = \text{var}\left\{ \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j) y_i}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \right\} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \sigma^2}{\left\{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2\right\}^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}.$$

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 variance inflation factor (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. I give a proof of Theorem 12.1 below based on the FWL Theorem.

Proof of Theorem 12.1: Let $\tilde{X}_j$ be the residual vector from the OLS fit of X_j on 1_n and other covariates, which has sample mean zero. 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

$$\text{var}(\hat{\beta}_j) = \frac{\sum_{i=1}^n \tilde{x}_{ij}^2 \sigma^2}{\left\{\sum_{i=1}^n \tilde{x}_{ij}^2\right\}^2} = \frac{\sigma^2}{\sum_{i=1}^n \tilde{x}_{ij}^2}.$$

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 other covariates, it is related to R_j^2 via

$$R_j^2 = 1 - \frac{\sum_{i=1}^n \tilde{x}_{ij}^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \Longrightarrow \sum_{i=1}^n \tilde{x}_{ij}^2 = (1 - R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2.$$

Combining the above two formulas, we have Theorem 12.1. □

12.3 Bias-variance tradeoff

Theorem 12.1 characterizes the variance of the OLS coefficient, but it does not characterize its bias. In general, with a more complex model, we have a higher chance to get closer to the true model $f(x_i)$, and then we have a higher chance to reduce the bias. However, a more complex model results in larger variances of the OLS coefficients. So we face a bias-variance tradeoff.

Consider a simple case where the true data generating process is linear:

$$y_i = \beta_1 + \beta_2 x_{i1} + \cdots + \beta_{s-1} x_{is} + \varepsilon_i. \tag{12.1}$$

Ideally, we want to use the model (12.1) with exactly s covariates. If we underfit the data using a short regression with $q < s$ and

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{i1} + \cdots + \tilde{\beta}_{q-1} x_{iq} + \tilde{\varepsilon}_i, \qquad (i = 1, \ldots, n) \tag{12.2}$$

then the OLS coefficients biased. If we increase the complexity of the model to overfit the data using a long regression with $p > s$ and

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i1} + \cdots + \hat{\beta}_{p-1} x_{ip} + \hat{\varepsilon}_i, \qquad (i = 1, \ldots, n) \tag{12.3}$$

then the OLS coefficients are unbiased. Theorem 12.1, however, shows that the OLS coefficients from the underfitted model (12.2) have smaller variances than those from the overfitted model (12.3).

Example 12.1 *In general we have a sequence of models with increasing complexity. For simplicity, we consider nested models containing* 1_n *and covariates*

$$\{X_1\} \subset \{X_1, X_2\} \subset \cdots \subset \{X_1, \ldots, X_p\}$$

in the following simulation setting. The true linear model is $y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i$ *with* $p = 40$ *but only the first 10 covariates have non-zero coefficients 1 and all other covariates have coefficients 0. Both the training and testing datasets have sample size* $n = 200$*, all covariates have IID* $\mathrm{N}(0, 1)$ *entries and the error terms have IID* $\mathrm{N}(0, 3^2)$ *entries. Figure 12.2 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.*

12.4 Model selection criteria

With a large number of covariates $X_1, \ldots, X_{\overline{p}}$, we want to select a model that has the best performance for prediction. In total, we have $2^{\overline{p}}$ possible models.

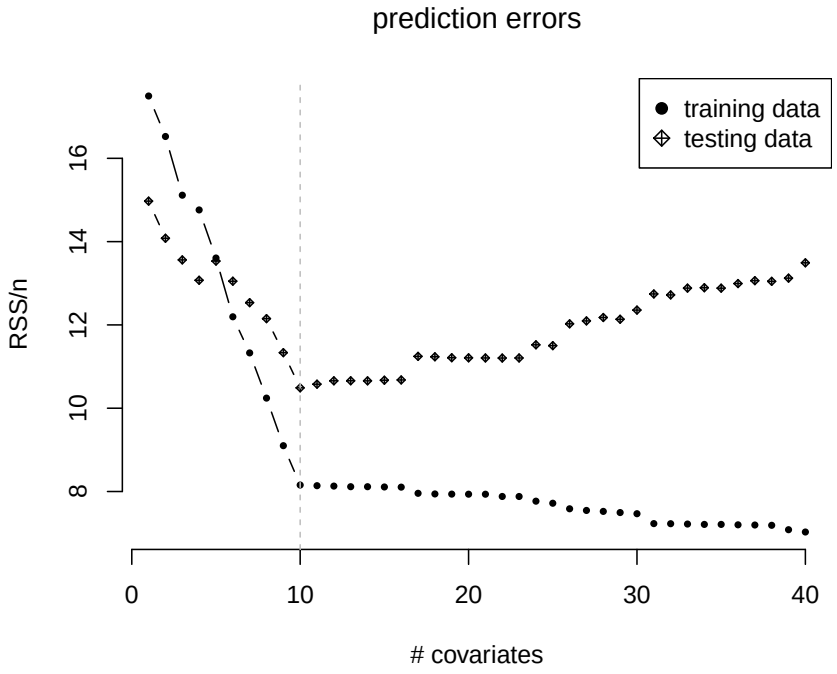


FIGURE 12.2: Bias-variance tradeoff in simulation

Which one is the best? What is the criterion for the best model? Ideally, we want to have the best prediction performance of our linear model in a new dataset. But 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.

12.4.1 RSS, R^2 and adjusted R^2

The first obvious criterion is the 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.

The adjusted R^2 takes into account the complexity of the model:

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{n-1}{n-p}(1-R^2) \\ &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-p)}{\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)} \\ &= 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}.\end{aligned}$$

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. The following theorem shows that $\bar{R}^2$ is closely related to the F statistic in testing two nested Gaussian linear models.

Theorem 12.2 *Test two nest Gaussian 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, and we can also compare the adjusted R^2's from these two models: $\bar{R}_1^2$ and $\bar{R}_2^2$. These are related via

$$F > 1 \Longleftrightarrow \bar{R}_1^2 < \bar{R}_2^2.$$

I leave the proof of the theorem as a homework problem. From Theorem 12.2, $\bar{R}^2$ does not necessarily favor the largest model. 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). So compared to the usual hypothesis testing based on the Gaussian linear model, $\bar{R}^2$ favors unnecessarily larger models.

12.4.2 Information criteria

Taking into account the model complexity, we can find the model with the smallest AIC or BIC, defined as

$$\text{AIC} = n \log \frac{\text{RSS}}{n} + 2p$$

and

$$\text{BIC} = n \log \frac{\text{RSS}}{n} + p \log n,$$

with full names "Akaike's information criterion " and "Bayes information criterion."

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.

12.4.3 Cross-validation (CV)

The first choice is the leave-one-out cross-validation based on the predicted residual:

$$\text{PRESS} = \sum_{i=1}^{n} \hat{\varepsilon}^2_{[-i]} = \sum_{i=1}^{n} \frac{\hat{\varepsilon}^2_{[i]}}{(1-h_{ii})^2}$$

which is called the predicted residual error sum of squares (PRESS) statistic.

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-valiation (GCV) criterion:

$$\text{GCV} = \sum_{i=1}^{n} \frac{\hat{\varepsilon}^2_{[i]}}{(1-p/n)^2} = \text{RSS} \times \left(1-\frac{p}{n}\right)^{-2}.$$

When $p/n \approx 0$, using a Taylor expansion, we can see that

$$\log \text{GCV} = \log \text{RSS} - 2\log\left(1-\frac{p}{n}\right) \approx \log \text{RSS} + \frac{2p}{n} = \text{AIC}/n + \log n,$$

so GCV is approximately equivalent to AIC with small p/n. With large p/n, they may have large difference.

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 general model without simple leave-one-out formulas, it is computationally intensive to obtain PRESS. The K-fold cross-validation is computationally more attractable. The best model has the smallest K-CV, computed as follows:

1. randomly shuffle the observations;
2. split the data into K fold;
3. for each fold k, use all other folds as the training data, and compute the predicted errors on fold k ($k = 1, \ldots, K$);
4. aggregate the prediction errors across K folds, denoted by K-CV.

12.4.4 Best subset and forward/backward selection

Given a model selection criterion, we can select the best model.

For small $\overline{p}$, we can enumerate all $2^{\overline{p}}$ models. The function `regsubsets` in

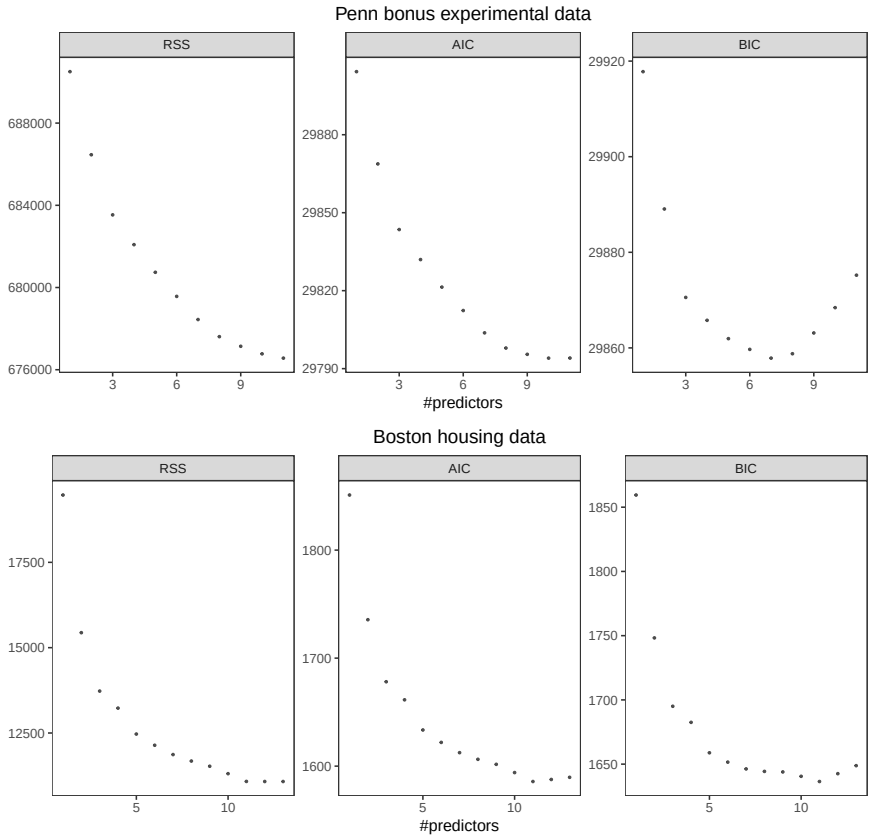


FIGURE 12.3: Best subset selection

the `R` package `leaps` implements this. Figure 12.3 shows the results of the best subset selection in two applications.

For large $\overline{p}$, we can use forward or backward regression. The functions `step` or `stepAIC` in the `MASS` package implement these.

12.5 Homework problems

12.1 Equivalence of F and $\bar{R}^2$

Prove Theorem 12.2.

12.2 Simulation with misspecified linear models

Replicated the simulation in Example 12.1 with correlated covariates and an outcome model with quadratic terms of covariates.

12.3 Best subset selection in `lalonde` *data*

Produce the figure similar to the ones in Figure 12.3 based on the `lalonde` data in the `Matching` package. Report the selected model based on AIC, BIC, PRESS, and GCV.

13

Ridge Regression

13.1 Introduction to the ridge estimator

The OLS estimator minimizes

$$\text{RSS}(b_0, b_1, \ldots, b_p) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2.$$

It has many nice properties, for example, it is BLUE under the Gauss–Markov model, and it follows a Normal distribution that allows for finite-sample exact inference under the Gaussian linear model. However, it can have the following problems which motivate the ridge estimator in this chapter.

First, from the formula

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Y,$$

if the columns of X are highly correlated, then $X^{\mathrm{T}}X$ will be near singular; more extremely, if $p > n$, then $X^{\mathrm{T}}X$ has rank smaller than or equal to n and thus is not invertible. So numerically, the OLS estimator can be unstable due to inverting $X^{\mathrm{T}}X$. Because $X^{\mathrm{T}}X$ must be positive semi-definite, its smallest eigen-value determines whether it is invertible or not. Hoerl and Kennard (1970) proposed the following ridge estimator as a modification of OLS:

$$\hat{\beta}^{\text{ridge}}(\lambda) = (X^{\mathrm{T}} X + \lambda I_p)^{-1} X^{\mathrm{T}} Y \qquad (\lambda > 0). \tag{13.1}$$

Because the smallest eigen-value of $X^{\mathrm{T}}X + \lambda I_p$ is larger than or equal to $\lambda > 0$, the ridge estimator is always well defined.

Second, from 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 and use

$$\hat{\beta}^{\text{ridge}}(\lambda) = \arg \min_{b_0, b_1, \ldots, b_p} \left\{ \text{RSS}(b_0, b_1, \ldots, b_p) + \lambda \sum_{j=1}^{p} b_j^2 \right\}. \tag{13.2}$$

In (13.2), 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 (13.2), we only penalize the slope parameters not the intercept. As a dual problem in optimization, we can also define the ridge estimator as

$$\begin{aligned} \hat{\beta}^{\text{ridge}}(t) = \arg \min_{b_0, b_1, \ldots, b_p} & \text{RSS}(b_0, b_1, \ldots, b_p) \\ \text{s.t. } & \sum_{j=1}^{p} b_j^2 \leq t. \end{aligned} \tag{13.3}$$

Definitions (13.2) and (13.3) are equivalent because for a given λ, we can always find a t such that the solutions from (13.2) and (13.3) are identical.

However, the ridge estimator has an obvious problem: it is not invariant to linear transformation 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:

$$n^{-1} \sum_{i=1}^{n} x_{ij} = 0, \qquad n^{-1} \sum_{i=1}^{n} x_{ij}^2 = 1, \qquad (j = 1, \ldots, p).$$

Because all covariates are centered at zero, the ridge estimator for the intercept, given any values of the slopes and the tuning parameter, is $\hat{\beta}_0^{\text{ridge}} = \bar{y}$. So if we center the outcomes at mean zero, then we can drop the intercept in the ridge estimators defined in (13.2) and (13.3). For descriptive simplicity, we assume these standardization of covariates and outcome, and drop the intercept from now on. Using the matrix form, the ridge estimator minimizes

$$(Y - Xb)^{\text{T}}(Y - Xb) + \lambda b^{\text{T}} b,$$

which is a quadratic function of b. From the first order condition, we have

$$\begin{aligned} & -2X^{\text{T}} \left\{ Y - X^{\text{T}} \hat{\beta}^{\text{ridge}}(\lambda) \right\} + 2\lambda \hat{\beta}^{\text{ridge}}(\lambda) = 0 \\ \Longrightarrow & \hat{\beta}^{\text{ridge}}(\lambda) = (X^{\text{T}} X + \lambda I_p)^{-1} X^{\text{T}} Y, \end{aligned}$$

which coincides with the definition in (13.1). We also have the second order condition

$$2X^{\text{T}} X + 2\lambda I_p \succ 0, \quad (\lambda > 0)$$

which verifies that the ridge estimator is indeed the minimizer. The predicted vector is

$$\hat{Y}^{\text{ridge}}(\lambda) = X\hat{\beta}^{\text{ridge}}(\lambda) = X(X^{\text{T}} X + \lambda I_p)^{-1} X^{\text{T}} Y = H(\lambda) Y,$$

where $H(\lambda) = X(X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}$ is the hat matrix for ridge. When $\lambda = 0$, it reduces to the hat matrix for the OLS; when $\lambda > 0$, it is not a projection matrix because $\{H(\lambda)\}^2 \neq H(\lambda)$.

13.2 Statistical properties

The Gauss–Markov Theorem shows that OLS 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 the ridge estimator improve OLS? I will discuss the statistical properties of the ridge estimator under the Gauss–Markov model. Define

$$S_\lambda = X^{\mathrm{T}}X + \lambda I_p,$$

which appears repeatedly below. So $X^{\mathrm{T}}X = S_0$ and the ridge estimator is $\hat{\beta}^{\mathrm{ridge}}(\lambda) = S_\lambda^{-1}X^{\mathrm{T}}Y$.

The ridge estimator is biased because

$$E\left\{\hat{\beta}^{\mathrm{ridge}}(\lambda)\right\} - \beta = S_\lambda^{-1}X^{\mathrm{T}}X\beta - \beta = S_\lambda^{-1}S_0\beta - \beta$$

is not zero in general. It has covariance matrix

$$\mathrm{cov}\left\{\hat{\beta}^{\mathrm{ridge}}(\lambda)\right\} = \sigma^2 S_\lambda^{-1}X^{\mathrm{T}}XS_\lambda^{-1} = \sigma^2 S_\lambda^{-1}S_0S_\lambda^{-1}.$$

Although the ridge estimator has larger bias than OLS, we will show that it has smaller variance. The mean squared error (MSE) is a measure capturing the bias-variance tradeoff:

$$\mathrm{MSE}(\lambda) = E\left[\left\{\hat{\beta}^{\mathrm{ridge}}(\lambda) - \beta\right\}^{\mathrm{T}}\left\{\hat{\beta}^{\mathrm{ridge}}(\lambda) - \beta\right\}\right].$$

Using the formula of the expectation of a quadratic form, we have

$$\mathrm{MSE}(\lambda) = \underbrace{\beta^{\mathrm{T}}(S_\lambda^{-1}S_0 - I_p)^{\mathrm{T}}(S_\lambda^{-1}S_0 - I_p)\beta}_{C_1} + \underbrace{\sigma^2\mathrm{trace}\left(S_\lambda^{-1}S_0S_\lambda^{-1}\right)}_{C_2}.$$

The following theorem simplifies C_1 and C_2.

Theorem 13.1 *Let the eigen-decomposition of* $S_0 = X^{\mathrm{T}}X$ *be* $S_0 = \Gamma\Lambda\Gamma^{\mathrm{T}}$ *where* Γ *is orthogonal and* $\Lambda = \mathrm{diag}\{\xi_1, \ldots, \xi_p\}$. *Then*

$$C_1 = \lambda^2 \sum_{j=1}^{p} \frac{\gamma_j^2}{(\xi_j + \lambda)^2}, \tag{13.4}$$

where $\gamma = \Gamma^{\mathrm{T}}\beta = (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$, *and*

$$C_2 = \sigma^2 \sum_{j=1}^{p} \frac{\xi_j}{(\xi_j + \lambda)^2}. \tag{13.5}$$

Proof of 13.1 T:he eigen-decomposition of $S_0 = X^{\mathrm{T}}X$ implies the eigen-decomposition of S_λ:

$$S_\lambda = X^{\mathrm{T}}X + \lambda I_p = \Gamma\Lambda\Gamma^{\mathrm{T}} + \lambda\Gamma\Gamma^{\mathrm{T}} = \Gamma(\Lambda + \lambda I_p)\Gamma^{\mathrm{T}}$$

so its inverse is

$$S_\lambda^{-1} = \{\Gamma(\Lambda + \lambda I_p)\Gamma^{\mathrm{T}}\}^{-1} = \Gamma(\Lambda + \lambda I_p)^{-1}\Gamma^{\mathrm{T}} \equiv \Gamma V_\lambda \Gamma^{\mathrm{T}},$$

where

$$V_\lambda = (\Lambda + \lambda I_p)^{-1} = \begin{pmatrix} \frac{1}{\xi_1+\lambda} & & \\ & \ddots & \\ & & \frac{1}{\xi_p+\lambda} \end{pmatrix}.$$

We first calculate C_1. It equals

$$\begin{aligned} C_1 &= \beta^{\mathrm{T}}(S_\lambda^{-1}S_0 - I_p)^{\mathrm{T}}(S_\lambda^{-1}S_0 - I_p)\beta \\ &= \beta^{\mathrm{T}}(\Gamma V_\lambda\Gamma^{\mathrm{T}}\Gamma\Lambda\Gamma^{\mathrm{T}} - \Gamma\Gamma^{\mathrm{T}})^{\mathrm{T}}(\Gamma V_\lambda\Gamma^{\mathrm{T}}\Gamma\Lambda\Gamma^{\mathrm{T}} - \Gamma\Gamma^{\mathrm{T}})\beta \\ &= \beta^{\mathrm{T}}\Gamma(V_\lambda\Lambda - I_p)^{\mathrm{T}}(V_\lambda\Lambda - I_p)\Gamma^{\mathrm{T}}\beta \\ &= \gamma^{\mathrm{T}}(V_\lambda\Lambda - I_p)^2\gamma, \end{aligned}$$

where

$$(V_\lambda\Lambda - I_p)^2 = \begin{pmatrix} \frac{\xi_1}{\xi_1+\lambda} - 1 & & \\ & \ddots & \\ & & \frac{\xi_p}{\xi_p+\lambda} - 1 \end{pmatrix}^2 = \begin{pmatrix} \left(\frac{\lambda}{\xi_1+\lambda}\right)^2 & & \\ & \ddots & \\ & & \left(\frac{\lambda}{\xi_p+\lambda}\right)^2 \end{pmatrix}.$$

So (13.4) holds.

We then calculate C_2. It equals

$$\begin{aligned} C_2 &= \sigma^2\mathrm{trace}\left(S_\lambda^{-1}S_0S_\lambda^{-1}\right) \\ &= \sigma^2\mathrm{trace}\left(\Gamma V_\lambda\Gamma^{\mathrm{T}}\Gamma\Lambda\Gamma^{\mathrm{T}}\Gamma V_\lambda\Gamma^{\mathrm{T}}\right) \\ &= \sigma^2\mathrm{trace}\left(\Gamma V_\lambda\Lambda V_\lambda\Gamma^{\mathrm{T}}\right) \\ &= \sigma^2\mathrm{trace}\left(V_\lambda\Lambda V_\lambda\Gamma^{\mathrm{T}}\Gamma\right) \\ &= \sigma^2\mathrm{trace}\left(V_\lambda\Lambda V_\lambda\right), \end{aligned}$$

where

$$V_\lambda\Lambda V_\lambda = \begin{pmatrix} \frac{\xi_1}{(\xi_1+\lambda)^2} & & \\ & \ddots & \\ & & \frac{\xi_p}{(\xi_p+\lambda)^2} \end{pmatrix}.$$

so (13.5) holds. □

Theorem 13.1 shows the bias-variance tradeoff for the ridge estimator. The MSE is

$$\text{MSE}(\lambda) = C_1 + C_2 = \lambda^2 \sum_{j=1}^{p} \frac{\gamma_j^2}{(\xi_j + \lambda)^2} + \sigma^2 \sum_{j=1}^{p} \frac{\xi_j}{(\xi_j + \lambda)^2}.$$

When $\lambda = 0$, the ridge estimator reduces to the OLS estimator: the bias is zero and the variance $\sigma^2 \sum_{j=1}^{p} \xi_j^{-1}$ 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 tradeoff.

13.3 Selection of the tuning parameter

13.3.1 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:

$$\begin{aligned} \frac{\partial \text{MSE}(\lambda)}{\partial \lambda} &= 2 \sum_{j=1}^{p} \gamma_j^2 \frac{\lambda}{\xi_j + \lambda} \frac{\xi_j + \lambda - \lambda}{(\xi_j + \lambda)^2} - 2\sigma^2 \sum_{j=1}^{p} \frac{\xi_j}{(\xi_j + \lambda)^3} = 0 \\ &\iff \lambda \sum_{j=1}^{p} \frac{\gamma_j^2 \xi_j}{(\xi_j + \lambda)^3} = \sigma^2 \sum_{j=1}^{p} \frac{\xi_j}{(\xi_j + \lambda)^3}. \end{aligned} \tag{13.6}$$

However, (13.6) is not directly useful because we do not know γ and σ^2. Three methods below try to solve (13.6) approximately.

Dempster et al. (1977) used OLS to construct unbiased estimator $\hat{\sigma}^2$ and $\hat{\gamma} = \Gamma^{\mathrm{T}} \hat{\beta}$, and then solve λ from

$$\lambda \sum_{j=1}^{p} \frac{\hat{\gamma}_j^2 \xi_j}{(\xi_j + \lambda)^3} = \hat{\sigma}^2 \sum_{j=1}^{p} \frac{\xi_j}{(\xi_j + \lambda)^3},$$

which is a nonlinear equation of λ.

Hoerl et al. (1975) assumed that $X^{\mathrm{T}} X = I_p$. Then $\xi_j = 1$ $(j = 1, \ldots, p)$ 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

$$\lambda_{\text{LW}} = p\hat{\sigma}^2/\hat{\beta}^{\text{T}}\Lambda\hat{\beta}$$

to weight the β_j's based on the eigenvalues of $X^{\text{T}}X$.

But all these methods requires 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 $p > n$.

13.3.2 Based on prediction

For prediction, we need slightly different criterion. Without estimating (β, σ^2), we can use leave-one-out cross-validation. The leave-one-out formula for ridge is below.

Theorem 13.2 *Define $\hat{\beta}(\lambda) = (X^{\text{T}}X + \lambda I_p)^{-1}X^{\text{T}}Y$ as the ridge coefficient (dropping the superscript "ridge" for simplicity), $\hat{\varepsilon}(\lambda)$ as the residual vector using the full data, and $h_{ii}(\lambda) = x_i^{\text{T}}(X^{\text{T}}X + \lambda I_p)^{-1}x_i$ as the (i,i)th diagonal element of $H(\lambda) = X(X^{\text{T}}X+\lambda I_p)^{-1}X^{\text{T}}$. Define $\hat{\beta}_{[-i]}(\lambda)$ as the ridge cofficient without observation i, and $\hat{\varepsilon}_{[-i]}(\lambda) = y_i - x_i^{\text{T}}\hat{\beta}_{[-i]}(\lambda)$ as the predicted residual. The leave-one-out formulas for ridge regression are*

$$\hat{\beta}_{[-i]}(\lambda) = \hat{\beta}(\lambda) - \{1 - h_{ii}(\lambda)\}^{-1}(X^{\text{T}}X + \lambda I_p)^{-1}x_i\hat{\varepsilon}_i(\lambda)$$

and

$$\hat{\varepsilon}_{[-i]}(\lambda) = \hat{\varepsilon}_i(\lambda)/\{1 - h_{ii}(\lambda)\}.$$

So the PRESS statistic for ridge is

$$\text{PRESS}(\lambda) = \sum_{i=1}^{n}\left\{\hat{\varepsilon}_{[-i]}(\lambda)\right\}^2 = \sum_{i=1}^{n}\frac{\{\hat{\varepsilon}_i(\lambda)\}^2}{\{1 - h_{ii}(\lambda)\}^2}.$$

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 $n^{-1}\text{trace}\{H(\lambda)\}$:

$$\text{GCV}(\lambda) = \frac{\sum_{i=1}^{n}\{\hat{\varepsilon}_i(\lambda)\}^2}{\left[1 - n^{-1}\text{trace}\{H(\lambda)\}\right]^2}.$$

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.

13.3.3 Numerical examples

We can use the following numerical example to illustrate this tradeoff. Figure 13.1 plots the bias, variance, and the MSE against λ.

In the first setting, we simulate uncorrelated covariates.

```
> library(MASS)
> n = 200
> p = 100
> beta = rep(1/sqrt(p), p)
> sig  = 1/2
> ## independent Normals
> X    = matrix(rnorm(n*p), n, p)
> ## standardize the covariates
> X    = scale(X)
> X    = X*sqrt(n/(n-1))
> Y    = as.vector(X%*%beta + rnorm(n, 0, sig))
> 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)
> pdf("biasvariancetradeoffridgeplot.pdf",
+     height = 10, width = 8.5)
> 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")
>
>
>
> ## ridge regression
> 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")
>
>
```

```
> ## 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)
```

In the second setting, we simulate 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))
> eigenxx = eigen(t(X)%*%X)
> xis     = eigenxx$values
> gammas  = t(eigenxx$vectors)%*%beta
>
> lambda.seq = seq(0, 800, 1)
> 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)
> 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")
>
>
>
> ## ridge regression
> 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$kLW, lty = 4, col = "grey")
> legend("right",
+        c("MSE", "GCV", "HKB", "LW"),
+        lty = 1:4, col = "grey", bty = "n")
>
>
>
> ## prediction
> X.new    = matrix(rnorm(n*p), n, p) + rnorm(n, 0, 0.5)
> X.new    = scale(X.new)
> X.new    = X.new*matrix(sqrt(n/(n-1)), n, p)
> Y.new    = as.vector(X.new%*%beta + rnorm(n, 0, sig))
> predict.error = Y.new - X.new%*%ridge.fit$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("correlated␣covariates", side = 1,
+       line = -28, outer = TRUE, font.main = 1, cex=1.5)
```

Figure 13.1 shows the bias-variance trade off. Overall, GCV works the best for selecting λ for prediction.

13.4 Computation of ridge

Consider the case with $n \geq p$, and the singular value decomposition of X is

$$X = UDV^{\mathrm{T}},$$

where $D \in \mathbb{R}^{p\times p}$ is a diagonal matrix containing all singular values of X, $U \in \mathbb{R}^{n\times p}$ has orthonormal columns such that $U^{\mathrm{T}}U = I_p$, and $V \in \mathbb{R}^{p\times p}$ is an orthogonal matrix with $VV^{\mathrm{T}} = V^{\mathrm{T}}V = I_p$. The ridge coefficient equals

$$\begin{aligned}
\hat{\beta}^{\mathrm{ridge}}(\lambda) &= (X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}Y \\
&= (VDU^{\mathrm{T}}UDV^{\mathrm{T}} + \lambda I_p)^{-1}VDU^{\mathrm{T}}Y \\
&= V(D^2 + \lambda I_p)^{-1}V^{\mathrm{T}}VDU^{\mathrm{T}}Y \\
&= V(D^2 + \lambda I_p)^{-1}DU^{\mathrm{T}}Y \\
&= V\mathrm{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)_{p\times p} U^{\mathrm{T}}Y,
\end{aligned}$$

and the predicted vector equals

$$\begin{aligned}
\hat{Y}(\lambda) &= X\hat{\beta}^{\mathrm{ridge}}(\lambda) \\
&= UDV^{\mathrm{T}}V\mathrm{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)_{p\times p} U^{\mathrm{T}}Y \\
&= UD\mathrm{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)_{p\times p} U^{\mathrm{T}}Y \\
&= U\mathrm{diag}\left(\frac{d_j^2}{d_j^2+\lambda}\right)_{p\times p} U^{\mathrm{T}}Y.
\end{aligned}$$

We have similar formulas for the case with $n < p$; see a homework problem. The above formulas allows us to compute the ridge coefficient and predictor vector for many values of λ without inverting each $X^{\mathrm{T}}X + \lambda I_p$.

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 $\bar{x}_1, \ldots, \bar{x}_p, \bar{y}$ be the means of the covariates and outcome, and let $\mathrm{sd}_j = \{n^{-1}\sum_{i=1}^n (x_{ij} - \bar{x}_i)^2\}^{1/2}$ be the standard deviation of the covariates which are report as `scales` in the output of `lm.ridge`. From the ridge coefficents $\{\hat{\beta}_1^{\mathrm{ridge}}(\lambda), \ldots, \hat{\beta}_p^{\mathrm{ridge}}(\lambda)\}$ based on the standardized variables, we can obtain the predicted values based on the original variables as

$$\hat{y}_i - \bar{y} = \hat{\beta}_1^{\mathrm{ridge}}(\lambda)(x_{i1} - \bar{x}_1)/\mathrm{sd}_1 + \cdots + \hat{\beta}_p^{\mathrm{ridge}}(\lambda)(x_{ip} - \bar{x}_p)/\mathrm{sd}_p$$

or, equivalently,

$$\hat{y}_i = \bar{y} - \hat{\beta}_1^{\mathrm{ridge}}(\lambda)\bar{x}_1/\mathrm{sd}_1 - \cdots - \hat{\beta}_p^{\mathrm{ridge}}(\lambda)\bar{x}_p/\mathrm{sd}_p + \hat{\beta}_1^{\mathrm{ridge}}(\lambda)/\mathrm{sd}_1 \times x_{i1} + \cdots + \hat{\beta}_p^{\mathrm{ridge}}(\lambda)/\mathrm{sd}_p \times x_{ip}.$$

13.5 Homework problems

13.1 Ridge coefficient as posterior mode under a Normal prior

Show that if

$$Y \sim \mathrm{N}(X\beta, \sigma^2), \quad \beta \sim \mathrm{N}(0, \tau^2 I_p),$$

then the posterior mode of $\beta \mid Y$ equals $\hat{\beta}^{\mathrm{ridge}}(\sigma^2/\tau^2)$.

13.2 Ridge as OLS with pseudo data

Show that $\hat{\beta}^{\mathrm{ridge}}(\lambda)$ equals the OLS coefficient of $\tilde{Y}$ on $\tilde{X}$ with augmented data

$$\tilde{Y} = \begin{pmatrix} Y \\ 0_p \end{pmatrix}, \qquad \tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda} I_p \end{pmatrix}.$$

13.3 Leave-one-out formulas for ridge

Use the result from the last problem to prove Theorem 13.2.

13.4 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:

$$\arg\min_{b \in \mathbb{R}^p} \left\{ (Y - Xb)^{\mathrm{T}} (Y - Xb) + \lambda b^{\mathrm{T}} Q b \right\}.$$

13.5 Degrees of freedom of ridge regression

For a predictor $\hat{Y}$ for Y, define the degrees of freedom of the predictor as $\sum_{i=1}^{n} \mathrm{cov}(y_i, \hat{y}_i)/\sigma^2$. Calculate the degrees of freedom of ridge regression in terms of the eigenvalues of $X^{\mathrm{T}}X$.

13.6 Extending the simulation in Figure 13.1

Re-run the simulation that generates Figure 13.1, and report the λ selected by Dempster et al. (1977)'s method, PRESS, and K-fold CV.

13.7 An equivalent form of ridge coefficient

Show that the ridge coefficient has two equivalent forms:

$$(X^{\mathrm{T}}X + \lambda I_p)^{-1} X^{\mathrm{T}} Y = X^{\mathrm{T}} (X X^{\mathrm{T}} + \lambda I_n)^{-1} Y.$$

The left-hand side involves inverting a $p \times p$ matrix, and it is more useful when $p < n$; the right-hand side involves inverting an $n \times n$ matrix, so it is more useful when $p > n$.

13.8 Computation of ridge with $n < p$

When $n < p$, X has singular value decomposition $X = UDV^{\mathrm{T}}$, 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 $UU^{\mathrm{T}} = U^{\mathrm{T}}U = I_n$, and $V \in \mathbb{R}^{p\times n}$ has orthonormal columns with $V^{\mathrm{T}}V = I_n$. Show that the ridge coefficient and the predicted value have the same form as the case with $n > p$.

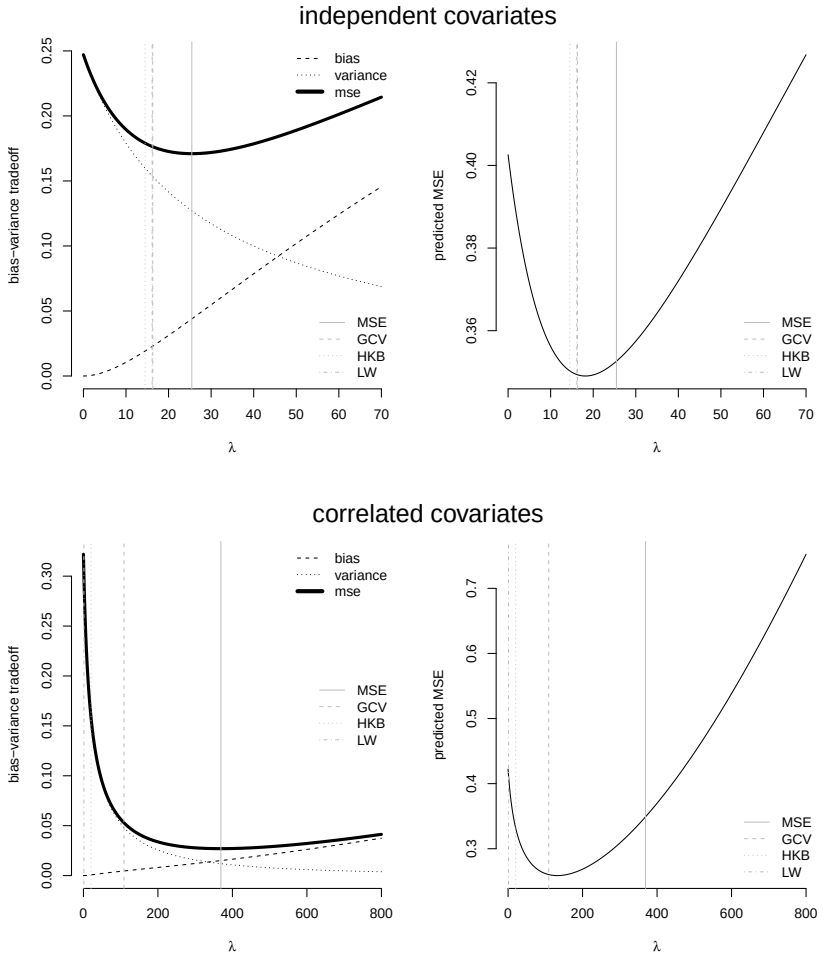


FIGURE 13.1: Bias-variance trade off in ridge regression

14

Lasso

14.1 Introduction to the lasso estimator

Ridge regression works well for prediction, but it may be difficult to interpret many small but non-zero coefficients. Tibshirani (1996) proposed to use the lasso, 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 the 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.

The lasso estimator is

$$\hat{\beta}^{\text{lasso}}(\lambda) = \arg \min_{b_0, b_1, \ldots, b_p} \left\{ \text{RSS}(b_0, b_1, \ldots, b_p) + \lambda \sum_{j=1}^{p} |b_j| \right\}, \tag{14.1}$$

or, by duality,

$$\begin{aligned} \hat{\beta}^{\text{lasso}}(t) = \arg \min_{b_0, b_1, \ldots, b_p} & \text{RSS}(b_0, b_1, \ldots, b_p) \\ \text{s.t.} & \sum_{j=1}^{p} |b_j| \leq t. \end{aligned} \tag{14.2}$$

The two forms of lasso are equivalent in the sense that for a given λ in (14.1), there exists a t such that the solution for (14.2) is identical to the solution for (14.1). We will focus on the form (14.1). Similar to the ridge estimator, the lasso is not invariant to the linear transformation of X. We proceed after standardizing the covariates and outcome in the same way in the ridge estimator. For the same reason, we can drop the intercept after standardization.

14.2 Comparing the lasso and the ridge

The ridge and lasso estimators are very similar: both minimize a penalized version of the residual sum of squares. Their difference is in the penalty term:

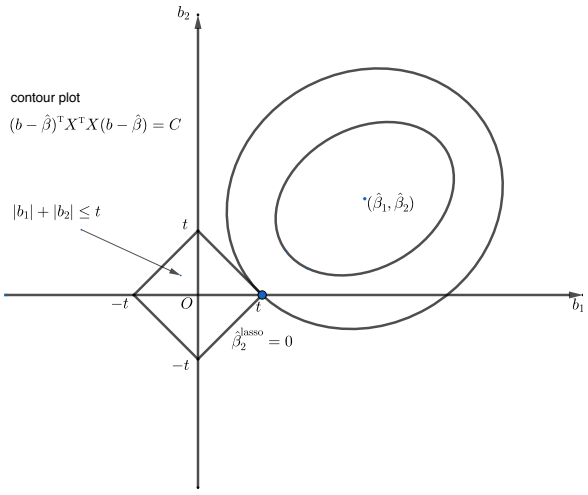


FIGURE 14.1: Lasso with a sparse solution

ridge uses an L_2 penalty, i.e., the L_2 norm of the coefficient, and lasso uses an L_1 penalty, i.e., the L_1 norm of the coefficient. Compared to ridge, lasso can give sparse solutions due to the non-smooth penalty term. That is, estimators of some coefficients are exactly zero.

We gain insight from the contour plot of the residual sum of squares as a function of b. With a well-defined $\hat{\beta}$, we have

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

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 $\sum_{j=1}^{p} |b_j| \le t$. If this happens, the intersection of the contour plot of $(Y - Xb)^{\mathrm{T}}(Y - Xb)$ and the border of the restriction region $\sum_{j=1}^{p} |b_j| \le t$ can be at some axis. For example, Figure 14.1 shows a case with $p = 2$, and the lasso estimator hit the x-axis, resulting in an 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 $(Y - Xb)^{\mathrm{T}}(Y - Xb)$ and the border of the restriction region is at an edge of the region. For example, Figure 14.2 shows a case with a non-sparse lasso solution.

In contrast, the restriction region of ridge is a circle, so the ridge solution does not hit any axis unless the original OLS coefficient is zero. Figure 14.3 shows the general ridge estimator.

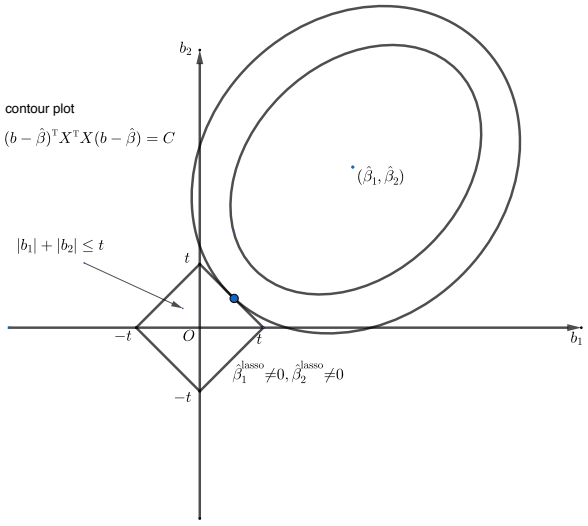


FIGURE 14.2: Lasso with a non-sparse solution

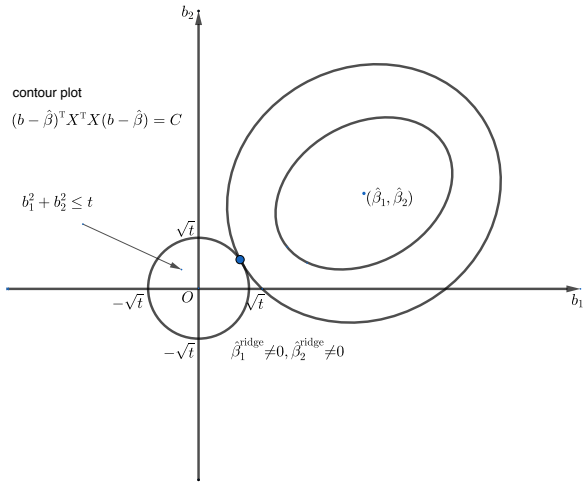


FIGURE 14.3: Ridge

14.3 Computing the lasso estimator via coordinate descent

14.3.1 The soft-thresholding lemma

Lemma 14.1 *Given b_0 and λ,*

$$\arg\min_{b\in R} \frac{1}{2}(b-b_0)^2 + \lambda|b| = \text{sign}(b_0)\,(|b_0|-\lambda)_+$$
$$= \begin{cases} b_0-\lambda, & \text{if } b_0 \geq \lambda, \\ 0 & \text{if } -\lambda \leq b_0 \leq \lambda, \\ b_0+\lambda & \text{if } b_0 \leq -\lambda, \end{cases}$$

where $\text{sign}(\cdot)$ *is the sign of a real number and* $(\cdot)_+ = \max(\cdot, 0)$ *is the positive part of a real number.*

The solution in Lemma 14.1 is a function of b_0 and λ, and we will use

$$S(b_0, \lambda) = \text{sign}(b_0)\,(|b_0|-\lambda)_+$$

from now on. For a given $\lambda > 0$, it is a function of b_0 illustrated by Figure 14.4. The proof of Lemma 14.1 is to solve a one-dimensional optimization problem, which is relegated as a homework problem.

14.3.2 Coordinate descent for the lasso

For a given $\lambda > 0$, we can use the following algorithm:

1. Standardize the covariates:

$$n^{-1}\sum_{i=1}^n x_{ij} = 0, \quad n^{-1}\sum_{i=1}^n x_{ij}^2 = 1, \qquad (j = 1, \ldots, p),$$

and center the outcome:

$$n^{-1}\sum_{i=1}^n y_i = 0.$$

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:

$$\min_{b_1,\ldots,b_p} \frac{1}{2n}\sum_{i=1}^n (y_i - b_1x_{i1} - \cdots - b_px_{ip})^2 + \lambda\sum_{j=1}^p |b_j|.$$

Initialize $\hat{\beta}$.

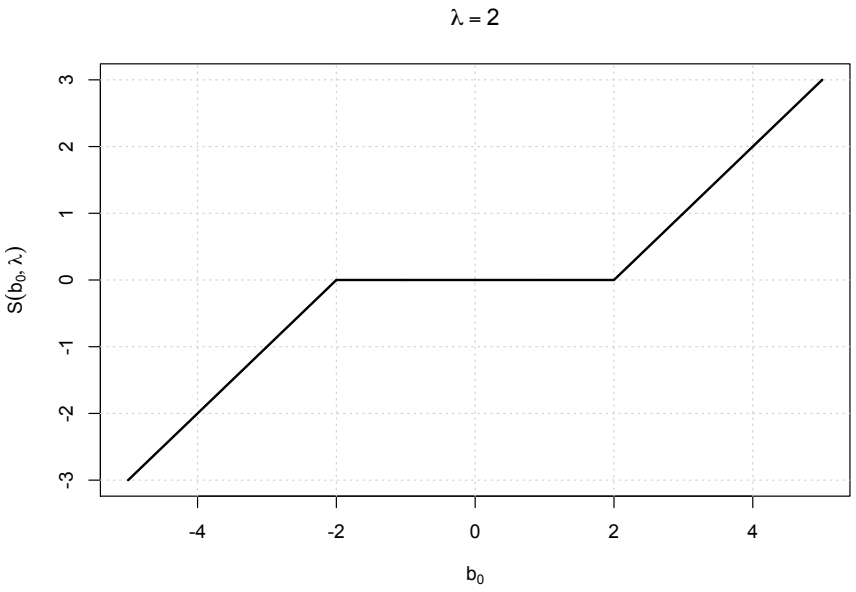


FIGURE 14.4: Soft-thresholding

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

$$\frac{1}{2n} \sum_{i=1}^{n} (r_{ij} - b_j x_{ij})^2 + \lambda |b_j|.$$

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 $\frac{1}{2}(b_j - \hat{\beta}_{j,0})^2 + \lambda |b_j|$. Using Lemma 14.1, we have

$$\hat{\beta}_j = S(\hat{\beta}_{j,0}, \lambda).$$

3. Iterative until convergence.

We can start with a large λ and all zero coefficients. We then gradually decrease λ, and for each λ, we apply the above algorithm. We finally select λ via K-fold cross-validation.

14.4 Example: comparing OLS, ridge and lasso

In the Boston housing data, OLS, ridge and lasso have similar performance in out-of-sample prediction. Lasso and ridge have similar coefficients. See the first panel of Figure 14.5.

```
> 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$medv
> dat = data.frame(yvector, xmatrix)
>
> ## linear regression
> bostonlm = lm(yvector ~ ., data = dat[trainindex, ])
> predicterror = dat$yvector[- trainindex] -
+                     predict(bostonlm, dat[- trainindex, ])
> mse.ols = sum(predicterror^2)/length(predicterror)
>
> ## ridge regression
> lambdas= seq(0, 5, 0.01)
> 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] 29.37365 29.07174 28.88161
```

But if we artificially add 200 columns of covariates of pure noise $\mathrm{N}(0, 1)$,

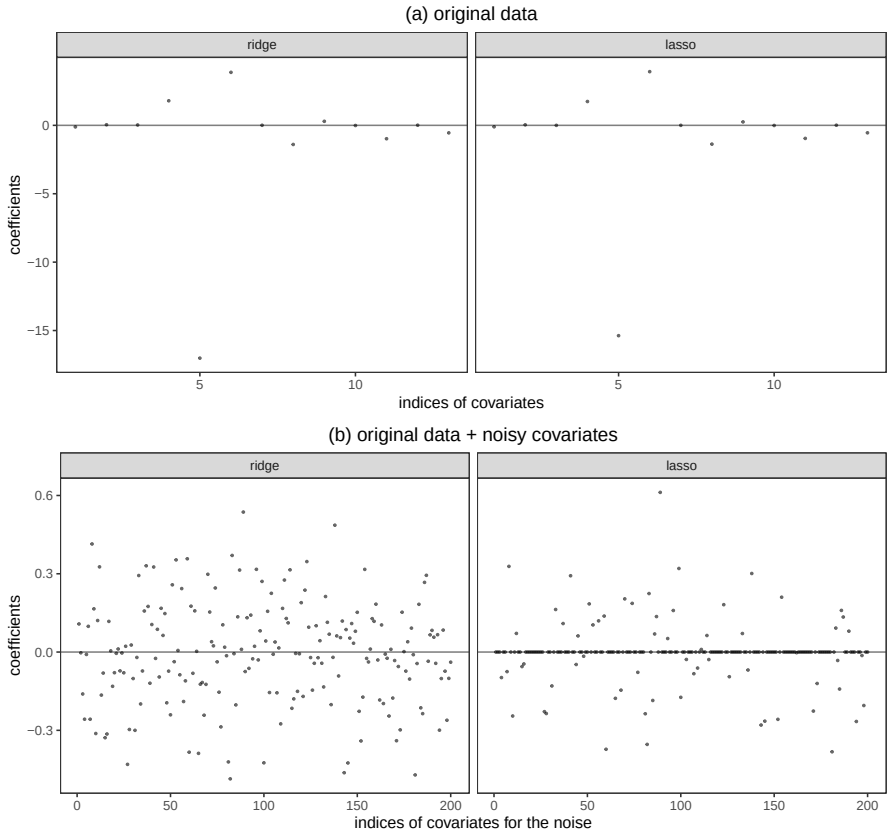


FIGURE 14.5: Comparing ridge, lasso, and elastic net

then ridge and lasso performs much better. Lasso can automatically shrink many coefficients to zero. See the second panel of Figure 14.5.

```
> ## 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, ])
> predicterror = dat$yvector[- trainindex] -
+   predict(bostonlm, dat[- trainindex, ])
> mse.ols = sum(predicterror^2)/length(predicterror)
>
> ## ridge regression
> lambdas= seq(100, 150, 0.01)
> 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
```

14.5 Other shrinkage estimators

A general class of shrinkage estimator is the bridge estimator (Frank and Friedman, 1993):

$$\hat{\beta}(\lambda) = \arg \min_{b_0, b_1, \ldots, b_p} \left\{ \text{RSS}(b_0, b_1, \ldots, b_p) + \lambda \sum_{j=1}^{p} |b_j|^q \right\}$$

or, by duality,

$$\hat{\beta}(t) = \arg \min_{b_0, b_1, \ldots, b_p} \text{RSS}(b_0, b_1, \ldots, b_p)$$
$$\text{s.t. } \sum_{j=1}^{p} |b_j|^q \leq t.$$

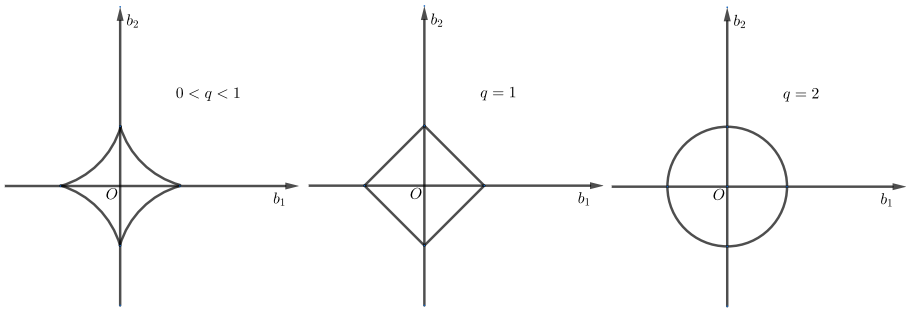
Figure 14.6 shows the constraints corresponding to different values of q.

Zou and Hastie (2005) proposed the elastic net which combines the penalty of lasso and ridge:

$$\hat{\beta}^{\text{enet}}(\lambda, \alpha) = \arg \min_{b_0, b_1, \ldots, b_p} \left[\text{RSS}(b_0, b_1, \ldots, b_p) + \lambda \sum_{j=1}^{p} \left\{ \alpha b_j^2 + (1-\alpha)|b_j| \right\} \right].$$

Figure 14.7 compares the constraints corresponding to ridge, lasso, and elastic net. Because the constraint of elastic net is not smooth, it encourages sparse solution as 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 called `glmnet`.



(a) $0 < q < 1$ (b) $q = 1$ (c) $q = 2$

FIGURE 14.6: Shrinkage estimators

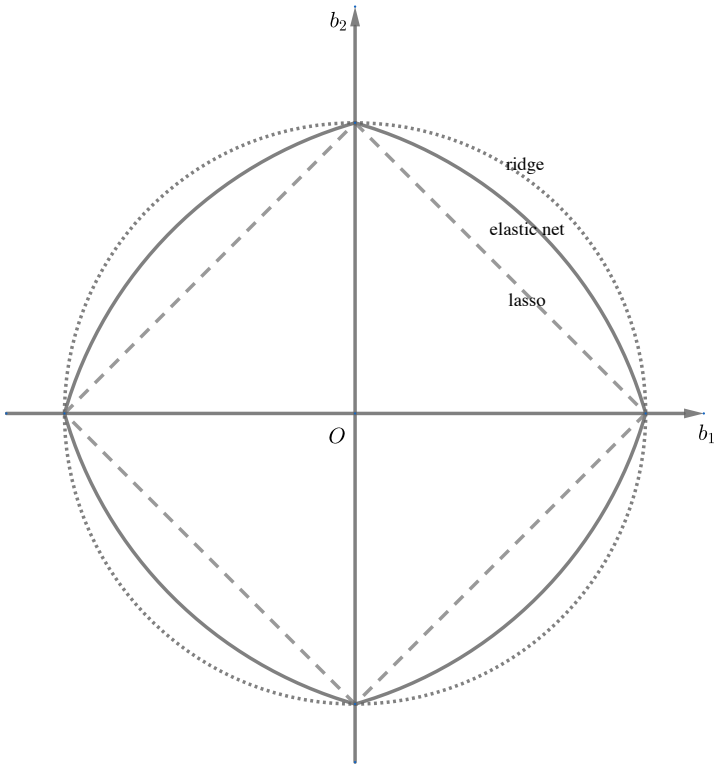


FIGURE 14.7: Comparing ridge, lasso, and elastic net

14.6 Homework problems

14.1 The soft-thresholding lemma

Prove Lemma 14.1.

14.2 Ridge and lasso with orthogonal design

Consider the special case with standardized and orthogonal design matrix:

$$X^{\mathrm{T}} 1_n = 0, \qquad X^{\mathrm{T}} X = I_p.$$

For a fixed $\lambda \geq 0$, find the explicit formulas of the following problems in terms of the least squares estimator $\hat{\beta}$ and λ:

$$\hat{\beta}^{\mathrm{ridge}}(\lambda) = \arg\min_{b \in \mathbb{R}^p} \left\{ \|Y - Xb\|^2 + \lambda \|b\|^2 \right\},$$
$$\hat{\beta}^{\mathrm{lasso}}(\lambda) = \arg\min_{b \in \mathbb{R}^p} \left\{ \|Y - Xb\|^2 + \lambda \|b\|_1 \right\},$$
$$\hat{\beta}^{\mathrm{subset}}(\lambda) = \arg\min_{b \in \mathbb{R}^p} \left\{ \|Y - Xb\|^2 + \lambda \|b\|_0 \right\},$$

where

$$\|b\|^2 = \sum_{j=1}^{p} b_j^2, \quad \|b\|_1 = \sum_{j=1}^{p} |b_j|, \quad \|b\|_0 = \sum_{j=1}^{p} 1(b_j \neq 0).$$

14.3 Coordinate descent for the elastic net

Give the detailed coordinate descent algorithm for the elastic net.

14.4 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 14.4. Add $p = 2n$ columns of covariates of random noise, and compare OLS, ridge, and lasso.

Part V

Transformation and weighting

15

Transformations in OLS

Transforming the outcome and covariates is an important topic in linear model. Carroll and Ruppert (1988) is a textbook on this topic. This chapter discusses some important special cases.

15.1 Transformation of the outcome

Although we can view

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i, \quad (i = 1, \ldots, n)$$

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 for the original outcome, but may perform well for certain transformation of the outcome.

15.1.1 Log transformation

With positive, especially heavy-tailed outcomes, a standard transformation is the log transformation. So we fit a linear model

$$\log y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i, \quad (i = 1, \ldots, n).$$

The interpretation of the coefficients changes a little bit. Because

$$\frac{\partial \log \hat{y}_i}{\partial x_{ij}} = \frac{\partial \hat{y}_i}{\hat{y}_i} \Big/ \partial x_{ij} = \hat{\beta}_j,$$

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. With the log transformation applied to both the outcome and covariate, $\hat{\beta}_j$ is the elasticity.

For a nonnegative outcome, we can modify the log transformation to $\log(y_i + 1)$.

15.1.2 Box–Cox transformation

Power transformation is another important class of transformations. 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

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{\mathrm{d}y^\lambda / \mathrm{d}\lambda}{1} = \lim_{\lambda \to 0} y^\lambda \log y = \log y,$$

so as a function of λ, $g_\lambda(y)$ is continuous at 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 Gaussian linear model:

$$Y_\lambda = \begin{pmatrix} y_{\lambda 1} \\ \vdots \\ y_{\lambda n} \end{pmatrix} = \begin{pmatrix} g_\lambda(y_1) \\ \vdots \\ g_\lambda(y_1) \end{pmatrix} \sim \mathrm{N}(X\beta, \sigma^2 I_n).$$

The density function of Y_λ is

$$f(Y_\lambda) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (Y_\lambda - X\beta)^{\mathrm{T}} (Y_\lambda - X\beta) \right\}.$$

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

$$f(Y) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (Y_\lambda - X\beta)^{\mathrm{T}} (Y_\lambda - X\beta) \right\} \prod_{i=1}^n y_i^{\lambda-1}.$$

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 $(Y_\lambda - X\beta)^{\mathrm{T}}(Y_\lambda - X\beta)$, i.e., we can run OLS of Y_λ on X to obtain

$$\hat{\beta}(\lambda) = (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}} Y_\lambda.$$

Given λ, maximizing the likelihood function is equivalent to first obtaining $\hat{\beta}(\lambda)$ and then obtaining $\hat{\sigma}^2(\lambda) = n^{-1} Y_\lambda (I_n - H) Y_\lambda$. The final step is to

maximize the profile likelihood as a function of λ:

$$L(\hat{\beta}(\lambda), \hat{\sigma}^2(\lambda), \lambda) = \left\{2\pi\hat{\sigma}^2(\lambda)\right\}^{-n/2} \exp\left\{-\frac{n\hat{\sigma}^2(\lambda)}{2\hat{\sigma}^2(\lambda)}\right\} \prod_{i=1}^{n} y_i^{\lambda-1}.$$

Dropping some constants, the log profile likelihood function of λ is

$$l_{\mathrm{P}}(\lambda) = -\frac{n}{2}\log\hat{\sigma}^2(\lambda) + (\lambda - 1)\sum_{i=1}^{n}\log y_i.$$

The `boxcox` function in the `R` package `MASS` plots $l_{\mathrm{P}}(\lambda)$, finds it maximizer $\hat{\lambda}$, and construct a 95% confidence interval $[\hat{\lambda}_{\mathrm{L}}, \hat{\lambda}_{\mathrm{U}}]$ based on the following asymptotic pivotal quantity

$$2\left\{l_{\mathrm{P}}(\hat{\lambda}) - l_{\mathrm{P}}(\lambda)\right\} \overset{\mathrm{a}}{\sim} \chi^2_1,$$

which holds by the Wilks' Theorem. In practice, we often use the λ values within $[\hat{\lambda}_{\mathrm{L}}, \hat{\lambda}_{\mathrm{U}}]$ that have more scientific meanings.

The the `jobs` data, $\lambda = 2$ seems a plausible value.

```
> library(MASS)
> library(mediation)
> pdf("boxcox_jobs.pdf", height = 4, width = 8.5)
> par(mfrow = c(1, 3))
> jobslm = lm(job_seek ~ treat + econ_hard + depress1 + sex + age + occp + marital +
+               nonwhite + educ + income, data = jobs)
> jobslm2 = lm(I(job_seek^2) ~ treat + econ_hard + depress1 + sex + age + occp + marital +
+                nonwhite + educ + income, data = jobs)
> boxcox(jobslm, lambda = seq(1.5, 3, 0.1), plotit = TRUE)
> 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)
```

The the Penn bonus experiment data, $\lambda = 0.3$ seems a plausible value. However, the residual plot does not seem Normal, making the Box–Cox transformation not very meaningful.

```
> penndata = read.table("Penn46_ascii.txt")
> pdf("boxcox_penn.pdf", height = 4, width = 8.5)
> 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)
```

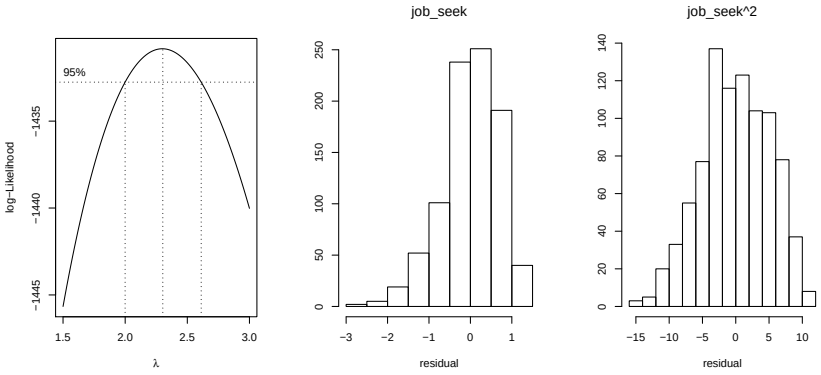


FIGURE 15.1: Box–Cox transformation in the `jobs` data

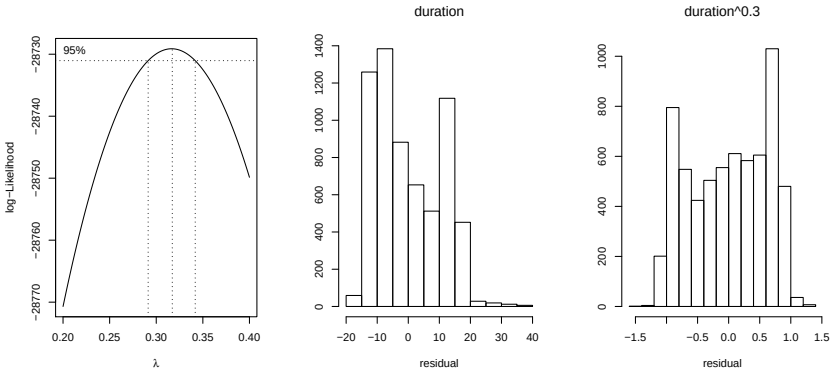


FIGURE 15.2: Box–Cox transformation in the Penn bonus experiment data

15.2 Transformation of the covariates

15.2.1 Dummy variable

If a covariate is a factor taking k discrete values, then we can create $k-1$ dummy variables indicating $k-1$ levels with the rest as the reference level.

15.2.2 Interaction

With two binary covariates $F_1, F_2 \in \{0,1\}^n$, we can fit an OLS:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 F_1 + \hat{\beta}_2 F_2 + \hat{\beta}_{12} F_1 \circ F_2 + \hat{\varepsilon},$$

where $F_1 \circ F_2$ denotes the component-wise product between the vectors F_1 and F_2. We can show that

$$\hat{\beta}_{12} = (\bar{y}_{11} - \bar{y}_{10}) - (\bar{y}_{01} - \bar{y}_{00}), \tag{15.1}$$

where $\bar{y}_{f_1 f_2}$ is the average value of the y_i's with F_1 equaling f_1 and F_2 equaling f_2. Practitioners also interpret the coefficient of the product term of two continuous variable as interaction. However, interaction is a subtle concept in statistics (Cox, 1984; Berrington de González and Cox, 2007). Below I will discuss two issues.

15.2.2.1 Removable interaction

The significance of the interaction term differs with y and $\log(y)$.

```
> 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
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> 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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

15.2.2.2 Main effect and interaction

In the OLS fit below, we observe significant main effects.

```
> ## 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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

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 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

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 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Based on the linear model with interaction

$$E(y_i \mid x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2},$$

better definitions of the main effects are

$$n^{-1} \sum_{i=1}^{n} \frac{\partial E(y_i \mid x_{i1}, x_{i2})}{\partial x_{i1}} = n^{-1} \sum_{i=1}^{n} (\beta_1 + \beta_{12} x_{i2}) = \beta_1 + \beta_{12} \bar{x}_2$$

and

$$n^{-1} \sum_{i=1}^{n} \frac{\partial E(y_i \mid x_{i1}, x_{i2})}{\partial x_{i2}} = n^{-1} \sum_{i=1}^{n} (\beta_2 + \beta_{12} x_{i1}) = \beta_2 + \beta_{12} \bar{x}_1.$$

So when the covariates are centered, 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

$$\frac{\partial^2 E(y_i \mid x_{i1}, x_{i2})}{\partial x_{i1} \partial x_{i2}} = \beta_{12}.$$

15.2.3 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

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 \cdots + \beta_p x_i^{p-1} + \varepsilon_i.$$

With two covariates, we can use

$$(1, x_{1i}, \ldots, x_{i1}^d, x_{i2}, \ldots, x_{i2}^l)$$

or

$$(1, x_{1i}, \ldots, x_{i1}^d, x_{i2}, \ldots, x_{i2}^l, x_{i1} x_{i2}, \ldots, x_{i1}^d x_{i2}^l).$$

Other basis expansion

$$y_i = f(x_i) + \varepsilon_i \cong \sum_{j=1}^{J} \beta_j S_j(x_i) + \varepsilon_i,$$

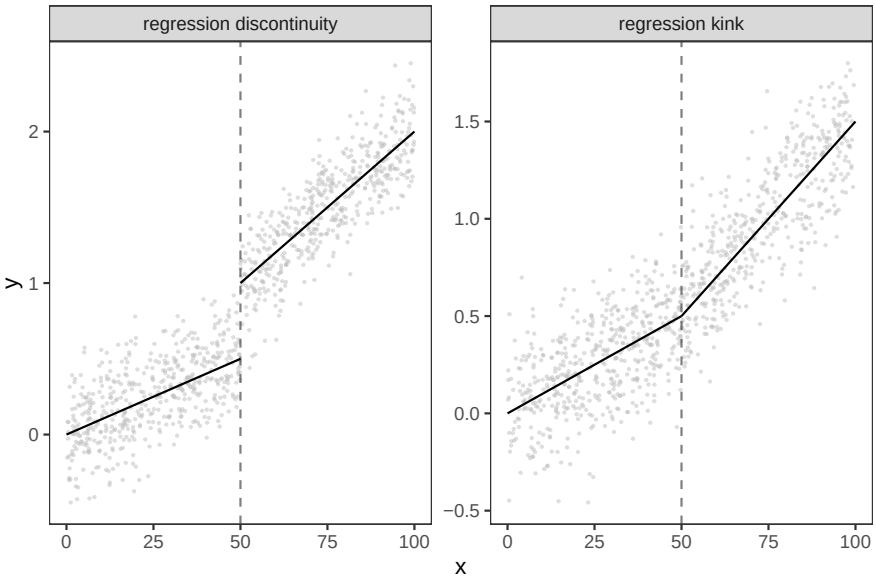


FIGURE 15.3: Regression discontinuity and kink

where the $S_j(x_i)$'s are basis functions.

The generalized additive model is an extension to the multivariate case:

$$\begin{aligned} y_i &= f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \varepsilon_i \\ &\cong \sum_{j=1}^{J_1} \beta_{1j} S_j(x_{i1}) + \cdots + \sum_{j=1}^{J_p} \beta_{pj} S_j(x_{ip}) + \varepsilon_i. \end{aligned}$$

In the `R` package `mgcv`, the function `gam` automatically fits generalized additive model.

15.2.4 Regression discontinuity and regression kink

The left panel of Figure 15.3 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

$$y_i = \begin{cases} \beta_1 + \beta_2 x_i + \varepsilon_i & x_i \le c, \\ (\beta_1 + \beta_3) + (\beta_2 + \beta_4) x_i + \varepsilon_i, & x_i > c. \end{cases}$$

Testing the discontinuity at c is equivalent to testing

$$(\beta_1 + \beta_3) + (\beta_2 + \beta_4)\, c = \beta_1 + \beta_2 c \Longleftrightarrow \beta_3 + \beta_4 c = 0.$$

If we center the covariates at c, then

$$y_i = \beta_1 + \beta_2(x_i - c) + \beta_3 1\,(x_i > c) + \beta_4(x_i - c)1\,(x_i > c) + \varepsilon_i$$

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 15.3 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:

$$y_i = \beta_1 + \beta_2 R_c(x_i) + \beta_3(x_i - c) + \varepsilon_i$$

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).

15.3 Homework problems

15.1 Interaction and difference-in-difference

Show (15.1). Note that it is a pure linear algebra fact.

15.2 Invariance of the interaction

In Section 15.2.2.2, the point estimate and standard error of the coefficient of the interaction term remains the same no matter whether we center the covariates or not. Prove that this result holds in general.

15.3 Piecewise linear regression

Generate data in the same way as the example in Section 15.2.3, and fit a continuous piecewise linear function with cutoff points $0, 0.2, 0.4, 0.6, 0.8, 1$.

16

Weighted Least Squares

16.1 Generalized least squares

We can extend the Gauss–Markov model to allow for a general covariance structure of the error term:

$$Y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, \quad \text{cov}(\varepsilon) = \sigma^2\Sigma. \tag{16.1}$$

where σ^2 is unknown and Σ is a known positive definite matrix. Two leading cases of generalized least squares are

$$\Sigma = \text{diag}\left\{w_1^{-1}, \ldots, w_n^{-1}\right\}, \tag{16.2}$$

which corresponds to a diagonal covariance matrix, and

$$\Sigma = \text{diag}\left\{\Sigma_1, \ldots, \Sigma_K\right\} \tag{16.3}$$

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 model (16.1), we can still use the OLS estimator $\hat{\beta} = (X^{\text{T}}X)^{-1}X^{\text{T}}Y$ which is unbiased

$$E(\hat{\beta}) = (X^{\text{T}}X)^{-1}X^{\text{T}}E(Y) = (X^{\text{T}}X)^{-1}X^{\text{T}}X\beta = \beta$$

with covariance matrix

$$\begin{aligned}\text{cov}(\hat{\beta}) &= \text{cov}\left\{(X^{\text{T}}X)^{-1}X^{\text{T}}Y\right\} \\ &= (X^{\text{T}}X)^{-1}X^{\text{T}}\text{cov}(Y)X(X^{\text{T}}X)^{-1} \\ &= \sigma^2(X^{\text{T}}X)^{-1}X^{\text{T}}\Sigma X(X^{\text{T}}X)^{-1}.\end{aligned} \tag{16.4}$$

The OLS estimator is BLUE under the Gauss–Markov model, but it is not under (16.1). Then what is the BLUE? We can transform (16.1) into the Gauss–Markov model by standardizing the error term:

$$\Sigma^{-1/2}Y = \Sigma^{-1/2}X\beta + \Sigma^{-1/2}\varepsilon.$$

Define $\tilde{Y} = \Sigma^{-1/2}Y, \tilde{X} = \Sigma^{-1/2}X$ and $\tilde{\varepsilon} = \Sigma^{-1/2}\varepsilon$. The model (16.1) reduces to

$$\tilde{Y} = \tilde{X}\beta + \tilde{\varepsilon}, \quad E(\tilde{\varepsilon}), \quad \text{cov}(\tilde{\varepsilon}) = \sigma^2 I_n,$$

which is the Gauss–Markov model for the transformed variables $\tilde{Y}$ and $\tilde{X}$. Using the Gauss–Markov Theorem, we know that the BLUE is

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

which is unbiased

$$E(\hat{\beta}_\Sigma) = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}E(Y) = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}X\beta = \beta$$

with covariance matrix

$$\begin{aligned}\mathrm{cov}(\hat{\beta}_\Sigma) &= \mathrm{cov}\left\{(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}Y\right\} \\ &= (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}\mathrm{cov}(Y)\Sigma^{-1}X(X^{\mathrm{T}}\Sigma^{-1}X)^{-1} \\ &= \sigma^2(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}\Sigma\Sigma^{-1}X(X^{\mathrm{T}}\Sigma^{-1}X)^{-1} \\ &= \sigma^2(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}. \end{aligned} \tag{16.5}$$

From the Gauss–Markov Theorem, we know that

$$\mathrm{cov}(\hat{\beta}_\Sigma) \preceq \mathrm{cov}(\hat{\beta}),$$

which, coupled with (16.4) and (16.5), implies the following pure linear algebra inequality:

Corollary 16.1 *We have*

$$(X^{\mathrm{T}}\Sigma^{-1}X)^{-1} \preceq (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\Sigma X(X^{\mathrm{T}}X)^{-1}.$$

This chapter focuses on the first covariance structure (16.2) and a later chapter will discuss the second (16.3). The Σ in (16.2) results in the weighted least squares (WLS) estimator

$$\hat{\beta}_w = \hat{\beta}_\Sigma = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}Y = \left(\sum_{i=1}^n w_i x_i x_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^n w_i x_i y_i.$$

From the derivation above, we can also write the WLS estimator as

$$\begin{aligned}\hat{\beta}_w &= \arg\min_b (Y_* - X_* b)^{\mathrm{T}}(Y - X'b) \\ &= \arg\min_b (Y - Xb)^{\mathrm{T}}\Sigma^{-1}(Y - Xb) \\ &= \arg\min_b \sum_{i=1}^n w_i (y_i - x_i^{\mathrm{T}} b)^2 \\ &= \arg\min_b \sum_{i=1}^n (y_{*i} - x_{*i}^{\mathrm{T}} b)^2,\end{aligned}$$

where $y_{*i} = w_i^{1/2} y_i$ and $x_{*i} = w_i^{1/2} x_i$. So WLS is equivalent to the OLS with transformed variables, with the weights inversely proportional to the variances of the errors.

16.2 Some special WLS

An important practical question is: where do these weights come from? The first two cases below have weights motivated by heteroskedasticity, and the last two cases are motivated by issues beyond heteroskedasticity.

16.2.1 Feasible generalized least squares

Assume that ε has mean zero and covariance diag $\left\{\sigma_1^2, \ldots, \sigma_n^2\right\}$. 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):

1. Run OLS of Y on X to obtain the residual vector $\hat{\varepsilon}$, and obtain the component-wise squared residual $\hat{\varepsilon}^2$;
2. Run OLS of $\log(\hat{\varepsilon}^2)$ on X to obtain the fitted values and exponentiate them to obtain $(\hat{\sigma}_i^2)_{i=1}^n$;
3. Run WLS of Y on X with weights $(\hat{\sigma}_i^{-2})_{i=1}^n$ to obtain

$$\hat{\beta}_{\text{FGLS}} = \left(\sum_{i=1}^n \hat{\sigma}_i^{-2} x_i x_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^n \hat{\sigma}_i^{-2} x_i y_i.$$

In the above step 2, we can change the model based on our understanding of the heteroskedasticity. 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
> dat.res$medv = log((ols.fit$residuals)^2)
> t.res.ols = lm(medv ~ ., data = dat.res)
> w.fgls = exp(-t.res.ols$fitted.values)
> fgls.fit = lm(medv ~ ., weights = w.fgls, data = BostonHousing)
> ols.fgls = cbind(summary(ols.fit)$coef[,1:3],
+                  summary(fgls.fit)$coef[,1:3])
> round(ols.fgls, 3)
            Estimate Std. Error t value Estimate Std. Error t value
(Intercept)   36.459      5.103   7.144    9.499      4.064    2.34
crim          -0.108      0.033  -3.287   -0.081      0.044   -1.82
zn             0.046      0.014   3.382    0.030      0.011    2.67
indus          0.021      0.061   0.334   -0.035      0.038   -0.92
chas1          2.687      0.862   3.118    1.462      1.119    1.31
nox          -17.767      3.820  -4.651   -7.161      2.784   -2.57
rm             3.810      0.418   9.116    5.675      0.364   15.59
age            0.001      0.013   0.052   -0.044      0.008   -5.50
dis           -1.476      0.199  -7.398   -0.927      0.139   -6.68
```

```
rad            0.306      0.066    4.613     0.170      0.051     3.31
tax           -0.012      0.004   -3.280    -0.010      0.002    -4.14
ptratio       -0.953      0.131   -7.283    -0.700      0.094    -7.45
b              0.009      0.003    3.467     0.014      0.002     6.54
lstat         -0.525      0.051  -10.347    -0.158      0.036    -4.38
```

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 is misspecified. Otherwise, both estimators are unbiased for the same true coefficient, and they should not be so different even in the presence of randomness.

Romano and Wolf (2017) highlighted the efficiency gain from the FGLS compared to 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, maybe it is not worth choosing FGLS over OLS.

16.2.2 Regression with aggregated data

In some case, (y_i, x_i) come from aggregated 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 `fultongen` data from the `R` package `ri`. It 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. It is naturally to use the weights defined by `n`, the total number of people. Figure 16.1 is the scatterplot. In this example, OLS and WLS give similar results although `n` varies a lot across precincts.

```
> library("ei")
> data(fultongen)
> ols.fit = lm(t ~ x, data = fultongen)
> wls.fit = lm(t ~ x, weights = n, data = fultongen)
> compare = cbind(summary(ols.fit)$coef[,1:3],
+                 summary(wls.fit)$coef[,1:3])
```

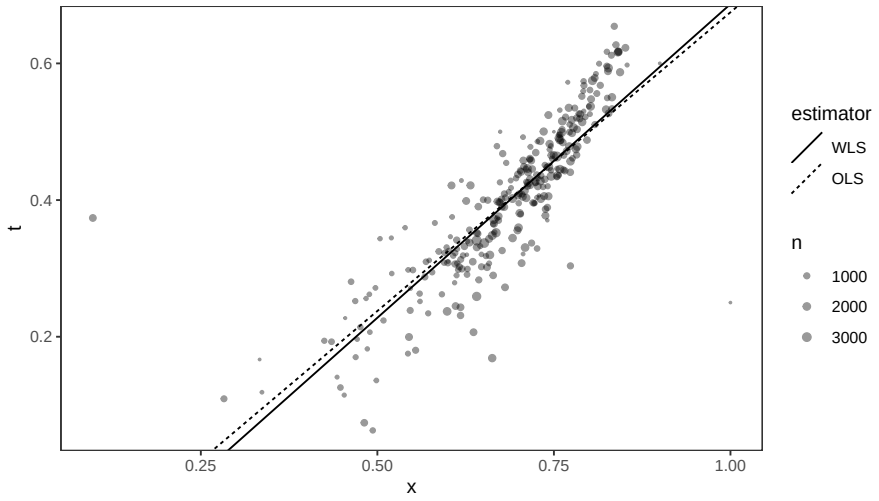


FIGURE 16.1: Fulton data

```
> round(compare, 4)
             Estimate Std. Error t value Estimate Std. Error t value
(Intercept)    -0.20      0.024    -8.3    -0.23      0.027    -8.5
x               0.87      0.035    25.2     0.92      0.038    23.8
```

In the above, we can interpret the coefficient of `x` as the precinct-level relationship between the fraction voting in 1994 and that in 1992. Political scientists are interested in using the aggregated data to infer the individual voting behaviors, which is a topic beyond this chapter. See King (2013) for more details.

16.2.3 Local linear regression

Calculus tells us that locally we can approximate any smooth function by a linear function even though the original function can be highly nonlinear. The left panel of Figure 16.2 shows that in the neighbor of $x = 0.4$, even a sin 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 linear 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 linear line $\hat{y}(x) = \hat{\alpha} + \hat{\beta}(x - x_0)$, the predicted value at $x = x_0$ is the intercept $\hat{\alpha}$.

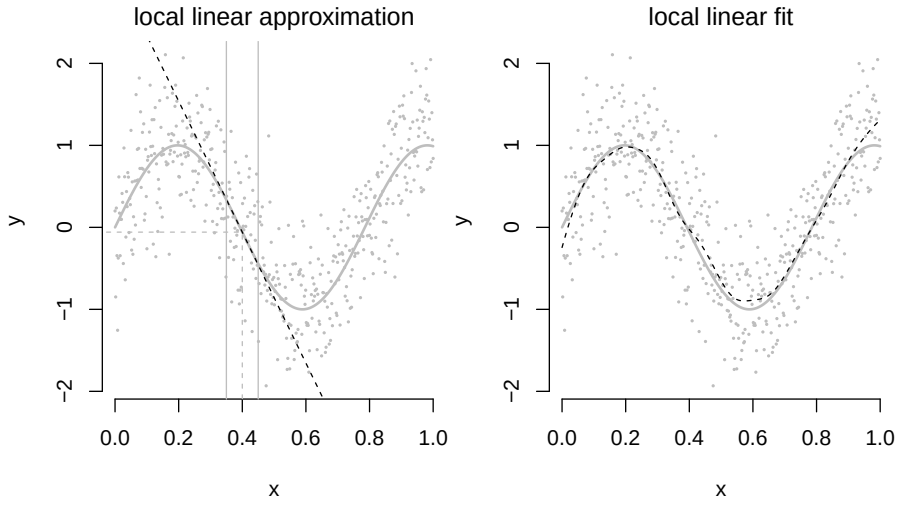


FIGURE 16.2: Local linear regression

Technically, $K(\cdot)$ can be any density function, and two canonical choices are the standard Normal density and the Epanechikov kernal $K(t) = 0.75(1 - t^2)1(|t| \leq 1)$. The choice of the kernel does not matter that much. It is much more crucial to choose the right bandwidth. With large bandwidth, we have poor linear approximation, leading to bias; with small banwidth, we have few data points, leading to large variance. In practice, we face a bias-variance tradeoff. In practice, we can either use cross-validation or other criterion to selection h.

In general, we can even fit a polynomial function locally, which is called local polynomial regression (Fan and Gijbels, 1996). In the `R` package `Kernsmooth`, the function `locpoly` fits local polynomial regression, and the function `dpill` selects h based Ruppert et al. (1995). The default specification of `locpoly` is the local linear regression.

```
> library("KernSmooth")
> n = 500
> x = seq(0, 1, length.out = n)
> fx= sin(8*x)
> y = fx + rnorm(n, 0, 0.5)
> 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)
```

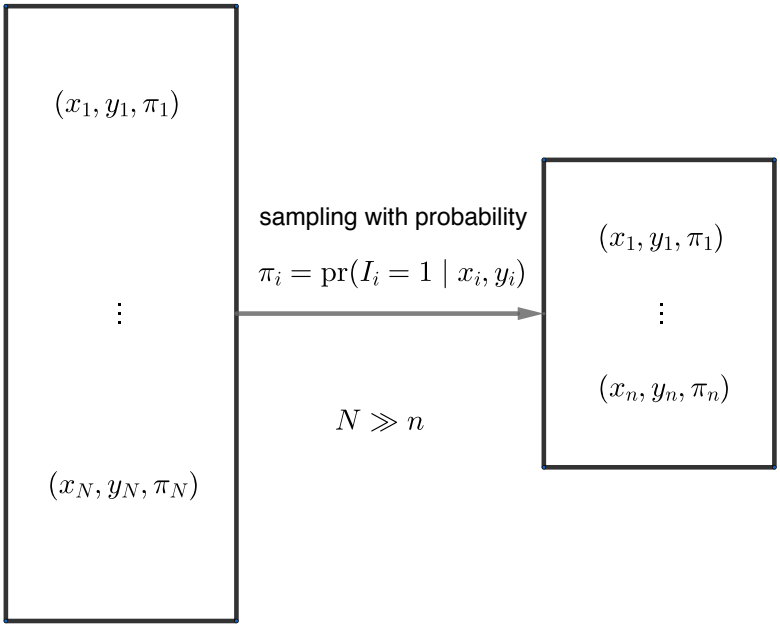


FIGURE 16.3: Survey sampling

16.2.4 Regression with survey data

Most discussions in this book are based on i.i.d. samples, or, at least, the sample represents the population of interest. Sometimes, survey samplers over sample some units and down sample some other units from a population of interest.

If we have the large population with size N, then the ideal OLS estimator is

$$\hat{\beta}_{\text{ideal}} = \left(\sum_{i=1}^{N} x_i x_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{N} x_i y_i.$$

However, we do not have all the data points in the large population, but sample each data point independently with probability

$$\pi_i = \mathrm{pr}(I_i = 1 \mid x_i, y_i),$$

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 estimator

$$\hat{\beta}_{1/\pi} = \left(\sum_{i=1}^{N} \frac{I_i}{\pi_i} x_i x_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{N} \frac{I_i}{\pi_i} x_i y_i = \left(\sum_{i=1}^{n} \pi_i^{-1} x_i x_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} \pi_i^{-1} x_i y_i,$$

with weights inversely proportional to the sampling probability. This inverse probability weighting estimator is reasonable because

$$E\left(\sum_{i=1}^{N}\frac{I_i}{\pi_i}x_i x_i^{\mathrm{T}} \mid X_N, Y_N\right) = \sum_{i=1}^{N} x_i x_i^{\mathrm{T}},$$
$$E\left(\sum_{i=1}^{N}\frac{I_i}{\pi_i}x_i y_i \mid X_N, Y_N\right) = \sum_{i=1}^{N} x_i y_i.$$

The inverse probability weighting estimators are called the Horvitz–Thompson estimators (Horvitz and Thompson, 1952), which are cornerstones of survey sampling.

```
> 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)
> compare = cbind(summary(ols.fit)$coef[,1:3],
+                 summary(wls.fit)$coef[,1:3])
> round(compare, 4)
            Estimate Std. Error t value Estimate Std. Error t value
(Intercept)   5.1667     0.1282    40.3   5.0740     0.1268    40.0
age          -0.0148     0.0067    -2.2  -0.0084     0.0067    -1.3
educ          0.1296     0.0066    19.7   0.1228     0.0065    18.8
exper2        0.0003     0.0001     2.2   0.0002     0.0001     1.3
black        -0.2467     0.0085   -29.2  -0.2574     0.0080   -32.0
```

16.3 Statistical inference with WLS

I will discuss statistical inference with the WLS estimator $\hat{\beta}_w$. Analogous to OLS, we can derive finite-sample exact inference based on a Gaussian linear model:

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2/w_i),$$

or, equivalently,

$$\tilde{y}_i = \tilde{x}_i^{\mathrm{T}}\beta + \tilde{\varepsilon}_i, \quad \tilde{\varepsilon}_i \sim \mathrm{N}(0, \sigma^2),$$

where $\tilde{y}_i = w_i^{1/2} y_i$ and $\tilde{x}_i = w_i^{1/2} x_i$. The `lm` function with `weights` reports the standard error, t-statistic, and p-value based on this model. This assumes that the weights fully capture the heteroskedasticity, which is unrealistic in many problems.

In addition, we can derive asymptotic inference based on the following heteroskedastic model

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i$$

where the ε_i's are independent with mean zero and variances σ_i^2 $(i = 1, \ldots, n)$. 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

$$\begin{aligned}
\hat{\beta}_w &= \left(\sum_{i=1}^{n} w_i x_i x_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} w_i x_i y_i \\
&= \left(\sum_{i=1}^{n} w_i x_i x_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} w_i x_i (x_i^{\mathrm{T}} \beta + \varepsilon_i) \\
&= \beta + \left(n^{-1} \sum_{i=1}^{n} w_i x_i x_i^{\mathrm{T}} \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} w_i x_i \varepsilon_i \right),
\end{aligned}$$

we can apply the law of large numbers to show that $\hat{\beta}_w$ is consistent for β and apply the central limit theorem to show that

$$\hat{\beta}_w \overset{\mathrm{a}}{\sim} \mathrm{N}\left(\beta, V_w\right),$$

where

$$V_w = n^{-1} \left(n^{-1} \sum_{i=1}^{n} w_i x_i x_i^{\mathrm{T}} \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} w_i^2 \sigma_i^2 x_i x_i^{\mathrm{T}} \right) \left(n^{-1} \sum_{i=1}^{n} w_i x_i x_i^{\mathrm{T}} \right)^{-1}.$$

We can used the generalized EHW robust covariance estimator to estimate the above asymptotic covariance:

$$\hat{V}_w = n^{-1} \left(n^{-1} \sum_{i=1}^{n} w_i x_i x_i^{\mathrm{T}} \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} w_i^2 \hat{\varepsilon}_{w,i}^2 x_i x_i^{\mathrm{T}} \right) \left(n^{-1} \sum_{i=1}^{n} w_i x_i x_i^{\mathrm{T}} \right)^{-1},$$

where $\hat{\varepsilon}_{w,i} = y_i - x_i^{\mathrm{T}} \hat{\beta}_w$ is the residual from the WLS. Note that in the sandwich covariance, w_i appears in the "bread" but w_i^2 appears in the "meat." 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 with weighted least squares. I leave the calculations of the EHW covariance estimators in previous examples as a homework problem.

16.4 Homework problem

16.1 Covariance of the generalized least squares estimator

Show that $X^{\mathrm{T}} \Sigma X$ in Section 16.1 is invertible if the columns of X are linearly independent and Σ is positive definite.

16.2 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 16.3 such that $X_k \in \mathbb{R}^{n_k \times p}$ and $Y_k \in \mathbb{R}^{n_k}$. Show that the generalized least squares estimator is

$$\hat{\beta}_\Sigma = \left(\sum_{k=1}^{K} X_k^{\mathrm{T}} \Sigma_k^{-1} X_k \right)^{-1} \left(\sum_{k=1}^{K} X_k^{\mathrm{T}} \Sigma_k^{-1} Y_k \right).$$

16.3 Difference-in-means with weights

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 $(i = 1, \ldots, n)$ equals $\bar{y}_{w,1} - \bar{y}_{w,0}$, where

$$\bar{y}_{w,1} = \frac{\sum_{i=1}^{n} w_i x_i y_i}{\sum_{i=1}^{n} w_i x_i}, \quad \bar{y}_{w,0} = \frac{\sum_{i=1}^{n} w_i (1 - x_i) y_i}{\sum_{i=1}^{n} w_i (1 - x_i)}$$

are the weighted averages of the outcome under treatment and control, respectively.

16.4 Asymptotic Normality of WLS and robust covariance estimator

Show that $\hat{\beta}_w$ is consistent and asymptotically Normal, and show that $n\hat{V}_w$ is consistent for the asymptotic covariance of $\sqrt{n}(\hat{\beta}_w - \beta)$. Specify the regularity conditions.

16.5 An infeasible generalized least squares estimator

Can we skip step 2 in Section 16.2.1 and directly apply the following WLS estimator:

$$\hat{\beta}_{\mathrm{IGLS}} = \left(\sum_{i=1}^{n} \hat{\varepsilon}_i^{-2} x_i x_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^{-2} x_i y_i$$

with $\hat{\varepsilon}_i = y_i - x_i^{\mathrm{T}} \hat{\beta}$. If so, give a theoretical justification; if not, give a counterexample (either theoretical or numerical).

16.6 FWL Theorem in WLS

Consider the WLS with weights w_i's. Show that $\hat{\beta}_{w,2}$ in the long WLS fit

$$Y = X_1 \hat{\beta}_{w,1} + X_2 \hat{\beta}_{w,2} + \hat{\varepsilon}_w$$

equals the coefficient of $\tilde{X}_{w,2}$ in the WLS fit of $\tilde{Y}_w$ 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, and $\tilde{Y}_w$ is the residual vector from the WLS of Y on X_1.

16.7 The sample version of Cochran's formula in WLS

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. Similar to Problem 6.5, show that

$$\tilde{\beta}_{w,2} = \hat{\beta}_{w,2} + \hat{\delta}_w\hat{\beta}_{w,1}.$$

16.8 EHW robust covariance estimator in WLS

We have shown in Section 16.1 that the coefficients from WLS are identical to those from OLS with transformed variables. Further show that the HC0 version of EHW covariance estimators are also identical.

16.9 Ridge with weights

Define the ridge regression with weights w_i's, and derive the formula for the ridge coefficient.

16.10 Coordinate descent algorithm in lasso with weights

Define the lasso with weights w_i's, and give the coordinate descent algorithm for it.

16.11 EHW standard errors in WLS

Report the EHW standard errors in the examples in Sections 16.2.1, 16.2.2, and 16.2.4.

Part VI

Generalized linear models

17

Logistic Regression for Binary Outcomes

Many applications have binary outcomes. This chapter discusses statistical models of binary outcomes, focusing on the logistic regression.

17.1 Regression with binary outcomes

17.1.1 Linear probability model

For simplicity, we can still use the linear model. For a binary outcome, it is also called the linear probability model:

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i, \qquad E(\varepsilon_i \mid x_i) = 0$$

because the conditional probability of y_i given x_i is a linear function of x_i:

$$\mathrm{pr}(y_i = 1 \mid x_i) = E(y_i \mid x_i) = x_i^{\mathrm{T}}\beta.$$

An advantage of this linear model is that the interpretation of the coefficient remains the same as linear models for general outcomes:

$$\frac{\partial \mathrm{pr}(y_i = 1 \mid x_i)}{\partial x_{ij}} = \beta_j,$$

that is, β_j measure the partial impact of x_{ij} on the probability of y_i. We may not believe that a linear model is the correct model for a binary outcome because the probability $\mathrm{pr}(y_i = 1 \mid x_i)$ on the lefthand side is bounded between zero and one, but the linear combination $x_i^{\mathrm{T}}\beta$ on the righthand side can be unbounded for general covariates and coefficient. Nevertheless, the OLS decomposition $y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i$ works for any $y_i \in \mathbb{R}$ so it is applicable for binary outcomes. 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.

17.1.2 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:

$$\text{pr}(y_i = 1 \mid x_i) = g(x_i^{\text{T}}\beta),$$

where $g(\cdot) : \mathbb{R} \to [0, 1]$ 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
logistic or 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	$g(z) = 1 - \exp(-e^z)$

In the above, the $g(z)$ for the logit model is the distribution function of the standard logistic distribution with density

$$g'(z) = \frac{e^z}{(1+e^z)^2} = g(z)\left\{1 - g(z)\right\}; \tag{17.1}$$

the $g(z)$ for the probit model 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 standard log-Weilbull distribution with density

$$g'(z) = \exp(z - e^z).$$

Figure 17.1 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 straightforward. We can also write the logit model as

$$\text{pr}(y_i = 1 \mid x_i) \equiv \pi(x_i, \beta) = \frac{e^{x_i^{\text{T}}\beta}}{1 + e^{x_i^{\text{T}}\beta}}, \tag{17.2}$$

or, equivalently,

$$\text{logit}\left\{\text{pr}(y_i = 1 \mid x_i)\right\} \equiv \log\frac{\text{pr}(y_i = 1 \mid x_i)}{1 - \text{pr}(y_i = 1 \mid x_i)} = x_i^{\text{T}}\beta,$$

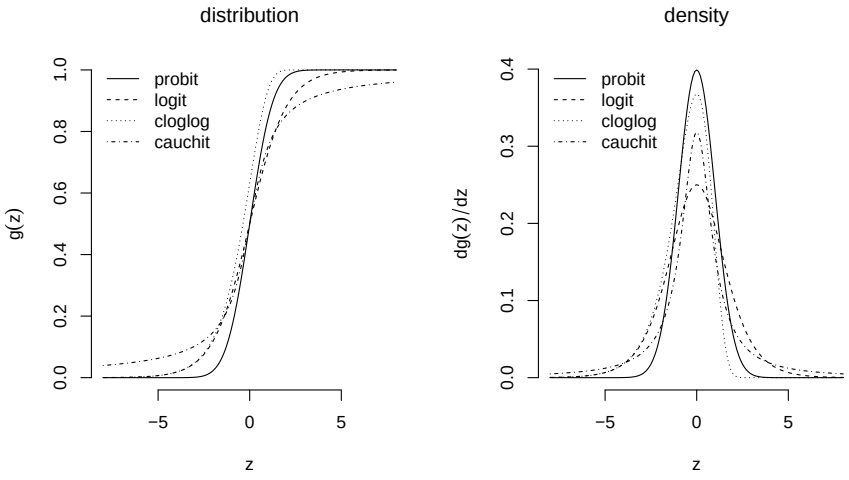


FIGURE 17.1: Distributions and densities corresponding to the link functions

where the logit function represents the log of the odds of y_i given x_i. Because y_i is a binary random variable, its probability completely determines its distribution. So we can also write the logit model as

$$y_i \mid x_i \sim \text{Bernoulli}\left(\frac{e^{x_i^{\mathrm{T}}\beta}}{1+e^{x_i^{\mathrm{T}}\beta}}\right).$$

We can interpret each coefficient β_j as the conditional log odds ratio:

$$\begin{aligned}\beta_j &= \text{logit}\left\{\text{pr}(y_i = 1 \mid \ldots, x_{ij}+1, \ldots)\right\} - \text{logit}\left\{\text{pr}(y_i = 1 \mid \ldots, x_{ij}, \ldots)\right\} \\ &= \log \frac{\text{pr}(y_i = 1 \mid \ldots, x_{ij}+1)}{1-\text{pr}(y_i = 1 \mid \ldots, x_{ij}+1)} - \log \frac{\text{pr}(y_i = 1 \mid \ldots, x_{ij}, \ldots)}{1-\text{pr}(y_i = 1 \mid \ldots, x_{ij}, \ldots)} \\ &= \log \left\{ \frac{\text{pr}(y_i = 1 \mid \ldots, x_{ij}+1, \ldots)}{1-\text{pr}(y_i = 1 \mid \ldots, x_{ij}+1, \ldots)} \Big/ \frac{\text{pr}(y_i = 1 \mid \ldots, x_{ij}, \ldots)}{1-\text{pr}(y_i = 1 \mid \ldots, x_{ij}, \ldots)} \right\},\end{aligned}$$

that is, the change of the log odds of y_i if we increase x_j by a unit holding other covariates unchanged.

17.2 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 is

$$L(\beta) = \prod_{i=1}^{n} f(y_i \mid x_i) = \prod_{i=1}^{n} \{\pi(x_i, \beta)\}^{y_i} \{1 - \pi(x_i, \beta)\}^{1-y_i}.$$

Under the logit form (17.2), the likelihood function simplifies to

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \left\{ \frac{\pi(x_i, \beta)}{1 - \pi(x_i, \beta)} \right\}^{y_i} \{1 - \pi(x_i, \beta)\} \\ &= \prod_{i=1}^{n} \left(e^{x_i^{\mathrm{T}}\beta} \right)^{y_i} \frac{1}{1 + e^{x_i^{\mathrm{T}}\beta}} \\ &= \prod_{i=1}^{n} \frac{e^{y_i x_i^{\mathrm{T}}\beta}}{1 + e^{x_i^{\mathrm{T}}\beta}}. \end{aligned}$$

The log-likelihood function is

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i^{\mathrm{T}}\beta - \log(1 + e^{x_i^{\mathrm{T}}\beta}) \right\},$$

the score function is

$$\begin{aligned} \frac{\partial \log L(\beta)}{\partial \beta} &= \sum_{i=1}^{n} \left(x_i y_i - \frac{x_i e^{x_i^{\mathrm{T}}\beta}}{1 + e^{x_i^{\mathrm{T}}\beta}} \right) \\ &= \sum_{i=1}^{n} x_i \left(y_i - \frac{e^{x_i^{\mathrm{T}}\beta}}{1 + e^{x_i^{\mathrm{T}}\beta}} \right) \\ &= \sum_{i=1}^{n} x_i \left(y_i - g(x_i^{\mathrm{T}}\beta) \right) \\ &= \sum_{i=1}^{n} x_i \left\{ y_i - \pi(x_i, \beta) \right\}, \end{aligned}$$

and the Hessian matrix

$$\begin{aligned} \frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} &= \left(\frac{\partial^2 \log L(\beta)}{\partial \beta_j \partial \beta_{j'}} \right)_{1 \le j, j' \le p} \\ &= -\sum_{i=1}^{n} x_i \frac{\partial g(x_i^{\mathrm{T}}\beta)}{\partial \beta^{\mathrm{T}}} \\ &\overset{(17.1)}{=} -\sum_{i=1}^{n} x_i x_i^{\mathrm{T}} g(x_i^{\mathrm{T}}\beta) \left\{ 1 - g(x_i^{\mathrm{T}}\beta) \right\} \\ &= -\sum_{i=1}^{n} \pi(x_i, \beta) \left\{ 1 - \pi(x_i, \beta) \right\} x_i x_i^{\mathrm{T}}. \end{aligned}$$

We can show that the Hessian matrix is negative semi-definite:

$$\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} \preceq 0, \tag{17.3}$$

and 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:

$$\sum_{i=1}^{n} x_i \left\{ y_i - \pi(x_i, \hat{\beta}) \right\} = \sum_{i=1}^{n} x_i \left(y_i - \frac{e^{x_i^{\mathrm{T}} \hat{\beta}}}{1 + e^{x_i^{\mathrm{T}} \hat{\beta}}} \right) = 0.$$

If we view $\pi(x_i, \hat{\beta})$ as the fitted probability value 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

$$\sum_{i=1}^{n} \left\{ y_i - \pi(x_i, \hat{\beta}) \right\} = 0 \Rightarrow n^{-1} \sum_{i=1}^{n} y_i = n^{-1} \sum_{i=1}^{n} \pi(x_i, \hat{\beta}),$$

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 formulas for the MLE. We usually use Newton's method to solve for it 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:

$$0 = \frac{\partial \log L(\hat{\beta})}{\partial \beta} \cong \frac{\partial \log L(\beta^{\mathrm{old}})}{\partial \beta} + \frac{\partial^2 \log L(\beta^{\mathrm{old}})}{\partial \beta \partial \beta^{\mathrm{T}}} (\beta - \beta^{\mathrm{old}}),$$

and then update

$$\beta^{\mathrm{new}} = \beta^{\mathrm{old}} - \left\{ \frac{\partial^2 \log L(\beta^{\mathrm{old}})}{\partial \beta \partial \beta^{\mathrm{T}}} \right\}^{-1} \frac{\partial \log L(\beta^{\mathrm{old}})}{\partial \beta}.$$

Using matrix form, we can gain more insight from Newton's method. Recall that

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix},$$

and define

$$\Pi^{\mathrm{old}} = \begin{pmatrix} \pi(x_1, \beta^{\mathrm{old}}) \\ \vdots \\ \pi(x_n, \beta^{\mathrm{old}}) \end{pmatrix}, \quad W^{\mathrm{old}} = \mathrm{diag}\left[\pi(x_i, \beta^{\mathrm{old}}) \left\{ 1 - \pi(x_i, \beta^{\mathrm{old}}) \right\} \right]_{i=1}^{n}.$$

Then

$$\frac{\partial \log L(\beta^{\text{old}})}{\partial \beta} = X^{\text{T}}(Y - \Pi^{\text{old}}),$$
$$\frac{\partial^2 \log L(\beta^{\text{old}})}{\partial \beta \partial \beta^{\text{T}}} = -X^{\text{T}} W^{\text{old}} X,$$

and Newton's method simplifies to

$$\begin{aligned}\beta^{\text{new}} &= \beta^{\text{old}} + (X^{\text{T}} W^{\text{old}} X)^{-1} X^{\text{T}}(Y - \Pi^{\text{old}}) \\ &= (X^{\text{T}} W^{\text{old}} X)^{-1} \left\{ X^{\text{T}} W^{\text{old}} X \beta^{\text{old}} + X^{\text{T}}(Y - \Pi^{\text{old}}) \right\} \\ &= (X^{\text{T}} W^{\text{old}} X)^{-1} X^{\text{T}} W^{\text{old}} Z^{\text{old}},\end{aligned}$$

where

$$Z^{\text{old}} = X\beta^{\text{old}} + (W^{\text{old}})^{-1}(Y - \pi^{\text{old}}).$$

So we can obtain β^{new} based on a weighted least squares fit of Z^{old} on X with weights W^{old}, which are the conditional variance of y_i given x_i at β^{old}. The `glm` function in `R` uses the Fisher Scoring algorithm, which is identical to Newton's method for the logit model. Sometimes, it is also called the iteratively reweighted least squares algorithm.

17.3 Statistics with the logit model

17.3.1 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

$$\hat{\beta} \overset{\text{a}}{\sim} \text{N}\left\{ \beta, \left(-\frac{\partial^2 \log L(\hat{\beta})}{\partial \beta \partial \beta^{\text{T}}} \right)^{-1} \right\} = \text{N}(\beta, X^{\text{T}} \hat{W} X),$$

where

$$\hat{W} = \text{diag}\left[\pi(x_i, \hat{\beta}) \left\{ 1 - \pi(x_i, \hat{\beta}) \right\} \right]_{i=1}^{n}.$$

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.

I use the data from Hirano et al. (2000) to illustrate logistic regression, where the main interest is the effect of an encourage of 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)
> assign.logit = glm(outcome ~ . - receive,
+                    family  = binomial(link = logit),
+                    data    = flu)
> summary(assign.logit)

Call:
glm(formula = outcome ~ . - receive, 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
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(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
```

17.3.2 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 $\hat{\pi}_i = g(x_i^{\mathrm{T}}\hat{\beta})$ for a unit with covariate value x_i, and we can easily dichotomize the fitted probability to predict the outcome itself by $\hat{y}_i = 1(\hat{\pi}_i \geq c)$, for example, with $c = 0.5$.

We can even quantify the uncertainty in the fitted probability based on a linear approximation. Based on

$$\hat{\pi}_i = g(x_i^{\mathrm{T}}\hat{\beta}) \cong g(x_i^{\mathrm{T}}\beta) + g'(x_i^{\mathrm{T}}\beta)x_i^{\mathrm{T}}(\hat{\beta}-\beta) = g(x_i^{\mathrm{T}}\beta) + g(x_i^{\mathrm{T}}\beta)\{1-g(x_i^{\mathrm{T}}\beta)\}x_i^{\mathrm{T}}(\hat{\beta}-\beta),$$

we can approximate the asymptotic variance of $\hat{\pi}_i$ by

$$[g(x_i^{\mathrm{T}}\beta)\{1 - g(x_i^{\mathrm{T}}\beta)\}]^2 x_i^{\mathrm{T}} X^{\mathrm{T}} \hat{W} X x_i.$$

We can use `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)
> 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
```

17.4 More on interpretations of the coefficients

17.4.1 Average partial effects

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 partial effect. For a continuous covariate x_{ij}, the average partial effect is defined as

$$\text{APE}_j = n^{-1} \sum_{i=1}^{n} \frac{\partial \text{pr}(y_i = 1 \mid x_i)}{\partial x_{ij}} = n^{-1} \sum_{i=1}^{n} g'(x_i^{\text{T}}\beta)\beta_j,$$

which reduces to the following form for logit model

$$\text{APE}_j = n^{-1} \sum_{i=1}^{n} \frac{\partial \text{pr}(y_i = 1 \mid x_i)}{\partial x_{ij}} = \beta_j \times n^{-1} \sum_{i=1}^{n} \pi(x_i, \beta) \left\{1 - \pi(x_i, \beta)\right\}.$$

For a binary covariate x_{ij}, the average partial effect is defined as

$$\text{APE}_j = n^{-1} \sum_{i=1}^{n} \{\text{pr}(y_i = 1 \mid \ldots, x_{ij} = 1, \ldots) - \text{pr}(y_i = 1 \mid \ldots, x_{ij} = 0, \ldots)\}$$

17.4.2 Difficulty of interpreting interaction

The interaction term is much more complicated as pointed out by Ai and Norton (2003). Consider the following model

$$\text{pr}(y_i = 1 \mid x_{i1}, x_{i2}) = g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}).$$

Define $z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}$. We have two ways to define the interaction effect: first,

$$n^{-1} \sum_{i=1}^{n} \frac{\partial \text{pr}(y_i = 1 \mid x_{i1}, x_{i2})}{\partial (x_{i1} x_{i2})} = n^{-1} \sum_{i=1}^{n} g'(z_i) \beta_{12}.$$

second,

$$\begin{aligned}
& n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \text{pr}(y_i = 1 \mid x_{i1}, x_{i2})}{\partial x_{i1} \partial x_{i2}} \\
&= n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial x_{i1}} \left\{ \frac{\partial \text{pr}(y_i = 1 \mid x_{i1}, x_{i2})}{\partial x_{i1}} \right\} \\
&= n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial x_{i1}} \left\{ g'(z_i)(\beta_1 + \beta_{12} x_{i2}) \right\} \\
&= n^{-1} \sum_{i=1}^{n} \left\{ g''(z_i)(\beta_2 + \beta_{12} x_{i1})(\beta_1 + \beta_{12} x_{i2}) + g'(z_i) \beta_{12} \right\};
\end{aligned}$$

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.

17.5 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.

```
> betacoef = c(lpmfit$coef[2],
+              probitfit$coef[2],
+              logitfit$coef[2],
+              cloglogfit$coef[2],
+              cauchitfit$coef[2])
> names(betacoef) = c("lpm", "probit", "logit", "cloglog", "cauchit")
> round(betacoef, 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    34    6
  1     7   53
> table(y, cauchitfit$fitted.values>0.5)

y   FALSE TRUE
  0    34    6
  1     7   53
```

Figure 17.2 shows the fitted probabilities versus the true probabilities $\text{pr}(y_i = 1 \mid x_i)$. 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 the coefficients from the logit model approximately equals those from the probit model multiplied by 1.7, a constant that

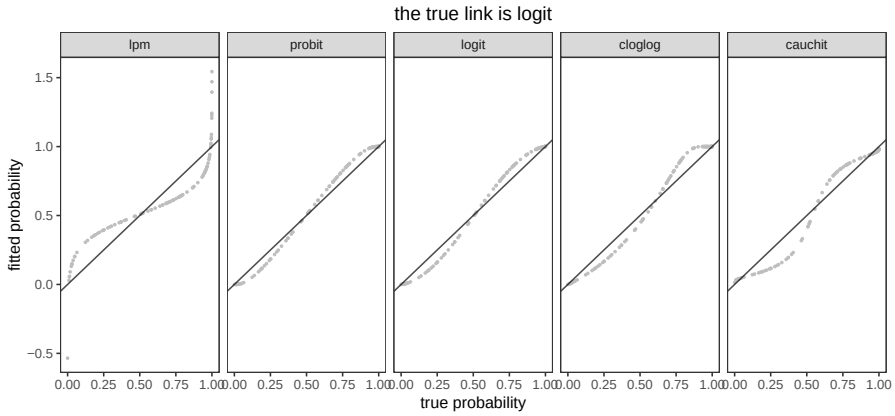


FIGURE 17.2: Comparing the fitted probabilities from different link functions

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
```

The minimum value is approximately 0.009. Therefore, the logit and probit link functions are extremely close after the scaling factor 1.7. However, $\min_b \max_y |g_{\text{logit}}(by) - g_*(y)|$ is much larger for the link function of cauchit and cloglog.

17.6 Extensions of the logistic regression

17.6.1 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:

$$\arg\min_b -\frac{1}{n}\sum_{i=1}^{n}\ell_i(b)+\lambda\sum_{j=1}^{p}|b_j|,$$

where $\ell_i(b)=y_i(b_0+b_1x_{i1}+\cdots+b_px_{ip})-\log(1+e^{b_0+b_1x_{i1}+\cdots+b_px_{ip}})$ is the log likelihood function. The `R` package `glmpath` 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.

17.6.2 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 $y_i \mid x_i$ but also for the case-control study with data from the conditional distribution $x_i \mid y_i$. 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) IID with

$$\text{pr}(s_i=1\mid x_i,y_i)=\text{pr}(s_i=1\mid y_i)=\begin{cases}p_1, & \text{if } y_i=1,\\ p_0, & \text{if } y_i=0.\end{cases}$$

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

$$\begin{aligned}
&\text{pr}(y_i=1\mid x_i,s_i=1)\\
&=\frac{\text{pr}(y_i=1\mid x_i)\text{pr}(s_i=1\mid x_i,y_i=1)}{\text{pr}(y_i=1\mid x_i)\text{pr}(s_i=1\mid x_i,y_i=1)+\text{pr}(y_i=0\mid x_i)\text{pr}(s_i=1\mid x_i,y_i=0)}\\
&=\frac{\frac{e^{\beta_0+x_i^{\mathrm{T}}\beta}}{1+e^{\beta_0+x_i^{\mathrm{T}}\beta}}p_1}{\frac{e^{\beta_0+x_i^{\mathrm{T}}\beta}}{1+e^{\beta_0+x_i^{\mathrm{T}}\beta}}p_1+\frac{1}{1+e^{\beta_0+x_i^{\mathrm{T}}\beta}}p_0}\\
&=\frac{e^{\beta_0+x_i^{\mathrm{T}}\beta}p_1}{e^{\beta_0+x_i^{\mathrm{T}}\beta}p_1+p_0}\\
&=\frac{e^{\beta_0+x_i^{\mathrm{T}}\beta}p_1/p_0}{e^{\beta_0+x_i^{\mathrm{T}}\beta}p_1/p_0+1}\\
&=\frac{e^{\delta+\beta_0+x_i^{\mathrm{T}}\beta}}{1+e^{\delta+\beta_0+x_i^{\mathrm{T}}\beta}},
\end{aligned}$$

where $\delta=\log(p_1/p_0)$. So 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. Kagan (2001) showed that the logistic link is the only one that enjoys this property.

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 disease. They used a case-control study from three cities (Mannitoba, Montreal, Ottawa) of Canada to investigate the potential association.

```
> 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
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(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
```

17.7 Other model formulations

17.7.1 Latent linear model

Let $y_i = 1(y_i^* \geq 0)$ where

$$y_i^* = x_i^{\mathrm{T}}\beta + \varepsilon_i$$

and $-\varepsilon_i$ has distribution function $g(\cdot)$ and is independent of x_i. From this latent linear model, we can verify that

$$\begin{aligned}\mathrm{pr}(y_i = 1 \mid x_i) &= \mathrm{pr}(y_i^* \geq 0 \mid x_i)\\ &= \mathrm{pr}(x_i^{\mathrm{T}}\beta + \varepsilon_i \geq 0 \mid x_i)\\ &= \mathrm{pr}(-\varepsilon_i \leq x_i^{\mathrm{T}}\beta \mid x_i)\\ &= g(x_i^{\mathrm{T}}\beta).\end{aligned}$$

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 coefficient in the models for binary data. It is a powerful way to generate models for mode complex data. We will see an example in the next chapter.

17.7.2 Inverse model

Assume that

$$y_i \sim \mathrm{Bernoulli}(q), \tag{17.4}$$

and

$$x_i \mid y_i = 1 \sim \mathrm{N}(\mu_1, \Sigma), \quad x_i \mid y_i = 0 \sim \mathrm{N}(\mu_0, \Sigma). \tag{17.5}$$

This is called the linear discriminant model. We can verify that $y_i \mid x_i$ follows a logit model as shown in the theorem below.

Theorem 17.1 *Under* (17.4) *and* (17.5)*, we have logit*$\{\mathrm{pr}(y_i = 1 \mid x_i)\} = \alpha + x_i^{\mathrm{T}}\beta$*, where*

$$\alpha = \log\frac{q}{1-q} - \frac{1}{2}\left(\mu_1^{\mathrm{T}}\Sigma^{-1}\mu_1 - \mu_0^{\mathrm{T}}\Sigma^{-1}\mu_0\right), \quad \beta = \Sigma^{-1}\left(\mu_1 - \mu_0\right).$$

I leave this as a homework problem. We can easily obtain the moment estimators for the unknown parameters under (17.4) and (17.5). Let $n_1 = \sum_{i=1}^n y_i$ and $n_0 = n - n_1$. The moment estimators are $\hat{q} = n_1/n$,

$$\hat{\mu}_1 = n_1^{-1}\sum y_i x_i, \quad \hat{\mu}_0 = n_0^{-1}\sum_{i=1}^n (1-y_i)x_i,$$

and

$$\hat{\Sigma} = \left[\sum_{i=1}^n y_i(x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^{\mathrm{T}} + \sum_{i=1}^n (1-y_i)(x_i - \hat{\mu}_0)(x_i - \hat{\mu}_0)^{\mathrm{T}}\right]/(n-2).$$

Based on Theorem 17.1, we can obtain estimates $\hat{\alpha}$ and $\hat{\beta}$ by replace the true parameters by their moment estimators. This gives us another way to fit the logistic model.

Efron (1975) compared these two methods. Since the linear discriminant model imposes stronger assumptions, the estimator based on Theorem 17.1 is more efficient. In contrast, the MLE of the logistic model is more robust because it does not impose the Normality assumption on x_i.

17.8 Homework problems

17.1 Negative semi-definiteness of the Hessian

Prove (17.3).

17.2 Two logistic regressions

Given data $(x_i, z_i, y_i)_{i=1}^n$ where x_i denotes the covariates, z_i denotes the binary treatment, and y_i denotes the binary outcome. We can fit two separate logistic regressions:

$$\text{logit}\{\text{pr}(y_i = 1 \mid z_i = 1, x_i)\} = \gamma_1 + x_i^{\mathrm{T}}\beta_1$$

and

$$\text{logit}\{\text{pr}(y_i = 1 \mid z_i = 0, x_i)\} = \gamma_0 + x_i^{\mathrm{T}}\beta_0$$

with the treated and control data, respectively. We can also fit a joint logistic regression using the pooled data:

$$\text{logit}\{\text{pr}(y_i = 1 \mid z_i, x_i)\} = \alpha_0 + \alpha_z z_i + x_i^{\mathrm{T}}\alpha_x + z_i x_i^{\mathrm{T}}\alpha_{zx}.$$

Let hats denote 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 $(\hat{\gamma}_1, \hat{\beta}_1, \hat{\gamma}_0, \hat{\beta}_0)$.

17.3 Logit and two Normals

Prove the result in Section 17.7.2.

17.4 Likelihood for Probit model

Write down the likelihood function for Probit model, and derive the steps for Newton's method.

17.5 Logit and linear discriminant analysis

Prove Theorem 17.1.

17.6 Logit and general exponential family

Efron (1975) pointed out an extension of Theorem 17.1. Show that under (17.4) and

$$f(x_i \mid y_i = y) = g(\theta_y, \eta)h(x_i, \eta)\exp(x_i^{\mathrm{T}}\theta_y), \quad (y = 0, 1)$$

with parameters $(\theta_1, \theta_0, \eta)$, we have $\mathrm{logit}\{\mathrm{pr}(y_i = 1 \mid x_i)\} = \alpha + x_i^{\mathrm{T}}\beta$. Find the formulas of α and β. As a sanity check, you can compare this problem with Theorem 17.1.

17.7 Empirical comparison of logistic regression and linear discriminant analysis

Compare the performance of logistic regression and linear discriminant analysis in terms of prediction accuracy. You should simulate at least three cases: (1) the model for linear discriminant analysis is correct; (2) the model for linear discriminant analysis is incorrect but the model for logistic regression is correct; (3) the model for logistic regression is incorrect.

17.8 Logit and other links

Compute the minimizer and minimum value of $\max_y |g_{\mathrm{logit}}(by) - g_*(y)|$ for $* =$ cauchit and cloglog.

17.9 Data analysis

Reanalyze the data in Section 17.6.2, stratifying the analysis based on `center`. Do the results vary significantly across centers?

18

Modeling Categorical Outcomes: Multinomial and Proportional Odds Logistic Regressions

Categorical outcomes are common in empirical research. The first type of categorical outcome is nominal. For example, the outcome denotes the preference of fruits (apply, orange, and pears) or transportation services (Uber, Lyft, or BART). The second type of categorical outcome is 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. There are two classes of models corresponding to the nominal and ordinal outcomes, respectively.

18.1 Multinomial distribution

A categorical random variable y taking values in $\{1, \ldots, K\}$ with probabilities $\text{pr}(y = k) = \pi_k$ $(k = 1, \ldots, K)$ is often called a multinomial distribution

$$y \sim \text{Multinomial}\left\{1; (\pi_1, \ldots, \pi_K)\right\}, \quad \sum_{k=1}^{K} \pi_k = 1. \tag{18.1}$$

We can easily calculate the mean and covariance matrix of y:

Proposition 18.1 *If y is the Multinomial random variable in (18.1), then $(1(y = 1), \ldots, 1(y = K - 1))$ has mean $(\pi_1, \ldots, \pi_{K-1})$ and covariance matrix*

$$\begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_{K-1} \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_1\pi_{K-1} & -\pi_2\pi_{K-1} & \cdots & \pi_{K-1}(1-\pi_{K-1}) \end{pmatrix}. \tag{18.2}$$

As a by product, we know that the matrix in (18.2) is positive semi-definite.

I leave the proof of Proposition 18.1 as a homework problem.

With independent samples of $(x_i, y_i)_{i=1}^n$, we want to model y_i based on covariates x_i:

$$y_i \mid x_i \sim \text{Multinomial}\left[1; \{\pi_1(x_i), \ldots, \pi_K(x_i)\}\right], \quad \sum_{k=1}^{K} \pi_k(x_i) = 1 \text{ for all } x_i.$$

We can write the probability mass function of $\text{pr}(y_i \mid x_i)$ as

$$\prod_{k=1}^{K} \{\pi_k(x_i)\}^{1(y_i=k)}.$$

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.

18.2 Multinomial logistic model for nominal outcomes

18.2.1 Modeling

Viewing category K as the reference level, we can model the ratio of the probabilities of categories k and K as

$$\log \frac{\pi_k(x_i)}{\pi_K(x_i)} = x_i^{\mathrm{T}} \beta_k \qquad (k = 1, \ldots, K-1)$$

which implies that

$$\pi_k(x_i) = \pi_K(x_i) e^{x_i^{\mathrm{T}} \beta_k} \qquad (k = 1, \ldots, K-1).$$

Due to the normalization, we have

$$\begin{aligned}
\textstyle\sum_{k=1}^{K} \pi_k(x_i) = 1 \quad &\Longrightarrow \quad \textstyle\sum_{k=1}^{K} \pi_K(x_i) e^{x_i^{\mathrm{T}} \beta_k} = 1 \\
&\Longrightarrow \quad \pi_K(x_i) \textstyle\sum_{k=1}^{K} e^{x_i^{\mathrm{T}} \beta_k} = 1 \\
&\Longrightarrow \quad \pi_K(x_i) = 1 \Big/ \textstyle\sum_{k=1}^{K} e^{x_i^{\mathrm{T}} \beta_k} \\
&\Longrightarrow \quad \pi_k(x_i) = \frac{e^{x_i^{\mathrm{T}} \beta_k}}{\sum_{k'=1}^{K} e^{x_i^{\mathrm{T}} \beta_{k'}}} \qquad (k = 1, \ldots, K-1).
\end{aligned}$$

A more compact form is

$$\pi_k(x_i) = \pi_k(x_i, \beta) = \frac{e^{x_i^{\mathrm{T}} \beta_k}}{\sum_{k'=1}^{K} e^{x_i^{\mathrm{T}} \beta_{k'}}}, \qquad (k = 1, \ldots, K) \tag{18.3}$$

where $\beta = (\beta_1, \ldots, \beta_{K-1})$ denotes the parameter with $\beta_K = 0$ for the reference category. Model 18.3 is called the multinomial logistic regression model.

Similar to the binary logistic regression model, we can interpret the coefficients as the conditional log odds ratio compared to the reference level:

$$\begin{aligned}\beta_{k,j} &= \log \frac{\pi_k(x_{i1}, \ldots, x_{ij}+1, \ldots, x_{ip})}{\pi_K(x_{i1}, \ldots, x_{ij}+1, \ldots, x_{ip})} - \log \frac{\pi_k(x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})}{\pi_K(x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})} \\ &= \log \left\{ \frac{\pi_k(x_{i1}, \ldots, x_{ij}+1, \ldots, x_{ip})}{\pi_K(x_{i1}, \ldots, x_{ij}+1, \ldots, x_{ip})} \Big/ \frac{\pi_k(x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})}{\pi_K(x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})} \right\}.\end{aligned}$$

18.2.2 MLE

The likelihood function for the multinomial logistic model is

$$\begin{aligned}L(\beta) &= \prod_{i=1}^{n} \prod_{k=1}^{K} \{\pi_k(x_i)\}^{1(y_i=k)} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{e^{x_i^{\mathrm{T}}\beta_k}}{\sum_{k'=1}^{J} e^{x_i^{\mathrm{T}}\beta_{k'}}} \right\}^{1(y_i=k)} \\ &= \prod_{i=1}^{n} \left[\left\{ \prod_{k=1}^{K} e^{1(y_i=k)x_i^{\mathrm{T}}\beta_k} \right\} \Big/ \sum_{k=1}^{J} e^{x_i^{\mathrm{T}}\beta_k} \right],\end{aligned}$$

and the log-likelihood function is

$$\log L(\beta) = \sum_{i=1}^{n} \left[\sum_{k=1}^{K} 1(y_i = k) x_i^{\mathrm{T}} \beta_k - \log \left\{ \sum_{k=1}^{K} e^{x_i^{\mathrm{T}}\beta_k} \right\} \right].$$

The score function is

$$\frac{\partial \log L(\beta)}{\partial \beta} = \begin{pmatrix} \frac{\partial \log L(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial \log L(\beta)}{\partial \beta_{K-1}} \end{pmatrix}$$

with

$$\begin{aligned}\frac{\partial \log L(\beta)}{\partial \beta_k} &= \sum_{i=1}^{n} \left\{ x_i 1(y_i = k) - \frac{x_i e^{x_i^{\mathrm{T}}\beta_k}}{\sum_{k'=1}^{K} e^{x_i^{\mathrm{T}}\beta_{k'}}} \right\} \\ &= \sum_{i=1}^{n} x_i \left\{ 1(y_i = k) - \frac{e^{x_i^{\mathrm{T}}\beta_k}}{\sum_{k'=1}^{K} e^{x_i^{\mathrm{T}}\beta_{k'}}} \right\} \\ &= \sum_{i=1}^{n} x_i \left\{ 1(y_i = k) - \pi_k(x_i, \beta) \right\}, \quad (k = 1, \ldots, K-1).\end{aligned}$$

The Hessian matrix is

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}} = \begin{pmatrix} \frac{\partial^2 \log L(\beta)}{\partial\beta_1\partial\beta_1^{\mathrm{T}}} & \frac{\partial^2 \log L(\beta)}{\partial\beta_1\partial\beta_2^{\mathrm{T}}} & \cdots & \frac{\partial^2 \log L(\beta)}{\partial\beta_1\partial\beta_{K-1}^{\mathrm{T}}} \\ \frac{\partial^2 \log L(\beta)}{\partial\beta_2\partial\beta_1^{\mathrm{T}}} & \frac{\partial^2 \log L(\beta)}{\partial\beta_2\partial\beta_2^{\mathrm{T}}} & \cdots & \frac{\partial^2 \log L(\beta)}{\partial\beta_2\partial\beta_{K-1}^{\mathrm{T}}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 \log L(\beta)}{\partial\beta_{K-1}\partial\beta_1^{\mathrm{T}}} & \frac{\partial^2 \log L(\beta)}{\partial\beta_{K-1}\partial\beta_2^{\mathrm{T}}} & \cdots & \frac{\partial^2 \log L(\beta)}{\partial\beta_{K-1}\partial\beta_{K-1}^{\mathrm{T}}} \end{pmatrix}$$

with the (k,k)th block

$$\begin{aligned} \frac{\partial^2 \log L(\beta)}{\partial\beta_k\partial\beta_k^{\mathrm{T}}} &= -\sum_{i=1}^n x_i \frac{\partial}{\partial\beta_k^{\mathrm{T}}} \left(\frac{e^{x_i^{\mathrm{T}}\beta_k}}{\sum_{k'=1}^K e^{x_i^{\mathrm{T}}\beta_{k'}}} \right) \\ &= -\sum_{i=1}^n x_i \frac{x_i^{\mathrm{T}} e^{x_i^{\mathrm{T}}\beta_k} \sum_{k'=1}^K e^{x_i^{\mathrm{T}}\beta_{k'}} - x_i^{\mathrm{T}} e^{x_i^{\mathrm{T}}\beta_k} e^{x_i^{\mathrm{T}}\beta_k}}{(\sum_{k'=1}^K e^{x_i^{\mathrm{T}}\beta_{k'}})^2} \\ &= -\sum_{i=1}^n x_i x_i^{\mathrm{T}} \frac{e^{x_i^{\mathrm{T}}\beta_k} \sum_{k'=1}^K e^{x_i^{\mathrm{T}}\beta_{k'}} - e^{x_i^{\mathrm{T}}\beta_k} e^{x_i^{\mathrm{T}}\beta_k}}{(\sum_{k'=1}^K e^{x_i^{\mathrm{T}}\beta_{k'}})^2} \\ &= -\sum_{i=1}^n \pi_k(x_i,\beta)\left\{1-\pi_k(x_i,\beta)\right\} x_i x_i^{\mathrm{T}},, \quad (k=1,\ldots,K-1) \end{aligned}$$

and the (k,k')th block

$$\begin{aligned} \frac{\partial^2 \log L(\beta)}{\partial\beta_k\partial\beta_{k'}^{\mathrm{T}}} &= -\sum_{i=1}^n x_i \frac{\partial}{\partial\beta_{k'}^{\mathrm{T}}} \left(\frac{e^{x_i^{\mathrm{T}}\beta_k}}{\sum_{k'=1}^K e^{x_i^{\mathrm{T}}\beta_{k'}}} \right) \\ &= -\sum_{i=1}^n x_i x_i^{\mathrm{T}} \frac{-e^{x_i^{\mathrm{T}}\beta_k} e^{x_i^{\mathrm{T}}\beta_{k'}}}{(\sum_{k'=1}^K e^{x_i^{\mathrm{T}}\beta_{k'}})^2} \\ &= \sum_{i=1}^n \pi_k(x_i,\beta)\pi_{k'}(x_i,\beta) x_i x_i^{\mathrm{T}} \quad (k \neq k' : k,k' = 1,\ldots,K-1). \end{aligned}$$

We can verify that the Hessian matrix is negative semi-definite based on Proposition 18.1, which is left as a homework problem.

In `R`, the function `multinom` in the `nnet` package use Newton's method to fit the multinomial logistic model. We can make inference about the parameters based on the asymptotic Normality of the MLE, and make prediction based on the fitted probabilities.

18.3 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. We want a

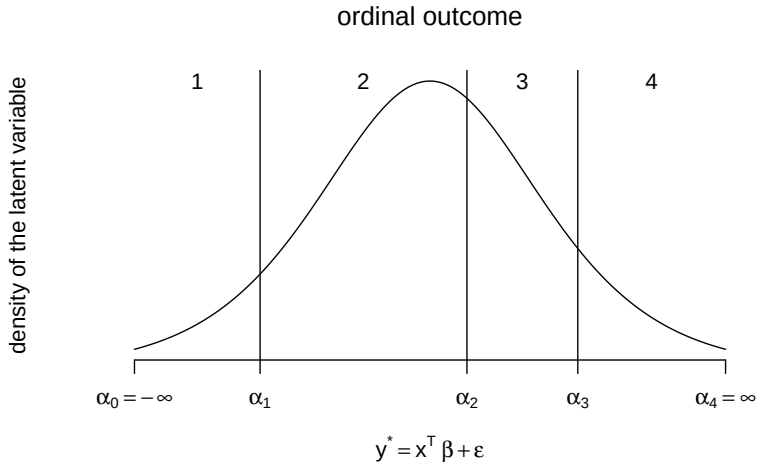


FIGURE 18.1: Latent variable representation of the ordinal outcome

model that has the property that

$$\pi_k(x_i) \leq \pi_{k+1}(x_i), \tag{18.4}$$

for all $k = 1, \ldots, K-1$ and all x_i. The multinomial logistic model does not satisfy (18.4) in general. Motivated by the latent linear representation of the binary logistic model, we imagine that the ordinal outcome arises from discretizing a continuous latent variable:

$$y_i^* = x_i^{\mathrm{T}}\beta + \varepsilon_i, \quad \mathrm{pr}(\varepsilon_i \leq z \mid x_i) = g(z) \tag{18.5}$$

and

$$y_i = k, \quad \text{if } \alpha_{k-1} \leq y_i^* \leq \alpha_k, \quad (k = 1, \ldots, K) \tag{18.6}$$

where

$$-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{K-1} < \alpha_K = \infty.$$

Figure 18.1 illustrates the data generating process with $K = 4$.

The unknown parameters are $(\beta, \alpha_1, \ldots, \alpha_{K-1})$. The distribution of the latent error term $g(\cdot)$ is known, for example, it can be logit or Normal. The former results in the proportional odds logistic model, and the latter results in the ordinal probit model. Based on the latent linear model, we can compute

$$\begin{aligned}
\mathrm{pr}(y_i \leq k \mid x_i) &= \mathrm{pr}(y_i^* \leq \alpha_k \mid x_i) \\
&= \mathrm{pr}(x_i^{\mathrm{T}}\beta + \varepsilon_i \leq \alpha_k \mid x_i) \\
&= \mathrm{pr}(\varepsilon_i \leq \alpha_k - x_i^{\mathrm{T}}\beta \mid x_i) \\
&= g(\alpha_k - x_i^{\mathrm{T}}\beta).
\end{aligned}$$

I will focus on the proportional odds logistic mode in this chapter. With this model, we have

$$\text{pr}(y_i \leq k \mid x_i) = \frac{e^{\alpha_k - x_i^{\mathrm{T}}\beta}}{1 + e^{\alpha_k - x_i^{\mathrm{T}}\beta}}$$

or

$$\text{logit pr}(y_i \leq k \mid x_i) = \log \frac{\text{pr}(y_i \leq k \mid x_i)}{\text{pr}(y_i > k \mid x_i)} = \alpha_k - x_i^{\mathrm{T}}\beta, \tag{18.7}$$

from which we can easily see that (18.4) holds. The model is has the "proportional odds" property because

$$\frac{\text{pr}(y_i \leq k \mid \ldots, x_{ij}+1, \ldots)}{\text{pr}(y_i > k \mid \ldots, x_{ij}+1, \ldots)} \Big/ \frac{\text{pr}(y_i \leq k \mid \ldots, x_{ij}, \ldots)}{\text{pr}(y_i > k \mid \ldots, x_{ij}, \ldots)} = e^{-\beta_j}$$

which is a positive constant not depending on k.

We cannot change the right-hand side of (18.7) to $\alpha_k - x_i^{\mathrm{T}}\beta_k$ because the general model cannot ensure (18.4). The sign of $x_i^{\mathrm{T}}\beta$ 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 (18.7).

The proportional odds model implies a quite complicate form of the probability for each category:

$$\text{pr}(y_i = k \mid x_i) = \frac{e^{\alpha_k - x_i^{\mathrm{T}}\beta}}{1 + e^{\alpha_k - x_i^{\mathrm{T}}\beta}} - \frac{e^{\alpha_{k-1} - x_i^{\mathrm{T}}\beta}}{1 + e^{\alpha_{k-1} - x_i^{\mathrm{T}}\beta}}, \quad (k = 1, \ldots, K).$$

So the likelihood function is

$$\begin{aligned} L(\beta, \alpha_1, \ldots, \alpha_{K-1}) &= \prod_{i=1}^{n} \prod_{k=1}^{K} \{\text{pr}(y_i = k \mid x_i)\}^{1(y_i = k)} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{e^{\alpha_k - x_i^{\mathrm{T}}\beta}}{1 + e^{\alpha_k - x_i^{\mathrm{T}}\beta}} - \frac{e^{\alpha_{k-1} - x_i^{\mathrm{T}}\beta}}{1 + e^{\alpha_{k-1} - x_i^{\mathrm{T}}\beta}} \right)^{1(y_i = k)}. \end{aligned}$$

The log likelihood function is concave (Pratt, 1981; Burridge, 1981), and it is strictly concave in most cases. The function `polr` function in the `R` package `MASS` computes the MLE of the proportional odds model using the BFGS algorithm. 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 as a homework problem. For more details of the Hessian matrix, see Agresti (2010), which is a textbook discussion on modeling ordinal data.

18.4 A case study

I use a small observational 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 or small volume hospital, where volume is defined by the number of patients with cardia cancer treated in recent years. I use the following variables: `HighVolDiagHosp` indicating whether a patient was diagnosed at a high volume hospital, `HighVolTreatHosp` indicating whether a patient was treated at a high volume hospital, `AgeAtDiagnosis` representing the age, `FromRuralArea` indicating whether the patient was from a rural area, and `YearsSurvivingAfterDiagnosis` representing the years of survival after diagnosis with three categories ("1", "2-4", "5+").

```
karolinska = read.table("karolinska.txt", header = TRUE)
karolinska = karolinska[, c("HighVolDiagHosp", "HighVolTreatHosp",
                            "AgeAtDiagnosis", "FromRuralArea",
                            "Male", "YearsSurvivingAfterDiagnosis")]
```

18.4.1 Binary logistic for the treatment

We have two choices of treatment: `HighVolDiagHosp` and `HighVolTreatHosp`.

```
> diagglm = glm(HighVolDiagHosp ~ AgeAtDiagnosis + FromRuralArea + Male,
+               data = karolinska, family = binomial(link = "logit"))
> summary(diagglm)

Call:
glm(formula = HighVolDiagHosp ~ AgeAtDiagnosis + FromRuralArea +
    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 **
AgeAtDiagnosis -0.03124    0.01481  -2.110 0.034854 *
FromRuralArea  -1.26322    0.34530  -3.658 0.000254 ***
Male           -0.97524    0.41303  -2.361 0.018216 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 219.03  on 157  degrees of freedom
Residual deviance: 195.69  on 154  degrees of freedom
AIC: 203.69

Number of Fisher Scoring iterations: 4

>
> treatglm = glm(HighVolTreatHosp ~ AgeAtDiagnosis + FromRuralArea + Male,
+               data = karolinska, family = binomial(link = "logit"))
> summary(treatglm)

Call:
```

```
glm(formula = HighVolTreatHosp ~ AgeAtDiagnosis + FromRuralArea +
    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 ***
AgeAtDiagnosis -0.06297    0.01890  -3.332 0.000862 ***
FromRuralArea  -1.28777    0.39572  -3.254 0.001137 **
Male           -0.74856    0.45285  -1.653 0.098329 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.04  on 157  degrees of freedom
Residual deviance: 167.21  on 154  degrees of freedom
AIC: 175.21

Number of Fisher Scoring iterations: 4
```

Both treatments are associated with the covariates. `HighVolTreatHosp` is more strongly associated with age. What is more, Rubin (2008) argued that `HighVolDiagHosp` is more random than `HighVolTreatHosp`, and may have weaker association with other hidden covariates. For each model below, I fit the data twice corresponding to two choices of treatment.

18.4.2 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$loneyear = (karolinska$YearsSurvivingAfterDiagnosis != "1")
> loneyearglm = glm(loneyear ~ HighVolDiagHosp + AgeAtDiagnosis + FromRuralArea + Male,
+                   data = karolinska, family = binomial(link = "logit"))
> summary(loneyearglm)

Call:
glm(formula = loneyear ~ HighVolDiagHosp + AgeAtDiagnosis + FromRuralArea +
    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
HighVolDiagHosp  0.13684    0.36586   0.374   0.7084
AgeAtDiagnosis  -0.00389    0.01411  -0.276   0.7829
FromRuralArea    0.33360    0.35798   0.932   0.3514
Male             0.86706    0.44034   1.969   0.0489 *
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

HighVolDiagHosp is not significant in the above regression.

```
> loneyearglm = glm(loneyear ~ HighVolTreatHosp + AgeAtDiagnosis +
+                     FromRuralArea + Male, data = karolinska, family = binomial(link = "logit"
> summary(loneyearglm)

Call:
glm(formula = loneyear ~ HighVolTreatHosp + AgeAtDiagnosis +
    FromRuralArea + 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 *
HighVolTreatHosp  1.417458   0.455603   3.111  0.00186 **
AgeAtDiagnosis    0.008725   0.014840   0.588  0.55655
FromRuralArea     0.633278   0.368525   1.718  0.08572 .
Male              1.079973   0.452191   2.388  0.01693 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

HighVolTreatHosp is significant in the above regression.

18.4.3 Multinomial logistic for the outcome

I then fit multinomial logistic models for the outcome with three categories.

```
> library(nnet)
> yearmultinom = multinom(YearsSurvivingAfterDiagnosis ~ HighVolDiagHosp +
+                           AgeAtDiagnosis + FromRuralArea + 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 = YearsSurvivingAfterDiagnosis ~ HighVolDiagHosp +
    AgeAtDiagnosis + FromRuralArea + Male, data = karolinska)

Coefficients:
    (Intercept) HighVolDiagHosp AgeAtDiagnosis FromRuralArea
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) HighVolDiagHosp AgeAtDiagnosis FromRuralArea
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
```

`HighVolDiagHosp` is not significant above.

```
> yearmultinom = multinom(YearsSurvivingAfterDiagnosis ~ HighVolTreatHosp +
+                           AgeAtDiagnosis + FromRuralArea + 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 = YearsSurvivingAfterDiagnosis ~ HighVolTreatHosp +
    AgeAtDiagnosis + FromRuralArea + Male, data = karolinska)

Coefficients:
    (Intercept) HighVolTreatHosp AgeAtDiagnosis FromRuralArea
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) HighVolTreatHosp AgeAtDiagnosis FromRuralArea
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
```

`HighVolTreatHosp` is significant above. We can also use `pred` to obtain the fitted probabilities of each category, for example,

```
> predict(yearmultinom, type = "probs")[1:5, ]
          1       2-4         5+
1 0.7046514 0.1952602 0.10008835
2 0.7064547 0.1940977 0.09944761
3 0.7589625 0.2187152 0.02232230
4 0.7046514 0.1952602 0.10008835
5 0.5312053 0.3190527 0.14974200
```

18.4.4 Proportional odds logistic for the outcome

The multinomial logisitic model does not reflect the ordering information of the outcome. For instance, in the regression with `HighVolDiagHosp`, the coefficient for level "2-4" is $-0.06973187 < 0$, but the coefficient for level "5+" is $0.64036289 > 0$, which means that `HighVolDiagHosp` decreases the probability of "2-4" but increase the probability of "5". However, this is illogical since a patient must first live longer than two years and then live longer than

five years. Nevertheless, it is not a severe problem in this cases study because those coefficients are not significant.

```
 library(MASS)
> yearpo = polr(YearsSurvivingAfterDiagnosis ~ HighVolDiagHosp + AgeAtDiagnosis +
+               FromRuralArea + Male, Hess = TRUE, data = karolinska)
> summary(yearpo)
Call:
polr(formula = YearsSurvivingAfterDiagnosis ~ HighVolDiagHosp +
    AgeAtDiagnosis + FromRuralArea + Male, data = karolinska,
    Hess = TRUE)

Coefficients:
                    Value Std. Error t value
HighVolDiagHosp  0.216755    0.35892  0.6039
AgeAtDiagnosis  -0.002881    0.01378 -0.2091
FromRuralArea    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
```

HighVolDiagHosp is not significant above.

```
> yearpo = polr(YearsSurvivingAfterDiagnosis ~ HighVolTreatHosp + AgeAtDiagnosis +
+                FromRuralArea + Male, Hess = TRUE, data = karolinska)
> summary(yearpo)
Call:
polr(formula = YearsSurvivingAfterDiagnosis ~ HighVolTreatHosp +
    AgeAtDiagnosis + FromRuralArea + Male, data = karolinska,
    Hess = TRUE)

Coefficients:
                    Value Std. Error t value
HighVolTreatHosp 1.399538    0.44518  3.1438
AgeAtDiagnosis   0.008032    0.01438  0.5584
FromRuralArea    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
```

HighVolTreatHosp is significant above. Again, we can use `pred` to obtain the fitted probabilities similar to the multinomial logistic model.

18.5 Homework problems

18.1 Covariance matrix of multinomial

Prove Proposition 18.1.

18.2 Hessian matrix in the multinomial logit model

Prove that the Hessian matrix of the log-likelihood function of the multinomial logit model is negative semi-definite.

18.3 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.

18.4 Score function of the proportional odds model

Derive the explicit formulas of the score function of the proportional odds model.

18.5 Ordered probit regression

If we choose $\varepsilon_i \mid x_i \sim \mathrm{N}(0,1)$ in (18.5), 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"`.

18.6 Case-control study and multinomial logistic model

Assume that

$$\mathrm{pr}(y_i = k \mid x_i) = \frac{e^{\alpha_k + x_i^{\mathrm{T}}\beta_k}}{\sum_{k'=1}^{K} e^{\alpha_{k'} + x_i^{\mathrm{T}}\beta_{k'}}}$$

with $\alpha_K = 0$ and $\beta_K = 0$, and

$$\mathrm{pr}(s_i = 1 \mid y_i = k, x_i) = \mathrm{pr}(s_i = 1 \mid y_i = k) = p_k$$

for $k = 1, \ldots, K$. Show that

$$\mathrm{pr}(y_i = k \mid x_i, s_i = 1) = \frac{e^{\alpha_k + \log p_k + x_i^{\mathrm{T}}\beta_k}}{\sum_{k'=1}^{K} e^{\alpha_{k'} + \log p_{k'} + x_i^{\mathrm{T}}\beta_{k'}}}.$$

19

Regression Models for Count Outcomes

19.1 Some random variables for counts

A random variable for counts can take values in $\{0, 1, 2, \ldots\}$. This type of variables are common in applied statistics. For example, it can represent how many times you visited the gym every week, how many lectures you missed in the linear model course, how many traffic accidents happened in certain area during certain period, etc. This chapter focuses on statistical modeling of those outcomes given covariates. Hilbe (2014) is a textbook focusing on count outcome regressions.

I first review four canonical choices of count random variables.

19.1.1 Poisson

A random variable y is Poisson(λ) if its probability mass function is

$$\text{pr}(y = k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad (k = 0, 1, 2, \ldots).$$

A Poisson(λ) random variable has the following properties:

Proposition 19.1 *If $y \sim$Poisson(λ), then*

$$E(y) = \text{var}(y) = \lambda.$$

Proposition 19.2 *If $y_1 \perp\!\!\!\perp y_2$ with $y_1 \sim$Poisson(λ_1) and $y_2 \sim$Poisson(λ_2), then*

$$\begin{aligned} y_1 + y_2 &\sim \text{Poisson}(\lambda_1 + \lambda_2), \\ y_1 \mid y_1 + y_2 = n &\sim \text{Binomial}\left(n, \tfrac{\lambda_1}{\lambda_1+\lambda_2}\right). \end{aligned}$$

Where does the Poisson random variable come from? One way to generate Poisson is through independent Bernoulli random variables. I will review Le Cam et al. (1960)'s theorem below without giving a proof.

Theorem 19.1 *Suppose X_i's are independent Bernoulli random variables with probabilities p_i's ($i = 1, \ldots, n$). 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

$$\sum_{k=0}^{\infty} \left| \text{pr}(S_n = k) - e^{-\lambda} \frac{\lambda^k}{k!} \right| \leq 2 \sum_{i=1}^{n} (\lambda/n)^2 = \lambda^2/n \to 0.$$

So the sum of IID Bernoulli random variables is approximately Poisson if the probability has order $1/n$. Based on Le Cam's Theorem, we can use Poisson as a model for the sum of many rare events.

19.1.2 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 variance larger than the mean. Negative-Binomial is an extension of Poisson that allows for overdispersion. The definition of Negative-Binomial below is different from its standard definition, but it is more natural as an extension of Poisson. Define y as a Negative-Binomial random variable, denoted by $\text{NB}(\mu, \theta)$ with $\mu > 0$ and $\theta > 0$, if

$$\begin{cases} y \mid \lambda & \sim \text{Poisson}(\lambda), \\ \lambda & \sim \text{Gamma}(\theta, \theta/\mu). \end{cases} \tag{19.1}$$

So Negative-Binomial is Poisson with a random Gamma intensity, that is, Negative-Binomial is a scale-mixture of Poisson. If $\theta \to \infty$, then λ is a point mass at μ and the Negative-Binomial reduces to $\text{Poisson}(\mu)$. We can verify that it has the following probability mass function, which is left as a homework problem.

Proposition 19.3 *The Negative-Binomial random variable defined in (19.1) has the probability mass function*

$$\text{pr}(y = k) = \frac{\Gamma(k + \theta)}{\Gamma(k + 1)\Gamma(\theta)} \left(\frac{\theta}{\mu + \theta} \right)^{\theta} \left(\frac{\mu}{\mu + \theta} \right)^{k}, \quad (k = 0, 1, 2, \ldots).$$

We can easily derive the mean and variance of Negative Binomial.

Proposition 19.4 *The Negative-Binomial random variable defined in (19.1) has moments*

$$E(y) = \mu, \quad \text{var}(y) = \mu + \frac{\mu^2}{\theta} > E(y).$$

Proof of Proposition 19.4: Recall that a $\text{Gamma}(\alpha, \beta)$ random variable has mean α/β and variance α/β^2. We have

$$E(y) = E\left\{E(y \mid \lambda)\right\} = E(\lambda) = \frac{\theta}{\theta/\mu} = \mu,$$

and

$$\text{var}(y) = E\left\{\text{var}(y \mid \lambda)\right\} + \text{var}\left\{E(y \mid \lambda)\right\} = E(\lambda) + \text{var}(\lambda) = \frac{\theta}{\theta/\mu} + \frac{\theta}{(\theta/\mu)^2} = \mu + \frac{\mu^2}{\theta}.$$

□

So the dispersion parameter θ controls the variance of Negative Binomial. With the same mean, Negative Binomial has larger variance than Poisson. Figure 19.1 further compares the log probability mass functions of Negative Binomial and Poisson. It shows that Negative Binomial has slightly higher probability at zero but much heavier tails than Poisson.

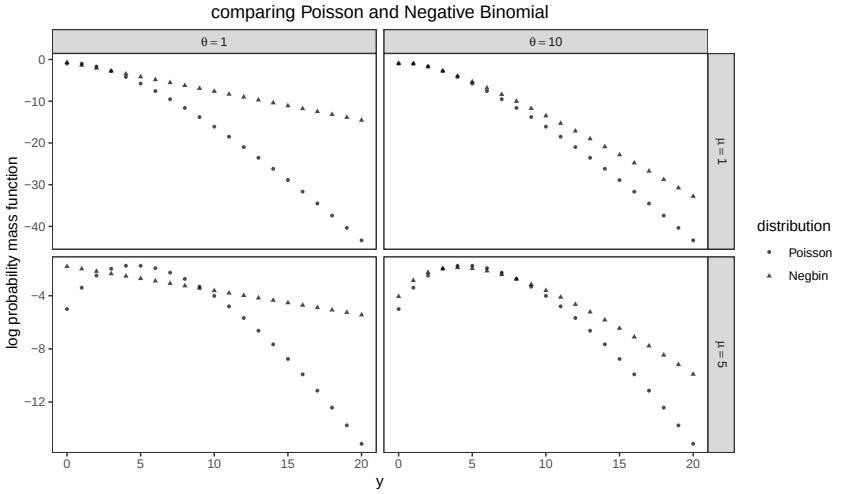


FIGURE 19.1: Comparing the log probabilities of Poisson and Negative Binomial with the same mean

19.1.3 Zero-inflated Poisson

A zero-inflated Poisson random variable y is a mixture of two components: a point mass at zero and a Poisson(λ) random variable, with probabilities p and $1-p$, respectively. So y has probability mass function

$$\text{pr}(y=k) = \begin{cases} p + (1-p)e^{-\lambda}, & \text{if } k = 0, \\ (1-p)e^{-\lambda}\frac{\lambda^k}{k!}, & \text{if } k = 1, 2, \ldots. \end{cases}$$

and moments below:

Proposition 19.5 *$E(y) = p\lambda$ and* $\text{var}(y) = (1-p)\lambda(1+p\lambda)$.

19.1.4 Zero-inflated Negative-Binomial

A zero-inflated Negative-Binomial random variable y is a mixture of two components: a point mass at zero and a NB(μ, θ) random variable, with probabil-

ities p and $1-p$, respectively. So y has probability mass function

$$\text{pr}(y=k) = \begin{cases} p + (1-p)\left(\frac{\theta}{\mu+\theta}\right)^{\theta}, & \text{if } k=0, \\ (1-p)\frac{\Gamma(k+\theta)}{\Gamma(k+1)\Gamma(\theta)}\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{k}, & \text{if } k=1,2,\ldots. \end{cases}$$

and moments below:

Proposition 19.6 *$E(y) = p\mu$ and* $\text{var}(y) = (1-p)\mu(1+\mu/\theta+p\mu)$.

19.2 Regression models for counts

To model a count outcome y_i given x_i, we can still use OLS. But 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 $y_i \mid x_i$ using the distributions reviewed in Section 19.1.

19.2.1 Poisson regression

Poisson regression assumes

$$\begin{cases} y_i \mid x_i & \sim \text{Poisson}(\lambda_i), \\ \lambda_i & = \lambda(x_i,\beta) = e^{x_i^{\mathrm{T}}\beta}. \end{cases}$$

So the mean and variance of $y_i \mid x_i$ are

$$E(y_i \mid x_i) = \text{var}(y_i \mid x_i) = e^{x_i^{\mathrm{T}}\beta}.$$

Because

$$\log E(y_i \mid x_i) = x_i^{\mathrm{T}}\beta,$$

this model is sometimes called the log-linear model, with the coefficient β_j interpreted as the conditional log mean ratio:

$$\log \frac{E(y_i \mid x_{i1},\ldots,x_{ij}+1,\ldots,x_{ip})}{E(y_i \mid x_{i1},\ldots,x_{ij},\ldots,x_{ip})} = \beta_j.$$

The likelihood function for independent Poisson random variables is

$$L(\beta) = \prod_{i=1}^{n} e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!} \propto \prod_{i=1}^{n} e^{-\lambda_i}\lambda_i^{y_i},$$

and omitting the constants, we can write the log-likelihood function as

$$\log L(\beta) = \sum_{i=1}^{n} (-\lambda_i + y_i \log \lambda_i) = \sum_{i=1}^{n} \left(-e^{x_i^{\mathrm{T}}\beta} + y_i x_i^{\mathrm{T}} \beta \right).$$

The score function is

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left(-x_i e^{x_i^{\mathrm{T}}\beta} + x_i y_i \right) = \sum_{i=1}^{n} x_i \left(y_i - e^{x_i^{\mathrm{T}}\beta} \right) = \sum_{i=1}^{n} x_i \left\{ y_i - \lambda(x_i, \beta) \right\},$$

and the Hession matrix is

$$\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} = -\sum_{i=1}^{n} x_i \frac{\partial}{\partial \beta^{\mathrm{T}}} \left(e^{x_i^{\mathrm{T}}\beta} \right) = -\sum_{i=1}^{n} e^{x_i^{\mathrm{T}}\beta} x_i x_i^{\mathrm{T}}$$

which is negative semi-definite. When the Hessian is negative definite, the MLE is unique. It must satisfy that

$$\sum_{i=1}^{n} x_i \left(y_i - e^{x_i^{\mathrm{T}}\hat{\beta}} \right) = \sum_{i=1}^{n} x_i \left\{ y_i - \lambda(x_i, \hat{\beta}) \right\} = 0.$$

We can solve this nonlinear equation using Newton's method:

$$\begin{aligned} \beta^{\mathrm{new}} &= \beta^{\mathrm{old}} - \left\{ \frac{\partial^2 \log L(\beta^{\mathrm{old}})}{\partial \beta \partial \beta^{\mathrm{T}}} \right\}^{-1} \frac{\partial \log L(\beta^{\mathrm{old}})}{\partial \beta} \\ &= \beta^{\mathrm{old}} - (X^{\mathrm{T}} W^{\mathrm{old}} X)^{-1} X^{\mathrm{T}} (Y - \Lambda^{\mathrm{old}}), \end{aligned}$$

where

$$X = \begin{pmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and

$$\Lambda^{\mathrm{old}} = \begin{pmatrix} \exp(x_1^{\mathrm{T}} \beta^{\mathrm{old}}) \\ \vdots \\ \exp(x_n^{\mathrm{T}} \beta^{\mathrm{old}}) \end{pmatrix}, \quad W^{\mathrm{old}} = \mathrm{diag} \left\{ \exp(x_i^{\mathrm{T}} \beta^{\mathrm{old}}) \right\}_{i=1}^{n}.$$

Similar to the derivation for the logit model, we can simplify Newton's method to

$$\beta^{\mathrm{new}} = (X^{\mathrm{T}} W^{\mathrm{old}} X)^{-1} X^{\mathrm{T}} W^{\mathrm{old}} Z^{\mathrm{old}},$$

where

$$Z^{\mathrm{old}} = X\beta^{\mathrm{old}} + (W^{\mathrm{old}})^{-1} (Y - \Lambda^{\mathrm{old}}).$$

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 based on

$$\hat{\beta} \overset{\text{a}}{\sim} \text{N}\left\{\beta, \left(-\frac{\partial^2 \log L(\hat{\beta})}{\partial\beta\partial\beta^{\text{T}}}\right)^{-1}\right\} = \text{N}\left\{\beta, (X^{\text{T}}\hat{W}X)^{-1}\right\},$$

where $\hat{W} = \text{diag}\left\{\exp(x_i^{\text{T}}\hat{\beta})\right\}_{i=1}^{n}$.

After obtaining the MLE, we can predict the mean $E(y_i \mid x_i)$ by $\hat{\lambda}_i = e^{x_i^{\text{T}}\hat{\beta}}$. Because Poisson regression is a fully parametrized model, we can also predict any other probability quantities involving $y_i \mid x_i$. For example, we can predict $\text{pr}(y_i = 0 \mid x_i)$ by $e^{-\hat{\lambda}_i}$, and $\text{pr}(y_i \geq 3 \mid x_i)$ by $1 - e^{-\hat{\lambda}_i}(1 + \hat{\lambda}_i + \hat{\lambda}_i^2/2)$.

19.2.2 Negative-Binomial regression

Negative-Binomial regression assumes

$$\begin{cases} y_i \mid x_i & \sim \text{NB}(\mu_i, \theta), \\ \mu_i & = e^{x_i^{\text{T}}\beta}, \end{cases}$$

so it has conditional mean $E(y_i \mid x_i) = e^{x_i^{\text{T}}\beta}$ and variance $\text{var}(y_i \mid x_i) = e^{x_i^{\text{T}}\beta}(1 + e^{x_i^{\text{T}}\beta}/\theta)$.

The log likelihood function for Negative-Binomial regression is

$$\begin{aligned} \log L(\beta, \theta) \quad = \quad & \sum_{i=1}^{n} \{\text{lgamma}(y_i + \theta) - \text{lgamma}(y_i + 1) - \text{lgamma}(\theta) \\ & + \theta \log\left(\frac{\theta}{\mu + \theta}\right) + y_i \log\left(\frac{\mu}{\mu + \theta}\right)\Big\}, \end{aligned}$$

where $\text{lgamma}(\cdot) = \log\Gamma(\cdot)$, coherent with the `R` function `lgamma`. The `glm.nb` in the `MASS` package iterates between β and θ, and uses Newton's method to update the parameters until convergence to the MLE. It reports standard errors based on the inverse of the Fisher information matrix.

19.2.3 Zero-inflated Poisson regression

Zero-inflated Poisson regression assumes that

$$y_i \mid x_i \sim \begin{cases} 0, & \text{with probability } p_i, \\ \text{Poisson}(\lambda_i), & \text{with probability } 1 - p_i, \end{cases}$$

where

$$p_i = \frac{e^{x_i^{\text{T}}\gamma}}{1 + e^{x_i^{\text{T}}\gamma}}, \quad \lambda_i = e^{x_i^{\text{T}}\beta}.$$

To avoid overparametrization, we can also restrict some coefficients to be zero. The `zeroinfl` function in the `R` package `pscl` can fit the zero-inflated Poisson regression.

19.2.4 Zero-inflated Negative-Binomial regression

Zero-inflated Negative-Binomial regression assumes that

$$y_i \mid x_i \sim \begin{cases} 0, & \text{with probability } p_i, \\ \text{NB}(\mu_i, \theta), & \text{with probability } 1 - p_i, \end{cases}$$

where

$$p_i = \frac{e^{x_i^{\mathrm{T}}\gamma}}{1 + e^{x_i^{\mathrm{T}}\gamma}}, \quad \mu_i = e^{x_i^{\mathrm{T}}\beta}.$$

To avoid overparametrization, we can also restrict some coefficients to be zero. The `zeroinfl` function in the `R` package `pscl` can fit the zero-inflated Negative-Binomial regression.

19.3 A case study

I will use the dataset from Royer et al. (2015) to illustrate the count outcome regressions. From the regression formula below, we are interested in the effect of two treatments `incentive_commit` and `incentive` on the the number of visits to the gym, controlling for two pretreatment covariates `target` and `member_gym_pre`.

```
> library("foreign")
> gym1 = read.dta("gym_treatment_exp_weekly.dta")
> f.reg = weekly_visit ~ incentive_commit + incentive + target + member_gym_pre
```

19.3.1 Linear, Poisson, and Negative-Binomial regressions

Each worker was observed over time, so we run regressions with the outcome data observed in each week. The following `R` code computes the linear regression coefficients, standard errors, and AICs.

```
> weekids             = sort(unique(gym1$incentive_week))
> lweekids            = length(weekids)
> coefincentivecommit = 1:lweekids
> coefincentive       = 1:lweekids
> seincentivecommit   = 1:lweekids
> seincentive         = 1:lweekids
> AIClm               = 1:lweekids
> for(i in 1:lweekids)
+ {
+         gymweek = gym1[which(gym1$incentive_week == weekids[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]
```

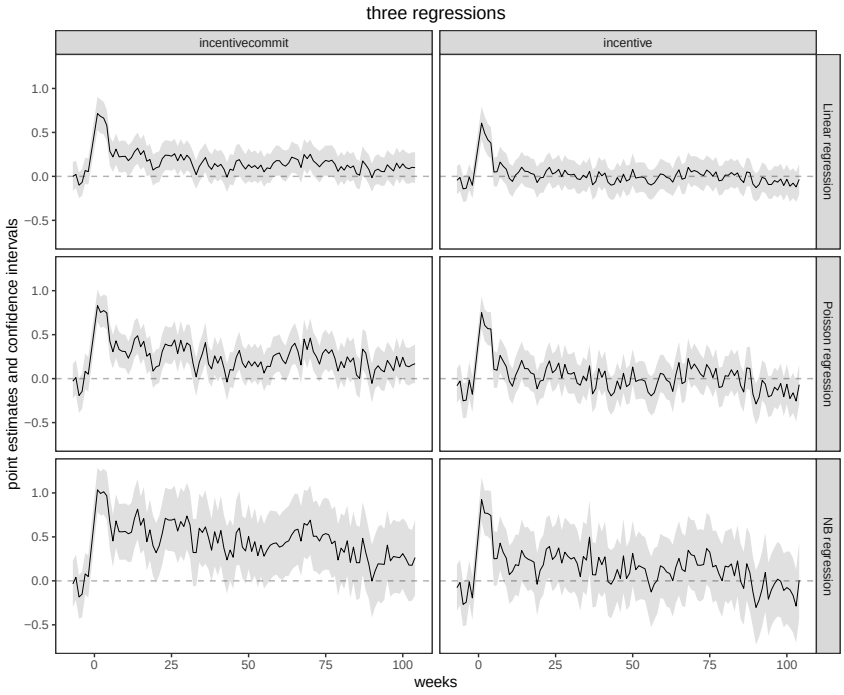


FIGURE 19.2: Linear, Poisson, and Negative-Binomial regressions

```
+          seincentive[i]          = regweekcoef[3, 2]
+
+          AIClm[i]                = AIC(regweek)
+ }
```

By changing the line with `lm` by

```
regweek = glm(f.reg, family = poisson(link = "log"), data = gymweek)
```

and

```
regweek = glm.nb(f.reg, data = gymweek)
```

we obtain the corresponding results from Poisson and Negative-Binomial regressions. Figure 19.2 compares the regression coefficients with the associated confidence intervals over time. Three regressions give very similar pattern: `incentive_commit` has both short-term and long-term effects, but `incentive` only has short-term effect.

The left panel of Figure 19.3 shows that variances are larger than the means for outcomes from all weeks, and the right panel of Figure 19.3 shows the point estimates and confidence intervals of θ from Negative-Binomial regressions. Overall, overdispersion seems an important feature of the data.

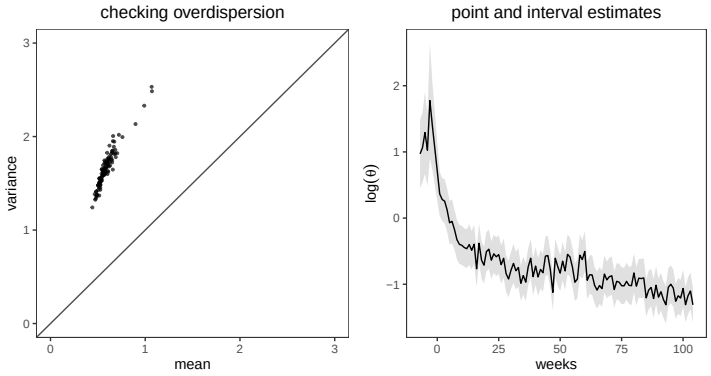


FIGURE 19.3: Overdispersion of the data

19.3.2 Zero-inflated regressions

Figure 19.4 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.

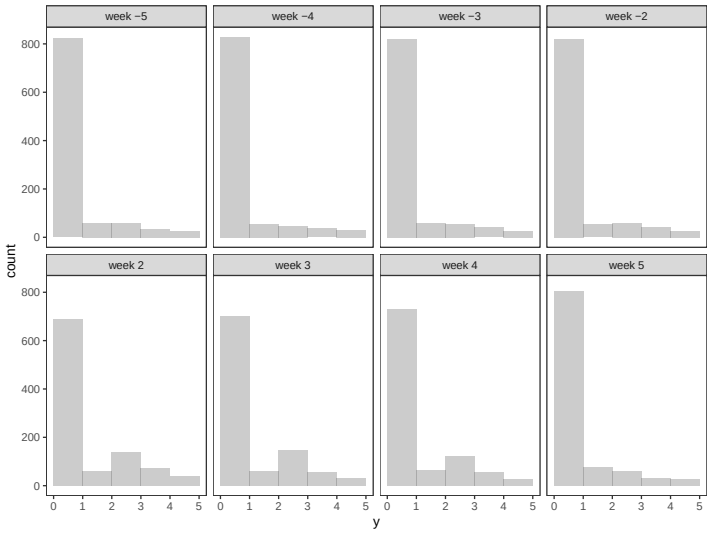


FIGURE 19.4: Zero-inflation of the data

The following R code fits zero-inflated Poisson regressions. The model has parameters for the zero component and parameters for the Poisson components.

```
> library("pscl")
```

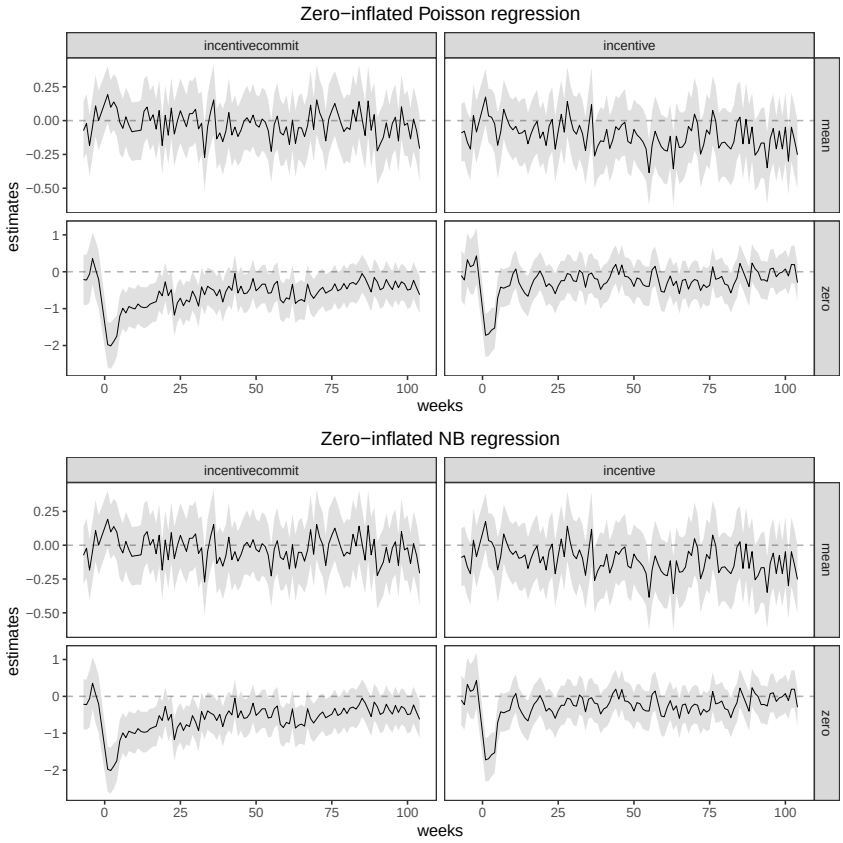


FIGURE 19.5: Zero-inflated regressions

```
> coefincentivecommit0 = coefincentivecommit
> coefincentive0       = coefincentive
> seincentivecommit0   = seincentivecommit
> seincentive0         = seincentive
> AIC0poisson          = AICnb
> for(i in 1:lweekids)
+ {
+   gymweek = gym1[which(gym1$incentive_week == weekids[i]), ]
+   regweek = zeroinfl(f.reg, dist = "poisson", data = gymweek)
+   regweekcoef = summary(regweek)$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)
```

we can fit the corresponding zero-inflated Negative-Binomial regressions. Figure 19.5 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 through 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. Figure 19.6 shows that zero-inflated Poisson regressions have the highest AICs, beating the zero-inflated Negative-Binomial regressions, which are more flexible but have more parameters to estimate.

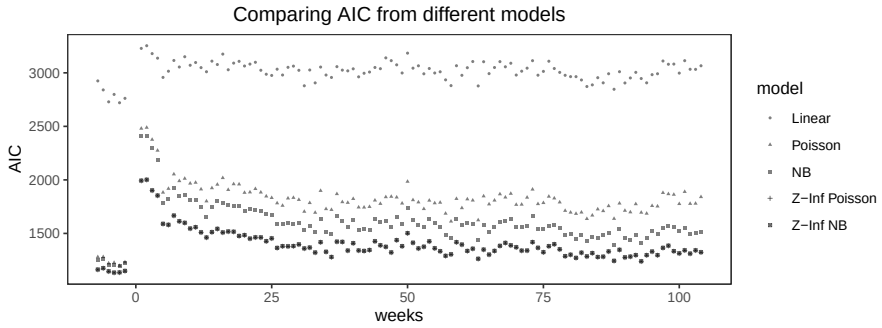


FIGURE 19.6: Comparing AICs from five regression models

19.4 Homework problems

19.1 Negative-Binomial probability mass function

Prove Proposition 19.3.

19.2 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. Implement Newton's method and compare it with `glm.nb`.

19.3 Another definition of Negative-Binomial

With IID Bernoulli(p) trials, a Negative-Binomial distribution, denoted by $y \sim \text{NB}'(r, p)$, as the number of success before the rth failure. Find the probability mass function of this random variable, and show that it is identical to the one in Proposition 19.3 with appropriately chosen (r, p). Note that this definition is more restrictive because r must be an integer.

19.4 Moments of Zero-inflated Poisson

Prove of Proposition 19.5.

19.5 Overdispersion and zero-inflation

Show 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 $p > 1/2$?

19.6 Moments of Zero-inflated Negative-Binomial

Prove of Proposition 19.6.

19.7 Poisson latent variable and the logistic model with the cloglog link

Assume that $y_i^* \mid x_i \sim \text{Poisson}(e^{x_i^{\mathrm{T}}\beta})$, and define $y_i = 1(y_i^* > 0)$ as the indicator that y_i^* is not zero. Show that $y_i \mid x_i$ follows a cloglog model, that is,

$$\text{pr}(y_i = 1 \mid x_i) = g(x_i^{\mathrm{T}}\beta),$$

where $g(z) = 1 - \exp(-e^z)$. So the cloglog model arises naturally from a latent Poisson model.

19.8 Likelihood for the Zero-inflated Poisson regression

Write down the likelihood function for the Zero-inflated Poisson model, and derive the steps for Newton's method.

19.9 Likelihood for the Zero-inflated Negative-Binomial regression

Write down the likelihood function for the Zero-inflated Negative-Binomial model, and derive the steps for Newton's method.

19.10 Prediction in the Zero-inflated Negative-Binomial regression

After obtaining the MLE $(\hat{\beta}, \hat{\gamma})$ and its asymptotic covariance matrix $\hat{V}$, predict the conditional mean $E(y_i \mid x_i)$, the conditional probability $\text{pr}(y_i = 0 \mid x_i)$, and the conditional probability $\text{pr}(y_i \geq 5 \mid x_i)$. What are the associated asymptotic standard errors?

19.11 Data analysis

Zeileis et al. (2008) give a tutorial on count outcome regressions using the dataset from Deb and Trivedi (1997). Replicate and extend their analysis based on the discussion in this chapter.

20

Generalized Linear Models, Restricted Mean Models, and the Sandwich Covariance Matrix

20.1 Generalized Linear Models

So far we have discussed the following models for independent observations $(y_i, x_i)_{i=1}^n$.

Example 20.1 *The Gaussian linear model for continuous outcomes assumes*

$$y_i \mid x_i \sim \mathrm{N}(\mu_i, \sigma^2), \quad \textit{with } \mu_i = x_i^{\mathrm{T}}\beta. \tag{20.1}$$

Example 20.2 *The logistic model for binary outcomes assumes*

$$y_i \mid x_i \sim \mathrm{Bernoulli}(\mu_i), \quad \textit{with } \mu_i = \frac{e^{x_i^{\mathrm{T}}\beta}}{1 + e^{x_i^{\mathrm{T}}\beta}}. \tag{20.2}$$

Example 20.3 *The Poisson model for count outcomes assumes*

$$y_i \mid x_i \sim \mathrm{Poisson}(\mu_i), \quad \textit{with } \mu_i = e^{x_i^{\mathrm{T}}\beta}. \tag{20.3}$$

Example 20.4 *The Negative-Binomial model for overdispersed count comes assumes*

$$y_i \mid x_i \sim \mathrm{NB}(\mu_i, \delta), \quad \textit{with } \mu_i = e^{x_i^{\mathrm{T}}\beta}, \tag{20.4}$$

where we use δ for the dispersion parameter θ avoid confusion because θ means something else below.

In the above models, μ_i denotes the conditional mean. This chapter will unify Examples 20.1–20.4 as generalized linear models (GLMs).

20.1.1 Exponential family

Consider a general conditional probability density or mass function:

$$f(y_i \mid x_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}, \tag{20.5}$$

where (θ_i, ϕ) are unknown parameters, and $\{b(\cdot), c(\cdot,\cdot)\}$ are known functions. The above conditional density (20.5) is called the natural exponential family with dispersion, where θ_i is the natural parameter depending on x_i, and ϕ is the dispersion parameter. Sometimes, $a(\phi) = 1$ and $c(y_i, \phi) = c(y_i)$, simplifying the conditional density to a natural exponential family. Examples 20.1–20.4 have a unified structure as (20.5), as detailed below.

Example 20.1 (continued) *Model (20.1) has conditional probability density function*

$$\begin{aligned} f(y_i \mid x_i; \mu_i, \sigma^2) &= (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2}\right\}, \end{aligned}$$

with

$$\theta_i = \mu_i, \quad b(\theta_i) = \frac{1}{2}\theta_i^2,$$

and

$$\phi = \sigma^2, \quad a(\phi) = \sigma^2 = \phi.$$

Example 20.2 (continued) *Model (20.2) has conditional probability mass function*

$$\begin{aligned} f(y_i \mid x_i; \mu_i) &= \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \left(\frac{\mu_i}{1-\mu_i}\right)^{y_i}(1-\mu_i) \\ &= \exp\left\{y_i \log\frac{\mu_i}{1-\mu_i} - \log\frac{1}{1-\mu_i}\right\}, \end{aligned}$$

with

$$\theta_i = \log\frac{\mu_i}{1-\mu_i} \Longleftrightarrow \mu_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}, \quad b(\theta_i) = \log\frac{1}{1-\mu_i} = \log(1+e^{\theta_i}),$$

and

$$a(\phi) = 1.$$

Example 20.3 (continued) *Model (20.3) has conditional probability mass function*

$$f(y_i \mid x_i; \mu_i) = e^{-\mu_i}\frac{\mu_i^{y_i}}{y_i!} = \exp\left\{y_i \log\mu_i - \mu_i - \log y_i!\right\},$$

with

$$\theta_i = \log\mu_i, \quad b(\theta_i) = \mu_i = e^{\theta_i},$$

and

$$a(\phi) = 1.$$

Example 20.4 (continued) *Model (20.4), for a fixed δ, has conditional probability mass function*

$$\begin{aligned} f(y_i \mid x_i; \mu_i) &= \frac{\Gamma(y_i+\delta)}{\Gamma(\delta+1)\Gamma(y_i+1)} \left(\frac{\mu_i}{\mu_i+\delta}\right)^{y_i} \left(\frac{\delta}{\mu_i+\delta}\right)^{\delta} \\ &= \exp\left\{ y_i \log\frac{\mu_i}{\mu_i+\delta} - \delta \log\frac{\mu_i+\delta}{\delta} \right. \\ &\quad \left. + \log\Gamma(y_i+\delta) - \log\Gamma(\delta+1) - \log\Gamma(y_i+1) \right\}, \end{aligned}$$

with

$$\theta_i = \log\frac{\mu_i}{\mu_i+\delta} \Longleftrightarrow \frac{\delta}{\mu_i+\delta} = 1 - e^{\theta_i}, \quad b(\theta_i) = \delta \log\frac{\mu_i+\delta}{\delta} = -\delta\log(1-e^{\theta_i}),$$

and

$$a(\phi) = 1.$$

The logit and Poisson models are simpler without the dispersion parameter. The Gaussian linear model has a dispersion parameter for the variance. The Negative-Binomial model is more complex: without fixing δ it does not belong to the exponential family with dispersion.

The exponential family (20.5) has nice properties derived from the classic Bartlett's identities. I first review Bartlett's identities:

Lemma 20.1 *Given a probability density or mass function $f(y \mid \theta)$ indexed by a scalar parameter θ, if we can change the order of expectation and differentiation, then*

$$E\left(\frac{\partial \log f(y \mid \theta)}{\partial\theta}\right) = 0, \quad E\left\{\left(\frac{\partial \log f(y \mid \theta)}{\partial\theta}\right)^2\right\} = E\left(-\frac{\partial^2 \log f(y \mid \theta)}{\partial\theta^2}\right).$$

This lemma is well-known in classic statistical theory for likelihood, and I give a simple proof below.

Proof 4 *Define $\ell(y \mid \theta) = \log f(y \mid \theta)$ as the log likelihood function, so $e^{\ell(y|\theta)}$ is the density satisfying*

$$\int e^{\ell(y|\theta)} \mathrm{d}y = \int f(y \mid \theta)\mathrm{d}y = 1$$

by the definition of a probability density function (we can replace the integral by summation for a probability mass function). Differentiate the above identity to obtain

$$\begin{aligned} & \frac{\partial}{\partial\theta}\int e^{\ell(y|\theta)}\mathrm{d}y = 0 \\ \Longrightarrow \quad & \int \frac{\partial}{\partial\theta} e^{\ell(y|\theta)}\mathrm{d}y = 0 \\ \Longrightarrow \quad & \int e^{\ell(y|\theta)} \frac{\partial}{\partial\theta}\ell(y \mid \theta)\mathrm{d}y = 0, \end{aligned}$$

which implies that $E\{\partial\ell(y \mid \theta)/\partial\theta\} = 0$. *Differentiate it twice to obtain*

$$\begin{aligned} & \frac{\partial}{\partial\theta}\int e^{\ell(y|\theta)}\frac{\partial}{\partial\theta}\ell(y \mid \theta)\mathrm{d}y = 0 \\ \Longrightarrow \quad & \int\left[e^{\ell(y|\theta)}\left\{\frac{\partial}{\partial\theta}\ell(y \mid \theta)\right\}^2 + e^{\ell(y|\theta)}\frac{\partial^2}{\partial\theta^2}\ell(y \mid \theta)\right]\mathrm{d}y = 0, \end{aligned}$$

which implies that $E[\{\partial\ell(y \mid \theta)/\partial\theta\}^2] = -E\{\partial^2\ell(y \mid \theta)/\partial\theta^2\}$.

Lemma 20.1 implies the moments of the exponential family (20.5).

Theorem 20.1 *The first two moments of (20.5) are*

$$E(y_i \mid x_i;\theta_i,\phi) \equiv \mu_i = b'(\theta_i), \quad \mathrm{var}(y_i \mid x_i;\theta_i,\phi) \equiv \sigma_i^2 = b''(\theta_i)a(\phi).$$

Proof 5 *The first two derivatives of the log conditional density are*

$$\frac{\partial\log f(y_i \mid x_i;\theta_i,\phi)}{\partial\theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}, \quad \frac{\partial^2\log f(y_i \mid x_i;\theta_i,\phi)}{\partial\theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)}.$$

Lemma 20.1 implies that

$$E\left\{\frac{y_i - b'(\theta_i)}{a(\phi)}\right\} = 0, \quad E\left[\left\{\frac{y_i - b'(\theta_i)}{a(\phi)}\right\}^2\right] = \frac{b''(\theta_i)}{a(\phi)},$$

which further imply the first two moments of y_i *given* x_i.

20.1.2 Generalized linear model

Section 20.1.1 is general, allowing the mean parameter μ_i to depend on x_i in an arbitrary way. This flexibility does not immediately generate a useful statistical procedure. To borrow information across observations, we assume that the relationship between y_i and x_i remain "stable" and can be captured by a fixed parameter β. A simple starting point is to use $x_i^{\mathrm{T}}\beta$ to approximate μ_i, which, however, works naturally only for continuous outcomes. For general outcome variables, we can link its mean and the linear combination of covariates by

$$\mu_i = \mu(x_i^{\mathrm{T}}\beta),$$

where $\mu(\cdot)$ is a known link function and β is an unknown parameter. This is called a GLM, which has the following components:

(C1) the conditional distribution (20.5) as an exponential family with dispersion;

(C2) the conditional mean $\mu_i = b'(\theta_i)$ and variance $\sigma_i^2 = b''(\theta_i)a(\phi)$ in Theorem 20.1;

(C3) the function linking the conditional mean and covariates $\mu_i = \mu(x_i^{\mathrm{T}}\beta)$.

Models (20.1)–(20.4) are the classical examples. Figure 20.1 illustrates the relationship among key quantities in a GLM. In particular,

$$\theta_i = (b')^{-1}(\mu_i) = (b')^{-1}\{\mu(x_i^{\mathrm{T}}\beta)\} \tag{20.6}$$

depends on x_i and β, with $(b')^{-1}$ indicating the inverse function of $b'(\cdot)$.

$$\begin{array}{ccccccc} x_i & \xrightarrow{\quad\beta\quad} & x_i^{\mathrm{T}}\beta & \xrightarrow{\quad\mu_i = \mu(x_i^{\mathrm{T}}\beta)\quad} & \mu_i & \xrightarrow{\quad\mu_i = b'(\theta_i)\quad} & \theta_i \\ & & & \Downarrow & & \Downarrow & \\ & & & \dfrac{\partial \mu_i}{\partial \beta} = \mu'(x_i^{\mathrm{T}}\beta)x_i & & \dfrac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) & \end{array}$$

FIGURE 20.1: Quantities in a GLM

The contribution of unit i to the log-likelihood function is

$$\ell_i = \log f(y_i \mid x_i; \beta, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi).$$

The contribution of unit i to the score function is

$$\frac{\partial \ell_i}{\partial \beta} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta},$$

where

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)},$$
$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{a(\phi)}{\sigma_i^2}$$

follow from Theorem 20.1. So

$$\frac{\partial \ell_i}{\partial \beta} = \frac{y_i - b'(\theta_i)}{\sigma_i^2}\frac{\partial \mu_i}{\partial \beta} = \frac{y_i - \mu_i}{\sigma_i^2}\frac{\partial \mu_i}{\partial \beta},$$

leading to the following score equation for the MLE:

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\sigma_i^2}\frac{\partial \mu_i}{\partial \beta} = 0, \tag{20.7}$$

or, more explicitly,

$$\sum_{i=1}^{n} \frac{y_i - \mu(x_i^{\mathrm{T}}\beta)}{\sigma_i^2}\mu'(x_i^{\mathrm{T}}\beta)x_i = 0$$

The general relationship (20.6) between θ_i and β is quite complicated. A more natural choice of $\mu(\cdot)$ is to cancel with $(b')^{-1}$ so that

$$\mu(\cdot) = b'(\cdot) \Longrightarrow \theta_i = x_i^{\mathrm{T}}\beta.$$

This link function $\mu(\cdot)$ is called the canonical link or the natural link, which leads to further simplifications:

$$\mu_i = b'(x_i^{\mathrm{T}}\beta) \Longrightarrow \frac{\partial \mu_i}{\partial \beta} = b''(x_i^{\mathrm{T}}\beta)x_i = b''(\theta_i)x_i = \frac{\sigma_i^2}{a(\phi)}x_i,$$

and

$$\begin{aligned} & \frac{\partial \ell_i}{\partial \beta} = \frac{y_i - \mu_i}{\sigma_i^2}\frac{\sigma_i^2}{a(\phi)}x_i = a(\phi)^{-1}x_i(y_i - \mu_i) \\ \Longrightarrow \quad & a(\phi)^{-1}\sum_{i=1}^{n} x_i(y_i - \mu_i) = 0 \\ \Longrightarrow \quad & \sum_{i=1}^{n} x_i\left(y_i - \mu_i\right) = 0. \end{aligned} \tag{20.8}$$

We have shown that the MLEs of models (20.1)–(20.3) all satisfy (20.8). However, the MLE of (20.4) does not because it does not use the natural link function resulting in $\mu(\cdot) \neq b'(\cdot)$:

$$\mu(*) = e^*, \quad b'(*) = \delta\frac{e^*}{1 - e^*}.$$

20.1.3 MLE and inference under a GLM

Using Bartlett's second identity in Lemma 20.1, we can write the expected Fisher information conditional on covariates as

$$\begin{aligned} \sum_{i=1}^{n} E\left(\frac{\partial \ell_i}{\partial \beta}\frac{\partial \ell_i}{\partial \beta^{\mathrm{T}}} \mid x_i\right) &= \sum_{i=1}^{n} E\left\{\left(\frac{y_i - \mu_i}{\sigma_i^2}\right)^2 \frac{\partial \mu_i}{\partial \beta}\frac{\partial \mu_i}{\partial \beta^{\mathrm{T}}} \mid x_i\right\} \\ &= \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\frac{\partial \mu_i}{\partial \beta}\frac{\partial \mu_i}{\partial \beta^{\mathrm{T}}} \\ &= \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left\{\mu'(x_i^{\mathrm{T}}\beta)\right\}^2 x_i x_i^{\mathrm{T}} \\ &= X^{\mathrm{T}}WX, \end{aligned}$$

where

$$W = \mathrm{diag}\left\{\frac{1}{\sigma_i^2}\left\{\mu'(x_i^{\mathrm{T}}\beta)\right\}^2\right\}_{i=1}^{n}.$$

With the canonical link, it further simplifies to

$$\sum_{i=1}^{n} E\left(\frac{\partial \ell_i}{\partial \beta}\frac{\partial \ell_i}{\partial \beta^{\mathrm{T}}} \mid x_i\right) = \sum_{i=1}^{n} E\left\{\left(\frac{y_i-\mu_i}{a(\phi)}\right)^2 x_i x_i^{\mathrm{T}} \mid x_i\right\}$$
$$= \{a(\phi)\}^{-2}\sum_{i=1}^{n}\sigma_i^2 x_i x_i^{\mathrm{T}}.$$

We can obtain the estimated covariance matrices by replacing the unknown parameters by their estimates. Now we review the estimated covariance matrices of the classical GLMs with canonical links.

Example 20.1 (continued) *In the Gaussian linear model, we have* $\hat{V} = \hat{\sigma}^2(X^{\mathrm{T}}X)^{-1}$ *with* σ^2 *estimated further by the residual sum of squares.*

Example 20.2 (continued) *In the binary logistic model, we have* $\hat{V} = (X^{\mathrm{T}}\hat{W}X)^{-1}$ *with* $\hat{W} = diag\{\hat{\pi}_i(1-\hat{\pi}_i)\}_{i=1}^n$, *where* $\hat{\pi}_i = e^{x_i^{\mathrm{T}}\hat{\beta}}/(1+e^{x_i^{\mathrm{T}}\hat{\beta}})$.

Example 20.3 (continued) *In the Poisson model, we have* $\hat{V} = (X^{\mathrm{T}}\hat{W}X)^{-1}$ *with* $\hat{W} = diag\{\hat{\lambda}_i\}_{i=1}^n$, *where* $\hat{\lambda}_i = e^{x_i^{\mathrm{T}}\hat{\beta}}$.

I relegate the derivation of the formula for the Negative-Binomial regression as a homework problem. It is a pure theoretical exercise since θ is usually unknown in practice .

20.2 Restricted mean model and sandwich covariance

The logit, Poisson and Negative-Binomial models are extensions of the Gaussian linear model. All of them are fully parametric models. However, we have also discussed OLS as a restricted mean model

$$E(y_i \mid x_i) = x_i^{\mathrm{T}}\beta$$

without imposing any additional assumptions (e.g., the variance) on the conditional distribution. The restricted mean model is a semiparametric model. Then a natural question is: what are the analogs of the restricted mean model for the binary and count models?

Binary outcome is too special, because the conditional mean determines the distribution. So if we assume that the conditional mean is $\mu_i = e^{x_i^{\mathrm{T}}\beta}/(1+e^{x_i^{\mathrm{T}}\beta})$, then conditional distribution must be Bernoulli(μ_i).

For other outcomes, the conditional mean cannot determine the conditional distribution. Nevertheless, if we assume

$$E(y_i \mid x_i) = \mu(x_i^{\mathrm{T}}\beta),$$

we can verify that

$$E\left\{\sum_{i=1}^{n}\frac{y_i-\mu(x_i^{\mathrm{T}}\beta)}{\tilde{\sigma}^2(x_i,\beta)}\frac{\partial\mu(x_i^{\mathrm{T}}\beta)}{\partial\beta}\right\}=E\left[E\left\{\sum_{i=1}^{n}\frac{y_i-\mu(x_i^{\mathrm{T}}\beta)}{\tilde{\sigma}^2(x_i,\beta)}\frac{\partial\mu(x_i^{\mathrm{T}}\beta)}{\partial\beta}\mid x_i\right\}\right]=0$$

for any $\tilde{\sigma}^2$ that can be a function of x_i, the true parameter β, and maybe ϕ. So we can estimate β by solving the estimating equation:

$$\sum_{i=1}^{n}\frac{y_i-\mu(x_i^{\mathrm{T}}\beta)}{\tilde{\sigma}^2(x_i,\beta)}\frac{\partial\mu(x_i^{\mathrm{T}}\beta)}{\partial\beta}=0. \tag{20.9}$$

If $\tilde{\sigma}^2(x_i,\beta)=\sigma^2(x_i)=\mathrm{var}(y_i\mid x_i)$, then the above estimating equation is the score equation derived from the GLM of an exponential family. If not, (20.9) is not a score function but it is still a valid estimating equation. In the latter case, $\tilde{\sigma}^2(x_i,\beta)$ is a "working" variance. This has important implications in practical data analysis. First, we can interpret the MLE from a GLM more broadly: it is also valid under a restricted mean model even if the conditional distribution is misspecified. Second, we can construct more general estimators beyond the MLEs from GLMs. However, we must address the issue of variance estimation since the inference based on the Fisher information matrix no longer works in general.

To simplify the notation, we assume $(x_i,y_i)_{i=1}^n$ are IID draws although we usually view the covariates as fixed. This additional assumption is innocuous as the final inferential procedures are identical. Applying the Sandwich Theorem in the Math appendix 3 to

$$w=(x,y),\quad m(w,b)=\frac{y-\mu(x^{\mathrm{T}}b)}{\tilde{\sigma}^2(x,b)}\frac{\partial\mu(x^{\mathrm{T}}b)}{\partial b},$$

we can derive the asymptotic distribution of the $\hat{\beta}$ from the estimating equation (20.9): $\sqrt{n}\left(\hat{\beta}-\beta\right)\to\mathrm{N}(0,B^{-1}MB^{-1})$ with

$$B=E\left\{\frac{1}{\tilde{\sigma}^2(x,\beta)}\frac{\partial\mu(x^{\mathrm{T}}\beta)}{\partial\beta}\frac{\partial\mu(x^{\mathrm{T}}\beta)}{\partial\beta^{\mathrm{T}}}\right\}, \tag{20.10}$$

$$M=E\left[\frac{\sigma^2(x)}{\{\tilde{\sigma}^2(x,\beta)\}^2}\frac{\partial\mu(x^{\mathrm{T}}\beta)}{\partial\beta}\frac{\partial\mu(x^{\mathrm{T}}\beta)}{\partial\beta^{\mathrm{T}}}\right]. \tag{20.11}$$

We can estimate the asymptotic variance by replacing B and M by their sample analogs, resulting in the following approximate distribution useful for statistical inference

$$\hat{\beta}\overset{\mathrm{a}}{\sim}\mathrm{N}(\beta,\hat{V}),$$

with $\hat{V} \equiv n^{-1}\hat{B}^{-1}\hat{M}\hat{B}^{-1}$, where

$$\hat{B} = n^{-1}\sum_{i=1}^{n} \frac{1}{\tilde{\sigma}^2(x_i,\hat{\beta})} \frac{\partial \mu(x_i^{\mathrm{T}}\hat{\beta})}{\partial \beta} \frac{\partial \mu(x_i^{\mathrm{T}}\hat{\beta})}{\partial \beta^{\mathrm{T}}},$$

$$\hat{M} = n^{-1}\sum_{i=1}^{n} \left\{ \frac{y_i - \mu(x_i^{\mathrm{T}}\hat{\beta})}{\tilde{\sigma}^2(x_i,\hat{\beta})} \right\}^2 \frac{\partial \mu(x_i^{\mathrm{T}}\hat{\beta})}{\partial \beta} \frac{\partial \mu(x_i^{\mathrm{T}}\hat{\beta})}{\partial \beta^{\mathrm{T}}}.$$

As a special case, when the GLM is correctly specified with $\sigma^2(x) = \tilde{\sigma}^2(x,\beta)$, then $B = M$ and the asymptotic variance reduces to the inverse of the Fisher information matrix discussed in Section 20.1.3.

Example 20.1 (continued) *In a working Gaussian linear model, we have* $\tilde{\sigma}^2(x_i,\beta) = \tilde{\sigma}^2$ *being constant and* $\partial\mu(x_i^{\mathrm{T}}\hat{\beta})/\partial\beta = x_i$. *So*

$$\hat{B} = n^{-1}\sum_{i=1}^{n} \frac{1}{\tilde{\sigma}^2} x_i x_i^{\mathrm{T}}, \quad \hat{M} = n^{-1}\sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{(\tilde{\sigma}^2)^2} x_i x_i^{\mathrm{T}}$$

with residual $\hat{\varepsilon}_i = y_i - x_i^{\mathrm{T}}\hat{\beta}$, *recovering the EHW variance estimator*

$$\hat{V} = \left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i^{\mathrm{T}}\right) \left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}.$$

Example 20.2 (continued) *In a working binary logistic model, we have* $\tilde{\sigma}^2(x_i) = \pi(x_i,\beta)\{1-\pi(x_i,\beta)\}$ *and* $\partial\mu(x_i^{\mathrm{T}}\hat{\beta})/\partial\beta = \pi(x_i,\beta)\{1-\pi(x_i,\beta)\}x_i$, *where* $\pi(x_i,\beta) = \mu(x_i^{\mathrm{T}}\beta) = e^{x_i^{\mathrm{T}}\beta}/(1+e^{x_i^{\mathrm{T}}\beta})$. *So*

$$\hat{B} = n^{-1}\sum_{i=1}^{n} \hat{\pi}_i(1-\hat{\pi}_i) x_i x_i^{\mathrm{T}}, \quad \hat{M} = n^{-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i^{\mathrm{T}}$$

with fitted mean $\hat{\pi}_i = e^{x_i^{\mathrm{T}}\hat{\beta}}/(1+e^{x_i^{\mathrm{T}}\hat{\beta}})$ *and residual* $\hat{\varepsilon}_i = y_i - \hat{\pi}_i$, *yielding a new covariance estimator*

$$\hat{V} = \left(\sum_{i=1}^{n} \hat{\pi}_i(1-\hat{\pi}_i) x_i x_i^{\mathrm{T}}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i^{\mathrm{T}}\right) \left(\sum_{i=1}^{n} \hat{\pi}_i(1-\hat{\pi}_i) x_i x_i^{\mathrm{T}}\right)^{-1}.$$

Example 20.3 (continued) *In a working Poisson model, we have* $\tilde{\sigma}^2(x_i) = \lambda(x_i,\beta)$ *and* $\partial\mu(x_i^{\mathrm{T}}\hat{\beta})/\partial\beta = \lambda(x_i,\beta)x_i$, *where* $\lambda(x_i,\beta) = \mu(x_i^{\mathrm{T}}\beta) = e^{x_i^{\mathrm{T}}\beta}$. *So*

$$\hat{B} = n^{-1}\sum_{i=1}^{n} \hat{\lambda}_i x_i x_i^{\mathrm{T}}, \quad \hat{M} = n^{-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i^{\mathrm{T}}$$

with fitted mean $\hat{\lambda}_i = e^{x_i^{\mathrm{T}}\hat{\beta}}$ *and residual* $\hat{\varepsilon}_i = y_i - \hat{\lambda}_i$, *yielding a new covariance estimator*

$$\hat{V} = \left(\sum_{i=1}^{n} \hat{\lambda}_i x_i x_i^{\mathrm{T}}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i^{\mathrm{T}}\right) \left(\sum_{i=1}^{n} \hat{\lambda}_i x_i x_i^{\mathrm{T}}\right)^{-1}.$$

Again, I relegate the derivation of the formulas for the Negative-Binomial regression as a homework problem. The R package `sandwich` implements the above covariance matrices (Zeileis, 2006).

20.3 Applications of the sandwich standard errors

20.3.1 Linear regression

In R, several functions can compute the EHW standard error: `hccm` in the `car` package, and `vcovHC` and `sandwich` in the `sandwich` package. The first two are special functions for OLS, and the third one works for general models. Below, we use these functions to compute various types of standard errors.

```
> library("car")
> library("sandwich")
> library("mlbench")
>
> ## linear regression
> data("BostonHousing")
> lm.boston = lm(medv ~ ., data = BostonHousing)
> dat.reg = data.frame(coef = coef(lm.boston),
+                      hmsk = diag(vcov(lm.boston))^(0.5),
+
+                      hccm0 = diag(hccm(lm.boston, type = "hc0"))^(0.5),
+                      sandwich0 = diag(sandwich(lm.boston, adjust = FALSE))^(0.5),
+                      vcovHC0 = diag(vcovHC(lm.boston, type = "HC0"))^(0.5),
+
+                      hccm1 = diag(hccm(lm.boston, type = "hc1"))^(0.5),
+                      sandwich1 = diag(sandwich(lm.boston, adjust = TRUE))^(0.5),
+                      vcovHC1 = diag(vcovHC(lm.boston, type = "HC1"))^(0.5),
+
+                      hccm3 = diag(hccm(lm.boston, type = "hc3"))^(0.5),
+                      vcovHC3 = diag(vcovHC(lm.boston, type = "HC3"))^(0.5))
> round(dat.reg[-1, ], 2)
          coef hmsk hccm0 sandwich0 vcovHC0 hccm1 sandwich1 vcovHC1 hccm3 vcovHC3
crim     -0.11 0.03  0.03      0.03    0.03  0.03      0.03    0.03
0.03    0.03
zn         0.05 0.01  0.01      0.01    0.01  0.01      0.01    0.01
0.01    0.01
indus      0.02 0.06  0.05      0.05    0.05  0.05      0.05    0.05
0.05    0.05
chas1      2.69 0.86  1.28      1.28    1.28  1.29      1.29    1.29
1.35    1.35
nox      -17.77 3.82  3.73      3.73    3.73  3.79      3.79    3.79
3.92    3.92
rm         3.81 0.42  0.83      0.83    0.83  0.84      0.84    0.84
0.89    0.89
age        0.00 0.01  0.02      0.02    0.02  0.02      0.02    0.02
0.02    0.02
dis       -1.48 0.20  0.21      0.21    0.21  0.21      0.21    0.21
0.22    0.22
```

```
rad        0.31 0.07  0.06      0.06     0.06  0.06      0.06     0.06
0.06     0.06
tax       -0.01 0.00  0.00      0.00     0.00  0.00      0.00     0.00
0.00     0.00
ptratio   -0.95 0.13  0.12      0.12     0.12  0.12      0.12     0.12
0.12     0.12
b          0.01 0.00  0.00      0.00     0.00  0.00      0.00     0.00
0.00     0.00
lstat     -0.52 0.05  0.10      0.10     0.10  0.10      0.10     0.10
0.10     0.10
```

In the above, `hmsk` denotes the variance estimator under the homoskedastic Gaussian linear model; the `sandwich` function can compute HC0 and HC1, corresponding to adjusting for the degrees of freedom or not; `hccm` and `vcovHC` can compute other HC standard errors.

20.3.2 Logistic regression

20.3.2.1 An application

In the flu shot example, two types of standard errors are rather similar.

```
> 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)
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
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(assign.logit, vcov = sandwich)

z test of coefficients:

              Estimate Std. Error z value  Pr(>|z|)
(Intercept) -2.1998145  0.4059386 -5.4191 5.991e-08 ***
assign      -0.1975283  0.1371785 -1.4399  0.149885
age         -0.0079859  0.0057053 -1.3997  0.161590
copd         0.3370371  0.1556781  2.1650  0.030391 *
dm           0.4543416  0.1394709  3.2576  0.001124 **
heartd       0.6761895  0.1521105  4.4454 8.774e-06 ***
race        -0.2429488  0.1430957 -1.6978  0.089544 .
renal        1.5195049  0.3659238  4.1525 3.288e-05 ***
```

```
sex          -0.2120954  0.1489435 -1.4240  0.154447
liverd        0.0989572  1.1411133  0.0867  0.930894
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

20.3.2.2 A misspecified logistic regression

Freedman (2006) discussed the following misspecified logistic regression. The discrepancy between the two types of standard errors is a warning of the misspecification of the conditional mean function because it determines the whole conditional distribution. In this case, it is not meaningful to interpret the coefficients.

```
> n = 100
> x = runif(n, 0, 10)
> prob.x = 1/(1 + exp(3*x - 0.5*x^2))
> y = rbinom(n, 1, prob.x)
> freedman.logit = glm(y ~ x, family = binomial(link = logit))
> summary(freedman.logit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.6764     1.3254  -5.037 4.72e-07 ***
x             1.1083     0.2209   5.017 5.25e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(freedman.logit, vcov = sandwich)

z test of coefficients:

            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.67641    2.46035 -2.7136 0.006656 **
x            1.10832    0.39672  2.7937 0.005211 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

20.3.3 Poisson regression

20.3.3.1 A correctly specified Poisson regression

I first generate data from a correctly specified Poisson regression. The two types of standard errors are very close.

```
> n = 1000
> x = rnorm(n)
> lambda.x = exp(x/5)
> y = rpois(n, lambda.x)
> pois.pois = glm(y ~ x, family = poisson(link = log))
> summary(pois.pois)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.004386   0.032117  -0.137    0.891
x            0.189069   0.031110   6.077 1.22e-09 ***
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(pois.pois, vcov = sandwich)

z test of coefficients:

              Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.0043862  0.0311957 -0.1406    0.8882
x            0.1890691  0.0299124  6.3208 2.603e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

20.3.3.2 A Negative-Binomial regression model

I then generate data from a Negative-Binomial regression model. The conditional mean function is still $E(y_i \mid x_i) = e^{x_i^\top \beta}$, so we can still use Poisson regression as a working model. The robust standard error doubles the classical standard error.

```
> library(MASS)
> theta = 0.2
> y = rnegbin(n, mu = lambda.x, theta = theta)
> nb.pois = glm(y ~ x, family = poisson(link = log))
> summary(nb.pois)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.07747    0.03315  -2.337   0.0194 *
x            0.13847    0.03241   4.272 1.94e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(nb.pois, vcov = sandwich)

z test of coefficients:

             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.077475   0.079431 -0.9754  0.32937
x            0.138467   0.061460  2.2530  0.02426 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Because the true model is the Negative-Binomial regression, we can use the correct model to fit the data. Theoretically, the MLE is the most efficient estimator. However, in this particular dataset, the robust standard error from Poisson regression is no larger than the one from Negative-Binomial regression. Moreover, the robust standard errors from the Poisson and Negative-Binomial regressions are very close.

```
> nb.nb = glm.nb(y ~ x)
> summary(nb.nb)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.08047    0.07336  -1.097   0.2727
x            0.16487    0.07276   2.266   0.0234 *
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(nb.nb, vcov = sandwich)

z test of coefficients:

             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.080467   0.079510  -1.012 0.311517
x            0.164869   0.063902   2.580 0.009879 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

20.3.3.3 Misspecification of the conditional mean

When the conditional mean function is misspecified, the Poisson and Negative-Binomial regressions give different point estimates, and it is unclear how to compare the standard errors.

```
> lambda.x = x^2
> y = rpois(n, lambda.x)
> wr.pois = glm(y ~ x, family = poisson(link = log))
> summary(wr.pois)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.03760    0.03245  -1.159 0.246457
x            0.11933    0.03182   3.751 0.000176 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(wr.pois, vcov = sandwich)

z test of coefficients:

             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.037604   0.053033 -0.7091   0.4783
x            0.119331   0.101126  1.1800   0.2380

>
> wr.nb = glm.nb(y ~ x)
There were 26 warnings (use warnings() to see them)
> summary(wr.nb)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.15984    0.05802   2.755  0.00587 **
x           -0.34622    0.05789  -5.981 2.22e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(wr.nb, vcov = sandwich)

z test of coefficients:

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.159837   0.061564  2.5963 0.009424 **
x           -0.346223   0.238124 -1.4540 0.145957
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Overall, for count outcome regression, it seems that Poisson regression suffices as long as we use the robust standard error. The Negative-Binomial is unlikely to offer more if only the conditional mean is of interest.

20.3.4 How robust are the robust standard errors?

Section 20.2 discussed the restricted mean model as an extension of the GLM, allowing for misspefication of the GLM while still preserving the conditional mean. We can extend the discussion to other parametric models. Huber (1967) started the literature on the statistical properties of the MLE in a misspecified model, and White (1982) addressed detailed inferential problems. Buja et al. (2019b) reviewed this topic recently.

The discussion in Section 20.2 is useful when the conditional mean is correctly specified. However, if we think the GLM is severely misspecified with a wrong conditional mean, then the robust sandwich standard errors are unlikely to be helpful because the MLE converges to a wrong parameter at the first place (Freedman, 2006).

20.4 Homework problems

20.1 MLE in GLMs with binary regressors

The MLEs of Models (20.1)–(20.3) does not have explicit formulas in general. But in the special case with x_i containing an intercept and a binary covariate, the MLEs do have simple formulas. Find them in terms of sample means of the outcomes. Does the MLE of Model (20.4) have explicit formulas with x_i containing an intercept and a binary covariate? If so, find it; if not, propose an estimator with explicit formula.

20.2 The sandwich variance formula

Verify (20.10) and (20.11).

20.3 Negative-Binomial covariance matrices

Assume that δ is known. Derive the estimated asymptotic covariance matrices of the MLE in the Negative-Binomial regression with $\mu_i = e^{x_i^\top \beta}$, one assuming a correctly specified model and the other allowing for misspefication of the model.

20.4 Robust standard errors in the Karolinska data

Report the robust standard errors in the case study of Section 4 in Lecture 18. For some models, the function `coeftest(*, vcov = sandwich)` does work. Alternatively, you can use the nonparametric bootstrap to obtain the robust standard errors.

20.5 Robust standard errors in the gym data

Report the robust standard errors in the case study of Section 3 in Lecture 19.

20.6 Additional reading

Freedman (2006).

21

Generalized Estimating Equation for Correlated Multivariate Data

In previous chapter, we dealt with cross-sectional data, that is, we observed n units at a particular time point, collecting various covariates and outcomes. In addition, we assume that these units are independent, and sometimes, we even assume they are IID draws. Many applications have correlated data. Two canonical examples are

(E1) repeated measurements of the same units over time, which are often called longitudinal data in biostatistics (Fitzmaurice et al., 2012) or panel data in econometrics (Wooldridge, 2010); and

(E2) clustered observations belonging to classrooms, villages, etc, which are common in cluster-randomized experiments in education (Schochet, 2013) and public heath (Turner et al., 2017a,b).

Many excellent textbooks treat this topic intensively. This chapter focuses on a simple yet powerful strategy, which is a natural extension of the GLM discussed in last chapter. It was initially proposed in Liang and Zeger (1986), the most cited paper published in *Biometrika* in the past one hundred years (Titterington, 2013). For simplicity, we will use the term "longitudinal data" for general correlated data.

21.1 Examples of correlated data

21.1.1 Longitudinal data

We have used the data from Royer et al. (2015) in lecture 19. Each worker's number of gym visits was repeatedly measured over more than 100 weeks. It is a standard longitudinal dataset. In lecture 19, we conducted analysis for each week, and in this lecture, we will accommodate the longitudinal nature of the data.

21.1.2 Clustered data: a neuroscience experiment

Moen et al. (2016) examined the effects of Pten knockdown and fatty acid delivery on soma size of neurons in the brain of a mouse. The useful variables for our analysis are the id of mouse `mouseid`, the fatty acid level `fa`, the Pten knockdown indicator `pten`, the outcome `somasize`, the number of neurons `numpten` and `numctrl` under Pten knockdown or not.

```
> Pten = read.csv("PtenAnalysisData.csv")
> head(Pten)
  mouseid fa pten somasize numctrl numpten  prop meanss_pten meanss_all
1       0  0    0    83.84       30       44 59.46        88.52
89.47
2       0  0    0    69.98       30       44 59.46        88.52
89.47
3       0  0    0    82.13       30       44 59.46        88.52
89.47
4       0  0    0    86.45       30       44 59.46        88.52
89.47
5       0  0    0    74.03       30       44 59.46        88.52
89.47
6       0  0    0    71.69       30       44 59.46        88.52
89.47
```

The three-way table below shows the treatment combinations for 14 mice, from which we can see that the Pten knockdown indicator varies within mice, but fatty acid level varies only between mice.

```
> table(Pten$mouseid, Pten$fa, Pten$pten)
, ,  = 0


     0  1  2
  0 30  0  0
  1 58  0  0
  2 18  0  0
  3  2  0  0
  4 56  0  0
  5  0 39  0
  6  0 33  0
  7  0 58  0
  8  0 60  0
  9  0  0 15
  10 0  0 27
  11 0  0  7
  12 0  0 34
  13 0  0 22

, ,  = 1


     0  1  2
  0 44  0  0
  1 68  0  0
  2 33  0  0
  3 11  0  0
```

```
4   76   0   0
5    0  55   0
6    0  55   0
7    0  75   0
8    0  92   0
9    0   0  34
10   0   0  29
11   0   0  20
12   0   0  53
13   0   0  38
```

21.1.3 Clustered data: a public health intervention

Poor sanitation leads to morbidity and mortality in developing countries. In 2012, Guiteras et al. (2015) conducted a cluster-randomized experiment in rural Bangladesh to evaluate the effectiveness of different policies on the use of hygienic latrines. To illustrate our theory, we use a subset of their original data and exclude the households not eligible to subsidies or with missing outcomes, resulting in 10125 households in total. The median, mean, and maximum of village size are 83, 119, and 500, respectively. We choose the outcome y_{it} as the binary indicator for whether the household (i, t) had access to a hygienic latrine or not, measured in June 2013, and covariate x_{it} as the access rate to hygienic latrines in the community that household (i, t) belonged to, measured in January 2012 before the experiment.

The useful variables below are `x`, `y`, `z`, and the village id `vid`.

```
> hygaccess = read.csv("hygaccess.csv")
> hygaccess = hygaccess[,c("r4_hyg_access", "treat_cat_1",
+                          "bl_c_hyg_access", "vid", "eligible")]
> hygaccess = hygaccess[which(hygaccess$eligible=="Eligible"&
+                               hygaccess$r4_hyg_access!="Missing"),]
> hygaccess$y = ifelse(hygaccess$r4_hyg_access=="Yes", 1, 0)
> hygaccess$z = hygaccess$treat_cat_1
> hygaccess$x = hygaccess$bl_c_hyg_access
```

21.2 Marginal model and the generalized estimating equation

We will extend the conditional mean model to deal with longitudinal data, where we observe outcome y_{it} and covariate x_{it} for each unit $i = 1, \ldots, n$ and time $t = 1, \ldots, n_i$. The n_i's can vary across units. When $n_i = 1$ for all units, we drop the time index and model the conditional mean as

$$E(y_i \mid x_i) = \mu(x_i^{\mathrm{T}}\beta),$$

and use the following estimating equation to estimate the parameter β:

$$\sum_{i=1}^{n} \frac{y_i - \mu(x_i^{\mathrm{T}}\beta)}{\tilde{\sigma}^2(x_i,\beta)} \frac{\partial \mu(x_i^{\mathrm{T}}\beta)}{\partial \beta} = 0, \tag{21.1}$$

where $\tilde{\sigma}^2(x_i,\beta)$ can be a misspecified conditional variance, usually motivated by a GLM. With an $n_i \times 1$ vector outcome and an $n_i \times p$ covariate matrix

$$Y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}, \quad X_i = \begin{pmatrix} x_{i1}^{\mathrm{T}} \\ \vdots \\ x_{in_i}^{\mathrm{T}} \end{pmatrix}, \quad (i = 1, \ldots, n) \tag{21.2}$$

we can extend the restricted mean model to

$$\begin{aligned} E(Y_i \mid X_i) &\equiv \begin{pmatrix} E(y_{i1} \mid X_i) \\ \vdots \\ E(y_{in_i} \mid X_i) \end{pmatrix} \\ &= \begin{pmatrix} E(y_{i1} \mid x_{i1}) \\ \vdots \\ E(y_{in_i} \mid x_{in_i}) \end{pmatrix} \qquad (21.3) \\ &= \begin{pmatrix} \mu(x_{i1}^{\mathrm{T}}\beta) \\ \vdots \\ \mu(x_{in_i}^{\mathrm{T}}\beta) \end{pmatrix} \qquad (21.4) \\ &\equiv \mu(X_i, \beta), \end{aligned}$$

where we have made two key assumptions in (21.3) and (21.4). Assumption (21.3) requires that the conditional mean of y_{it} depends only on x_{it} but not any other x_{is} with $s \neq t$. Assumption (21.4) requires that the relationship between x_{it} and y_{it} is stable across time with time-invariant functional form $\mu(\cdot)$ and parameter β. The model assumptions in (21.3) and (21.4) are really strong, and I defer the critiques to the end of this chapter. Nevertheless, the marginal model attracts practitioners for

(A1) its similarity to GLM and the restricted mean model, and

(A2) its simplicity of requiring only specification of the marginal conditional means, not the whole joint distribution.

The advantage (A1) facilitates the interpretation of the coefficient, and the advantage (A2) is crucial because of the lack of familiar multivariate distributions in statistics except the multivariate Normal. The generalized estimating equation (GEE) for β is the vector form of (21.1):

$$\sum_{i=1}^{n} \underbrace{\frac{\partial \mu^{\mathrm{T}}(X_i,\beta)}{\partial \beta}}_{p \times n_i} \underbrace{\tilde{V}^{-1}(X_i,\beta)}_{n_i \times n_i} \underbrace{\{Y_i - \mu(X_i,\beta)\}}_{n_i \times 1} = \underbrace{0}_{p \times 1}, \tag{21.5}$$

where (21.5) has a similar form as (21.1) with three terms organized to match the dimension so that matrix multiplications are well-defined:

(GEE1) the last term

$$Y_i - \mu(X_i, \beta) = \begin{pmatrix} E(y_{i1} \mid x_{i1}) - \mu(x_{i1}^{\mathrm{T}}\beta) \\ \vdots \\ E(y_{in_i} \mid x_{in_i}) - \mu(x_{in_i}^{\mathrm{T}}\beta) \end{pmatrix}$$

represent the residual vector,

(GEE2) the second term is the inverse of $\tilde{V}(X_i, \beta)$, a working covariance matrix of the conditional distribution of Y_i given X_i which may be misspecified:

$$\tilde{V}(X_i, \beta) \neq V(X_i) \equiv \mathrm{cov}(Y_i \mid X_i).$$

It is relatively easy to specify the working variance $\tilde{\sigma}^2(x_{it}, \beta)$ for each marginal component, for example, based on the marginal GLM. So the key is to specify the $n_i \times n_i$ dimensional correlation matrix R_i to obtain

$$\tilde{V}(X_i, \beta) = \mathrm{diag}\{\tilde{\sigma}^2(x_{it}, \beta)\}_{i=1}^{n_i} R_i \mathrm{diag}\{\tilde{\sigma}^2(x_{it}, \beta)\}_{i=1}^{n_i}.$$

We assume that R_i's are given now, and will discuss how to choose them in a later section.

(GEE3) the first term is the partial derivative of an $1 \times n_i$ row vector $\mu^{\mathrm{T}}(X_i, \beta) = (\mu(x_{i1}^{\mathrm{T}}\beta), \ldots, \mu(x_{in_i}^{\mathrm{T}}\beta))$ with respect to a $p \times 1$ column vector $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$:

$$\begin{aligned} \frac{\partial \mu^{\mathrm{T}}(X_i, \beta)}{\partial \beta} &= \left(\frac{\mu(x_{i1}^{\mathrm{T}}\beta)}{\partial \beta}, \ldots, \frac{\mu(x_{in_i}^{\mathrm{T}}\beta)}{\partial \beta} \right) \\ &= \begin{pmatrix} \frac{\partial \mu(x_{i1}^{\mathrm{T}}\beta)}{\partial \beta_1} & \cdots & \frac{\partial \mu(x_{in_i}^{\mathrm{T}}\beta)}{\partial \beta_1} \\ \vdots & & \vdots \\ \frac{\partial \mu(x_{i1}^{\mathrm{T}}\beta)}{\partial \beta_p} & \cdots & \frac{\partial \mu(x_{in_i}^{\mathrm{T}}\beta)}{\partial \beta_p} \end{pmatrix}, \end{aligned}$$

which is a $p \times n_i$ matrix, denoted by $D_i^{\mathrm{T}}(\beta)$.

21.3 Statistical inference with GEE

21.3.1 Computation using the Gauss–Newton method

We can use Newton's method to solve the GEE (21.5). However, calculating the derivative of the left-hand side of (21.5) involves calculating the second

order derivative of $\partial\mu^{\mathrm{T}}(X_i,\beta)/\partial\beta$. A simpler alternative without calculating the second order derivative is the Gauss–Newton method. We use the following approximation:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \frac{\partial\mu^{\mathrm{T}}(X_i,\beta)}{\partial\beta}\tilde{V}^{-1}(X_i)\left\{Y_i-\mu(X_i,\beta)\right\} \\
&= \sum_{i=1}^{n} D_i^{\mathrm{T}}(\beta^{\mathrm{old}})\tilde{V}^{-1}(X_i,\beta^{\mathrm{old}})\left[\left\{Y_i-\mu(X_i,\beta^{\mathrm{old}})\right\}-D_i(\beta^{\mathrm{old}})(\beta-\beta^{\mathrm{old}})\right] \\
&= \sum_{i=1}^{n} D_i^{\mathrm{T}}(\beta^{\mathrm{old}})\tilde{V}^{-1}(X_i,\beta^{\mathrm{old}})\left\{Y_i-\mu(X_i,\beta^{\mathrm{old}})\right\} \\
&\quad -\sum_{i=1}^{n} D_i^{\mathrm{T}}(\beta^{\mathrm{old}})\tilde{V}^{-1}(X_i,\beta^{\mathrm{old}})D_i(\beta^{\mathrm{old}})(\beta-\beta^{\mathrm{old}}).
\end{aligned}$$

So given β^{old}, we update it as

$$\begin{aligned}
\beta^{\mathrm{new}} &= \beta^{\mathrm{old}} + \left\{\sum_{i=1}^{n} D_i^{\mathrm{T}}(\beta^{\mathrm{old}})\tilde{V}^{-1}(X_i,\beta^{\mathrm{old}})D_i(\beta^{\mathrm{old}})\right\}^{-1} \\
&\quad \times \sum_{i=1}^{n} D_i^{\mathrm{T}}(\beta^{\mathrm{old}})\tilde{V}^{-1}(X_i,\beta^{\mathrm{old}})\left\{Y_i-\mu(X_i,\beta^{\mathrm{old}})\right\}, \qquad (21.6)
\end{aligned}$$

which can be written as WLS.

21.3.2 Asymptotic inference

The asymptotic distribution of $\hat{\beta}$ follows from the sandwich lemma directly. We can verify that $\sqrt{n}(\hat{\beta}-\beta)\to \mathrm{N}(0,B^{-1}MB^{-1})$ in distribution where

$$\begin{aligned}
B &= E\left\{n^{-1}\sum_{i=1}^{n} D_i^{\mathrm{T}}(\beta)\tilde{V}^{-1}(X_i,\beta)D_i(\beta)\right\}, \\
M &= E\left\{n^{-1}\sum_{i=1}^{n} D_i^{\mathrm{T}}(\beta)\tilde{V}^{-1}(X_i,\beta)V(X_i)\tilde{V}^{-1}(X_i,\beta)D_i(\beta)\right\}.
\end{aligned}$$

After obtaining $\hat{\beta}$ and the residual vector $\hat{\varepsilon}_i = Y_i-\mu(X_i,\hat{\beta})$ for unit i $(i=1,\ldots,n)$, we can conduct asymptotic inference based on the Normal approximation

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

where

$$\begin{aligned}
\hat{B} &= n^{-1}\sum_{i=1}^{n} D_i^{\mathrm{T}}(\hat{\beta})\tilde{V}^{-1}(X_i,\hat{\beta})D_i(\hat{\beta}), \\
\hat{M} &= n^{-1}\sum_{i=1}^{n} D_i^{\mathrm{T}}(\hat{\beta})\tilde{V}^{-1}(X_i,\hat{\beta})\hat{\varepsilon}_i\hat{\varepsilon}_i^{\mathrm{T}}\tilde{V}^{-1}(X_i,\hat{\beta})D_i(\hat{\beta}).
\end{aligned}$$

This covariance estimator proposed by Liang and Zeger (1986), is robust to the misspecification of the marginal variances and the correlation structure as long as the conditional mean of Y_i given X_i is correctly specified.

21.3.3 Implementation: choice of the working covariance matrix

We have not discussed the choice of the working correlation matrix R_i. Different choices do not affect the consistency but affect the efficiency of $\hat{\beta}$. A simple starting point is the independent working correlation matrix $R_i = I_{n_i}$. Under this correlation matrix, the GEE reduces to

$$\sum_{i=1}^{n} \left(\frac{\mu(x_{i1}^{\mathrm{T}}\beta)}{\partial\beta}, \ldots, \frac{\mu(x_{in_i}^{\mathrm{T}}\beta)}{\partial\beta} \right) \begin{pmatrix} \tilde{\sigma}^{-2}(x_{i1},\beta) & & \\ & \ddots & \\ & & \tilde{\sigma}^{-2}(x_{in_i},\beta) \end{pmatrix} \times \begin{pmatrix} E(y_{i1} \mid x_{i1}) - \mu(x_{i1}^{\mathrm{T}}\beta) \\ \vdots \\ E(y_{in_i} \mid x_{in_i}) - \mu(x_{in_i}^{\mathrm{T}}\beta) \end{pmatrix} = 0,$$

or, more transparently,

$$\sum_{i=1}^{n} \sum_{t=1}^{n_i} \frac{y_{it} - \mu(x_{it}^{\mathrm{T}}\beta)}{\tilde{\sigma}^2(x_{it},\beta)} \frac{\partial\mu(x_{it}^{\mathrm{T}}\beta)}{\partial\beta} = 0. \tag{21.7}$$

This is simply the estimating equation of a conditional mean model treating all data points (i,t) as independent observations. This implies that the point estimate assuming independence across all data points is still consistent, although we must change the standard error as in Section 21.3.2.

With this simple starting point, we have a consistent yet inefficient estimate of β, and then we can compute the residuals. The correlation among the residuals contains information about the true covariance matrix. With small and equal n_i's, we can estimate the conditional covariance without imposing any structure based on the residuals. Using the estimated the covariance matrix, we can update the GEE estimate to improve efficiency. This leads a a two-step procedure.

An important intermediate case is the motivated by the exchangeability of the data points within the same unit i, so the working covariance matrix is $\tilde{V}(X_i,\beta) = \mathrm{diag}\{\tilde{\sigma}(x_{it})\}_{i=1}^{n_i} R_i(\rho) \mathrm{diag}\{\tilde{\sigma}(x_{it})\}_{i=1}^{n_i}$, where

$$R_i(\rho) = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

We can estimate ρ based on the residuals from the first step.

The above three choices of the working covariance matrix are called "`independent`", "`unstructured`", and "`exchangeable`" in the "`corstr`" parameter of the function `gee` in the `gee` package in `R`. This function also contains other choices proposed by Liang and Zeger (1986).

A carefully chosen working covariance matrix can lead to efficiency gain compared to the simple independent covariance matrix. A fully efficient estimator requires a correctly specified working covariance matrix. This is often an infeasible goal, and what is more, the conditional covariance $\text{cov}(Y_i \mid X_i)$ is a nuisance parameter if the conditional mean is the main parameter of interest. In practice, the independent working covariance suffices in many applications despite its potential efficiency loss. This is similar to the use of OLS in the presence of heteroskedasticity in linear models. Section 21.4 focuses on the independent working covariance, which is common in econometrics. Section 21.6 gives further justifications of this simple strategy.

21.4 A special case: cluster-robust standard error

I will discuss linear and logistic regressions in this section, and leave the technical details of Poisson regression as a homework problem. Stack the Y_i's and X_i's in (21.2) together to obtain

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix},$$

which are the N dimensional outcome vector and $N \times p$ covariate matrix, where $N = \sum_{i=1}^{n} n_i$.

21.4.1 OLS

An importance special case is the marginal linear model with an independent working covariance matrix and homoskedasticity, resulting in the following estimating equation:

$$\sum_{i=1}^{n} \sum_{t=1}^{n_i} x_{it}(y_{it} - x_{it}^{\mathrm{T}}\beta) = 0.$$

So the point estimator is nothing but the pooled OLS using all data points:

$$\begin{aligned}\hat{\beta} &= \left(\sum_{i=1}^{n}\sum_{t=1}^{n_i} x_{it}x_{it}^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{n_i} x_{it}y_{it} \\ &= \left(\sum_{i=1}^{n} X_i^{\mathrm{T}}X_i\right)^{-1}\sum_{i=1}^{n} X_i^{\mathrm{T}}Y_i \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y.\end{aligned}$$

The three forms of $\hat{\beta}$ above are identical: the first one is based on N observations, the second one is based on n independent units, and the last one is based on matrix form with pooled data. Although the point estimate is identical to the case with independent data points, we must adjust for the standard error according to Section 21.3.2. From

$$D_i(\beta) = \begin{pmatrix} x_{i1}^{\mathrm{T}} \\ \vdots \\ x_{in_i}^{\mathrm{T}} \end{pmatrix} = X_i,$$

we can verify that

$$\hat{\mathrm{cov}}(\hat{\beta}) = \left(\sum_{i=1}^{n} X_i^{\mathrm{T}}X_i\right)^{-1}\sum_{i=1}^{n} X_i^{\mathrm{T}}\hat{\varepsilon}_i\hat{\varepsilon}_i^{\mathrm{T}}X_i\left(\sum_{i=1}^{n} X_i^{\mathrm{T}}X_i\right)^{-1},$$

where $\hat{\varepsilon}_i = Y_i - X_i\hat{\beta} = (\hat{\varepsilon}_{i1},\ldots,\hat{\varepsilon}_{in_i})^{\mathrm{T}}$ is the residual vector of unit i. This is called the cluster-robust standard error in econometrics, which is often much larger than the Eicker-Huber–White heteroskedasticity-robust standard error assuming independence of observations (i,t):

$$\hat{\mathrm{cov}}_{\mathrm{EHW}}(\hat{\beta}) = \left(\sum_{i=1}^{n}\sum_{t=1}^{n_i} x_{it}x_{it}^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{n_i}\hat{\varepsilon}_{it}^2 x_{it}x_{it}^{\mathrm{T}}\left(\sum_{i=1}^{n}\sum_{t=1}^{n_i} x_{it}x_{it}^{\mathrm{T}}\right)^{-1}.$$

Note that

$$X^{\mathrm{T}}X = \sum_{i=1}^{n} X_i^{\mathrm{T}}X_i = \sum_{i=1}^{n}\sum_{t=1}^{n_i} x_{it}x_{it}^{\mathrm{T}},$$

so the bread matrices in $\hat{\mathrm{cov}}(\hat{\beta})$ and $\hat{\mathrm{cov}}_{\mathrm{EHW}}(\hat{\beta})$ are identical. The only difference is due to the meat matrices:

$$\sum_{i=1}^{n} X_i^{\mathrm{T}}\hat{\varepsilon}_i\hat{\varepsilon}_i^{\mathrm{T}}X_i = \sum_{i=1}^{n}\left(\sum_{t=1}^{n_i}\hat{\varepsilon}_{it}x_{it}\right)\left(\sum_{t=1}^{n_i}\hat{\varepsilon}_{it}x_{it}\right)^{\mathrm{T}} \neq \sum_{i=1}^{n}\sum_{t=1}^{n_i}\hat{\varepsilon}_{it}^2 x_{it}x_{it}^{\mathrm{T}}$$

in general.

21.4.2 Logistic regression

For binary outcomes, we can use marginal logistic models with an independent working covariance matrix, resulting in the following estimating equation:

$$\sum_{i=1}^{n}\sum_{t=1}^{n_i} x_{it}\left\{y_{it}-\pi(x_{it},\beta)\right\}=0$$

where $\pi(x_{it},\beta)=e^{x_{it}^{\mathrm{T}}\beta}/(1+e^{x_{it}^{\mathrm{T}}\beta})$. So the point estimator is the pooled logistic regression using all data points, but we must adjust for the standard error according to Section 21.3.2. From

$$D_i(\beta)=\begin{pmatrix}\pi(x_{i1},\beta)\{1-\pi(x_{i1},\beta)\}x_{i1}^{\mathrm{T}}\\ \vdots\\ \pi(x_{in_i},\beta)\{1-\pi(x_{in_i},\beta)\}x_{in_i}^{\mathrm{T}}\end{pmatrix}=\tilde{V}(X_i,\beta)X_i,$$

with $\tilde{V}(X_i,\beta)=\mathrm{diag}\{\pi(x_{it},\beta)\{1-\pi(x_{it},\beta)\}\}_{t=1}^{n_i}$, we can verify that

$$\hat{B}=n^{-1}\sum_{i=1}^{n}X_i^{\mathrm{T}}\hat{V}_iX_i,\quad \hat{M}=n^{-1}\sum_{i=1}^{n}X_i^{\mathrm{T}}\hat{\varepsilon}_i\hat{\varepsilon}_i^{\mathrm{T}}X_i,$$

where $\hat{\varepsilon}_i=(\hat{\varepsilon}_{i1},\ldots,\hat{\varepsilon}_{in_i})^{\mathrm{T}}$ with residual $\hat{\varepsilon}_{it}=y_{it}-e^{x_{it}\hat{\beta}}/(1+e^{x_{it}\hat{\beta}})$, and $\hat{V}_i=\mathrm{diag}\{\pi(x_{it},\hat{\beta})\{1-\pi(x_{it},\hat{\beta})\}\}_{t=1}^{n_i}$. So the cluster robust covariance estimator for logistic regression is

$$\hat{\mathrm{cov}}(\hat{\beta})=\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}\hat{V}_iX_i\right)^{-1}\sum_{i=1}^{n}X_i^{\mathrm{T}}\hat{\varepsilon}_i\hat{\varepsilon}_i^{\mathrm{T}}X_i\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}\hat{V}_iX_i\right)^{-1}.$$

21.5 Application

21.5.1 Longitudinal data

We will use the `gee` package for all analyses below. The regression formula `f.reg` will remain the same although other parameters may vary.

```
> library("gee")
> library("foreign")
> gym1 = read.dta("gym_treatment_exp_weekly.dta")
> f.reg = weekly_visit ~ incentive_commit + incentive + target + member_gym_pre
```

Using all data, we find significant effect of `incentive_commit` but insignificant effect of `incentive`.

```
 normal.gee = gee(f.reg, id = id,
+                 family = gaussian,
+                 corstr = "independence",
```

```
+                data = gym1)
> normal.gee = summary(normal.gee)$coef
> normal.gee
                 Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)      -0.69005   0.011136 -61.968     0.08672  -7.9572
incentive_commit  0.15666   0.008358  18.745     0.06376   2.4569
incentive         0.01022   0.008275   1.235     0.05910   0.1729
target            0.62666   0.007465  83.949     0.06773   9.2527
member_gym_pre    1.14919   0.007077 162.375     0.06252  18.3801
```

However, this pooled analysis can be misleading because we have seen from the analysis before that the treatments have no effects in the pre-experimental periods and smaller effects in the long term. A pooled analysis can dilute the short term effects, missing the treatment effect heterogeneity across time. This can be fixed by the following subgroup analysis based on time.

```
> normal.gee1 = gee(f.reg, id = id,
+                   subset = (incentive_week<0),
+                   family = gaussian,
+                   corstr = "independence",
+                   data = gym1)
> normal.gee1 = summary(normal.gee1)$coef
> normal.gee1
                  Estimate Naive S.E.  Naive z Robust S.E.  Robust z
(Intercept)      -0.879374    0.04230 -20.7868     0.08739 -10.06224
incentive_commit -0.004241    0.03175  -0.1336     0.06243  -0.06794
incentive        -0.073884    0.03144  -2.3502     0.06223  -1.18728
target            0.742675    0.02836  26.1887     0.06701  11.08301
member_gym_pre    1.601569    0.02689  59.5664     0.06600  24.26763
>
>
> normal.gee2 = gee(f.reg, id = id,
+                    subset = (incentive_week>0&incentive_week<15),
+                    family = gaussian,
+                    corstr = "independence",
+                    data = gym1)
> normal.gee2 = summary(normal.gee2)$coef
> normal.gee2
                 Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)       -0.7925    0.03275 -24.194     0.08982   -8.823
incentive_commit   0.3662    0.02458  14.898     0.06895    5.311
incentive          0.1744    0.02434   7.166     0.06457    2.701
target             0.6735    0.02196  30.674     0.07159    9.408
member_gym_pre     1.4138    0.02082  67.914     0.06727   21.018
>
> normal.gee3 = gee(f.reg, id = id,
+                    subset = (incentive_week>=15),
+                    family = gaussian,
+                    corstr = "independence",
+                    data = gym1)
> normal.gee3 = summary(normal.gee3)$coef
> normal.gee3
                  Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)      -0.661500   0.012222  -54.13     0.09028  -7.3273
incentive_commit  0.134789   0.009173   14.69     0.06676   2.0189
incentive        -0.009716   0.009082   -1.07     0.06142  -0.1582
```

```
target            0.611635   0.008193   74.66     0.07042   8.6860
member_gym_pre    1.077874   0.007768  138.77     0.06494  16.5967
```

Changing the `family` parameter to `poisson(link = log)`, we can fit a log linear model with independent Poisson covariance. Figure 21.1 shows the point estimates and confidence intervals based on the regressions above. The confidence intervals based on the cluster-robust standard errors are much wider than those based on the EHW standard errors. Without dealing with clustering, the confidence intervals are too narrow and give wrong inference .

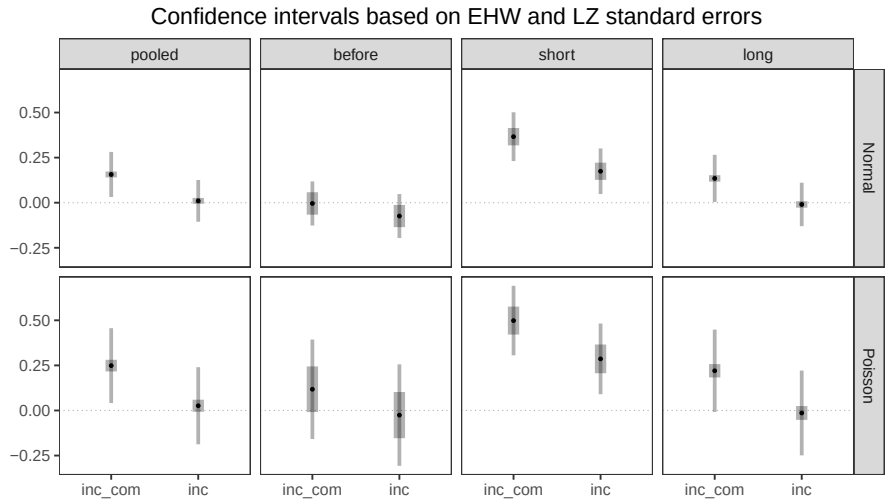


FIGURE 21.1: GEE analysis of the gym data

21.5.2 Clustered data: a neuroscience experiment

The original study was interested in the potential interaction between two treatments, so I always include the interaction term in the regression model.

From the simple specification below, `pten` has significant effect, but `fa` and the interactions are not significant.

```
 Pten.gee = gee(somasize ~ factor(fa)*pten,
+                id = mouseid,
+                family = gaussian,
+                corstr = "independence",
+                data = Pten)
> summary(Pten.gee)$coef
                Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)       93.106      1.594 58.4216       3.059  30.4374
factor(fa)1        3.756      2.175  1.7268       3.174   1.1836
factor(fa)2        6.907      2.551  2.7078       5.407   1.2774
pten              11.039      2.082  5.3016       2.200   5.0166
factor(fa)1:pten   8.727      2.834  3.0795       5.023   1.7373
```

```
factor(fa)2:pten   -2.904      3.270 -0.8881       3.554  -0.8173
>
>
> Pten.gee = gee(somasize ~ factor(fa)*pten,
+                id = mouseid,
+                family = gaussian,
+                corstr = "exchangeable",
+                data = Pten)
> summary(Pten.gee)$coef
                Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)       90.900      3.532 25.7376       2.701  33.6535
factor(fa)1        4.921      5.115  0.9621       2.914   1.6889
factor(fa)2        6.408      5.066  1.2649       5.904   1.0853
pten              11.501      1.979  5.8120       2.190   5.2515
factor(fa)1:pten   8.807      2.688  3.2766       5.050   1.7439
factor(fa)2:pten  -1.525      3.113 -0.4898       2.703  -0.5641
```

Including two covariates, we have the following results. The covariates are predictive to the outcome, changing the significance level of the main effect of `fa`. The interaction terms between `pten` and `fa` are not significant either.

```
> Pten.gee = gee(somasize ~ factor(fa)*pten + numctrl + numpten,
+                id = mouseid,
+                family = gaussian,
+                corstr = "independence",
+                data = Pten)
> summary(Pten.gee)$coef
                Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)      81.9422      2.791 29.3602      4.0917   20.026
factor(fa)1       6.2267      2.237  2.7835      4.2429    1.468
factor(fa)2      14.8956      2.657  5.6053      4.1839    3.560
pten             12.3771      2.020  6.1272      2.2477    5.507
numctrl           0.8721      0.120  7.2672      0.3028    2.880
numpten          -0.4843      0.101 -4.7948      0.2381   -2.034
factor(fa)1:pten  7.7498      2.744  2.8240      5.1064    1.518
factor(fa)2:pten -2.9629      3.166 -0.9359      3.3105   -0.895
>
>
> Pten.gee = gee(somasize ~ factor(fa)*pten + numctrl + numpten,
+                id = mouseid,
+                family = gaussian,
+                corstr = "exchangeable",
+                data = Pten)
> summary(Pten.gee)$coef
                Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)      85.3316     5.2872 16.1393      5.4095  15.7745
factor(fa)1       5.4952     4.2761  1.2851      4.0207   1.3667
factor(fa)2      12.2174     4.1669  2.9320      4.2363   2.8840
pten             11.8044     1.9718  5.9865      2.1946   5.3789
numctrl           0.9326     0.2867  3.2527      0.3479   2.6810
numpten          -0.5678     0.2504 -2.2674      0.2772  -2.0482
factor(fa)1:pten  8.5137     2.6777  3.1795      5.0612   1.6821
factor(fa)2:pten -1.7755     3.0995 -0.5728      2.7547  -0.6445
```

From the regressions above, we observe that (1) two choices of the covariance matrix do not lead to fundamental differences; and (2) without using the cluster-robust standard error, the results can be misleading.

21.5.3 Clustered data: a public health intervention

We first fit simple GEE without using the covariate.

```
> hygaccess.gee = gee(y ~ z, id = vid,
+                     family = binomial(link = logit),
+                     corstr = "independence",
+                     data = hygaccess)
> summary(hygaccess.gee)$coef
                    Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)          -0.7568    0.04439 -17.049      0.1763  -4.2924
zLPP Only             0.1551    0.06657   2.330      0.2301   0.6741
zLPP+Subsidy          0.7562    0.05503  13.742      0.2027   3.7313
zLPP+Subsidy+Supply   0.7344    0.05444  13.490      0.2010   3.6546
zSupply Only          0.3568    0.07364   4.846      0.3091   1.1544
>
> hygaccess.gee = gee(y ~ z, id = vid,
+                     family = binomial(link = logit),
+                     corstr = "exchangeable",
+                     data = hygaccess)
> summary(hygaccess.gee)$coef
                    Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)          -0.7799     0.1314 -5.9371      0.1522  -5.1235
zLPP Only             0.1638     0.2042  0.8021      0.2290   0.7152
zLPP+Subsidy          0.7789     0.1500  5.1926      0.1790   4.3524
zLPP+Subsidy+Supply   0.7348     0.1506  4.8798      0.1760   4.1753
zSupply Only          0.2690     0.2207  1.2187      0.3011   0.8931
```

Without adjusting for the covariates, treatment levels "zLPP+Subsidy" and "zLPP+Subsidy+Supply" are significant. The `exchangeable` working covariance matrix does seems to improve the estimated precision.

We then fit GEE with a covariate.

```
> hygaccess.gee = gee(y ~ z + x, id = vid,
+                     family = binomial(link = logit),
+                     corstr = "independence",
+                     data = hygaccess)
> summary(hygaccess.gee)$coef
                    Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)          -1.7526    0.06174 -28.386      0.1398  -12.538
zLPP Only             0.2277    0.06833   3.332      0.1393    1.635
zLPP+Subsidy          0.6850    0.05645  12.133      0.1191    5.749
zLPP+Subsidy+Supply   0.7389    0.05578  13.246      0.1361    5.430
zSupply Only          0.3614    0.07514   4.810      0.2426    1.490
x                     2.0488    0.08209  24.957      0.2158    9.492
>
> hygaccess.gee = gee(y ~ z + x, id = vid,
+                     family = binomial(link = logit),
+                     corstr = "exchangeable",
+                     data = hygaccess)
> summary(hygaccess.gee)$coef
                    Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)          -1.7976     0.1324 -13.575      0.1541  -11.667
zLPP Only             0.3038     0.1781   1.705      0.1946    1.561
zLPP+Subsidy          0.7227     0.1316   5.491      0.1271    5.688
zLPP+Subsidy+Supply   0.8547     0.1327   6.441      0.1247    6.855
zSupply Only          0.3236     0.1911   1.693      0.2398    1.350
```

```
x                    1.9497     0.1128  17.286     0.1947   10.016
```

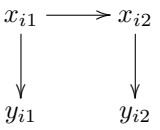
Covariate adjustment improves efficiency, and makes the choice of the working covariance matrix less important.

21.6 Critiques on the key assumptions

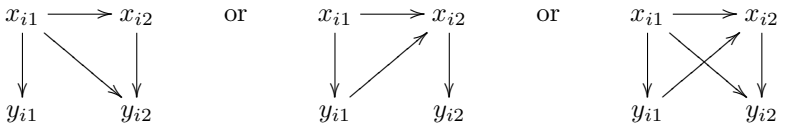
Consider the simple case with $n_i = 2$ for all i below.

21.6.1 Assumption (21.3)

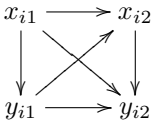
Assumption (21.3) holds automatically if $x_{it} = x_i$ is time-invariant. With time-varying covariates, it effectively rules of dynamics between x and y. Assumption (21.3) holds in the following data generating process:



It does not hold if the lagged x affects y or the lagged y affects x:



With more complex data generating processes, Assumption (21.3) does not hold in general:



Liang and Zeger (1986) assumed fixed covariates, ruling out the dynamics of x. Sullivan Pepe and Anderson (1994) pointed out the importance of Assumption (21.3) in GEE with random time-varying covariates. Sullivan Pepe and Anderson (1994) also showed that with an independent working covariance matrix, we can drop Assumption (21.3) as long as the marginal conditional mean is correctly specified: $E(y_{it} \mid x_{it}) = \mu(x_{it}^{\mathrm{T}}\beta)$. This somewhat justifies the use of the independent working covariance matrix even through it can result in efficiency loss when Assumption (21.3) holds.

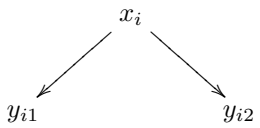
21.6.2 Assumption (21.4)

Assumption (21.4) requires "stable" relationship between x and y across time. For clustered data, we can justify this assumption by the exchangeability of the units within clusters. However, it is much harder to interpret or justify it for longitudinal data with outcome dynamics.

We consider linear structural equations with a scalar time-invariant covariate. Without direct dependence of y_{i2} on y_{i1}, the data generating process

$$\begin{aligned} y_{i1} &= \alpha_1 + \beta x_i + \varepsilon_{i1}, \\ y_{i2} &= \alpha_2 + \beta x_i + \varepsilon_{i2}, \end{aligned}$$

corresponding to the graph



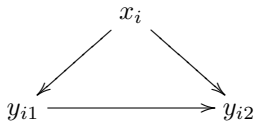
has conditional expectations $E(y_{it} \mid x_i) = \beta x_i$ if

$$E(\varepsilon_{it} \mid x_i) = 0. \tag{21.8}$$

However, with direct dependence of y_{i2} on y_{i1}, the data generating process

$$\begin{aligned} y_{i1} &= \alpha_1 + \beta x_i + \varepsilon_{i1}, \\ y_{i2} &= \alpha_2 + \gamma y_{i1} + \delta x_i + \varepsilon_{i2}, \end{aligned}$$

corresponding to the graph



has conditional expectations $E(y_{i1} \mid x_i) = \beta x_i$ but $E(y_{i2} \mid x_i) = (\beta\gamma + \delta)x_i$ if (21.8) holds. The stability assumption requires $\beta = \beta\gamma + \delta$, which is a strange restriction on the model parameters.

With time-varying covariates, this issue become even more subtle because Assumption (21.3) is unlikely to hold at the first place.

21.7 Homework problems

21.1 Gauss–Newton method to compute GEE

Write the Gauss–Newton update (21.6) as a WLS problem.

21.2 Sandwich asymptotic covariance matrix for GEE

Verify the formulas of B and M in Section 21.3.2.

21.3 Cluster-robust standard error in OLS with a constant binary regressor

Consider a special case with $x_{it} = (1, x_i)$ and $x_i \in \{0, 1\}$ for $i = 1, \ldots, n$, and view "1" as treatment and "0" as control. Show that the coefficient of x_i in the pooled OLS fit of y_{it} on $x_{it} = (1, x_i)$ equals $\hat{\tau} = \bar{y}_1 - \bar{y}_0$ where

$$\bar{y}_1 = \sum_{i=1}^{n} \sum_{t=1}^{n_i} x_i y_{it} / N_1, \quad \bar{y}_0 = \sum_{i=1}^{n} \sum_{t=1}^{n_i} (1 - x_i) y_{it} / N_0,$$

with $N_1 = \sum_{i=1}^{n} n_i x_i$ and $N_0 = \sum_{i=1}^{n} n_i (1 - x_i)$ denoting the total number of observations under treatment and control, respectively. Show further that the cluster-robust standard error of $\hat{\tau}$ equals the square root of

$$\sum_{i=1}^{n} x_i R_i^2 / N_1^2 + \sum_{i=1}^{n} (1 - x_i) R_i^2 / N_0^2,$$

where

$$R_i = \begin{cases} \sum_{t=1}^{n_i} (y_{it} - \bar{y}_1), & \text{if } x_i = 1, \\ \sum_{t=1}^{n_i} (y_{it} - \bar{y}_0), & \text{if } x_i = 0. \end{cases}$$

21.4 Cluster-robust standard error for Poisson regression

Similar to Sections 21.4.1 and 21.4.2, derive the cluster-robust standard error for Poisson regression.

21.5 Independent working covariance matrix

Verify that (21.7) is an unbiased estimating equation as long as $E(y_{it} \mid x_{it}) = \mu(x_{it}^{\mathrm{T}} \beta)$ holds, even without Assumption (21.3).

21.6 Data analysis

Re-analyze the data from Royer et al. (2015) using exchangeable working covariance matrix. Compare the corresponding results with Figure 21.1.

Part VII

Beyond modeling the conditional mean

22

Quantile Regression

22.1 From the mean to the quantile

For a random variable y, we can define its mean as

$$E(y) = \arg\min_{\mu} E\left\{(y-\mu)^2\right\}.$$

With IID data $(y_i)_{i=1}^n$, we can compute the sample mean

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

which enjoys the CLT:

$$\sqrt{n}(\bar{y}-\mu) \to \mathrm{N}(0,\sigma^2)$$

in distribution if the variance $\sigma^2 = \mathrm{var}(y)$ is finite.

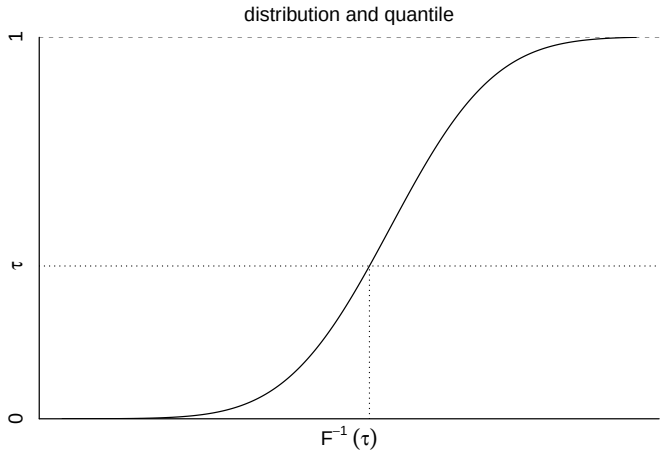


FIGURE 22.1: CDF and quantile

However, the mean can miss important information about y. How about other features of the outcome y? Quantiles can characterize the distribution of

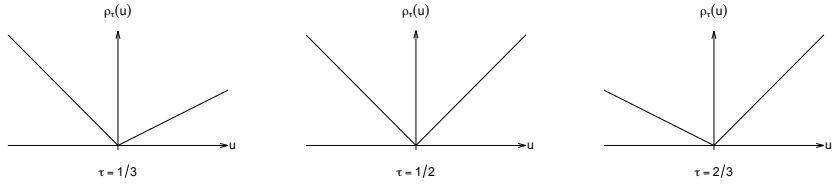


FIGURE 22.2: Check function

y. For a random variable y, we can define its distribution function as $F(c) = \text{pr}(y \leq c)$ and its τth quantile as

$$F^{-1}(\tau) = \inf \{q : F(q) \geq \tau\}.$$

This defines a quantile function $F^{-1} : [0, 1] \to \mathbb{R}$. If the distribution function is strictly monotone, then the quantile function reduces to the inverse of the distribution function, and the τ-th quantile solves $\tau = \text{pr}(y \leq q)$ as an equation of q. See Figure 22.1. For simplicity, this chapter focuses on the case with a monotone distribution function. The definition above formulate the mean as the minimizer of an objective function. Similarly, we can define quantiles in an equivalent way below.

Proposition 22.1 *With a monotone distribution function and positive density at the τth quantile, we have*

$$F^{-1}(\tau) = \arg\min_q E \{\rho_\tau(y - q)\},$$

where

$$\rho_\tau(u) = u \{\tau - 1(u < 0)\} = \begin{cases} u\tau, & \text{if } u \geq 0, \\ -u(1 - \tau), & \text{if } u < 0, \end{cases}$$

is the check function (the name comes from its shape; see Figure 22.2). In particular, the median of y is

$$\text{median}(y) = F^{-1}(0.5) = \arg\min_q E \{|y - q|\}.$$

Proof of Proposition 22.1: To simplify the proof, we further assume that y has density function $f(\cdot)$. We will use Leibniz's integral rule:

$$\frac{\text{d}}{\text{d}x} \left\{ \int_{a(x)}^{b(x)} f(x, t)\text{d}t \right\} = f(x, b(x))b'(x) - f(x, a(x))a'(x) + \int_{a(x)}^{b(x)} \frac{\partial f(x, t)}{\partial x} \text{d}t.$$

We can write

$$E \{\rho_\tau(y - q)\} = \int_{-\infty}^{q} (\tau - 1)(c - q)f(c)\text{d}c + \int_{q}^{\infty} \tau(c - q)f(c)\text{d}c.$$

To minimize it over q, we can solve the first order condition

$$\frac{\partial E\left\{\rho_\tau(y-q)\right\}}{\partial q} = (1-\tau)\int_{-\infty}^{q} f(c)\mathrm{d}c - \tau\int_{q}^{\infty} f(c)\mathrm{d}c = 0.$$

So

$$(1-\tau)\mathrm{pr}(y \leq q) - \tau\left\{1 - \mathrm{pr}(y \leq q)\right\} = 0 \Longrightarrow \tau = \mathrm{pr}(y \leq q),$$

implying that the τth quantile satisfies the first order condition. The second order condition ensures it is the minimizer:

$$\frac{\partial^2 E\left\{\rho_\tau(y-q)\right\}}{\partial q^2}\Big|_{q=F^{-1}(\tau)} = f\left\{F^{-1}(\tau)\right\} > 0.$$

□

The empirical distribution function is $\hat{F}(c) = n^{-1}\sum_{i=1}^{n} 1(y_i \leq c)$, which is a step function, increasing but not strictly monotone. With Proposition 22.1, we can easily define the sample quantile as

$$\hat{F}^{-1}(\tau) = \arg\min_q n^{-1}\sum_{i=1}^{n}\rho_\tau(y_i - q),$$

which may not be unique even though the population quantile is. We can view $\hat{F}^{-1}(\tau)$ as a set containing all minimizers, and with large samples the values in the set do not differ much. Similar to the sample mean, the sample quantile also enjoys a CLT.

Theorem 22.1 *Assume* $(y_i)_{i=1}^n \overset{\text{IID}}{\sim} y$ *with distribution function* $F(\cdot)$ *that is strictly increasing and density function* $f(\cdot)$ *that is positive at the* τ*th quantile. The sample quantile is consistent for the true quantile and is asymptotically Normal:*

$$\sqrt{n}\left\{\hat{F}^{-1}(\tau) - F^{-1}(\tau)\right\} \to \mathrm{N}\left(0, \frac{\tau(1-\tau)}{\left[f\left\{F^{-1}(\tau)\right\}\right]^2}\right)$$

in distribution. In particular, the sample median satisfies

$$\sqrt{n}\left\{\hat{F}^{-1}(0.5) - \mathrm{median}(y)\right\} \to \mathrm{N}\left(0, \frac{1}{4\left[f\left\{\mathrm{median}(y)\right\}\right]^2}\right)$$

in distribution.

Proof of Theorem 22.1: I use the sandwich lemma to give a heuristic proof; see Van der Vaart (2000, chapter 21) for a rigorous proof. Based on the first order condition in Proposition 22.1, the population quantile solves

$$E\{m_\tau(y-q)\} = 0,$$

and the sample quantile solves

$$n^{-1}\sum_{i=1}^{n} m_\tau(y_i - q) = 0,$$

where the check function has partial derivative with respect to u except for the point 0:

$$m_\tau(u) = (\tau - 1)1(u < 0) + \tau 1(u > 0) = \tau - 1(u \leq 0).$$

We only need to find the bread and meat matrices, which are scalars now:

$$\begin{aligned} B &= \frac{\partial E\{m_\tau(y-q)\}}{\partial q}\Big|_{q=F^{-1}(\tau)} \\ &= \frac{\partial E\{\tau - 1(y \leq q)\}}{\partial q}\Big|_{q=F^{-1}(\tau)} \\ &= -\frac{\partial F(q)}{\partial q}\Big|_{q=F^{-1}(\tau)} \\ &= -f\{F^{-1}(\tau)\}, \end{aligned}$$

and

$$\begin{aligned} M &= E\left[\{m_\tau(y-q)\}^2\right]\Big|_{q=F^{-1}(\tau)} \\ &= E\left[\{\tau - 1(y \leq q)\}^2\right]\Big|_{q=F^{-1}(\tau)} \\ &= E\left[\tau^2 + 1(y \leq q) - 21(y \leq q)\tau\right]\Big|_{q=F^{-1}(\tau)} \\ &= \tau^2 + \tau - 2\tau^2 \\ &= \tau(1-\tau). \end{aligned}$$

Therefore, $\sqrt{n}\{\hat{F}^{-1}(\tau) - F^{-1}(\tau)\}$ converges to Normal with mean zero and variance $M/B^2 = \tau(1-\tau)/[f\{F^{-1}(\tau)\}]^2$. □

To conduct statistical inference for the quantile $F^{-1}(\tau)$, we need to estimate the density of y at the τth quantile to obtain the estimated standard error of $\hat{F}^{-1}(\tau)$. Alternatively, we can use the bootstrap to obtain the estimated standard error. We will discuss the inference of quantiles in R in Section 22.4.

22.2 From the conditional mean to the conditional quantile

With an explanatory variable x for outcome y, we can define the conditional mean as

$$E(y \mid x) = \arg\min_{m(\cdot)} E\left[\{y - m(x)\}^2\right].$$

We can use a linear function $x^{\mathrm{T}}\beta$ to approximate the conditional mean with the population OLS coefficient

$$\beta = \arg\min_b E\left\{(y - x^{\mathrm{T}}b)^2\right\} = \left\{E(xx^{\mathrm{T}})\right\}^{-1} E(xy),$$

and the sample OLS coefficient

$$\hat{\beta} = \left(n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{T}}\right)^{-1}\left(n^{-1}\sum_{i=1}^n x_i y_i\right).$$

We have discussed intensively the statistical properties of $\hat{\beta}$. Motivated by Proposition 22.1, we can define the conditional quantile function as

$$F^{-1}(\tau \mid x) = \arg\min_{q(\cdot)} E\left[\rho_\tau\left\{y - q(x)\right\}\right].$$

We can use a linear function $x^{\mathrm{T}}\beta(\tau)$ to approximate the conditional quantile function with

$$\beta(\tau) = \arg\min_b E\left\{\rho_\tau(y - x^{\mathrm{T}}b)\right\}$$

called the τth population regression quantile, and

$$\hat{\beta}(\tau) = \arg\min_b n^{-1}\sum_{i=1}^n \rho_\tau(y_i - x_i^{\mathrm{T}}b) \tag{22.1}$$

called the τth sample regression quantile. As a special case, when $\tau = 0.5$, we have the regression median:

$$\hat{\beta}(0.5) = \arg\min_b n^{-1}\sum_{i=1}^n |y_i - x_i^{\mathrm{T}}b|,$$

which is also called the least absolute deviations (LAD).

Koenker and Bassett Jr (1978) started the literature under a correctly specified conditional quantile model:

$$F^{-1}(\tau \mid x) = x^{\mathrm{T}}\beta(\tau);$$

Angrist et al. (2006) discussed quantile regression under misspecification, viewing it as a best linear approximation to the true conditional quantile function. This chapter will focus on the statistical properties of the sample regression quantiles. I follow Angrist et al. (2006) to discuss statistical inference allowing for misspecification.

Before that, we first comment on the population regression quantiles based on some generative models. Below assume that v_i IID independent of covariates, with mean zero and distribution $g(c) = \mathrm{pr}(v_i \le c)$.

Example 22.1 *Under the linear model* $y_i = x_i^{\mathrm{T}}\beta + \sigma v_i$*, we can verify that*

$$E(y_i \mid x_i) = x_i^{\mathrm{T}}\beta$$

and

$$F^{-1}(\tau \mid x_i) = x_i^{\mathrm{T}}\beta + \sigma g^{-1}(\tau).$$

Therefore, with the first regressor being 1*,we have*

$$\beta_1(\tau) = \beta_1 + \sigma g^{-1}(\tau), \quad \beta_j(\tau) = \beta_j, \quad (j = 2, \ldots, p).$$

In this case, both the true conditional mean and quantile functions are linear, and the population regression quantiles are constant across τ *except for the intercept.*

Example 22.2 *Under a heteroskedastic linear model* $y_i = x_i^{\mathrm{T}}\beta + (x_i^{\mathrm{T}}\gamma)v_i$ *with* $x_i^{\mathrm{T}}\gamma > 0$ *for all* x_i*'s, we can verify that*

$$E(y_i \mid x_i) = x_i^{\mathrm{T}}\beta$$

and

$$F^{-1}(\tau \mid x_i) = x_i^{\mathrm{T}}\beta + x_i^{\mathrm{T}}\gamma g^{-1}(\tau).$$

Therefore,

$$\beta(\tau) = \beta + \gamma g^{-1}(\tau).$$

In this case, both the true conditional quantile functions are linear, and all coordinates of the population regression quantiles vary with τ*.*

Example 22.3 *Under the transformed linear model* $\log y_i = x_i^{\mathrm{T}}\beta + \sigma v_i$*, we can verify that*

$$E(y_i \mid x_i) = \exp(x_i^{\mathrm{T}}\beta) M_v(\sigma),$$

where $M_v(t) = E(e^{tv})$ *is the moment generating function of* v*, and*

$$F^{-1}(\tau \mid x_i) = \exp\left\{x_i^{\mathrm{T}}\beta + \sigma g^{-1}(\tau)\right\}.$$

In this case, both the true conditional mean and quantile functions are log-linear in covariates.

22.3 Sample regression quantiles

22.3.1 Computation

The regression quantiles (22.1) do not have explicit formulas in general, and we need to solve the optimization problem numerically. Motivated by the

piece-wise linear feature of the check function, we decompose $y_i - x_i^{\mathrm{T}}\beta$ into the difference between its positive part and negative part:

$$y_i - x_i^{\mathrm{T}}\beta = u_i - v_i,$$

where

$$u_i = \max(y_i - x_i^{\mathrm{T}}\beta, 0), \quad v_i = -\min(y_i - x_i^{\mathrm{T}}\beta, 0).$$

So the objective function simplifies to the summation of

$$\rho_\tau(y_i - x_i^{\mathrm{T}}\beta) = \tau u_i + (1-\tau)v_i,$$

which is simply a linear function of the u_i's and v_i's. Of course, these u_i's and v_i's are not arbitrary because they must satisfy the constraints by the data. Using the notation

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \quad v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

finding the τth regression quantile is equivalent to a linear programming problem with linear objective function and linear constraints:

$$\begin{aligned} &\min_{b,u,v} \tau 1_n^{\mathrm{T}} u + (1-\tau) 1_n^{\mathrm{T}} v, \\ &\text{s.t. } Y = Xb + u - v, \\ &\text{and } u \succeq 0, v \succeq 0. \end{aligned}$$

The function `rq` in the `R` package `quantreg` computes the regression quantiles with various choices of methods.

22.3.2 Asymptotic inference

Similar to the sample quantiles, the regression quantiles are also consistent for the population regression quantiles and asymptotically Normal. So we can conduct asymptotic inference. We summarize the results in the following theorem (Angrist et al., 2006).

Theorem 22.2 *Assume* $(y_i, x_i)_{i=1}^n \overset{\text{IID}}{\sim} (y, x)$. *Under some regularity conditions, we have*

$$\sqrt{n}\left\{\hat{\beta}(\tau) - \beta(\tau)\right\} \to \mathrm{N}(0, B^{-1}MB^{-1})$$

in distribution, where

$$B = E\left[f_{y|x}\left\{x^{\mathrm{T}}\beta(\tau)\right\} xx^{\mathrm{T}}\right], \quad M = E\left[\left\{\tau - 1\left(y - x^{\mathrm{T}}\beta(\tau) \le 0\right)\right\}^2 xx^{\mathrm{T}}\right],$$

with $f_{y|x}(\cdot)$ *denoting the conditional density of* y *given* x.

Proof of Theorem 22.2: The population regression quantile solves

$$E\left\{m_\tau(y - x^{\mathrm{T}}b)x\right\} = 0,$$

and the sample regression quantile solves

$$n^{-1}\sum_{i=1}^{n} m_\tau(y_i - x_i^{\mathrm{T}}b)x_i = 0.$$

The consistency and asymptotic Normality follows from the sandwich theorem, and we only need to calculate the explicit forms of B and M. Let $F_{y|x}(\cdot)$ and $f_{y|x}(\cdot)$ be the conditional distribution and density functions. We have

$$\begin{aligned} E\left\{m_\tau(y - x^{\mathrm{T}}b)x\right\} &= E\left[\left\{\tau - 1(y - x^{\mathrm{T}}b \le 0)\right\}x\right] \\ &= E\left[\left\{\tau - F_{y|x}(x^{\mathrm{T}}b)\right\}x\right], \end{aligned}$$

so

$$\frac{\partial E\left\{m_\tau(y - x^{\mathrm{T}}b)x\right\}}{\partial b} = -E\left\{f_{y|x}(x^{\mathrm{T}}b)xx^{\mathrm{T}}\right\}.$$

This implies the formula of B. The formula of M follows from

$$\begin{aligned} M &= E\left\{m_\tau^2(y - x^{\mathrm{T}}\beta(\tau))xx^{\mathrm{T}}\right\} \\ &= E\left[\left\{\tau - 1(y - x^{\mathrm{T}}\beta(\tau) \le 0)\right\}^2 xx^{\mathrm{T}}\right]. \end{aligned}$$

□

Based on Theorem 22.2, we can estimate the asymptotic covariance matrix of $\hat{\beta}(\tau)$ by $n^{-1}\hat{B}^{-1}\hat{M}\hat{B}^{-1}$, where

$$\hat{M} = n^{-1}\sum_{i=1}^{n}\left\{\tau - 1\left(y_i - x_i^{\mathrm{T}}\hat{\beta}(\tau) \le 0\right)\right\}^2 x_i x_i^{\mathrm{T}}$$

and

$$\hat{B} = (2nh)^{-1}\sum_{i=1}^{n} 1\left\{|y_i - x_i^{\mathrm{T}}\hat{\beta}(\tau)| \le h\right\}x_i x_i^{\mathrm{T}}$$

for a carefully chosen h. Powell (1991)'s theoretical result suggest to use h satisfying condition $h = O(n^{-1/3})$, and the `quantreg` package in `R` chooses a specific h that satisfies this condition.

22.4 Numerical examples

22.4.1 Sample quantiles

We can use the `quantile` function to obtain the sample quantiles. However, it does not report standard errors. Instead, we can use the `rq` function to

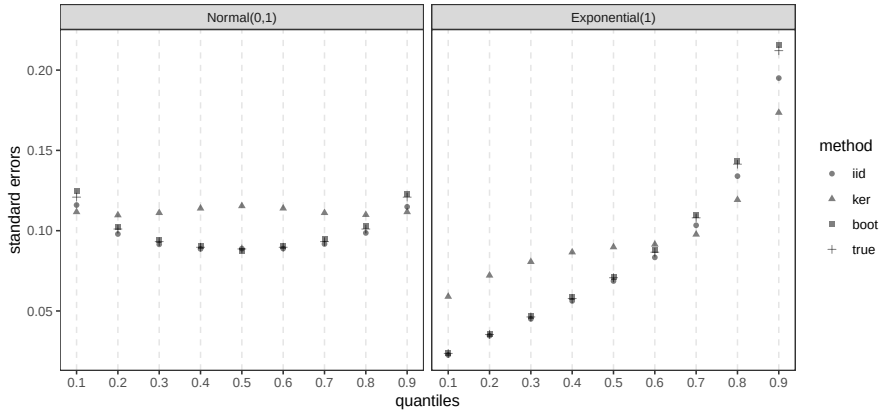


FIGURE 22.3: Standard errors for sample quantiles

compute sample quantiles by regressing the outcome on constant 1. These two functions may return different sample quantiles when they are not unique. The difference is often small with large sample sizes.

I use the following simulation to compare various methods for standard error estimation. The first data generating process has a standard Normal outcome.

```
library(quantreg)
mc= 2000
n = 200
taus = (1:9)/10
get.se = function(x){x$coef[1,2]}
q.normal = replicate(mc,{
  y  = rnorm(n)
  qy = rq(y~1, tau = taus)
  se.iid = summary(qy, se = "iid")
  se.ker = summary(qy, se = "ker")
  se.boot= summary(qy, se = "boot")

  qy = qy$coef
  se.iid = sapply(se.iid, get.se)
  se.ker = sapply(se.ker, get.se)
  se.boot= sapply(se.boot, get.se)

  c(qy, se.iid, se.ker, se.boot)
})
```

I also run the same simulation but replace the Normal outcome by Exponential: `y = rexp(n)`. Figure 22.3 compares the estimated standard errors with the true asymptotic standard error in Theorem 22.1. Bootstrap works the best, and the one involving kernel estimation of the density seems biased.

22.4.2 OLS versus LAD

I will use simulation to compare OLS and LAD. In `rq`, the default value is `tau=0.5`, fitting the LAD. The first data generating process is a Gaussian linear model:

```
x = rnorm(n)
simu.normal = replicate(mc, {
  y = 1 + x + rnorm(n)
  c(lm(y~x)$coef[2], rq(y~x)$coef[2])
})
```

The second data generating process replaces the error term to a Laplace distribution:

```
simu.laplace = replicate(mc, {
  y = 1 + x + rexp(n) - rexp(n)
  c(lm(y~x)$coef[2], rq(y~x)$coef[2])
})
```

Note that the difference between two independent Exponential has the same distribution as Laplace; I leave this as a homework problem. OLS is the MLE under a Gaussian linear model, and LAD is the MLE under a linear model with independent Laplace errors.

The third data generating process replaces the error term to standard Exponential:

```
simu.exp = replicate(mc, {
  y = 1 + x + rexp(n)
  c(lm(y~x)$coef[2], rq(y~x)$coef[2])
})
```

The fourth data generating process has $y_i = 1 + e_i x_i$ with e_i IID Exponential, so

$$E(y_i \mid x_i) = 1 + x_i, \quad \text{var}(y_i \mid x_i) = x_i^2,$$

which is a heteroskedastic linear model, and

$$F^{-1}(0.5 \mid x_i) = 1 + \text{median}(e_i)x_i = 1 + (\log 2)x_i,$$

which is a linear quantile model. The coefficients are different in the conditional mean and quantile functions.

```
x = abs(x)
simu.x = replicate(mc, {
  y = 1 + rexp(n)*x
  c(lm(y~x)$coef[2], rq(y~x)$coef[2])
})
```

Figure 22.4 compares OLS and LAD under the above four data generating processes. With Normal errors, OLS is more efficient; with Laplace errors, LAD is more efficient. This confirms the theory of MLE. With Exponential errors, LAD is also more efficient that OLS. Under the four data generating process, LAD is more efficient than OLS, but these two estimators target different parameters which make the comparison of standard errors not very meaningful.

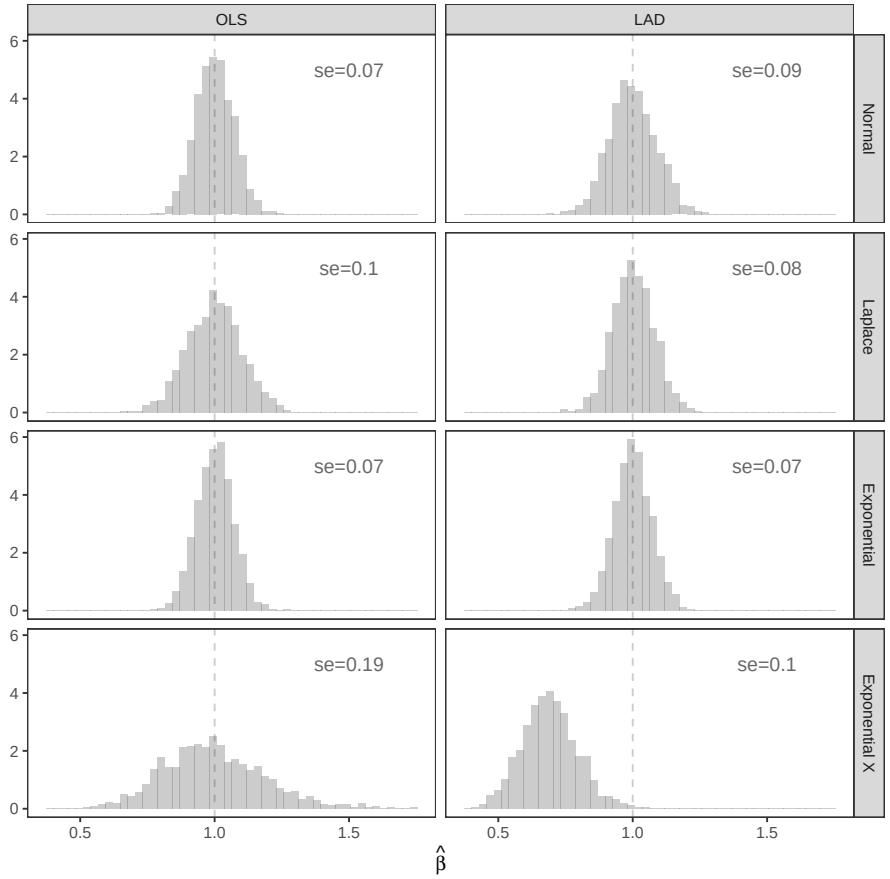


FIGURE 22.4: Regression quantiles

22.5 Application

22.5.1 Parents' and children's heights

I revisit Galton's data that were used at the beginning of this course. The following code gives the coefficients for quantiles 0.1 to 0.9.

```
> library("HistData")
> taus      = (1:9)/10
> qr.galton = rq(childHeight ~ midparentHeight,
+                tau = taus,
+                data = GaltonFamilies)
> coef.galton = qr.galton$coef
```

Figure 22.5 shows the quantile regression lines, which are almost parallel

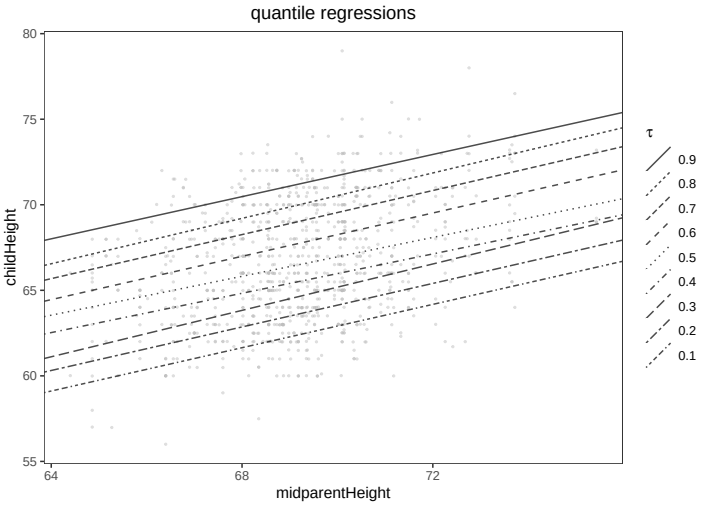


FIGURE 22.5: Galton's data

with different intercepts. In Galton's data, x and y are very close to a bivariate Normal distribution. Theoretically, we can verify that with bivariate Normal (x, y), the conditional quantile function $F^{-1}(\tau \mid x)$ is linear in x with the same slope. See a homework problem.

22.5.2 U.S. wage structure

Angrist et al. (2006) used quantile regression to study the U.S. wage structure. They used census data in 1980, 1990 and 2000 to fit quantile regressions on log weekly wage on education and other variables. The following code gives the coefficients for quantile regressions with τ equaling 0.1 to 0.9. I repeated the regressions with data from three years. Due to the large sample size, I use the m-of-n bootstrap with $m = 500$.

```
> library(foreign)
> census80 = read.dta("census80.dta")
> census90 = read.dta("census90.dta")
> census00 = read.dta("census00.dta")
> f.reg = logwk ~ educ + exper + exper2 + black
>
> m.boot   = 500
> rq80     = rq(f.reg, data = census80, tau = taus)
> rqlist80 = summary(rq80, se = "boot",
+                    bsmethod= "xy", mofn = m.boot)
> rq90     = rq(f.reg, data = census90, tau = taus)
> rqlist90 = summary(rq90, se = "boot",
+                    bsmethod= "xy", mofn = m.boot)
> rq00     = rq(f.reg, data = census00, tau = taus)
> rqlist00 = summary(rq00, se = "boot",
```

```
+                       bsmethod= "xy", mofn = m.boot)
```

Figure 22.6 shows the coefficient of `educ` across years and across quantiles. In 1980, the coefficients are nearly constant cross quantiles, showing no evidence of heterogeneity of the return of education. In 1990, the return of education increases across all quantiles, but it increases more at the upper quantiles. In 2000, the return of education decreases at the lower quantile and increases at the upper quantiles, showing more dramatic heterogeneity across quantiles.

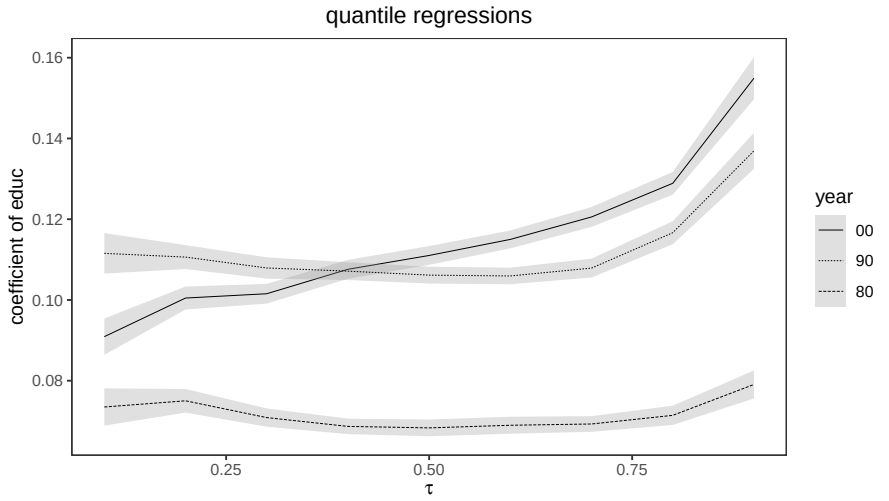


FIGURE 22.6: Angrist et al. (2006)'s data

22.6 Homework problems

22.1 Exponential mean and median

If y is an Exponential(λ) random variable with density $\lambda e^{-\lambda y}$ for $y > 0$ with $\lambda > 0$, find its mean and median.

22.2 Laplace as difference of two independent Exponentials

Assume that y_1 and y_2 are two IID Exponential(λ). Show that $y = y_1 - y_2$ is Laplace with density

$$\frac{\lambda}{2} \exp(-\lambda |c|), \quad -\infty < c < \infty.$$

22.3 Quantile regression with a binary regressor

For $i = 1, \ldots, n$, the first 1/3 observations have $x_i = 1$ and the last 2/3 observations have $x_i = 0$; $y_i \mid x_i = 1$ follows an Exponential(1), and $y_i \mid x_i = 0$ follows an Exponential(2). Find

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

and the joint asymptotic distribution.

22.4 Interquartile range and estimation

The interquartile range of a random variable y equals the difference between its 75% and 25% quantiles. Based on IID data $(y_i)_{i=1}^n$, write a function to estimate the interquartile range and the corresponding standard error using the bootstrap. Use simulated data to evaluate the finite sample properties of the point estimate (e.g., bias and variance) and the 95% confidence interval (e.g. coverage rate and length).

22.5 Conditional quantile function in bivariate Normal

Show that if (x, y) follows a bivariate Normal, the conditional quantile function of y given x is linear in x with the same slope across all τ.

22.6 Quantile range and variance

A symmetric random variable y satisfies $y \sim -y$. Define the $1 - \alpha$ quantile range of a symmetric random variable y as the interval of its $\alpha/2$ and $1 - \alpha/2$ quantiles. Given two symmetric random variables y_1 and y_2, show that if the $1 - \alpha$ quantile range of y_1 is wider than that of y_2 for all α, then $\text{var}(y_1) \geq \text{var}(y_2)$. Does the converse of the statement hold? If so, give a proof; if not, give a counterexample.

22.7 Weighted quantile regression and application

Many real data contain weights due to sampling. For example, in Angrist et al. (2006)'s data, `perwt` is the sampling weight. Define the weighted quantile regression problem theoretically and re-analyze Angrist et al. (2006)'s data with weights. Note that similar to `lm` and `glm`, the quantile regression function `rq` also has a parameter `weights`.

23

Modeling Time-to-Event Data

23.1 Examples

Time-to-event data are common in biomedical and social sciences. Statistical analysis of time-to-event data is called survival analysis in biostatistics and duration analysis in econometrics. The former name comes from the biomedical applications where the outcome denotes the survival time or the recurrent of certain disease. The latter name comes from the economic applications where the outcome denotes the weeks unemployed or days until arrest after incarceration. See Kalbfleisch and Prentice (2011) for more biomedical applications and Heckman and Singer (1984) for economic applications. Freedman (2008) gave a concise and critical introduction to survival analysis.

23.1.1 Survival analysis

The Combined Pharmacotherapies and Behavioral Interventions study evaluated the efficacy of medication, behavioral therapy, and their combination for treatment of alcohol dependence (Anton et al., 2006). Between January 2001 and January 2004, $n = 1224$ recently alcohol-abstinent volunteers were randomized to receive medical management with 16 weeks of naltrexone (100mg daily) or placebo, with or without a combined behavioral intervention. It was a 2×2 factorial experiment. The outcome of interest is the time to the first day of heavy drinking and other endpoints. I adopt the data from Lin et al. (2016).

```
> COMBINE = read.table("combine_data.txt", header = TRUE)[, -1]
> head(COMBINE)
  AGE GENDER    T0_PDA NALTREXONE THERAPY   site relapse futime
1  31   male  3.333333          1       0 site_0       0    112
2  41 female 16.666667          1       1 site_0       1      8
3  44   male 73.333333          0       1 site_0       1     20
4  65   male 10.000000          1       0 site_0       0    112
5  39   male  0.000000          0       1 site_0       1      4
6  56   male 13.333333          0       0 site_0       1      1
```

`NALTREXONE` and `THERAPY` are two treatment indicators. `futime` is the follow-up time, which is censored if `relapse` equals 0. For those censored observations, `futime` equals 112, so it seems administrative censoring. Figure 23.1 shows the

histograms of `futime` in four treatment groups. A larger number of patients have censored outcomes. Other variables are covariates.

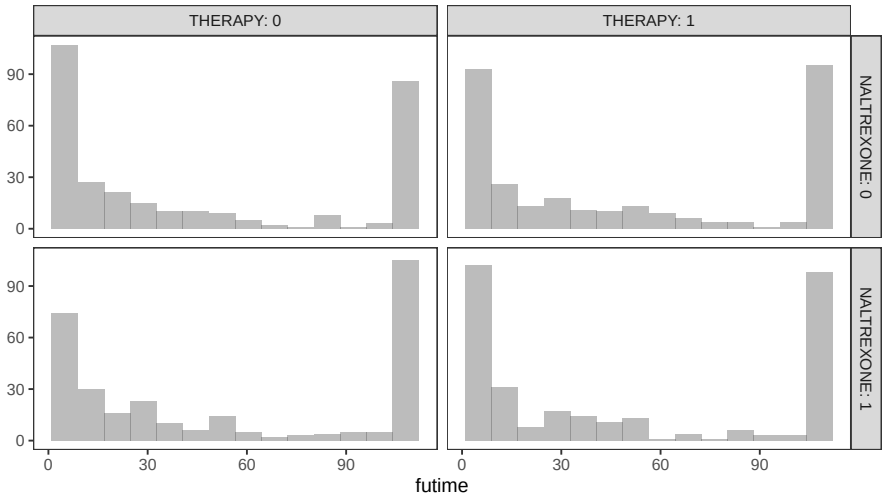


FIGURE 23.1: Histograms of the time-to-event in the data from Lin et al. (2016)

23.1.2 Duration analysis

Carpenter (2002) asked the question: Why does the U.S. Food and Drug Administration approve some drugs more quickly than other? With data of 450 drugs reviewed from 1977 to 2000, he studied the dependence of review times on various covariates, including political influence, wealth of the richest organization representing the disease, media coverage, etc. I use the version of data analyzed in Keele (2010). The outcome `acttime` is censored indicated by `censor`. The original paper contains more detailed explanations of the variables.

```
> fda <- read.dta("fda.dta")
> names(fda)
 [1] "acttime"  "censor"   "hcomm"    "hfloor"   "scomm"
 [6] "sfloor"   "prespart" "demhsmaj" "demsnmaj" "orderent"
[11] "stafcder" "prevgenx" "lethal"   "deathrt1" "hosp01"
[16] "hospdisc" "hhosleng" "acutediz" "orphdum"  "mandiz01"
[21] "femdiz01" "peddiz01" "natreg"   "natregsq" "wpnoavg3"
[26] "vandavg3" "condavg3" "_st"      "_d"       "_t"
[31] "_t0"      "caseid"
```

An obvious feature of time-to-event data is that the outcome is non-negative. This can be easily dealt with by the log transformation. However, the outcomes may be censored, resulting in inadequate tail information. So modeling the mean can be challenging because it often involves extrapolation in the right tail.

23.2 Time-to-event data

Let $T \geq 0$ denote the outcome of interest. We can characterize a non-negative continuous T using its density $f(t)$, distribution function $F(t)$, survival function $S(t) = 1 - F(t) = \text{pr}(T > t)$, and hazard function

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \text{pr}(t \leq T < t + \Delta t \mid T \geq t)/\Delta t.$$

Within a small time interval $[t, t + \Delta t]$, we have approximation

$$\text{pr}(t \leq T < t + \Delta t \mid T \geq t) \cong \lambda(t)\Delta t,$$

so the hazard function denotes the death rate within a small interval conditioning on surviving up to time t. The hazard function is commonly used to describe a positive random variable, which can determine the whole distribution.

Proposition 23.1 *For a non-negative continuous random variable T,*

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\text{d}}{\text{d}t} \log S(t),$$
$$S(t) = \exp\left\{ -\int_0^t \lambda(s)\text{d}s \right\}.$$

Proof of Proposition 23.1: By definition,

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{\text{pr}(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{\text{pr}(T \geq t)} = \lim_{\Delta t \downarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} = \frac{f(t)}{S(t)}.$$

We can further write the above equation as

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\text{d}S(t)/\text{d}t}{S(t)} = -\frac{\text{d}}{\text{d}t} \log S(t),$$

so using the Newton–Leibniz formula we have

$$\text{d} \log S(t) = -\lambda(t)\text{d}t \Longrightarrow \log S(t) = -\int_0^t \lambda(s)\text{d}s + \alpha.$$

The constant α must be zero because $\log S(0) = 0$. So $\log S(t) = -\int_0^t \lambda(s)\text{d}s$, giving the final result. □

Example 23.1 (Exponential) *An Exponential(λ) random variable T has density $f(t) = \lambda e^{-\lambda t}$, survival function $S(t) = e^{-\lambda t}$, and constant hazard function $\lambda(t) = \lambda$. An important feature of an Exponential random variable is its memoryless property:*

$$\text{pr}(T \geq t + c \mid T \geq c) = \frac{\text{pr}(T \geq t + c)}{\text{pr}(T \geq c)} = \frac{e^{-\lambda(t+c)}}{e^{-\lambda c}} = e^{-\lambda t} = \text{pr}(T \geq t),$$

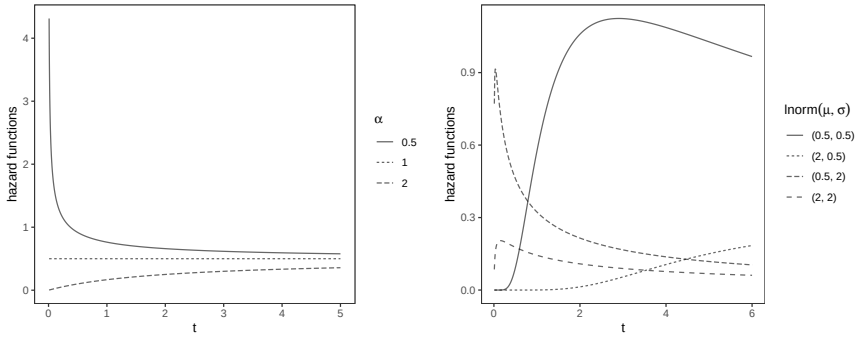


FIGURE 23.2: Left: Gamma(α, β) hazard functions; Right: Log-Normal(μ, σ^2) hazard functions

that is, the probability of surviving another t time is always the same no matter how long the existing survival time is.

Example 23.2 (Gamma) *An Gamma(α, β) random variable T has density $f(t) = \beta^\alpha t^{\alpha-1} e^{-\beta t}/\Gamma(\alpha)$. When $\alpha = 1$, it reduces to Exponential(β) with a constant hazard function. In general, the survival function and hazard function do not have simple forms, but we can use* `dgamma` *and* `pgamma` *to compute them numerically. The left panel of Figure 23.2 plots the hazard functions of Gamma(α, β). When $\alpha < 1$, the hazard function is decreasing; when $\alpha > 1$, the hazard function is increasing.*

Example 23.3 (Log-Normal) *A Log-Normal random variable T equals exponential of $\mathrm{N}(\mu, \sigma^2)$. The right panel of Figure 23.2 plots the hazard functions with four different parameter combinations.*

Example 23.4 (Weibull) *Weibull distribution has many different parametrizations. Here I follow the* `R` *function* `dweibull`*, which has a* `shape` *parameter $a > 0$ and* `scale` *parameter $b > 0$ A Weibull(a, b) random variable T can be generated by*

$$T = bZ^{1/a}, \tag{23.1}$$

where Z is Exponential(1). We can verify that T has density function

$$f(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{a-1} \exp\left\{-\left(\frac{t}{b}\right)^a\right\},$$

survival function

$$S(t) = 1 - \exp\left\{-\left(\frac{t}{b}\right)^a\right\},$$

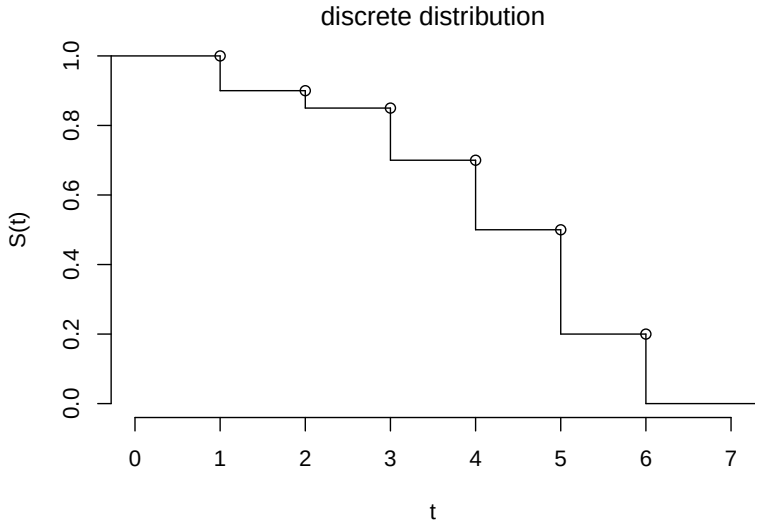


FIGURE 23.3: Discrete survival function with masses $(0.1, 0.05, 0.15, 0.2, 0.3, 0.2)$ at $(1, 2, 3, 4, 5, 6)$

and hazard function

$$\lambda(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{a-1}.$$

So when $a = 1$, Weibull reduces to Exponential with constant hazard function. When $a > 1$, the hazard function increases; when $a < 1$, the hazard function decreases.

We can characterize a discrete positive random variable $T \in \{t_1, t_2, \ldots\}$ by its probability mass function $f(t_k) = \text{pr}(T = t_k)$, distribution function $F(t) = \sum_{k:t_k \le t} f(t_k)$, survival function $S(t) = \sum_{k:t_k > t} f(t_k)$, and discrete hazard function

$$\lambda_k = \text{pr}(T = t_k \mid T \ge t_k) = \frac{f(t_k)}{S(t_k-)},$$

where $S(t_k-)$ denotes the left limit of the function $S(t)$ at t_k. Figure 23.3 shows an example of survival function for a discrete random variable, which shows that $S(t)$ is a step function and right-continuous with left limits.

The discrete hazard and survival function have the following connection which will be useful for the next section.

Proposition 23.2 *For a non-negative discrete random variable T, its survival function is a step function determined by*

$$S(t) = \text{pr}(T > t) = \prod_{k:t_k \le t} (1 - \lambda_k).$$

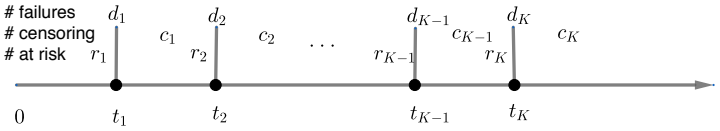


FIGURE 23.4: Data structure for the Kaplan–Meier curve

Note that $S(t)$ is a step function decreasing at each t_k because λ_k is probability and thus bounded between zero and one.

Proof of Proposition 23.2: By definition,

$$1 - \lambda_k = 1 - \text{pr}(T = t_k \mid T \geq t_k) = \text{pr}(T > t_k \mid T \geq t_k)$$

is the probability of surviving longer than t_k conditional on surviving at least as long as t_k. We can verify Proposition 23.2 within each interval of the t_k's. For example, if $t < t_1$, then $S(t) = \text{pr}(T > t) = 1$; if $t_1 \leq t < t_2$, then $S(t) = \text{pr}(T > t_1) = \text{pr}(T > t_1 \mid T \geq t_1) = 1 - \lambda_1$; if $t_2 \leq t < t_3$, then

$$S(t) = \text{pr}(T > t_2) = \text{pr}(T > t_2 \mid T \geq t_2)\text{pr}(T > t_1) = (1 - \lambda_2)(1 - \lambda_1).$$

We can also verify other values of $S(t)$ by induction. □

23.3 Kaplan–Meier survival curve

Figure 23.4 shows the common data structure in survival analysis:

(S1) $t_1, \ldots, t_K$ are the death times, and $d_1, \ldots, d_K$ are the corresponding number of deaths;

(S2) $r_1, \ldots, r_K$ are the number of patients at risk, that is, r_1 patients are not dead or censored right before time t_1, and so on;

(S3) $c_1, \ldots, c_K$ are the number of censored patients within interval $[t_1, t_2], \ldots, [d_K, \infty)$.

Kaplan and Meier (1958) proposed the following simple estimator for the survival function: estimate the discrete hazard function at the failure times $\{t_1, \ldots, t_K\}$ as $\hat{\lambda}_k = d_k / r_k$ $(k = 1, \ldots, K)$ and the survival function as

$$\hat{S}(t) = \prod_{k: t_k \leq t} (1 - \hat{\lambda}_k),$$

which is called the Kaplan–Meier or the product-limit estimator of the survival function.

At each failure time t_k, we view d_k as the result of r_k Bernoulli trials with probability λ_k. So $\hat{\lambda}_k = d_k/r_k$ has variance $\lambda_k(1-\lambda_k)/r_k$ which can be estimated by

$$\hat{\text{var}}(\hat{\lambda}_k) = \hat{\lambda}_k(1-\hat{\lambda}_k)/r_k.$$

We can estimate the variance of the survival function using the delta method. We can approximate the variance of

$$\log \hat{S}(t) = \sum_{k:t_k \le t} \log(1-\hat{\lambda}_k) \cong \sum_{k:t_k \le t} \log(1-\lambda_k) - \sum_{k:t_k \le t} (1-\lambda_k)^{-1}(\hat{\lambda}_k - \lambda_k)$$

by

$$\begin{aligned} \hat{\text{var}}\left\{\log \hat{S}(t)\right\} &= \sum_{k:t_k \le t} (1-\lambda_k)^{-2} \hat{\text{var}}(\hat{\lambda}_k) \\ &= \sum_{k:t_k \le t} (1-\hat{\lambda}_k)^{-2} \hat{\lambda}_k(1-\hat{\lambda}_k)/r_k \\ &= \sum_{k:t_k \le t} \frac{d_k}{r_k(r_k-d_k)}, \end{aligned}$$

which is called Greenwood's formula. A hidden assumption above is the independence of the $\hat{\lambda}_k$'s. This assumption cannot be justified. However, deeper theory of counting processes shows that Greenwood's formula holds even without the independence (Fleming and Harrington, 2011).

Based on Greenwood's formula, we can construct a confidence interval for $\log S(t)$:

$$\log \hat{S}(t) \pm z_\alpha \sqrt{\hat{\text{var}}\left\{\log \hat{S}(t)\right\}},$$

which implies a confidence interval for $S(t)$. However, this interval can be outside of range $[0,1]$. A better transformation is log-log:

$$v(t) = \log\{-\log S(t)\}, \quad \hat{v}(t) = \log\left\{-\log \hat{S}(t)\right\}.$$

We can approximate the variance of

$$\hat{v}(t) \cong \log\{-\log S(t)\} - \frac{1}{\log S(t)}\left\{\log \hat{S}(t) - \log S(t)\right\}$$

by

$$\frac{\hat{\text{var}}\left\{\log \hat{S}(t)\right\}}{\left\{\log \hat{S}(t)\right\}^2}.$$

Based on this formula and Greenwood's formula above, we can construct a confidence interval for $v(t)$:

$$\log\left\{-\log \hat{S}(t)\right\} \pm z_\alpha \sqrt{\hat{\text{var}}\left\{\log \hat{S}(t)\right\}} / \log \hat{S}(t),$$

which implies another confidence interval for $S(t)$.

The `survfit` function in the `R` package `survival` can fit Kaplan–Meier curves. Figure 23.5 plots four curves based on the combination of `NALTREXONE` and `THERAPY` using the data of Lin et al. (2016). I do not show the confidence intervals due to the large overlap.

```
> km4groups = survfit(Surv(futime, relapse) ~ NALTREXONE + THERAPY,
+                     data = COMBINE)
> plot(km4groups, bty = "n", col = 1:4,
+      xlab = "t", ylab = "survival␣functions")
> legend("topright",
+        c("NALTREXONE=0,␣THERAPY=0",
+          "NALTREXONE=0,␣THERAPY=1",
+          "NALTREXONE=1,␣THERAPY=0",
+          "NALTREXONE=1,␣THERAPY=1"),
+        col = 1:4, lty = 1, bty = "n")
```

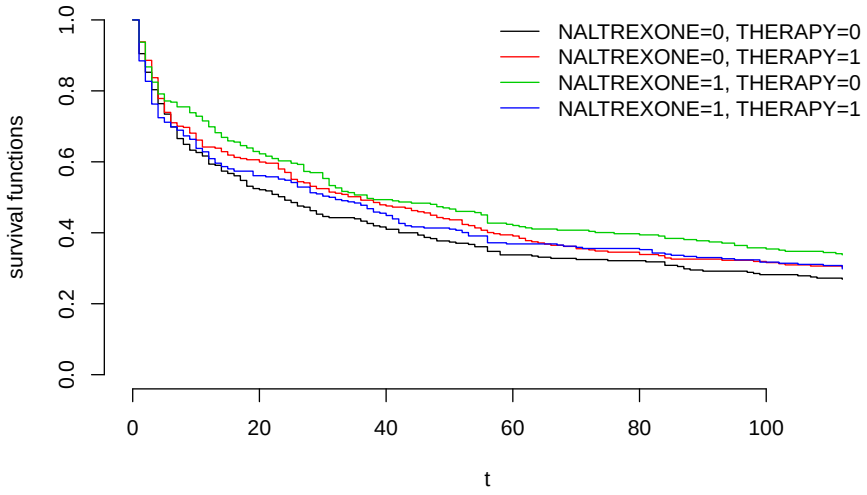


FIGURE 23.5: Lin et al. (2016)'s data

The above discussion on the Kaplan–Meier curve is rather heuristic. More fundamentally, what is the underlying censoring mechanism that ensures the possibility that the distribution of the survival time can be recovered by the observed data? It turns out that we have implicitly assumed that the survival time and the censoring time are independent. Homework problem 23.1 gives a theoretical statement.

23.4 Cox model for time-to-event outcome

23.4.1 Modeling and interpretation

Another important problem to model the dependence of the survival time T on covariates x. The major challenge is that the survival time is often censored. Let C_i be the censoring time of unit i, and we can only observe the minimum value of the survival time and the censoring time. So the observed data are $(x_i, y_i, \delta_i)_{i=1}^n$, where

$$y_i = \min(T_i, C_i), \quad \delta_i = 1(T_i \leq C_i)$$

are the event time and the censoring indicator, respectively. A key assumption is that the censoring mechanism is noninformative:

Assumption 23.1 (noninformative censoring) $T_i \perp\!\!\!\perp C_i \mid x_i$.

Cox (1972) proposed to model on the conditional hazard function

$$\lambda(t \mid x) = \lim_{\Delta t \downarrow 0} \text{pr}(t \leq T < t + \Delta t \mid T \geq t, x)/\Delta t = \frac{f(t \mid x)}{S(t \mid x)},$$

with the canonical specification

$$\lambda(t \mid x) = \lambda_0(t) \exp(x^{\mathrm{T}}\beta), \tag{23.2}$$

or

$$\log \lambda(t \mid x) = \log \lambda_0(t) + x^{\mathrm{T}}\beta.$$

Unlike other regression models, x does not contain the intercept in (23.2). If the first component of x is 1, then we can write

$$\lambda(t \mid x) = \lambda_0(t) \exp(x_1\beta_1 + \cdots + x_1\beta_p) = \lambda_0(t)e^{\beta_1} \exp(x_2\beta_2 + \cdots + x_1\beta_p)$$

and redefine $\lambda_0(t)e^{\beta_1}$ as another unknown function. With an intercept, we cannot identify $\lambda_0(t)$ and β_1 separately. So we drop the intercept to avoid the identifiability issue.

From the log linear form of the conditional hazard function, we have

$$\log \frac{\lambda(t \mid x')}{\lambda(t \mid x)} = (x' - x)^{\mathrm{T}}\beta,$$

so each coordinate of β measures the log conditional hazard ratio holding other covariates constant. Because of this, (23.2) is called the proportional hazards model. A positive β_j suggests a "positive" effect on the hazard function and thus a "negative" effect on the survival time itself. Consider a special case

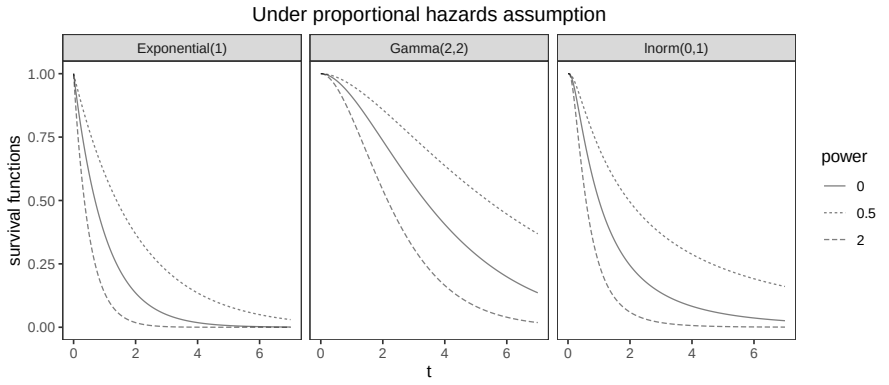


FIGURE 23.6: Proportional hazards assumption with different baseline survival functions

with a binary x_i, then the proportional hazards assumption implies that $\lambda(t \mid 1) = \gamma\lambda(t \mid 0)$ with $\gamma = \exp(\beta)$, and therefore the survival functions satisfy

$$S(t \mid 1) = \exp\left\{-\int_0^t \lambda(u \mid 1)\mathrm{d}u\right\} = \exp\left\{-\gamma\int_0^t \lambda(u \mid 0)\mathrm{d}u\right\} = \{S(t \mid 0)\}^\gamma,$$

which is a power transformation. Qualitatively, we have the following two cases:

(PH1) $\beta < 0$: so the hazard ratio $\gamma = \exp(\beta) < 1$ and $S(t \mid 1) \geq S(t \mid 0)$ for all t, which implies longer survival time under treatment;

(PH2) $\beta > 0$: so the hazard ratio $\gamma = \exp(\beta) > 1$ and $S(t \mid 1) \leq S(t \mid 0)$ for all t, which implies shorter survival time under treatment.

Figure 23.6 shows some survival functions satisfying the proportional hazards assumption, none of which cross each other within the interval $t \in (0, \infty)$. When the two survival functions cross, the proportional hazards assumption does not hold.

Theoretically, we can allow the covariates to be time-dependent, that is, $x_i(t)$ can depend on t and thus is a stochastic process. However, the interpretation of the coefficient becomes challenging (Fisher and Lin, 1999). This chapter focuses on the simple case with time-invariant covariates.

23.4.2 Partial likelihood

The likelihood function is rather complicated, which depends on an unknown function $\lambda_0(t)$. Assuming no ties, Cox (1972) proposed to use the partial like-

lihood function to estimate β:

$$L(\beta) = \prod_{k=1}^{K} \frac{\exp(x_k^{\mathrm{T}}\beta)}{\sum_{l \in R(t_k)} \exp(x_l^{\mathrm{T}}\beta)},$$

where x_k is the covariate value of the failure at time t_k, and $R(t_k)$ contains the indices of the units at risk at time t_k, i.e., the units not censored or failed at time t_k.

Freedman (2008) gives a heuristic explanation of the partial likelihood based on the following results.

Theorem 23.1 *If $T_1, \ldots, T_n$ are independent with hazard function $\lambda_i(t)$, then their minimum value $\underline{T} = \min_{1 \le i \le n} T_i$ has hazard function $\sum_{i=1}^{n} \lambda_i(t)$. Moreover, if $\lambda_i(t) = c_i\lambda(t)$, then*

$$\mathrm{pr}(T_i = \underline{T}) = \frac{c_i}{\sum_{i'=1}^{n} c_{i'}}.$$

Proof of Theorem 23.1: The survival function of $\underline{T}$ is

$$\mathrm{pr}(\underline{T} > t) = \mathrm{pr}(T_1 > t, \ldots, T_n > t) = \prod_{i=1}^{n} S_i(t),$$

so Proposition 23.1 implies that its hazard function is

$$-\frac{\mathrm{d}}{\mathrm{d}t} \log \mathrm{pr}(\underline{T} > t) = \sum_{i=1}^{n} \left\{ -\frac{\mathrm{d}}{\mathrm{d}t} \log S_i(t) \right\} = \sum_{i=1}^{n} \lambda_i(t).$$

So the first conclusion follows.

As a by product of the above proof, the density of $\underline{T}$ is $\sum_{i=1}^{n} \lambda_i(t) \prod_{i=1}^{n} S_i(t)$ based in Proposition 23.1. It must have integral one; with $\lambda_i(t) = c_i\lambda(t)$, this implies

$$\left(\sum_{i=1}^{n} c_i \right) \int_0^\infty \lambda(t) \prod_{i=1}^{n} S_i(t)\mathrm{d}t = 1. \tag{23.3}$$

Therefore, we have

$$\begin{aligned}
\mathrm{pr}(T_i = \underline{T}) &= \mathrm{pr}\{T_i \le T_{i'} \text{ for all } i' \ne i\} \\
&= \int_0^\infty \prod_{i' \ne i} S_{i'}(t) f_i(t)\mathrm{d}t \\
&= \int_0^\infty \prod_{i'=1}^{n} S_{i'}(t) \lambda_i(t)\mathrm{d}t \\
&= c_i \int_0^\infty \lambda(t) \prod_{i'=1}^{n} S_{i'}(t)\mathrm{d}t \\
&= c_i \Big/ \sum_{i=1}^{n} c_{i'},
\end{aligned}$$

where the last equality holds due to (23.3). □

Theorem 23.1 explains each of the K components in the partial likelihood. At time t_k, the units in $R(t_k)$ are all at risk, and unit k fails, assuming no ties. The probability that unit k has the smallest failure time among units in $R(t_k)$ is

$$\frac{\exp(x_k^{\mathrm{T}}\beta)}{\sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta)}$$

from Theorem 23.1. The product in the partial likelihood is based on the independence of the events at the K failure times, which is more difficult to justify. A full justification relies on deeper theory of counting processes (Fleming and Harrington, 2011) or semiparametric statistics (Tsiatis, 2007).

The log-likelihood function is

$$\log L(\beta) = \sum_{k=1}^{K}\left\{x_k^{\mathrm{T}}\beta - \log\sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta)\right\},$$

and the score function is

$$\frac{\partial \log L(\beta)}{\partial\beta} = \sum_{k=1}^{K}\left\{x_k - \frac{\sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta)x_l}{\sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta)}\right\}.$$

Define

$$\pi_\beta(l \mid R_k) = \exp(x_l^{\mathrm{T}}\beta)/\sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta), \quad (l\in R(t_k))$$

which sum to one, so they induce a probability measure leading to expectation $E_\beta(\cdot \mid R_k)$ and covariance $\mathrm{cov}_\beta(\cdot \mid R_k)$. With this notation, the score function simplifies to

$$\frac{\partial \log L(\beta)}{\partial\beta} = \sum_{k=1}^{K}\{x_k - E_\beta(x \mid R_k)\},$$

where $E_\beta(x \mid R_k) = \sum_{l\in R(t_k)}\pi_l(\beta \mid R_k)x_l$; the Hessian matrix simplifies to

$$\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}} = -\sum_{k=1}^{K}\mathrm{cov}_\beta(x \mid R_k) \preceq 0,$$

where $\mathrm{cov}_\beta(x \mid R_k)$ equals

$$\frac{\sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta)x_l x_l^{\mathrm{T}}\sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta) - \sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta)x_l\sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta)x_l^{\mathrm{T}}}{\left\{\sum_{l\in R(t_k)}\exp(x_l^{\mathrm{T}}\beta)\right\}^2}$$
$$= \sum_{l\in R(t_k)}\pi_\beta(l \mid R_k)x_l x_l^{\mathrm{T}} - \sum_{l\in R(t_k)}\pi_\beta(l \mid R_k)x_l\sum_{l\in R(t_k)}\pi_\beta(l \mid R_k)x_l^{\mathrm{T}}.$$

The `coxph` function in the `R` package `survival` uses Newton's method to compute the maximizer $\hat\beta$ of the partial likelihoof function, and uses the inverse of the observed Fisher information to approximate its asymptotic variance.

23.4.3 Examples

Using Lin et al. (2016)'s data, we have the following results.

```
> cox.fit <- coxph(Surv(futime, relapse) ~ NALTREXONE*THERAPY +
+                    AGE + GENDER + TO_PDA + site,
+                  data=COMBINE)
> summary(cox.fit)
Call:
coxph(formula = Surv(futime, relapse) ~ NALTREXONE * THERAPY +
    AGE + GENDER + TO_PDA + site, data = COMBINE)

  n= 1226, number of events= 856

                       coef exp(coef)  se(coef)      z Pr(>|z|)
NALTREXONE         -0.249719  0.779020  0.097690 -2.556  0.01058 *
THERAPY            -0.167050  0.846158  0.096102 -1.738  0.08217 .
AGE                -0.015540  0.984580  0.003559 -4.366 1.27e-05 ***
GENDERmale         -0.140621  0.868818  0.075368 -1.866  0.06207 .
TO_PDA              0.002550  1.002553  0.001368  1.863  0.06242 .
sitesite_1         -0.091853  0.912239  0.167261 -0.549  0.58290
sitesite_10        -0.227185  0.796774  0.175427 -1.295  0.19531
sitesite_2          0.121236  1.128892  0.160052  0.757  0.44876
sitesite_3         -0.084483  0.918987  0.161121 -0.524  0.60004
sitesite_4         -0.471612  0.623996  0.175203 -2.692  0.00711 **
sitesite_5         -0.128286  0.879602  0.161782 -0.793  0.42780
sitesite_6         -0.240563  0.786185  0.161958 -1.485  0.13745
sitesite_7          0.372004  1.450639  0.157616  2.360  0.01827 *
sitesite_8          0.067700  1.070045  0.160876  0.421  0.67388
sitesite_9          0.267373  1.306528  0.154911  1.726  0.08435 .
NALTREXONE:THERAPY  0.337539  1.401495  0.137441  2.456  0.01405 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

`NALTREXONE` has a significant negative log hazard ratio, but `THERAPY` has a nonsignificant negative log hazard ratio. More interestingly, their interaction `NALTREXONE:THERAPY` has a significant positive log hazard ratio. This suggests that combining `NALTREXONE` and `THERAPY` is worse than using `NALTREXONE` alone to delay the first time of heavy drinking and other endpoints. This is also coherent with the survival curves in Figure 23.5, in which the best Kaplan–Meier curve corresponds to `NALTREXONE=1, THERAPY=0`.

Using Keele (2010)'s data, we have the following results:

```
> cox.fit <- coxph(Surv(acttime, censor) ~
+                    hcomm + hfloor + scomm + sfloor +
+                    prespart + demhsmaj + demsnmaj +
+                    prevgenx + lethal +
+                    deathrt1 + acutediz + hosp01  +
+                    hospdisc  + hhosleng +
+                    mandiz01 + femdiz01 + peddiz01 + orphdum +
+                    natreg + I(natreg^2) + vandavg3 + wpnoavg3 +
+                    condavg3 + orderent + stafcder,
+                  data=fda)
> summary(cox.fit)
Call:
coxph(formula = Surv(acttime, censor) ~ hcomm + hfloor + scomm +
```

```
    sfloor + prespart + demhsmaj + demsnmaj + prevgenx + lethal +
    deathrt1 + acutediz + hosp01 + hospdisc + hhosleng + mandiz01 +
    femdiz01 + peddiz01 + orphdum + natreg + I(natreg^2) + vandavg3 +
    wpnoavg3 + condavg3 + orderent + stafcder, data = fda)

  n= 408, number of events= 262

                   coef  exp(coef)   se(coef)      z Pr(>|z|)
hcomm         3.642e-01  1.439e+00  2.951e+00  0.123 0.901775
hfloor        7.944e+00  2.819e+03  8.173e+00  0.972 0.331071
scomm         4.716e-01  1.603e+00  1.898e+00  0.248 0.803771
sfloor        2.604e+00  1.352e+01  2.370e+00  1.099 0.271877
prespart      8.038e-01  2.234e+00  3.042e-01  2.643 0.008226 **
demhsmaj      1.363e+00  3.909e+00  1.917e+00  0.711 0.476890
demsnmaj      1.217e+00  3.377e+00  5.606e-01  2.171 0.029940 *
prevgenx     -9.915e-04  9.990e-01  7.779e-04 -1.275 0.202459
lethal        7.872e-02  1.082e+00  2.378e-01  0.331 0.740605
deathrt1      6.537e-01  1.923e+00  2.435e-01  2.685 0.007253 **
acutediz      1.994e-01  1.221e+00  2.262e-01  0.882 0.377896
hosp01        4.280e-02  1.044e+00  2.495e-01  0.172 0.863768
hospdisc     -1.238e-06  1.000e+00  5.278e-07 -2.345 0.019002 *
hhosleng     -1.273e-02  9.874e-01  1.988e-02 -0.640 0.521891
mandiz01     -1.177e-01  8.889e-01  3.800e-01 -0.310 0.756711
femdiz01      9.032e-01  2.468e+00  3.497e-01  2.583 0.009799 **
peddiz01     -3.401e-02  9.666e-01  5.112e-01 -0.067 0.946968
orphdum       5.540e-01  1.740e+00  2.109e-01  2.626 0.008630 **
natreg       -2.221e-02  9.780e-01  8.282e-03 -2.682 0.007318 **
I(natreg^2)   1.029e-04  1.000e+00  4.567e-05  2.253 0.024276 *
vandavg3     -2.014e-02  9.801e-01  1.536e-02 -1.311 0.189802
wpnoavg3      5.220e-03  1.005e+00  1.426e-03  3.660 0.000252 ***
condavg3      9.628e-03  1.010e+00  2.271e-02  0.424 0.671637
orderent     -1.810e-02  9.821e-01  8.147e-03 -2.222 0.026296 *
stafcder      8.013e-04  1.001e+00  7.986e-04  1.003 0.315719
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

23.4.4 Log-rank test as a score test from Cox model

A standard problem in clinical trials is to compare survival times under treatment and control. Assume no ties in the failure times, and let x denote the binary indicator for treatment. Under the proportional hazards assumption, the control group has hazard $\lambda_0(t)$, and treatment group has hazard $\lambda_1(t) = \lambda_0(t)e^{\beta}$. We are interested in testing the null hypothesis

$$\beta = 0 \Leftrightarrow \lambda_1(t) = \lambda_0(t) \Leftrightarrow S_1(t) = S_0(t).$$

Under the null hypothesis, the score function reduces to

$$\frac{\partial \log L(\beta)}{\partial \beta}\Big|_{\beta=0} = \sum_{k=1}^{K} \{x_k - E_{\beta=0}(x \mid R_k)\} = \sum_{k=1}^{K} \left(x_k - \frac{r_{k1}}{r_k}\right),$$

because

$$E_{\beta=0}(x \mid R_k) = \frac{\sum_{l \in R(t_k)} x_l}{\sum_{l \in R(t_k)} 1} = \frac{r_{k1}}{r_k}$$

equaling the ratio of the number of treated units at risk r_{k1} over the number of units at risk r_k, at time t_k. The Fisher information at the null is

$$-\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}\Big|_{\beta=0} = \sum_{k=1}^{K} \mathrm{cov}_{\beta=0}(x \mid R_k) = \sum_{k=1}^{K} \frac{r_{k1}}{r_k}\left(1 - \frac{r_{k1}}{r_k}\right).$$

The score test for classical parametric models relies on

$$\frac{\partial \log L(\beta)}{\partial\beta}\Big|_{\beta=0} \overset{\mathrm{a}}{\sim} \mathrm{N}\left(0, -\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}\Big|_{\beta=0}\right),$$

which follows from Bartlett's identity and CLT. Applying this fact to Cox's model, we have

$$\mathrm{LR} = \frac{\sum_{k=1}^{K}\left(x_k - \frac{r_{k1}}{r_k}\right)}{\sqrt{\sum_{k=1}^{K} \frac{r_{k1}}{r_k}\left(1 - \frac{r_{k1}}{r_k}\right)}} \overset{\mathrm{a}}{\sim} \mathrm{N}(0,1).$$

So we reject the null at level α if $|\mathrm{LR}|$ is larger than the upper $1-\alpha/2$ quantile of standard Normal. This is almost identical to the log-rank test (Mantel, 1966).

The `survdiff` function in the `survival` package implements various tests including the log rank test as a special case. Below, I use the `gehan` dataset in the `MASS` package to illustrate the log rank test. The data were from a matched-pair experiment of 42 leukaemia patients (Gehan, 1965). Treated units received the drug 6-mercaptopurine, and the rest are controls. For illustration purpose, I ignore the pair indicators.

```
> library(MASS)
> head(gehan)
  pair time cens   treat
1    1    1    1 control
2    1   10    1    6-MP
3    2   22    1 control
4    2    7    1    6-MP
5    3    3    1 control
6    3   32    0    6-MP
> survdiff(Surv(time, cens) ~ treat,
+          data = gehan)
Call:
survdiff(formula = Surv(time, cens) ~ treat, data = gehan)

               N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP    21        9     19.3      5.46      16.8
treat=control 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

The treatment was quite effective, yielding extremely small p-value even with moderate sample size. It is also clear from the Kaplan–Meier curves in Figure 23.7 and the results from fitting the Cox proportional hazards model.

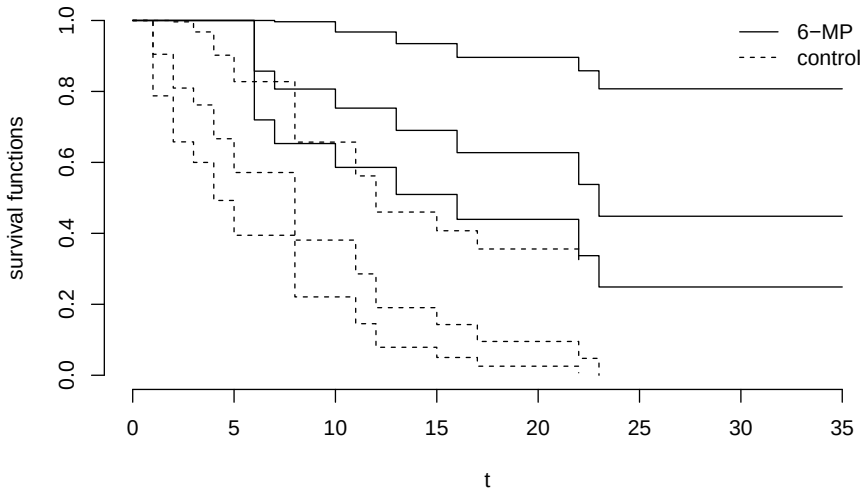


FIGURE 23.7: Kaplan–Meier curves with 95% confidence intervals based on Gehan (1965)'s data

```
> cox.gehan = coxph(Surv(time, cens) ~ treat,
+                   data = gehan)
> summary(cox.gehan)
Call:
coxph(formula = Surv(time, cens) ~ treat, data = gehan)

  n= 42, number of events= 30

               coef exp(coef) se(coef)     z Pr(>|z|)
treatcontrol 1.5721    4.8169   0.4124 3.812 0.000138 ***
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

             exp(coef) exp(-coef) lower .95 upper .95
treatcontrol     4.817     0.2076     2.147      10.81

Concordance= 0.69  (se = 0.041 )
Likelihood ratio test= 16.35  on 1 df,   p=5e-05
Wald test            = 14.53  on 1 df,   p=1e-04
Score (logrank) test = 17.25  on 1 df,   p=3e-05
```

Log rank test is a standard tool in survival analysis. However, what it delivers is just a special case of the Cox proportional hazards model. The p-value from the log rank test is close to the p-value from the score test of the

Cox proportional hazards model with only a binary treatment indicator. The latter can also adjust for other pretreatment covariates.

23.5 Critiques on survival analysis

Kaplan–Meier curves and Cox proportional hazards model are standard tools for analyzing medical data with censored survival time. They are among the most commonly-used methods in medical journals. Kaplan and Meier (1958) and Cox (1972) are two of the most cited papers in statistics.

Freedman (2008) criticized these two standard tools. Both rely on the critical assumption of noninformative censoring that censoring and survival time are independent or conditionally independent given covariates. When censoring is due to administrative constraints, this may be a plausible assumption. The data from Lin et al. (2016) is a convincing example of noninformative censoring. However, many other studies have more complex censoring mechanisms, for example, one may drop out the study, and another may be killed by an irrelevant cause. Cox's model relies on an additional assumption of proportional hazards. This particular functional form allows for interpreting the coefficients as log conditional hazard ratios if the model is correctly specified. However, its interpretation becomes obscure when the model is mis-specified. Two survival curves based on Lin et al. (2016) 's data cross, which make the proportional hazards assumption dubious.

Hernán (2010) offered a more fundamental critique on hazard-based survival analysis. For example, in a randomized treatment-control experiment, the hazard ratio at time t is the ratio of the instantaneous probability of death conditioning on the event that the patients have survived up to time t:

$$\frac{\lim_{\Delta t \downarrow 0} \mathrm{pr}(t \leq T < t + \Delta t \mid x = 1, T \geq t)/\Delta t}{\lim_{\Delta t \downarrow 0} \mathrm{pr}(t \leq T < t + \Delta t \mid x = 0, T \geq t)/\Delta t}$$

This ratio is difficult to interpret because patients who have survived up to time t can be quite different in treatment and control groups, especially when the treatment is effective. Even though patients are randomly assigned at baseline, the survivors up to time t are not. Hernán (2010) suggested focusing on the comparison of the survival functions themselves.

23.6 Homework problems

23.1 Identifiability of the survival time under independent censoring

Assume the survival time T and censoring time C are continuous and independent random variables. But we can only observe $y = \min(T, C)$ and $\delta = 1(T \leq C)$. Show that the hazard function of T can be identified by the following formula:

$$\lambda_T(t) = \frac{\text{pr}(y = t, \delta = 1)}{\text{pr}(y \geq t)}.$$

Remark: When the survival time T and censoring time C are continuous random variables, a similar but more complex formula holds.

23.2 Exponential random variables

Assume that $T_i \sim \text{Exponential}(\lambda_i)$ are independent $(i = 1, \ldots, n)$. Show that

$$\text{pr}\{T_i = \min(T_1, \ldots, T_n)\} = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_n}.$$

This is a special case of Theorem 23.1, but deriving it directly can give more insights.

23.3 Log-Normal regression model

If $\log T = x^{\text{T}}\beta + \sigma \text{N}(0, 1)$, find the conditional hazard of T given x. Does it satisfy the proportional hazards assumption? Based on $(y_i, x_i, \delta_i)_{i=1}^n$, what is the likelihood function? Compare it with the partial likelihood function.

23.4 Weibull random variable

Using (23.1) to show the formulas of density, survival, and hazard functions.

23.5 Weibull regression model

Assume that $T \mid x \sim \text{Weibull}(a, b = e^{x^{\text{T}}\beta})$. Show that

$$\log T = x^{\text{T}}\beta + u$$

with $u \perp\!\!\!\perp x$, and find the distribution of u. Show $E(T \mid x)$ is log-linear in x, and $E(\log T \mid x)$ is linear in x. Does it satisfy the proportional hazards assumption? Based on $(y_i, x_i, \delta_i)_{i=1}^n$, what is the likelihood function? Compare it with the partial likelihood function.

23.6 Invariance of the proportional hazards model

Assume that $T \mid x$ follows a proportional hazards model, show that any nonnegative and strictly increasing transformation $g(T) \mid x$ also follows a proportional hazards model.

Part VIII

Appendices

A1

Linear Algebra

A1.1 Basics of vectors and matrices

All vectors are column vectors as in R unless stated otherwise. Let the superscript "T" denote the transpose of a vector or matrix.

Euclidean space

The n-dimensional Euclidean space $\mathbb{R}^n$ is a set of all n-dimensional vectors equipped with an inner product: $\langle x, y\rangle = x^{\mathrm{T}}y = \sum_{i=1}^{n} x_i y_i$ where $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ and $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ are two n-dimensional vectors. We say that x and y are orthogonal, denoted by $x \perp y$, if $\langle x, y\rangle = 0$. The length of a vector x is defined as $\|x\| = \sqrt{\langle x, x\rangle} = \sqrt{x^{\mathrm{T}}x}$. The Cauchy–Schwarz inequality states that

$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|.$$

Column space of a matrix

Given an $n \times m$ matrix $A = (A_1, \ldots, A_m)$, we define its column space as

$$\mathcal{C}(A) = \{\alpha_1 A_1 + \cdots + \alpha_m A_m : a_1, \ldots, a_m \in \mathbb{R}\},$$

which is the set of all linear combinations of the column vectors of A.

Inverse of a matrix

Let I_n be the $n \times n$ identity matrix. An $n \times n$ matrix A is nonsingular if there exists an $n \times n$ matrix B such that $AB = BA = I_n$. We call B the inverse of A, denoted by A^{-1}.

Some special matrices

An $n \times n$ matrix A is symmetric if $A^{\mathrm{T}} = A$. An $n \times n$ matrix is orthogonal if $A^{\mathrm{T}}A = AA^{\mathrm{T}} = I_n$, that is $A^{\mathrm{T}} = A^{-1}$. An $n \times n$ diagonal matrix A has zero off-diagonal elements, denoted by $A = \mathrm{diag}\{a_{11}, \ldots, a_{nn}\}$.

Eigenvalues and eigenvectors

For an $n \times n$ matrix A, if there exists a pair of n-dimensional vector x and a scalar λ such that $Ax = \lambda x$, then we call λ an eigenvalue and x an eigenvector of A. From the definition, eigenvalue and eigenvector come always in pair. The following eigen-decomposition theorem is important for real symmetric matrices.

Theorem A1.1 *If A is an $n \times n$ symmetric matrix, then there exists an orthogonal matrix P such that*

$$P^{\mathrm{T}}AP = diag\{\lambda_1, \ldots, \lambda_n\},$$

where the λ's are the n eigenvalues of A, and the column vectors of $P = (\gamma_1, \cdots, \gamma_n)$ are the corresponding eigenvectors.

If we write the eigen-decomposition as

$$AP = P\mathrm{diag}\{\lambda_1, \ldots, \lambda_n\} \Longleftrightarrow A(\gamma_1, \cdots, \gamma_n) = (\lambda_1\gamma_1, \cdots, \lambda_n\gamma_n),$$

then (λ_i, γ_i) must be a pair of eigenvalue and eigenvector. Moreover, the eigen-decomposition in Theorem A1.1 is unique up to the permutation of the columns of P and the corresponding λ_i's.

Corollary A1.1 *If $P^{\mathrm{T}}AP = diag\{\lambda_1, \ldots, \lambda_n\}$, then*

$$A = Pdiag\{\lambda_1, \ldots, \lambda_n\}P^{\mathrm{T}}, \quad A^k = A \cdot A \cdots A = Pdiag\{\lambda_1^k, \ldots, \lambda_n^k\}P^{\mathrm{T}};$$

if the eigenvalues of A are nonzero, then

$$A^{-1} = Pdiag\{1/\lambda_1, \ldots, 1/\lambda_n\}P^{\mathrm{T}}.$$

The eigen-decomposition is also useful for defining the square root of an $n \times n$ symmetric matrix. In particular, if the eigenvalues of A are nonnegative, then we can define

$$A^{1/2} = P\mathrm{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\}P^{\mathrm{T}}$$

By definition $A^{1/2}$ is a symmetric matrix satisfying $A^{1/2}A^{1/2} = A$. There are other definitions of square root of a symmetric matrix, but we adopt this form.

From Theorem A1.1, we can write A as

$$\begin{aligned} A &= P\mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}P^{\mathrm{T}} \\ &= (\gamma_1, \cdots, \gamma_n)\mathrm{diag}\{\lambda_1, \ldots, \lambda_n\} \begin{pmatrix} \gamma_1^{\mathrm{T}} \\ \vdots \\ \gamma_n^{\mathrm{T}} \end{pmatrix} \\ &= \sum_{i=1}^{n} \lambda_i \gamma_i \gamma_i^{\mathrm{T}}. \end{aligned}$$

Theorem A1.2 *(Rayleigh quotient and eigenvalues) For an $n \times n$ symmetric matrix A, let $r(x) = x^{\mathrm{T}}Ax/x^{\mathrm{T}}x$ be the Rayleigh quotient of x. The maximum and minimum eigenvalues of A are*

$$\lambda_{\max}(A) = \max_{x \neq 0} r(x), \qquad \lambda_{\min}(A) = \min_{x \neq 0} r(x).$$

Rank and determinant

We will only use the rank and determinant for a real symmetric matrix although they are well-defined for general matrices. For an $n \times n$ symmetric matrix, its rank equals the number of non-zero eigenvalues and its determinant equals the product of all eigenvalues. The matrix A is of full rank if all its eigenvalues are non-zero, which implies that its rank equals n and its determinant is non-zero.

Quadratic form

For an $n \times n$ symmetric matrix $A = (a_{ij})$ and an n-dimensional vector x, we can define the quadratic form as

$$x^{\mathrm{T}}Ax = \langle x, Ax \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

We always consider a symmetric matrix in the quadratic form without loss of generality. Otherwise, we can symmetrize A as $(A+A^{\mathrm{T}})/2$ without changing the value of the quadratic form because

$$x^{\mathrm{T}}Ax = x^{\mathrm{T}} \left(\frac{A + A^{\mathrm{T}}}{2} \right) x.$$

We call A positive semi-definite, denoted by $A \succeq 0$, if $x^{\mathrm{T}}Ax \geq 0$ for all x; we call A positive definite, denoted by $A \succ 0$, if $x^{\mathrm{T}}Ax > 0$ for all nonzero x.

Theorem A1.3 *For a symmetric matrix A, it is positive semi-definite if and only if all its eigenvalues are nonnegative, and it is positive definite if and only if all its eigenvalues are positive.*

We can also define the partial order between matrices. We call $A \succeq B$ if and only if $A - B \succeq 0$, and we call $A \succ B$ if and only if $A - B \succ 0$. This is important in statistics because we often compare the efficiency of estimators based on their covariance matrices.

Trace

The trace of an $n \times n$ matrix $A = (a_{ij})$ is the sum of all its diagonal elements, denoted by

$$\mathrm{trace}(A) = \sum_{i=1}^{n} a_{ii}.$$

The trace operator has two important properties that can sometimes help to simplify calculations.

Proposition A1.1 *trace*(AB) = *trace*(BA) *as long as* AB *and* BA *are both square matrices.*

We can verify Proposition A1.1 by the definitions. It can be particularly useful if the the dimension of BA is much lower than the dimension of AB. For example, if both $A = (a_1, \ldots, a_n)^{\mathrm{T}}$ and $B = (b_1, \ldots, b_n)$ are vectors, then $\mathrm{trace}(AB) = \mathrm{trace}(BA) = \langle B^{\mathrm{T}}, A\rangle = \sum_{i=1}^n a_i b_i$.

Proposition A1.2 *The trace of an* $n \times n$ *symmetric matrix* A *equals the sum of its eigenvalues:* *trace*$(A) = \sum_{i=1}^n \lambda_i$.

Proof of Proposition : It follows from the eigen-decomposition and Proposition A1.1. Let $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$, and we have

$$\mathrm{trace}(A) = \mathrm{trace}(P\Lambda P^{\mathrm{T}}) = \mathrm{trace}(\Lambda P^{\mathrm{T}} P) = \mathrm{trace}(\Lambda) = \sum_{i=1}^n \lambda_i.$$

□

Projection matrix

An $n \times n$ symmetric matrix H is a projection matrix if $H^2 = H$. Based on the eigen-decomposition $H = \sum_{i=1}^n \lambda_i \gamma_i \gamma_i^{\mathrm{T}}$, we have

$$\begin{aligned} H^2 = H &\Longrightarrow \sum_{i=1}^n \lambda_i^2 \gamma_i \gamma_i^{\mathrm{T}} = \sum_{i=1}^n \sum_{i=1}^n \lambda_i \gamma_i \gamma_i^{\mathrm{T}} \\ &\Longrightarrow \sum_{i=1}^n (\lambda_i^2 - \lambda_i) \gamma_i \gamma_i^{\mathrm{T}} = 0 \\ &\Longrightarrow \lambda_i^2 - \lambda_i = 0, \qquad (i = 1, \ldots, n) \end{aligned}$$

which implies that the eigenvalues of H are either 1 or 0. So the trace of H equals its rank:

$$\mathrm{trace}(H) = \mathrm{rank}(H).$$

Why is this a reasonable definition of a "projection matrix"? Or, why must a projection matrix satisfy $H^{\mathrm{T}} = H$ and $H^2 = H$? First, it is reasonable to require that $Hx_1 = x_1$ for any $x_1 \in \mathcal{C}(H)$, the column space of H. Since $x_1 = H\alpha$ for some α, we indeed have $Hx_1 = H(H\alpha) = H^2\alpha = H\alpha = x_1$ because of the property $H^2 = H$. Second, it is reasonable to require that $x_1 \perp x_2$ for any vector $x_1 = H\alpha \in \mathcal{C}(H)$ and x_2 such that $Hx_2 = 0$. So we need $\alpha^{\mathrm{T}} H^{\mathrm{T}} x_2 = 0$ which is true if $H = H^{\mathrm{T}}$. Therefore, the two conditions are natural for the definition of a projection matrix.

Cholesky decomposition

An $n \times n$ positive semi-definite matrix A can be decomposed as $A = LL^{\mathrm{T}}$ where L is an $n \times n$ lower triangular matrix with non-negative diagonal elements. Take an arbitrary orthogonal matrix Q, we have $A = LQQ^{\mathrm{T}}L^{\mathrm{T}} = CC^{\mathrm{T}}$ where $C = LQ$. So we can decompose a positive semi-definite matrix A as $A = CC^{\mathrm{T}}$, but this decomposition is not unique.

Polar decomposition

A square matrix A can be decomposed as $A = (AA^{\mathrm{T}})^{1/2}\Gamma$ where Γ is an orthogonal matrix.

A1.2 Vector calculus

If $f(x)$ is a function from $\mathbb{R}^p$ to $\mathbb{R}$, then we use the notation

$$\frac{\partial f(x)}{\partial x} \equiv \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_p} \end{pmatrix}$$

for the component-wise partial derivative, which must have the same dimension as x. For example, for a linear function $f(x) = x^{\mathrm{T}}a = a^{\mathrm{T}}x$ with $a, x \in \mathbb{R}^p$, we have

$$\frac{\partial a^{\mathrm{T}}x}{\partial x} = \begin{pmatrix} \frac{\partial a^{\mathrm{T}}x}{\partial x_1} \\ \vdots \\ \frac{\partial a^{\mathrm{T}}x}{\partial x_p} \end{pmatrix} = \begin{pmatrix} \frac{\partial \sum_{j=1}^p a_j x_j}{\partial x_1} \\ \vdots \\ \frac{\partial \sum_{j=1}^p a_j x_j}{\partial x_p} \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = a;$$

for a quadratic function $f(x) = x^{\mathrm{T}}Ax$ with a symmetric $A \in \mathbb{R}^{p \times p}$ and $x \in \mathbb{R}^p$, we have

$$\frac{\partial x^{\mathrm{T}}Ax}{\partial x} = \begin{pmatrix} \frac{\partial x^{\mathrm{T}}Ax}{\partial x_1} \\ \vdots \\ \frac{\partial x^{\mathrm{T}}Ax}{\partial x_p} \end{pmatrix} = \begin{pmatrix} \frac{\partial \sum_{i=1}^p \sum_{j=1}^p a_{ij} x_i x_j}{\partial x_1} \\ \vdots \\ \frac{\partial \sum_{i=1}^p \sum_{j=1}^p a_{ij} x_i x_j}{\partial x_p} \end{pmatrix} = \begin{pmatrix} 2a_{11}x_1 + \cdots + 2a_{1p}x_p \\ \vdots \\ 2a_{p1}x_1 + \cdots + 2a_{pp}x_p \end{pmatrix} = 2Ax.$$

These are two important rules of vector calculus used in this class, summarized below.

Proposition A1.3 *We have*

$$\frac{\partial a^{\mathrm{T}}x}{\partial x} = a, \quad \frac{\partial x^{\mathrm{T}}Ax}{\partial x} = 2Ax.$$

We can also extend the definition to vector functions. If $f(x) = (f_1(x), \ldots, f_q(x))^{\mathrm{T}}$ is a function from $\mathbb{R}^p$ to $\mathbb{R}^q$, then we use the notation

$$\frac{\partial f(x)}{\partial x} \equiv \left(\frac{\partial f_1(x)}{\partial x}, \cdots, \frac{\partial f_q(x)}{\partial x} \right) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_q(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(x)}{\partial x_p} & \cdots & \frac{\partial f_q(x)}{\partial x_p} \end{pmatrix},$$

which is a $p \times q$ matrix with rows corresponding to the elements of x and the columns corresponding to the elements of $f(x)$. We can easily extend the first result of Proposition A1.3.

Proposition A1.4 *For $B \in \mathbb{R}^{q \times p}$ and $x \in \mathbb{R}^p$, we have*

$$\frac{\partial Bx}{\partial x} = B.$$

A1.3 Homework problems

A1.1 Inverse of a block matrix

Show that

$$\begin{aligned} &\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} \\ &= \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix} \\ &= \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}, \end{aligned}$$

provided that all the inverses of matrices exist. The two forms of the inverse imply the Woodbury formula:

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1},$$

which further implies the Sherman–Morrison formula:

$$(A + uv^{\mathrm{T}})^{-1} = A^{-1} - (1 + v^{\mathrm{T}}A^{-1}u)^{-1}A^{-1}uv^{\mathrm{T}}A^{-1},$$

where A is an invertible square matrix and u and v are two column vectors.

A1.2 Vector calculus

What is the formula for $\partial x^{\mathrm{T}}Ax/\partial x$ if A is not symmetric in Proposition A1.3?

A2

Random Variables

Let "IID" denote "independent and identically distributed", "$\overset{\text{IID}}{\sim}$" denote a sequence of random variables that are IID with some common distribution, and "$\perp\!\!\!\perp$" denote independence between random variables. Define Euler's Gamma function as

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x, \qquad (z > 0).$$

A2.1 Some important univariate random variables

A2.1.1 Normal, χ^2, t and F

The standard Normal random variable $Z \sim \mathrm{N}(0,1)$ has density

$$f(z) = (2\pi)^{-1/2} \exp\left(-z^2/2\right).$$

A Normal random variable X has mean μ and variance σ^2 if $X = \mu + \sigma Z$. We can show that X has density

$$f(x) = (2\pi)^{1/2} \exp\left\{-(x-\mu)^2/(2\sigma^2)\right\}.$$

A chi-squared random variable with degrees of freedom n, denoted by $Q_n \sim \chi^2_n$, can be represented as

$$Q_n = \sum_{i=1}^{n} Z_i^2,$$

where $Z_i \overset{\text{IID}}{\sim} \mathrm{N}(0,1)$. We can show that its density is

$$f_n(q) = q^{n/2} \exp(-q/2) \Big/ \left\{2^{n/2}\Gamma(n/2)\right\}, \qquad (q > 0). \tag{A2.1}$$

We can verify that the above density (A2.1) is well-defined even if we change the integer n to be an arbitrary positive real number ν, and call the corresponding random variable Q_ν a chi-squared random variable with degrees of freedom ν, denoted by $Q_\nu \sim \chi^2_\nu$.

A t random variable with degrees of freedom ν can be represented as

$$t_\nu = \frac{Z}{\sqrt{Q_\nu/\nu}}$$

where $Z \sim \mathrm{N}(0,1), Q_\nu \sim \chi^2_\nu$ and $Z \perp\!\!\!\perp Q_\nu$.

An F random variable with degrees of freedom (r,s) can be represented as

$$F = \frac{Q_r/r}{Q_s/s}$$

where $Q_r \sim \chi^2_r, Q_s \sim \chi^2_s$ and $Q_r \perp\!\!\!\perp Q_s$.

A2.1.2 Beta–Gamma duality

The Gamma(α,β) random variable with $\alpha,\beta > 0$ has density

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad (x > 0). \tag{A2.2}$$

The Beta(α,β) random variable with $\alpha,\beta > 0$ has density

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad (0 < x < 1).$$

These two random variables are closely related as shown in the following theorem.

Theorem A2.1 *If $X \sim Gamma(\alpha,\theta), Y \sim Gamma(\beta,\theta)$ and $X \perp\!\!\!\perp Y$, then*

1. $X + Y \sim Gamma(\alpha+\beta,\theta)$,
2. $X/(X+Y) \sim Beta(\alpha,\beta)$,
3. $X + Y \perp\!\!\!\perp X/(X+Y)$.

Another simple but useful fact is that χ^2 is a special Gamma random variable. Comparing the densities in (A2.1) and (A2.2), we obtain the following result.

Proposition A2.1 $\chi^2_n \sim \mathrm{Gamma}(n/2, 1/2)$.

We can also calculate the moments of the Gamma and Beta distributions.

Proposition A2.2 *If $X \sim \mathrm{Gamma}(\alpha,\beta)$, then*

$$E(X) = \frac{\alpha}{\beta}, \qquad E(X^2) = \frac{\alpha}{\beta^2}.$$

Proposition A2.3 *If $X \sim Beta(\alpha,\beta)$, then*

$$E(X) = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}.$$

I leave the proofs of the above results as homework problems.

A2.2 Multivariate distributions

A random vector $(X_1, \ldots, X_n)^{\mathrm{T}}$ is a vector consisting of n random variables. If all components are continuous, we can define its joint density $f_{X_1 \cdots X_n}(x_1, \ldots, x_n)$.

For a random vector $\binom{X}{Y}$ with X and Y possibly being vectors, if it has joint density $f_{XY}(x, y)$ then we can obtain the marginal distribution of X as $f_X(x) = \int f_{XY}(x, y)\mathrm{d}y$ and define the conditional density as

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)} \qquad \text{if } f_X(x) \neq 0.$$

Based on the conditional density, we can define the conditional expectation of any function of Y as

$$E\{g(Y) \mid X = x\} = \int g(y) f_{Y|X}(y \mid x)\mathrm{d}y$$

and the conditional variance as

$$\mathrm{var}\{g(Y) \mid X = x\} = E\left[\{g(Y)\}^2 \mid X = x\right] - \left[E\{g(Y) \mid X = x\}\right]^2.$$

In the above definitions, the conditional mean and variance are both function of x. Sometimes, we also use the notation $E\{g(Y) \mid X\}$ and $\mathrm{var}\{g(Y) \mid X\}$, which are functions of the random variable X and are thus random variables.

Two important laws of conditional expectation and variance are below.

Theorem A2.2 *The law of total expectation states that*

$$E\{g(Y)\} = E\left[E\{g(Y) \mid X\}\right]$$

Theorem A2.3 *The law of total variance or analysis of variance states that*

$$\mathrm{var}\{g(Y)\} = E\left[\mathrm{var}\{g(Y) \mid X\}\right] + \mathrm{var}\left[E\{g(Y) \mid X\}\right]$$

Independence

Random variables $(X_1, \ldots, X_n)$ are mutually independent if

$$f_{X_1 \cdots X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Note that in this definition, each of $(X_1, \ldots, X_n)$ can be vectors. We have the following rules under independence.

Proposition A2.4 *If $X \perp\!\!\!\perp Y$, then $h(X) \perp\!\!\!\perp g(Y)$ for any functions $h(\cdot)$ and $g(\cdot)$.*

Proposition A2.5 *If $X \perp\!\!\!\perp Y$, then*

$$\begin{aligned} f_{XY}(x,y) &= f_X(x) f_Y(y), \\ f_{Y|X}(y \mid x) &= f_Y(y), \\ E\{g(Y) \mid X\} &= E\{g(Y)\}, \\ E\{g(Y)h(X)\} &= E\{g(Y)\} E\{h(X)\}. \end{aligned}$$

Expectations of random vectors or random matrices

For a random matrix $W = (W_{ij})$, we define $E(W) = (E(W_{ij}))$. For constant matrices A and C, we can verify that

$$\begin{aligned} E(AW + C) &= AE(W) + C, \\ E(AWC) &= AE(W)C. \end{aligned}$$

Covariance between two random vectors

If $W \in \mathbb{R}^r$ and $Y \in \mathbb{R}^s$, then their covariance

$$\text{cov}(W, Y) = E\left[\{W - E(W)\}\{Y - E(Y)\}^{\mathrm{T}}\right]$$

is an $r \times s$ matrix. As a special case, $\text{cov}(Y) = \text{cov}(Y, Y) = E\left[\{Y - E(Y)\}\{Y - E(Y)\}^{\mathrm{T}}\right]$. For a scalar random variable, $\text{cov}(Y) = \text{var}(Y)$.

Proposition A2.6 *For $A \in \mathbb{R}^{r \times n}, Y \in \mathbb{R}^n$ and $C \in \mathbb{R}^r$, $\text{cov}(AY + C) = A\text{cov}(Y)A^{\mathrm{T}}$.*

Using Proposition A2.6, we can verify that for any n-dimensional random vector, $\text{cov}(Y) \succeq 0$ because for all $x \in \mathbb{R}^n$,

$$x^{\mathrm{T}}\text{cov}(Y)x = \text{cov}(x^{\mathrm{T}}Y) = \text{var}(x^{\mathrm{T}}Y) \geq 0.$$

Proposition A2.7 *For two random vectors W and Y, we have*

$$\text{cov}(AW + C, BY + D) = A\text{cov}(W, Y)B^{\mathrm{T}}$$

and

$$\text{cov}(AW+BY) = A\text{cov}(W)A^{\mathrm{T}}+B\text{cov}(W)B^{\mathrm{T}}+A\text{cov}(W,Y)B^{\mathrm{T}}+B\text{cov}(Y,W)A^{\mathrm{T}}.$$

A2.3 Multivariate Normal and its properties

I use a generative definition of multivariate Normal random vector. First, Z is a standard Normal random vector if $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ with the $Z_i \overset{\text{IID}}{\sim} \text{N}(0, 1)$.

Given a mean vector μ and a positive semi-definite covariance matrix Σ, define a Normal random vector $Y \sim \mathrm{N}(\mu, \Sigma)$ with mean μ and covariance Σ as

$$Y = \mu + AZ, \tag{A2.3}$$

where A satisfies $\Sigma = AA^{\mathrm{T}}$. We can verify that $\mathrm{cov}(Y) = \Sigma$, so indeed Σ is its covariance matrix. If $\Sigma \succ 0$, then we can verify that Y has density

$$f_Y(y) = (2\pi)^{-n/2} \left\{\det(\Sigma)\right\}^{-1/2} \exp\left\{-(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)/2\right\}.$$

We can easily verify the following result by calculating the density.

Proposition A2.8 *If $Z \sim N(0, I_n)$ and Γ is an orthogonal matrix, then $\Gamma Z \sim N(0, I_n)$.*

In the definition (A2.3), we do not require Σ be positive definite. However, this definition has a subtle issue of uniqueness. Although the decomposition $\Sigma = AA^{\mathrm{T}}$ is not unique, the resulting distribution $Y = \mu + AZ$ is unique. We can verify this using the Polar decomposition: Because $A = \Sigma^{1/2}\Gamma$ where Γ is an orthogonal matrix, we have $Y = \mu + \Sigma^{1/2}\Gamma Z = \mu + \Sigma^{1/2}\tilde{Z}$ where $\tilde{Z} = \Gamma Z$ is a standard Normal random vector by Proposition A2.8.

Theorem A2.4 *Assume that*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right).$$

Then $Y_1 \perp\!\!\!\perp Y_2$ if and only if $\Sigma_{12} = 0$.

I leave the proof of the theorem as a homework problem.

Proposition A2.9 *If $Y \sim N(\mu,\Sigma)$, then $BY + C \sim N(B\mu + C, B\Sigma B^{\mathrm{T}})$, that is, any linear transformation of a Normal random vector is also a Normal random vector.*

Proof of Proposition A2.9: By definition, $Y = \mu + \Sigma^{1/2}Z$ where Z is the standard Normal random vector, we have

$$\begin{aligned} BY + c &= B(\mu + \Sigma^{1/2}Z) + C \\ &= B\mu + C + B\Sigma^{1/2}Z \\ &\sim \mathrm{N}(B\mu + C, B\Sigma^{1/2}\Sigma^{1/2\mathrm{T}}B^{\mathrm{T}}) \\ &\sim \mathrm{N}(B\mu + C, B\Sigma B^{\mathrm{T}}). \end{aligned}$$

□

Theorem A2.5 *Assume that*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right).$$

1. The marginal distributions are Normal:

$$Y_1 \sim \mathrm{N}\left(\mu_1, \Sigma_{11}\right),$$
$$Y_2 \sim \mathrm{N}\left(\mu_2, \Sigma_{22}\right).$$

2. If $\Sigma_{22} \succ 0$, then the conditional distribution is Normal:

$$Y_1 \mid Y_2 = y_2 \sim \mathrm{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right);$$

Y_2 is independent of the residual

$$Y_1 - \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2) \sim \mathrm{N}\left(\mu_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

I list some other results of multivariate Normal below:

Proposition A2.10 *Assume $Y \sim N(\mu, \sigma^2 I_n)$. If $AB^{\mathrm{T}} = 0$, then $AY \perp\!\!\!\perp BY$.*

Proposition A2.11 *Assume that $(Y_1, Y_2)^{\mathrm{T}}$ follows a bivariate Normal distribution*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right),$$

where ρ is the correlation coefficient defined as

$$\rho = \frac{\mathrm{cov}(Y_1, Y_2)}{\sqrt{\mathrm{var}(Y_1)\mathrm{var}(Y_2)}}.$$

Then the conditional distribution is

$$Y_1 \mid Y_2 = y_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right).$$

Proposition A2.12 *If $(Y_1, Y_2)^{\mathrm{T}}$ follows a bivariate Normal distribution*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

then $Y_1 + Y_2 \perp\!\!\!\perp Y_1 - Y_2$.

A2.4 Quadratic Forms of Random Vectors

Given a random vector Y and a symmetric matrix A, we can define the quadratic form $Y^{\mathrm{T}}AY$, which is a random variable playing important roles in statistics. The first theorem is about its mean.

Theorem A2.6 *If Y has mean μ and covariance Σ, then*

$$E(Y^{\mathrm{T}}AY) = trace(A\Sigma) + \mu^{\mathrm{T}}A\mu.$$

Proof of Theorem A2.6: The proof relies on the following two basic facts.

- $E(YY^{\mathrm{T}}) = \mathrm{cov}(Y) + E(Y)E(Y^{\mathrm{T}}) = \Sigma + \mu\mu^{\mathrm{T}}$.
- For an $n \times n$ symmetric random matrix $W = (w_{ij})$, $E\{\mathrm{trace}(W)\} = E\left(\sum_{i=1}^{n} w_{ii}\right) = \sum_{i=1}^{n} E(w_{ii}) = \mathrm{trace}\{E(W)\}$.

The conclusion follows from

$$\begin{aligned}
E(Y^{\mathrm{T}}AY) &= E\{\mathrm{trace}(Y^{\mathrm{T}}AY)\} \\
&= E\{\mathrm{trace}(AYY^{\mathrm{T}})\} \\
&= \mathrm{trace}\{E(AYY^{\mathrm{T}})\} \\
&= \mathrm{trace}\left\{AE(YY^{T})\right\} \\
&= \mathrm{trace}\{A(\Sigma + \mu\mu^{\mathrm{T}})\} \\
&= \mathrm{trace}\left(A\Sigma + A\mu\mu^{T}\right) \\
&= \mathrm{trace}(A\Sigma) + \mathrm{trace}(\mu^{\mathrm{T}}A\mu) \\
&= \mathrm{trace}(A\Sigma) + \mu^{\mathrm{T}}A\mu.
\end{aligned}$$

□

The variance of the quadratic form is much more complicated for a general random vector. For multivariate Normal random vector, we have the following formula.

Theorem A2.7 *If $Y \sim N(\mu, \Sigma)$, then*

$$\mathrm{var}(Y^{\mathrm{T}}AY) = 2trace(A\Sigma A\Sigma) + 4\mu^{\mathrm{T}}A\Sigma A\mu.$$

From its definition, χ^2_n is the summation of the squares of n IIID standard Normal random variables. It is closely related to quadratic forms of multivariate Normals.

Theorem A2.8 *1. If $Y \sim N(\mu, \Sigma)$ is an n-dimensional random vector with $\Sigma \succ 0$, then*

$$(Y-\mu)^{\mathrm{T}}\Sigma^{-1}(Y-\mu) \sim \chi^2_n.$$

2. If $Y \sim N(0, I_n)$ and H is a projection matrix of rank K, then

$$Y^{\mathrm{T}}HY \sim \chi^2_K.$$

3. If $Y \sim N(0, H)$ where H is a projection matrix of rank K, then

$$Y^{\mathrm{T}}Y \sim \chi^2_K.$$

Proof of Theorem A2.8:

1. By definition $Y = \mu + \Sigma^{1/2} Z$ where Z is a standard Normal random vector, then

$$(Y-\mu)^{\mathrm{T}} \Sigma^{-1} (Y-\mu) = Z^{\mathrm{T}} \Sigma^{1/2} \Sigma^{-1} \Sigma^{1/2} Z = Z^{\mathrm{T}} Z \sim \chi^2_n.$$

2. Using the eigen-decomposition of the projection matrix

$$H = P \mathrm{diag}\{1, \ldots, 1, 0, \ldots, 0\} P^{\mathrm{T}}$$

with K 1's in the diagonal matrix, we have

$$\begin{aligned} Y^{\mathrm{T}} H Y &= Y^{\mathrm{T}} P \mathrm{diag}\{1, \ldots, 1, 0, \ldots, 0\} P^{\mathrm{T}} Y \\ &= Z^{\mathrm{T}} \mathrm{diag}\{1, \ldots, 1, 0, \ldots, 0\} Z, \end{aligned}$$

where $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}} = P^{\mathrm{T}} Y \sim \mathrm{N}(0, P^{\mathrm{T}} P) = \mathrm{N}(0, I_n)$ is a standard Normal random vector. So

$$Y^{\mathrm{T}} H Y = \sum_{i=1}^{K} Z_i^2 \sim \chi^2_K.$$

3. Writing $Y = H^{1/2} Z$ where Z is a standard Normal random vector, we have

$$Y^{\mathrm{T}} Y = Z^{\mathrm{T}} H^{1/2} H^{1/2} Z = Z^{\mathrm{T}} H Z \sim \chi^2_K$$

using the second result.

□

A2.5 Homework problems

A2.1 Beta-Gamma duality

Prove Theorem A2.1.

A2.2 Gamma and Beta moments

Prove Propositions A2.2 and A2.3.

A2.3 Independence and uncorrelatedness in multivariate Normal

Prove Theorem A2.4.

A2.4 Transformation of bivariate Normal

Prove Proposition A2.12.

A2.5 Independence of linear and quadratic function of multivariate Normal

Assume $Y \sim \mathrm{N}(\mu, \sigma^2 I_n)$. For a p dimensional vector a and two $p \times p$ symmetric matrices A and B, show that

1. if $a^{\mathrm{T}} A = 0$, then $a^{\mathrm{T}} Y \perp\!\!\!\perp Y^{\mathrm{T}} A Y$;
2. if $AB = BA = 0$, then $Y^{\mathrm{T}} A Y \perp\!\!\!\perp Y^{\mathrm{T}} B Y$.

Hint: For any $m \times n$ matrix A, we can always find an $n \times m$ matrix A^- such that $AA^-A = A$, where A^- is called a generalized inverse of A, and it may not be unique in general. When A is invertible, then $A^- = A^{-1}$ is unique.

A2.6 Independence of sample mean and variance of IID Normals

If $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then $\bar{X} \perp\!\!\!\perp S^2$, where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

A2.7 Variance of a quadratic form of Normal

Prove Theorem A2.7.

A2.8 Random permutations

Let $c = (c_1, \ldots, c_n)^{\mathrm{T}}$ and $a = (a_1, \ldots, a_n)^{\mathrm{T}}$ be two n-dimensional vectors. Define $\bar{c} = n^{-1} \sum_{i=1}^n c_i$ and $S_c^2 = \sum_{i=1}^n (c_i - \bar{c})^2$ for c, and similarly, define $\bar{a}$ and S_a^2 for a. Let $A = (a_{ij})$ be an $n \times n$ positive semi-definite matrix. Let $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be a random permutation on integers $\{1, \ldots, n\}$, so $c_\pi = (c_{\pi(1)}, \ldots, c_{\pi(n)})$ is a random permutation on the entries $(c_1, \ldots, c_n)$.

1. Show that the mean and covariance matrix of c_π are
$$E(c_\pi) = \bar{c} 1_n, \quad \mathrm{cov}(c_\pi) = (n-1)^{-1} S_c^2 C_n,$$
where $C_n = I_n - n^{-1} 1_n 1_n^{\mathrm{T}}$.
2. Show that the mean and variance of $a^{\mathrm{T}} c_\pi$ are
$$E(a^{\mathrm{T}} c_\pi) = n \bar{a} \bar{c}, \quad \mathrm{var}(a^{\mathrm{T}} c_\pi) = (n-1)^{-1} S_c^2 S_a^2.$$
Note that these two formulas are symmetric in a and c since $a^{\mathrm{T}} c_\pi$ and $c^{\mathrm{T}} a_\pi$ have the same distribution.
3. Show that the mean of $c_\pi^{\mathrm{T}} A c_\pi$ is
$$\begin{aligned} E(c_\pi^{\mathrm{T}} A c_\pi) &= \bar{c}^2 1_n^{\mathrm{T}} A 1_n + (n-1)^{-1} S_c^2 \mathrm{trace}(AC_n) \\ &= \bar{c}^2 \|A\|_1 + (n-1)^{-1} S_c^2 \{\mathrm{trace}(A) - n^{-1} \|A\|_1\}, \end{aligned}$$
where $\|A\|_1 = \sum_{i=1}^n \sum_{j=1}^n a_{ij}$ is the sum of all entries of A.
4. Show that the mean of $S_{c_\pi}^2 = \sum_{i=1}^n (c_\pi - \bar{c})^2$ is
$$E(S_{c_\pi}^2) = S_c^2.$$

A3

Limiting Theorems and Basic Asymptotics

I review the basics of limiting theorems and asymptotic analyses that are useful for this course. See Newey and McFadden (1994) and Van der Vaart (2000) for in-depth discussions.

A3.1 Convergence in probability and distribution

Definition 1 *Random vectors $Z_n \in \mathbb{R}^K$ converge to Z in probability, denoted by $Z_n \to Z$ in probability, if for all $c > 0$,*

$$\text{pr}\left\{\|Z_n - Z\| > c\right\} \to 0, \quad n \to \infty.$$

This definition incorporates the classic definition of convergence of non-random vectors:

Proposition A3.1 *If non-random vectors $Z_n \to Z$, the convergence also holds in probability.*

Convergence in probability for random vectors is equivalent to element-wise convergence because of the following result:

Proposition A3.2 *If $Z_n \to Z$ and $W_n \to W$ in probability, then $(Z_n, W_n) \to (Z, W)$ in probability.*

For an IID sequence of random vectors, we have the following Khintchine weak law of large numbers:

Proposition A3.3 *If $Z_1, \ldots, Z_n$ are IID with mean $\mu \in \mathbb{R}^K$, then $n^{-1}\sum_{i=1}^n Z_i \to \mu$ in probability.*

A more elementary tool is Markov's inequality:

$$\text{pr}\left\{\|Z_n - Z\| > c\right\} \leq E\left\{\|Z_n - Z\|\right\}/c \tag{A3.1}$$

or

$$\text{pr}\left\{\|Z_n - Z\| > c\right\} \leq E\left\{\|Z_n - Z\|^2\right\}/c^2. \tag{A3.2}$$

Inequality (A3.1) is useful if $E\left\{\|Z_n - Z\|\right\}$ converges to zero, and inequality (A3.2) is useful if $E\left\{\|Z_n - Z\|^2\right\}$ converges to zero. The latter gives a

standard tool for establishing convergence in probability by showing that the covariance matrix converges to zero.

Proposition A3.4 *If random vectors $Z_n \in \mathbb{R}^K$ have mean zero and covariance $\mathrm{cov}(Z_n) = a_n C_n$ where $a_n \to 0$ and $C_n \to C < \infty$, then $Z_n \to 0$ in probability.*

Proof of Proposition A3.4: Using (A3.2), we have

$$\begin{aligned}
\mathrm{pr}\left\{\|Z_n\| > c\right\} &\le c^{-2} E\left\{\|Z_n\|^2\right\} \\
&= c^{-2} E\left(Z_n^{\mathrm{T}} Z_n\right) \\
&= c^{-2} \mathrm{trace}\left\{E\left(Z_n Z_n^{\mathrm{T}}\right)\right\} \\
&= c^{-2} \mathrm{trace}\left\{\mathrm{cov}(Z_n)\right\} \\
&= c^{-2} a_n \mathrm{trace}(C_n) \to 0,
\end{aligned}$$

which implies that $Z_n \to 0$ in probability. □

For example, we usually use Proposition A3.4 to show the weak law of large numbers for the sample mean of independent random variables $\bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$. If we can show that

$$\mathrm{cov}(\bar{Z}_n) = n^{-2} \sum_{i=1}^n \mathrm{cov}(Z_i) \to 0, \tag{A3.3}$$

then $\bar{Z}_n - n^{-1} \sum_{i=1}^n E(Z_i) \to 0$ in probability. The condition in (A3.3) holds if $n^{-1} \sum_{i=1}^n \mathrm{cov}(Z_i)$ converges to a constant matrix.

Definition 2 *Random vectors $Z_n \in \mathbb{R}^K$ converge to Z in distribution, if for all every continuous point z of the function $t \to \mathrm{pr}(Z \le t)$,*

$$\mathrm{pr}(Z_n \le z) \to \mathrm{pr}(Z \le z), \quad n \to \infty.$$

When the limit is a constant, we have equivalence of convergences in probability and distribution:

Proposition A3.5 *If c is a non-random vector, $Z_n \to c$ in probability is equivalent to $Z_n \to c$ in distribution.*

For IID sequences of random vectors, we have the Lindeberg–Lévy CLT:

Proposition A3.6 *If random vectors $Z_1, \ldots, Z_n$ are IID with mean μ and covariance Σ, then $n^{1/2}(\bar{Z}_n - \mu) = n^{-1/2} \sum_{i=1}^n (Z_i - \mu) \to \mathrm{N}(0, \Sigma)$ in distribution.*

The more general Lindeberg–Feller CLT holds for independent sequences of random vectors:

Proposition A3.7 *For each n, let $Z_{n1}, \ldots, Z_{n,k_n}$ be independent random vectors with finite variances such that*

(LF1) $\sum_{i=1}^{k_n} E\left[\|Z_{ni}\|^2 1\left\{\|Z_{ni}\| > c\right\}\right] \to 0$ *for every* $c > 0$*;*

(LF2) $\sum_{i=1}^{k_n} \text{cov}(Z_{ni}) \to \Sigma$.

Then $\sum_{i=1}^{k_n} \left\{Z_{ni} - E(Z_{ni})\right\} \to \text{N}(0, \Sigma)$ *in distribution.*

Condition (LF2) often holds by proper standardization, and the key is to verify Condition (LF1). Condition (LF1) is general but it looks cumbersome. In many cases, we impose a stronger condition that is easier to verify:

(LF'1) $\sum_{i=1}^{k_n} E\|Z_{ni}\|^{2+\delta} \to 0$ for some $\delta > 0$.

We can show that 1 implies that 1:

$$\begin{aligned}\sum_{i=1}^{k_n} E\left[\|Z_{ni}\|^2 1\left\{\|Z_{ni}\| > c\right\}\right] &= \sum_{i=1}^{k_n} E\left[\|Z_{ni}\|^{2+\delta}\|Z_{ni}\|^{-\delta} 1\left\{\|Z_{ni}\|^{\delta} > c^{\delta}\right\}\right] \\ &\leq \sum_{i=1}^{k_n} E\|Z_{ni}\|^{2+\delta} c^{-\delta} \to 0.\end{aligned}$$

Condition (LF'1) is called the Lyapunov condition.

A3.2 Tools for proving convergence in probability and distribution

The first tool is the continuous mapping theorem:

Proposition A3.8 *Let* $f : \mathbb{R}^K \to \mathbb{R}^L$ *be continuous except on a set* O *with* $\text{pr}(Z \in O) = 0$*. Then* $Z_n \to Z$ *implies* $f(Z_n) \to f(Z)$ *in probability (and in distribution).*

The second tool is Slutsky's Theorem:

Proposition A3.9 *Let* Z_n *and* W_n *be random vectors. If* $Z_n \to Z$ *in distribution, and* $W_n \to c$ *in probability (or in distribution) for a constant* c*, then*

1. $Z_n + W_n \to Z + c$ *in distribution;*
2. $W_n Z_n \to cZ$ *in distribution;*
3. $W_n^{-1} Z_n \to c^{-1} Z$ *in distribution if* $c \neq 0$.

The third tool is the delta-method. I will present a special case below for asymptotically Normal random vectors. Heuristically, it states that if T_n is asymptotically Normal, then any function of T_n is also asymptotically Normal. This is true because any function is a locally linear function by the first order Taylor expansion.

Proposition A3.10 *Let $f(z)$ be a function from $\mathbb{R}^p$ to $\mathbb{R}^q$, and $\partial f(z)/\partial z \in \mathbb{R}^{p\times q}$ be the partial derivative matrix. If $\sqrt{n}(Z_n-\theta) \to \mathrm{N}(\mu,\Sigma)$ in distribution, then*

$$\sqrt{n}\{f(Z_n)-f(\theta)\} \to \mathrm{N}\left(\frac{\partial f(\theta)}{\partial z^{\mathrm{T}}}\mu, \frac{\partial f(\theta)}{\partial z^{\mathrm{T}}}\Sigma\frac{\partial f(\theta)}{\partial z}\right)$$

in distribution.

Proof of Proposition A3.10: I will give an informal proof. Using Taylor expansion, we have

$$\sqrt{n}\{f(Z_n)-f(\theta)\} \cong \frac{\partial f(\theta)}{\partial z^{\mathrm{T}}}\sqrt{n}(Z_n-\theta),$$

which is a linear transformation of $\sqrt{n}(Z_n-\theta)$. Because $\sqrt{n}(Z_n-\theta) \cong \mathrm{N}(\mu,\Sigma)$, we have

$$\sqrt{n}\{f(Z_n)-f(\theta)\} \cong \frac{\partial f(\theta)}{\partial z^{\mathrm{T}}}\mathrm{N}(\mu,\Sigma) = \mathrm{N}\left(\frac{\partial f(\theta)}{\partial z^{\mathrm{T}}}\mu, \frac{\partial f(\theta)}{\partial z^{\mathrm{T}}}\Sigma\frac{\partial f(\theta)}{\partial z}\right).$$

□

A3.3 M-estimation

Theorem A3.1 *Assume that $\{w_i\}_{i=1}^n$ are independent observations. The true parameter $\beta \in \mathbb{R}^p$ is the unique solution of*

$$E\{\bar{m}(W,b)\} = 0,$$

where $\bar{m}(W,b) = n^{-1}\sum_{i=1}^n m(w_i,b)$. The estimator $\hat{\beta} \in \mathbb{R}^p$ is the solution of

$$\bar{m}(W,b) = 0,$$

Under some regularity conditions, $\hat{\beta}$ is consistent for β and asymptotically Normal with

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \to \mathrm{N}(0, B^{-1}MB^{-\mathrm{T}})$$

in distribution, where

$$B = -n^{-1}\sum_{i=1}^n \frac{\partial E\{m(w_i,\beta)\}}{\partial b^{\mathrm{T}}} = -n^{-1}\sum_{i=1}^n \left(\begin{array}{ccc} \frac{\partial E\{m(w_i,\beta)\}}{\partial b_1} & \cdots & \frac{\partial E\{m(w_i,\beta)\}}{\partial b_p} \end{array} \right)$$

and

$$M = n^{-1}\sum_{i=1}^n \mathrm{cov}\{m(w_i,\beta)\}$$

are both $p \times p$ matrices.

Proof of Theorem A3.1: I give a "physics" proof; see Newey and McFadden (1994) for a rigorous proof. When I use approximations, I mean the error terms are of higher orders under some regularity conditions. The consistency follows from swapping the order of "solving equation" and "taking the limit based on law of large numbers":

$$\begin{aligned}\lim_{n\to\infty}\hat{\beta} &= \lim_{n\to\infty}\{\text{solve } \bar{m}(W,b)=0\}\\ &= \text{solve } \left\{\lim_{n\to\infty}\bar{m}(W,b)=0\right\}\\ &= \text{solve } [E\{\bar{m}(W,b)\}=0]\\ &= \beta.\end{aligned}$$

The asymptotic Normality follows from three steps. First, from Taylor expansion

$$0=\bar{m}(W,\hat{\beta})\cong\bar{m}(W,\beta)+\frac{\partial\bar{m}(W,\beta)}{\partial b}(\hat{\beta}-\beta)$$

we obtain

$$\sqrt{n}\left(\hat{\beta}-\beta\right)\cong\left\{-\frac{\partial\bar{m}(W,\beta)}{\partial b}\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}m(w_i,\beta)\right\}.$$

Second, the law of large numbers ensures that

$$-\frac{\partial\bar{m}(W,\beta)}{\partial b}\to-\frac{\partial E\{\bar{m}(W,\beta)\}}{\partial b^{\mathrm{T}}}=B$$

in probability, and the CLT ensures that $n^{-1/2}\sum_{i=1}^{n}m(w_i,\beta)\to\ \mathrm{N}(0,M)$ in distribution. Finally, Slutsky's Theorem implies that $\sqrt{n}\left(\hat{\beta}-\beta\right)\ \to$ $\mathrm{N}(0,B^{-1}MB^{-\mathrm{T}})$. in distribution. □

Corollary A3.1 *Assume that* $\{w_i\}_{i=1}^{n}$ *are IID with the same distribution as* w*. The true parameter* $\beta\in\mathbb{R}^p$ *is the unique solution of*

$$E\{m(w,b)\}=0,$$

and the estimator $\hat{\beta}\in\mathbb{R}^p$ *is the solution of*

$$\bar{m}(W,b)=0,$$

The same conclusion holds with

$$B=\frac{\partial E\{m(w,\beta)\}}{\partial b^{\mathrm{T}}}=\left(\begin{array}{ccc}\frac{\partial E\{m(w,\beta)\}}{\partial b_1} & \cdots & \frac{\partial E\{m(w,\beta)\}}{\partial b_p}\end{array}\right)$$

and

$$M=E\{m(w,\beta)m(w,\beta)^{\mathrm{T}}\}.$$

Example A3.1 *An an important application, we can derive the asymptotic properties of the MLE* $\hat{\theta}$ *under IID sampling from a parametric model* $y_1, \ldots, y_n \overset{\text{IID}}{\sim} f(y \mid \theta)$*. The MLE satisfies the following estimating equation:*

$$E\left\{\frac{\partial \log f(y \mid \theta)}{\partial \theta}\right\} = 0, \tag{A3.4}$$

which is Bartlett's first identity. Under regularity conditions, $\sqrt{n}(\hat{\theta} - \theta)$ *converges in distribution to Normal with mean zero and covariance* $B^{-1}MB^{-1}$*, where*

$$B = -\frac{\partial}{\partial \theta^{\mathrm{T}}} E\left\{\frac{\partial \log f(y \mid \theta)}{\partial \theta}\right\} = E\left\{-\frac{\partial^2 \log f(y \mid \theta)}{\partial \theta \partial \theta^{\mathrm{T}}}\right\}$$

is called the Fisher information matrix, denoted by $I(\theta)$*, and*

$$M = E\left\{\frac{\partial \log f(y \mid \theta)}{\partial \theta}\frac{\partial \log f(y \mid \theta)}{\partial \theta^{\mathrm{T}}}\right\}$$

is sometimes also called the Fisher information matrix, denoted by $J(\theta)$*.*

If the model is correct, Bartlett's second identity ensures that

$$I(\theta) = J(\theta), \tag{A3.5}$$

and therefore $\sqrt{n}(\hat{\theta} - \theta)$ *converges in distribution to Normal with mean zero and covariance* $I(\theta)^{-1} = J(\theta)^{-1}$*. So a covariance matrix estimator for the MLE is* $I_n(\hat{\theta})^{-1}$ *or* $J_n(\hat{\theta})^{-1}$*, where*

$$I_n(\hat{\theta}) = -\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i \mid \hat{\theta})}{\partial \theta^2}$$

and

$$J_n(\hat{\theta}) = \sum_{i=1}^{n} \frac{\partial \log f(y_i \mid \hat{\theta})}{\partial \theta}\frac{\partial \log f(y_i \mid \hat{\theta})}{\partial \theta^{\mathrm{T}}}.$$

If the model is incorrect, $I(\theta)$ *can be different from* $J(\theta)$ *but the sandwich covariance still holds. So a covariance matrix estimator for the MLE under misspecification is* $I_n(\hat{\theta})^{-1}J_n(\hat{\theta})I_n(\hat{\theta})^{-1}$*.*

A3.4 Parametric tests

Wald, LRT, Score

A3.5 Homework problems

A3.1 Bartlett's identities

Verify (A3.4) and (A3.5).

A3.2 Function of the MLE

Derive the asymptotic distribution of $g(\hat{\theta})$ where $\hat{\theta}$ is the MLE in Example A3.1.

A3.3 A misspecified Exponential model

Assume that $y_1, \ldots, y_n \overset{\text{IID}}{\sim}$ Exponential distribution with mean μ. Find the MLE of μ and its asymptotic variance estimators under correctly specified and incorrectly specified model.

Bibliography

Agresti, A. (2010). *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons.

Ai, C. and Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics letters*, 80(1):123–129.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Anton, R. F., O'Malley, S. S., Ciraulo, D. A., Cisler, R. A., Couper, D., Donovan, D. M., Gastfriend, D. R., Hosking, J. D., Johnson, B. A., LoCastro, J. S., et al. (2006). Combined pharmacotherapies and behavioral interventions for alcohol dependence: the combine study: a randomized controlled trial. *Jama*, 295(17):2003–2017.

Berrington de González, A. and Cox, D. R. (2007). Interpretation of interaction: A review. *Ann. Appl. Statist.*, 1:371–385.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. (2019a). Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, 34:523–544.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., Zhao, L., et al. (2019b). Models as approximations ii: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565.

Burridge, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):41–45.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6):2453–2483.

Carpenter, D. P. (2002). Groups, the media, agency waiting costs, and fda drug approval. *American Journal of Political Science*, pages 490–505.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and weighting in regression*, volume 30. CRC Press.

Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Supplement to the Journal of the Royal Statistical Society*, 5(2):171–176.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19:15–18.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cox, D. R. (1984). Interaction. *International Statistical Review*, 52:1–24.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of applied Econometrics*, 12(3):313–336.

Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, 72(357):77–91.

DiCiccio, C. J., Romano, J. P., and Wolf, M. (2019). Improving weighted least squares inference. *Econometrics and Statistics*, 10:96–119.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70:892–898.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 59–82. Berkeley, CA: University of California Press.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66. CRC Press.

Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh by Oliver and Boyd, 1st edition.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9:1218–1228.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, 37:152–155.

Freedman, D. A. (2006). On the so-called "huber sandwich estimator" and "robust standard errors". *The American Statistician*, 60(4):299–302.

Freedman, D. A. (2008). Survival analysis: A primer. *The American Statistician*, 62(2):110–119.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4).

Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, 1:387–401.

Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya*, 37(3):117–132.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–224.

Gelman, A. and Park, D. K. (2009). Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician*, 63(1):1–8.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Guiteras, R., Levinsohn, J., and Mobarak, A. M. (2015). Encouraging sanitation investment in the developing world: a cluster-randomized trial. *Science*, 348(6237):903–906.

Heckman, J. J. and Singer, B. (1984). Econometric duration analysis. *Journal of Econometrics*, 24(1-2):63–132.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1):13.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19:285–292.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1:69–88.

Hoerl, A. E., Kannard, R. W., and Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, 4(2):105–123.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Cam, L. M. L. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. Berkeley, California: University of California Press.

Kagan, A. (2001). A note on the logistic link function. *Biometrika*, 88(2):599–601.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Keele, L. (2010). Proportionally difficult: testing for nonproportional hazards in cox models. *Political Analysis*, 18(2):189–205.

King, G. (2013). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton: Princeton University Press.

King, G. and Roberts, M. E. (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, 23(2):159–179.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, pages 33–50.

Lawless, J. F. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics-theory and Methods*, 5(4):307–323.

Le Cam, L. et al. (1960). An approximation theorem for the poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Lin, D.-Y., Gong, J., Gallo, P., Bunn, P. H., and Couper, D. (2016). Simultaneous inference on treatment effects in survival studies with factorial designs. *Biometrics*, 72(4):1078–1085.

Long, J. S. and Ervin, L. H. (1998). Correcting for heteroscedasticity with heteroscedasticity consistent standard errors in the linear regression model: Small sample considerations. *Indiana University, Bloomington, IN*, 47405.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58:993–1010.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325.

Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):115–126.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, 50:163–170.

Miller, R. G. J. (1974). An unbalanced jackknife. *The Annals of Statistics*, 2:880–891.

Moen, E. L., Fricano-Kugler, C. J., Luikart, B. W., and O'Malley, A. J. (2016). Analyzing clustered data: why and how to account for multiple observations nested within a study participant? *Plos one*, 11(1).

Newey, K. and McFadden, D. (1994). Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pages 2112–2245.

Powell, J. L. (1991). Estimation of monotonic regression models under quantile restrictions. *Nonparametric and semiparametric methods in Econometrics*, pages 357–384.

Pratt, J. W. (1981). Concavity of the log likelihood. *Journal of the American Statistical Association*, 76(373):103–106.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11:68–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43(3/4):353–360.

Romano, J. P. and Wolf, M. (2017). Resurrecting weighted least squares. *Journal of Econometrics*, 197(1):1–19.

Royer, H., Stehr, M., and Sydnor, J. (2015). Incentives, commitments, and habit formation in exercise: evidence from a field experiment with workers at a fortune-500 company. *American Economic Journal: Applied Economics*, 7(3):51–84.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2:808–840.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270.

Samarani, S., Mack, D. R., Bernstein, C. N., Iannello, A., Debbeche, O., Jantchou, P., Faure, C., Deslandres, C., Amre, D. K., and Ahmad, A. (2019). Activating killer-cell immunoglobulin-like receptor genes confer risk for crohn's disease in children and adults of the western european descent: Findings based on case-control studies. *PloS one*, 14(6).

Schochet, P. Z. (2013). Estimators for clustered education rcts using the neyman model for causal inference. *Journal of Educational and Behavioral Statistics*, 38(3):219–238.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica sinica*, 7:221–242.

Sims, C. A. (2010). But economics is not an experimental science. *Journal of Economic Perspectives*, 24:59–68.

Sullivan Pepe, M. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in statistics-simulation and computation*, 23(4):939–951.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Titterington, D. (2013). *Biometrika* highlights from volume 28 onwards. *Biometrika*, pages 17–73.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. New York: Springer.

Tukey, J. (1958). Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29:614.

Turner, E. L., Li, F., Gallis, J. A., Prague, M., and Murray, D. M. (2017a). Review of recent methodological developments in group-randomized trials: part 1—design. *American Journal of Public Health*, 107:907–915.

Turner, E. L., Prague, M., Gallis, J. A., Li, F., and Murray, D. M. (2017b). Review of recent methodological developments in group-randomized trials: part 2—analysis. *American Journal of Public Health*, 107:1078–1086.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.

Weber, N. C. (1986). The jackknife and heteroskedasticity: Consistent variance estimation for regression models. *Economics Letters*, 20(2):161–163.

White, H. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838.

White, H. (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, 21:149–170.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, pages 1–25.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT press.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14(4):1261–1295.

Yule, G. U. (1907). On the theory of correlation for any number of variables, treated by a new system of notation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 79:182–193.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in r. *Journal of statistical software*, 27(8):1–25.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.