Chapter 3

The Algebra of Least Squares

3.1 Introduction

In this chapter we introduce the popular least-squares estimator. Most of the discussion will be algebraic, with questions of distribution and inference defered to later chapters.

3.2 Random Samples

In Section 2.18 we derived and discussed the best linear predictor of y given $\boldsymbol{x}$ for a pair of random variables $(y, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^k$, and called this the linear projection model. We are now interested in **estimating** the parameters of this model, in particular the projection coefficient

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x}y\right).$$

We can estimate $\boldsymbol{\beta}$ from observational data which includes joint measurements on the variables $(y, \boldsymbol{x})$. For example, supposing we are interested in estimating a wage equation, we would use a dataset with observations on wages (or weekly earnings), education, experience (or age), and demographic characteristics (gender, race, location). One possible dataset is the Current Population Survey (CPS), a survey of U.S. households which includes questions on employment, income, education, and demographic characteristics.

Notationally we wish to emphasize when we are discussing observations. Typically in econometrics we denote observations by appending a subscript i which runs from 1 to n, thus the i^{th} observation is $(y_i, \boldsymbol{x}_i)$, and n denotes the sample size. The dataset is then $\{(y_i, \boldsymbol{x}_i);\ i = 1, ..., n\}$.

From the viewpoint of empirical analysis, a dataset is a array of numbers often organized as a table, where the columns of the table correspond to distinct variables and the rows correspond to distinct observations. For empirical analysis, the dataset and observations are fixed in the sense that they are numbers presented to the researcher. For statistical analysis we need to view the dataset as random, or more precisely as a realization of a random process. For cross-sectional studies, the most common approach is to treat the individual observations as independent draws from an underlying population F. When the observations are realizations of independent and identically distributed random variables, we say that the data is a random sample.

Assumption 3.2.1 *The observations* $\{(y_1, \boldsymbol{x}_1), ..., (y_i, \boldsymbol{x}_i), ..., (y_n, \boldsymbol{x}_n)\}$ *are a random sample.*

With a random sample, the ordering of the data is irrelevant. There is nothing special about any specific observation or ordering. You can permute the order of the observations and no information is gained or lost.

As most economic data sets are not literally the result of a random experiment, the random sampling framework is best viewed as an approximation rather than being literally true.

The linear projection model applies to the random observations $(y_i, \boldsymbol{x}_i)$. This means that the probability model for the observations is the same as that described in Section 2.18. We can write the model as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{3.1}$$

where the linear projection coefficient $\boldsymbol{\beta}$ is defined as

$$\boldsymbol{\beta} = \underset{b\in\mathbb{R}^k}{\operatorname{argmin}} S(\boldsymbol{b}), \tag{3.2}$$

the minimizer of the expected squared error

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2, \tag{3.3}$$

and has the explicit solution

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x_i}\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i y_i\right). \tag{3.4}$$

3.3 Sample Means

Consider the intercept-only model

$$\begin{aligned} y_i &= \mu + e_i \\ \mathbb{E}(e_i) &= 0. \end{aligned}$$

In this case the regression parameter is the unconditional mean $\mu = \mathbb{E}(y_i)$.

The standard estimator of a population mean is the sample mean, namely

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The sample mean is the empirical analog of the population mean, and is the conventional estimator in the lack of other information about μ or the distribution of y. We call $\widehat{\mu}$ the **moment estimator** for μ.

Indeed, whenever we have a parameter which can be written as the expectation of a function of random variables, a natural estimator of the parameter is the moment estimator, which is the sample mean of the corresponding function of the observations. For example, for $\mu_2 = \mathbb{E}(y_i^2)$ the moment estimator is $\widehat{\mu}_2 = \dfrac{1}{n}\sum_{i=1}^n y_i^2$, and for $\theta = \mathbb{E}(y_{1i}y_{2i})$ the moment estimator is $\widehat{\theta} = \dfrac{1}{n}\sum_{i=1}^n y_{1i}y_{2i}$.

3.4 Least Squares Estimator

The linear projection coefficient $\boldsymbol{\beta}$ is defined in (3.2) as the minimizer of the expected squared error $S(\boldsymbol{\beta})$ defined in (3.3). For given $\boldsymbol{\beta}$, the expected squared error is the expectation of the squared error $\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$. The moment estimator of $S(\boldsymbol{\beta})$ is the sample average:

$$\begin{aligned} S_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\ &= \frac{1}{n}SSE_n(\boldsymbol{\beta}) \end{aligned} \tag{3.5}$$

where

$$SSE_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

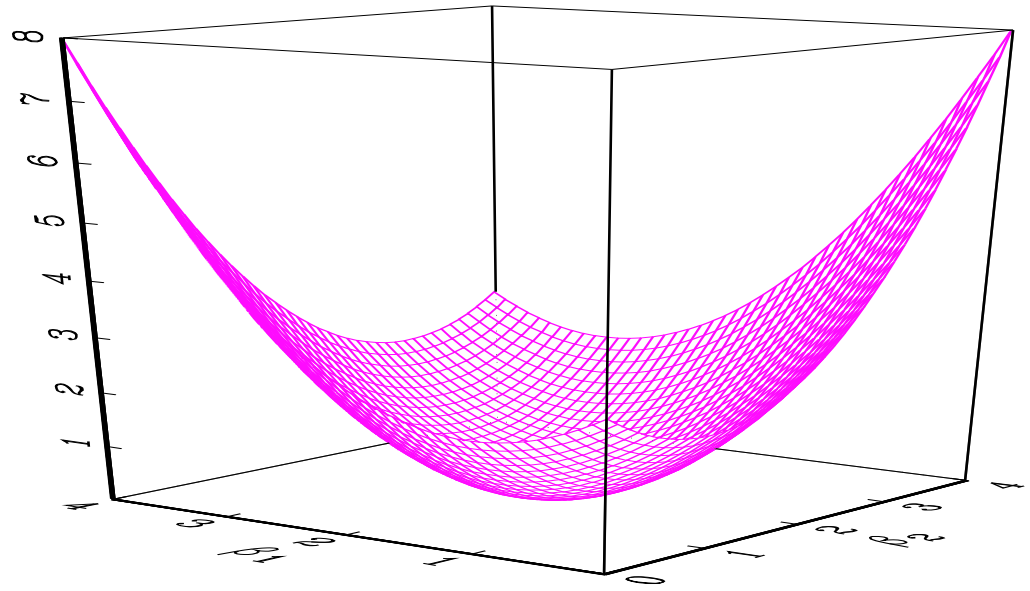

Figure 3.1: Sum-of-Squared Errors Function

is called the **sum-of-squared-errors** function.

Since $S_n(\boldsymbol{\beta})$ is a sample average, we can interpret it as an estimator of the expected squared error $S(\boldsymbol{\beta})$. Examining $S_n(\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$ therefore is informative about how $S(\boldsymbol{\beta})$ varies with $\boldsymbol{\beta}$. The projection coefficient coeffient minimizes $S(\boldsymbol{\beta})$, an analog estimator minimizes (3.5):

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{\operatorname{argmin}} S_n(\boldsymbol{\beta}).$$

Alternatively, as $S_n(\boldsymbol{\beta})$ is a scale multiple of $SSE_n(\boldsymbol{\beta})$, we may equivalently define $\widehat{\boldsymbol{\beta}}$ as the minimizer of $SSE_n(\boldsymbol{\beta})$. Hence $\widehat{\boldsymbol{\beta}}$ is commonly called the **least-squares (LS) (or ordinary least squares (OLS)) estimator** of $\boldsymbol{\beta}$. Here, as is common in econometrics, we put a hat "^" over the parameter $\boldsymbol{\beta}$ to indicate that $\widehat{\boldsymbol{\beta}}$ is a sample estimate of $\boldsymbol{\beta}$. This is a helpful convention, as just by seeing the symbol $\widehat{\boldsymbol{\beta}}$ we can immediately interpret it as an estimator (because of the hat), and as an estimator of a parameter labelled $\boldsymbol{\beta}$. Sometimes when we want to be explicit about the estimation method, we will write $\widehat{\boldsymbol{\beta}}_{\text{ols}}$ to signify that it is the OLS estimator. It is also common to see the notation $\widehat{\boldsymbol{\beta}}_n$, where the subscript "$n$" indicates that the estimator depends on the sample size n.

It is important to understand the distinction between population parameters such as $\boldsymbol{\beta}$ and sample estimates such as $\widehat{\boldsymbol{\beta}}$. The population parameter $\boldsymbol{\beta}$ is a non-random feature of the population while the sample estimate $\widehat{\boldsymbol{\beta}}$ is a random feature of a random sample. $\boldsymbol{\beta}$ is fixed, while $\widehat{\boldsymbol{\beta}}$ varies across samples.

To visualize the quadratic function $S_n(\boldsymbol{\beta})$, Figure 3.1 displays an example sum-of-squared errors function $SSE_n(\boldsymbol{\beta})$ for the case $k = 2$. The least-squares estimator $\widehat{\boldsymbol{\beta}}$ is the the pair $(\widehat{\beta}_1, \widehat{\beta}_2)$ minimizing this function.

3.5 Solving for Least Squares with One Regressor

For simplicity, we start by considering the case $k = 1$ so that the coefficient β is a scalar. Then the sum of squared errors is a simple quadratic

$$SSE_n(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2$$
$$= \left(\sum_{i=1}^{n} y_i^2\right) - 2\beta\left(\sum_{i=1}^{n} x_i y_i\right) + \beta^2\left(\sum_{i=1}^{n} x_i^2\right).$$

The OLS estimator $\widehat{\beta}$ minimizes this function. From elementary algebra we know that the minimizer of the quadratic function $a - 2bx + cx^2$ is $x = b/c$. Thus the minimizer of $SSE_n(\beta)$ is

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{3.6}$$

The intercept-only model is the special case $x_i = 1$. In this case we find

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} 1 y_i}{\sum_{i=1}^{n} 1^2} = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}, \tag{3.7}$$

the sample mean of y_i. Here, as is common, we put a bar "$-$" over y to indicate that the quantity is a sample mean. This calculation shows that the OLS estimator in the intercept-only model is the sample mean.

3.6 Solving for Least Squares with Multiple Regressors

We now consider the case with $k \geq 1$ so that the coefficient $\boldsymbol{\beta}$ is a vector.

To solve for $\widehat{\boldsymbol{\beta}}$, expand the SSE function to find

$$SSE_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i^2 - 2\boldsymbol{\beta}' \sum_{i=1}^{n} \boldsymbol{x}_i y_i + \boldsymbol{\beta}' \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \boldsymbol{\beta}.$$

This is a quadratic expression in the vector argument $\boldsymbol{\beta}$. The first-order-condition for minimization of $SSE_n(\boldsymbol{\beta})$ is

$$0 = \frac{\partial}{\partial \boldsymbol{\beta}} SSE_n(\widehat{\boldsymbol{\beta}}) = -2\sum_{i=1}^{n} \boldsymbol{x}_i y_i + 2\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}. \tag{3.8}$$

We have written this using a single expression, but it is actually a system of k equations with k unknowns (the elements of $\widehat{\boldsymbol{\beta}}$).

The solution for $\widehat{\boldsymbol{\beta}}$ may be found by solving the system of k equations in (3.8). We can write this solution compactly using matrix algebra. Inverting the $k \times k$ matrix $\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$ we find an explicit formula for the least-squares estimator

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right). \tag{3.9}$$

This is the natural estimator of the best linear projection coefficient $\boldsymbol{\beta}$ defined in (3.2), and can also be called the linear projection estimator.

We see that (3.9) simplifies to the expression (3.6) when $k = 1$. The expression (3.9) is a notationally simple generalization but requires a careful attention to vector and matrix manipulations.

Alternatively, equation (3.4) writes the projection coefficient $\boldsymbol{\beta}$ as an explicit function of the population moments $\boldsymbol{Q}_{xy}$ and $\boldsymbol{Q}_{xx}$. Their moment estimators are the sample moments

$$\widehat{\boldsymbol{Q}}_{xy} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i$$
$$\widehat{\boldsymbol{Q}}_{xx} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'.$$

The moment estimator of $\boldsymbol{\beta}$ replaces the population moments in (3.4) with the sample moments:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{Q}}_{xx}^{-1}\widehat{\boldsymbol{Q}}_{xy} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right) \\ &= \left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right)\end{aligned}$$

which is identical with (3.9).

Least Squares Estimation

Definition 3.6.1 *The* ***least-squares estimator*** $\widehat{\boldsymbol{\beta}}$ *is*

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\operatorname{argmin}} S_n(\boldsymbol{\beta})$$

where

$$S_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

and has the solution

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right).$$

Adrien-Marie Legendre

The method of least-squares was first published in 1805 by the French mathematician Adrien-Marie Legendre (1752-1833). Legendre proposed least-squares as a solution to the algebraic problem of solving a system of equations when the number of equations exceeded the number of unknowns. This was a vexing and common problem in astronomical measurement. As viewed by Legendre, (3.1) is a set of n equations with k unknowns. As the equations cannot be solved exactly, Legendre's goal was to select $\boldsymbol{\beta}$ to make the set of errors as small as possible. He proposed the sum of squared error criterion, and derived the algebraic solution presented above. As he noted, the first-order conditions (3.8) is a system of k equations with k unknowns, which can be solved by "ordinary" methods. Hence the method became known as **Ordinary Least Squares** and to this day we still use the abbreviation OLS to refer to Legendre's estimation method.

3.7 Illustration

We illustrate the least-squares estimator in practice with the data set used to generate the estimates from Chapter 2. This is the March 2009 Current Population Survey, which has extensive information on the U.S. population. This data set is described in more detail in Section 3.19. For this illustration, we use the sub-sample of married (spouse present) black female wages earners with 12 years potential work experience. This sub-sample has 20 observations. Let y_i be log wages and $\boldsymbol{x}_i$ be years of education and an intercept. Then

$$\sum_{i=1}^{n} \boldsymbol{x}_i y_i = \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix},$$

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix},$$

and

$$\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} = \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix}$$

Thus

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix} \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix}$$

$$= \begin{pmatrix} 0.155 \\ 0.698 \end{pmatrix}. \tag{3.10}$$

We often write the estimated equation using the format

$$\widehat{\log(Wage)} = 0.155\ education + 0.698. \tag{3.11}$$

An interpretation of the estimated equation is that each year of education is associated with an 16% increase in mean wages.

Equation (3.11) is called a **bivariate regression** as there are only two variables. A **multivariate regression** has two or more regressors, and allows a more detailed investigation. Let's take an example similar to (3.11) but include all levels of experience. This time, we use the sub-sample of single (never married) asian men, which has 268 observations. Including as regressors years of potential work experience (*experience*) and its square ($experience^2/100$) (we divide by 100 to simplify reporting), we obtain the estimates

$$\widehat{\log(Wage)} = 0.143\ education + 0.036\ experience - 0.071\ experience^2/100 + 0.575. \tag{3.12}$$

These estimates suggest a 14% increase in mean wages per year of education, holding experience constant.

3.8 Least Squares Residuals

As a by-product of estimation, we define the **fitted value**

$$\hat{y}_i = \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}$$

and the **residual**

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}. \tag{3.13}$$

Sometimes $\hat{y}_i$ is called the predicted value, but this is a misleading label. The fitted value $\hat{y}_i$ is a function of the entire sample, including y_i, and thus cannot be interpreted as a valid prediction of y_i. It is thus more accurate to describe $\hat{y}_i$ as a *fitted* rather than a *predicted* value.

Note that $y_i = \hat{y}_i + \hat{e}_i$ and

$$y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \hat{e}_i. \tag{3.14}$$

We make a distinction between the **error** e_i and the **residual** $\hat{e}_i$. The error e_i is unobservable while the residual $\hat{e}_i$ is a by-product of estimation. These two variables are frequently mislabeled, which can cause confusion.

Equation (3.8) implies that

$$\sum_{i=1}^{n} \boldsymbol{x}_i\hat{e}_i = \mathbf{0}. \tag{3.15}$$

To see this by a direct calculation, using (3.13) and (3.9),

$$\begin{aligned}
\sum_{i=1}^{n} \boldsymbol{x}_i\hat{e}_i &= \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right) \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i y_i \\
&= \mathbf{0}.
\end{aligned}$$

When $\boldsymbol{x}_i$ contains a constant, an implication of (3.15) is

$$\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0. \tag{3.16}$$

Thus the residuals have a sample mean of zero and the sample correlation between the regressors and the residual is zero. These are algebraic results, and hold true for all linear regression estimates.

3.9 Model in Matrix Notation

For many purposes, including computation, it is convenient to write the model and statistics in matrix notation. The linear equation (2.26) is a system of n equations, one for each observation. We can stack these n equations together as

$$\begin{aligned}
y_1 &= \boldsymbol{x}_1'\boldsymbol{\beta} + e_1 \\
y_2 &= \boldsymbol{x}_2'\boldsymbol{\beta} + e_2 \\
&\vdots \\
y_n &= \boldsymbol{x}_n'\boldsymbol{\beta} + e_n.
\end{aligned}$$

Now define

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}, \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Observe that $\boldsymbol{y}$ and $\boldsymbol{e}$ are $n \times 1$ vectors, and $\boldsymbol{X}$ is an $n \times k$ matrix. Then the system of n equations can be compactly written in the single equation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{3.17}$$

Sample sums can be written in matrix notation. For example

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \boldsymbol{X}'\boldsymbol{X}$$
$$\sum_{i=1}^{n} \boldsymbol{x}_i y_i = \boldsymbol{X}'\boldsymbol{y}.$$

Therefore the least-squares estimator can be written as

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}\right). \tag{3.18}$$

The matrix version of (3.14) and estimated version of (3.17) is

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}},$$

or equivalently the residual vector is

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Using the residual vector, we can write (3.15) as

$$\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}. \tag{3.20}$$

Using matrix notation we have simple expressions for most estimators. This is particularly convenient for computer programming, as most languages allow matrix notation and manipulation.

Important Matrix Expressions

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}\right) \\ \hat{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ \boldsymbol{X}'\hat{\boldsymbol{e}} &= \boldsymbol{0}. \end{aligned}$$

Early Use of Matrices

The earliest known treatment of the use of matrix methods to solve simultaneous systems is found in Chapter 8 of the Chinese text *The Nine Chapters on the Mathematical Art*, written by several generations of scholars from the 10th to 2nd century BCE.

3.10 Projection Matrix

Define the matrix

$$\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'.$$

Observe that

$$\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}.$$

This is a property of a **projection matrix**. More generally, for any matrix $\boldsymbol{Z}$ which can be written as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Gamma}$ for some matrix $\boldsymbol{\Gamma}$ (we say that $\boldsymbol{Z}$ lies in the **range space** of $\boldsymbol{X}$), then

$$\boldsymbol{P}\boldsymbol{Z} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{Z}.$$

As an important example, if we partition the matrix $\boldsymbol{X}$ into two matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ so that

$$\boldsymbol{X} = \left[\boldsymbol{X}_1 \quad \boldsymbol{X}_2\right],$$

then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$. (See Exercise 3.7.)

The matrix $\boldsymbol{P}$ is **symmetric** and **idempotent**[1]. To see that it is symmetric,

$$\begin{aligned}
\boldsymbol{P}' &= \left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)' \\
&= \left(\boldsymbol{X}'\right)'\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)'\left(\boldsymbol{X}\right)' \\
&= \boldsymbol{X}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)'\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{X}\left(\left(\boldsymbol{X}\right)'\left(\boldsymbol{X}'\right)'\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{P}.
\end{aligned}$$

To establish that it is idempotent, the fact that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ implies that

$$\begin{aligned}
\boldsymbol{P}\boldsymbol{P} &= \boldsymbol{P}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{P}.
\end{aligned}$$

The matrix $\boldsymbol{P}$ has the property that it creates the fitted values in a least-squares regression:

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \hat{\boldsymbol{y}}.$$

Because of this property, $\boldsymbol{P}$ is also known as the "hat matrix".

A special example of a projection matrix occurs when $\boldsymbol{X} = \boldsymbol{1}$ is an n-vector of ones. Then

$$\begin{aligned}
\boldsymbol{P}_1 &= \boldsymbol{1}\left(\boldsymbol{1}'\boldsymbol{1}\right)^{-1}\boldsymbol{1}' \\
&= \frac{1}{n}\boldsymbol{1}\boldsymbol{1}'.
\end{aligned}$$

Note that

$$\begin{aligned}
\boldsymbol{P}_1\boldsymbol{y} &= \boldsymbol{1}\left(\boldsymbol{1}'\boldsymbol{1}\right)^{-1}\boldsymbol{1}'\boldsymbol{y} \\
&= \boldsymbol{1}\bar{y}
\end{aligned}$$

creates an n-vector whose elements are the sample mean $\bar{y}$ of y_i.

[1] A matrix $\boldsymbol{P}$ is symmetric if $\boldsymbol{P}' = \boldsymbol{P}$. A matrix $\boldsymbol{P}$ is idempotent if $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$. See Appendix A.8.

The i'th diagonal element of $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ is

$$h_{ii} = \boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i \tag{3.21}$$

which is called the **leverage** of the i'th observation.

Some useful properties of the the matrix $\boldsymbol{P}$ and the leverage values h_{ii} are now summarized.

Theorem 3.10.1

$$\sum_{i=1}^{n} h_{ii} = \operatorname{tr}\boldsymbol{P} = k \tag{3.22}$$

and

$$0 \leq h_{ii} \leq 1. \tag{3.23}$$

To show (3.22),

$$\begin{aligned}\operatorname{tr}\boldsymbol{P} &= \operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\\ &= \operatorname{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\right)\\ &= \operatorname{tr}\left(\boldsymbol{I}_k\right)\\ &= k.\end{aligned}$$

See Appendix A.4 for definition and properties of the trace operator. The proof of (3.23) is defered to Section 3.21.

3.11 Orthogonal Projection

Define

$$\begin{aligned}\boldsymbol{M} &= \boldsymbol{I}_n - \boldsymbol{P}\\ &= \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\end{aligned}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Note that

$$\boldsymbol{M}\boldsymbol{X} = \left(\boldsymbol{I}_n - \boldsymbol{P}\right)\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{P}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}.$$

Thus $\boldsymbol{M}$ and $\boldsymbol{X}$ are orthogonal. We call $\boldsymbol{M}$ an **orthogonal projection matrix** or an **annihilator matrix** due to the property that for any matrix $\boldsymbol{Z}$ in the range space of $\boldsymbol{X}$ then

$$\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{Z} - \boldsymbol{P}\boldsymbol{Z} = \boldsymbol{0}.$$

For example, $\boldsymbol{M}\boldsymbol{X}_1 = \boldsymbol{0}$ for any subcomponent $\boldsymbol{X}_1$ of $\boldsymbol{X}$, and $\boldsymbol{M}\boldsymbol{P} = \boldsymbol{0}$ (see Exercise 3.7).

The orthogonal projection matrix $\boldsymbol{M}$ has many similar properties with $\boldsymbol{P}$, including that $\boldsymbol{M}$ is symmetric ($\boldsymbol{M}' = \boldsymbol{M}$) and idempotent ($\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$). Similarly to (3.22) we can calculate

$$\operatorname{tr}\boldsymbol{M} = n - k. \tag{3.24}$$

(See Exercise 3.9.) While $\boldsymbol{P}$ creates fitted values, $\boldsymbol{M}$ creates least-squares residuals:

$$\boldsymbol{M}\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{e}}. \tag{3.25}$$

As discussed in the previous section, a special example of a projection matrix occurs when $\boldsymbol{X} = \mathbf{1}$ is an n-vector of ones, so that $\boldsymbol{P}_1 = \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'$. Similarly, set

$$\begin{aligned}\boldsymbol{M}_1 &= \boldsymbol{I}_n - \boldsymbol{P}_1 \\ &= \boldsymbol{I}_n - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'.\end{aligned}$$

While $\boldsymbol{P}_1$ creates a vector of sample means, $\boldsymbol{M}_1$ creates demeaned values:

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{y} - \mathbf{1}\bar{y}.$$

For simplicity we will often write the right-hand-side as $\boldsymbol{y} - \bar{y}$. The i'th element is $y_i - \bar{y}$, the **demeaned** value of y_i.

We can also use (3.25) to write an alternative expression for the residual vector. Substituting $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ into $\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{y}$ and using $\boldsymbol{M}\boldsymbol{X} = \mathbf{0}$ we find

$$\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}\right) = \boldsymbol{M}\boldsymbol{e} \tag{3.26}$$

which is free of dependence on the regression coefficient $\boldsymbol{\beta}$.

3.12 Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}e_i^2$ is a moment, so a natural estimator is a moment estimator. If e_i were observed we would estimate σ^2 by

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{3.27}$$

However, this is infeasible as e_i is not observed. In this case it is common to take a two-step approach to estimation. The residuals $\hat{e}_i$ are calculated in the first step, and then we substitute $\hat{e}_i$ for e_i in expression (3.27) to obtain the feasible estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2. \tag{3.28}$$

In matrix notation, we can write (3.27) and (3.28) as

$$\tilde{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e}$$

and

$$\hat{\sigma}^2 = n^{-1}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}. \tag{3.29}$$

Recall the expressions $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{e}$ from (3.25) and (3.26). Applied to (3.29) we find

$$\begin{aligned}\hat{\sigma}^2 &= n^{-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \\ &= n^{-1}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{M}\boldsymbol{y} \\ &= n^{-1}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y} \\ &= n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}\end{aligned}$$

the third equality since $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

An interesting implication is that

$$\begin{aligned}\tilde{\sigma}^2 - \hat{\sigma}^2 &= n^{-1}\boldsymbol{e}'\boldsymbol{e} - n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} \\ &= n^{-1}\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} \\ &\geq 0.\end{aligned}$$

The final inequality holds because $\boldsymbol{P}$ is positive semi-definite and $\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e}$ is a quadratic form. This shows that the feasible estimator $\hat{\sigma}^2$ is numerically smaller than the idealized estimator (3.27).

3.13 Analysis of Variance

Another way of writing (3.25) is

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}. \tag{3.30}$$

This decomposition is **orthogonal**, that is

$$\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = (\boldsymbol{P}\boldsymbol{y})'(\boldsymbol{M}\boldsymbol{y}) = \boldsymbol{y}'\boldsymbol{P}\boldsymbol{M}\boldsymbol{y} = 0.$$

It follows that

$$\boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + 2\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

or

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} \hat{e}_i^2.$$

Subtracting $\bar{y}$ from both sizes of (3.30) we obtain

$$\boldsymbol{y} - \mathbf{1}\bar{y} = \hat{\boldsymbol{y}} - \mathbf{1}\bar{y} + \hat{\boldsymbol{e}}$$

This decomposition is also orthogonal when $\boldsymbol{X}$ contains a constant, as

$$(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})'\hat{\boldsymbol{e}} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} - \bar{y}\mathbf{1}'\hat{\boldsymbol{e}} = 0$$

under (3.16). It follows that

$$(\boldsymbol{y} - \mathbf{1}\bar{y})'(\boldsymbol{y} - \mathbf{1}\bar{y}) = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})'(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

or

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

This is commonly called the **analysis-of-variance** formula for least squares regression.

A commonly reported statistic is the **coefficient of determination** or **R-squared**:

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

It is often described as the fraction of the sample variance of y_i which is explained by the least-squares fit. R^2 is a crude measure of regression fit. We have better measures of fit, but these require a statistical (not just algebraic) analysis and we will return to these issues later. One deficiency with R^2 is that it increases when regressors are added to a regression (see Exercise 3.16) so the "fit" can be always increased by increasing the number of regressors.

3.14 Regression Components

Partition

$$\boldsymbol{X} = [\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}.$$

Then the regression model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}. \tag{3.31}$$

The OLS estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ is obtained by regression of $\boldsymbol{y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ and can be written as

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}. \tag{3.32}$$

We are interested in algebraic expressions for $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$.

The algebra for the estimator is identical as that for the population coefficients as presented in Section 2.21.

Partition $\widehat{\boldsymbol{Q}}_{xx}$ and $\widehat{\boldsymbol{Q}}_{xy}$ as

$$\widehat{\boldsymbol{Q}}_{xx} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2 \\ \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}$$

and similarly $\boldsymbol{Q}_{xy}$

$$\widehat{\boldsymbol{Q}}_{xy} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1y} \\ \widehat{\boldsymbol{Q}}_{2y} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{y} \\ \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.4)

$$\widehat{\boldsymbol{Q}}_{xx}^{-1} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix}^{-1} \overset{def}{=} \begin{bmatrix} \widehat{\boldsymbol{Q}}^{11} & \widehat{\boldsymbol{Q}}^{12} \\ \widehat{\boldsymbol{Q}}^{21} & \widehat{\boldsymbol{Q}}^{22} \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot1}^{-1} \end{bmatrix} \tag{3.33}$$

where $\widehat{\boldsymbol{Q}}_{11\cdot2} = \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}$ and $\widehat{\boldsymbol{Q}}_{22\cdot1} = \widehat{\boldsymbol{Q}}_{22} - \widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{12}$.

Thus

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} \\ &= \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot1}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1y} \\ \widehat{\boldsymbol{Q}}_{2y} \end{bmatrix} \\ &= \begin{pmatrix} \widehat{\boldsymbol{Q}}_{11\cdot2}^{-1}\widehat{\boldsymbol{Q}}_{1y\cdot2} \\ \widehat{\boldsymbol{Q}}_{22\cdot1}^{-1}\widehat{\boldsymbol{Q}}_{2y\cdot1} \end{pmatrix} \end{aligned}$$

Now

$$\begin{aligned} \widehat{\boldsymbol{Q}}_{11\cdot2} &= \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1 \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1 \end{aligned}$$

where

$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'$$

is the orthogonal projection matrix for $\boldsymbol{X}_2$. Similarly $\widehat{\boldsymbol{Q}}_{22\cdot1} = \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2$ where

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$$

is the orthogonal projection matrix for $\boldsymbol{X}_1$. Also

$$\begin{aligned}\widehat{\boldsymbol{Q}}_{1y\cdot 2} &= \widehat{\boldsymbol{Q}}_{1y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{2y} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{y} - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{y} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}\end{aligned}$$

and $\widehat{\boldsymbol{Q}}_{2y\cdot 1} = \dfrac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}$.

Therefore

$$\widehat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}\right) \tag{3.34}$$

and

$$\widehat{\boldsymbol{\beta}}_2 = \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}\right). \tag{3.35}$$

These are algebraic expressions for the sub-coefficient estimates from (3.32).

3.15 Residual Regression

As first recognized by Frisch and Waugh (1933), expressions (3.34) and (3.35) can be used to show that the least-squares estimators $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ can be found by a two-step regression procedure.

Take (3.35). Since $\boldsymbol{M}_1$ is idempotent, $\boldsymbol{M}_1 = \boldsymbol{M}_1\boldsymbol{M}_1$ and thus

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_2 &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}\right) \\ &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{y}\right) \\ &= \left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\left(\widetilde{\boldsymbol{X}}_2'\tilde{\boldsymbol{e}}_1\right)\end{aligned}$$

where

$$\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$$

and

$$\tilde{\boldsymbol{e}}_1 = \boldsymbol{M}_1\boldsymbol{y}.$$

Thus the coefficient estimate $\widehat{\boldsymbol{\beta}}_2$ is algebraically equal to the least-squares regression of $\tilde{\boldsymbol{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$. Notice that these two are $\boldsymbol{y}$ and $\boldsymbol{X}_2$, respectively, premultiplied by $\boldsymbol{M}_1$. But we know that multiplication by $\boldsymbol{M}_1$ is equivalent to creating least-squares residuals. Therefore $\tilde{\boldsymbol{e}}_1$ is simply the least-squares residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$, and the columns of $\widetilde{\boldsymbol{X}}_2$ are the least-squares residuals from the regressions of the columns of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$.

We have proven the following theorem.

Theorem 3.15.1 ***Frisch-Waugh-Lovell***
In the model (3.31), the OLS estimator of $\boldsymbol{\beta}_2$ and the OLS residuals $\hat{\boldsymbol{e}}$ may be equivalently computed by either the OLS regression (3.32) or via the following algorithm:

1. *Regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$, obtain residuals $\tilde{\boldsymbol{e}}_1$;*
2. *Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, obtain residuals $\widetilde{\boldsymbol{X}}_2$;*
3. *Regress $\tilde{\boldsymbol{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$, obtain OLS estimates $\widehat{\boldsymbol{\beta}}_2$ and residuals $\hat{\boldsymbol{e}}$.*

In some contexts, the FWL theorem can be used to speed computation, but in most cases there is little computational advantage to using the two-step algorithm.

This result is a direct analogy of the coefficient representation obtained in Section 2.22. The result obtained in that section concerned the population projection coefficients, the result obtained here concern the least-squares estimates. The key message is the same. In the least-squares regression (3.32), the estimated coefficient $\widehat{\boldsymbol{\beta}}_2$ numerically equals the regression of $\boldsymbol{y}$ on the regressors $\boldsymbol{X}_2$, only after the regressors $\boldsymbol{X}_1$ have been linearly projected out. Similarly, the coefficient estimate $\widehat{\boldsymbol{\beta}}_1$ numerically equals the regression of $\boldsymbol{y}$ on the regressors $\boldsymbol{X}_1$, after the regressors $\boldsymbol{X}_2$ have been linearly projected out. This result can be very insightful when intrepreting regression coefficients.

A common application of the FWL theorem, which you may have seen in an introductory econometrics course, is the demeaning formula for regression. Partition $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ where $\boldsymbol{X}_1 = \mathbf{1}$ is a vector of ones and $\boldsymbol{X}_2$ is a matrix of observed regressors. In this case,

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'.$$

Observe that

$$\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2 = \boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2$$

and

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{y} - \overline{\boldsymbol{y}}$$

are the "demeaned" variables. The FWL theorem says that $\widehat{\boldsymbol{\beta}}_2$ is the OLS estimate from a regression of $y_i - \overline{y}$ on $\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2$:

$$\widehat{\boldsymbol{\beta}}_2 = \left(\sum_{i=1}^{n}\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)'\right)^{-1}\left(\sum_{i=1}^{n}\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)\left(y_i - \overline{y}\right)\right).$$

Thus the OLS estimator for the slope coefficients is a regression with demeaned data.

Ragnar Frisch

Ragnar Frisch (1895-1973) was co-winner with Jan Tinbergen of the first Nobel Memorial Prize in Economic Sciences in 1969 for their work in developing and applying dynamic models for the analysis of economic problems. Frisch made a number of foundational contributions to modern economics beyond the Frisch-Waugh-Lovell Theorem, including formalizing consumer theory, production theory, and business cycle theory.

3.16 Prediction Errors

The least-squares residual $\hat{e}_i$ are not true prediction errors, as they are constructed based on the full sample including y_i. A proper prediction for y_i should be based on estimates constructed using only the other observations. We can do this by defining the **leave-one-out** OLS estimator of $\boldsymbol{\beta}$ as that obtained from the sample of $n-1$ observations *excluding* the i'th observation:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(-i)} &= \left(\frac{1}{n-1}\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j'\right)^{-1}\left(\frac{1}{n-1}\sum_{j\neq i}\boldsymbol{x}_j y_j\right) \\ &= \left(\boldsymbol{X}_{(-i)}'\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}_{(-i)}\boldsymbol{y}_{(-i)}. \end{aligned} \tag{3.36}$$

Here, $\boldsymbol{X}_{(-i)}$ and $\boldsymbol{y}_{(-i)}$ are the data matrices omitting the i'th row. The leave-one-out predicted value for y_i is

$$\tilde{y}_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)},$$

and the **leave-one-out residual** or **prediction error** or **prediction residual** is

$$\tilde{e}_i = y_i - \tilde{y}_i.$$

A convenient alternative expression for $\widehat{\boldsymbol{\beta}}_{(-i)}$ (derived in Section 3.21) is

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} - (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \tag{3.37}$$

where h_{ii} are the leverage values as defined in (3.21).

Using (3.37) we can simplify the expression for the prediction error:

$$\begin{aligned} \tilde{e}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \\ &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + (1 - h_{ii})^{-1} \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \\ &= \hat{e}_i + (1 - h_{ii})^{-1} h_{ii} \hat{e}_i \\ &= (1 - h_{ii})^{-1} \hat{e}_i. \end{aligned} \tag{3.38}$$

To write this in vector notation, define

$$\begin{aligned} \boldsymbol{M}^* &= (\boldsymbol{I}_n - \text{diag}\{h_{11}, .., h_{nn}\})^{-1} \\ &= \text{diag}\{(1 - h_{11})^{-1}, .., (1 - h_{nn})^{-1}\}. \end{aligned} \tag{3.39}$$

Then (3.38) is equivalent to

$$\widetilde{\boldsymbol{e}} = \boldsymbol{M}^* \widehat{\boldsymbol{e}}. \tag{3.40}$$

A convenient feature of this expression is that it shows that computation of the full vector of prediction errors $\widetilde{\boldsymbol{e}}$ is based on a simple linear operation, and does not really require n separate estimations.

One use of the prediction errors is to estimate the out-of-sample mean squared error

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} \tilde{e}_i^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} (1 - h_{ii})^{-2} \hat{e}_i^2. \end{aligned} \tag{3.41}$$

This is also known as the **sample mean squared prediction error**. Its square root $\tilde{\sigma} = \sqrt{\tilde{\sigma}^2}$ is the **prediction standard error**.

3.17 Influential Observations

Another use of the leave-one-out estimator is to investigate the impact of **influential observations**, sometimes called **outliers**. We say that observation i is influential if its omission from the sample induces a substantial change in a parameter estimate of interest.

For illustration, consider Figure 3.2 which shows a scatter plot of random variables (y_i, x_i). The 25 observations shown with the open circles are generated by $x_i \sim U[1, 10]$ and $y_i \sim N(x_i, 4)$. The 26^{th} observation shown with the filled circle is $x_{26} = 9$, $y_{26} = 0$. (Imagine that $y_{26} = 0$ was incorrectly recorded due to a mistaken key entry.) The Figure shows both the least-squares fitted line from the full sample and that obtained after deletion of the 26^{th} observation from the sample. In this example we can see how the 26^{th} observation (the "outlier") greatly tilts the least-squares

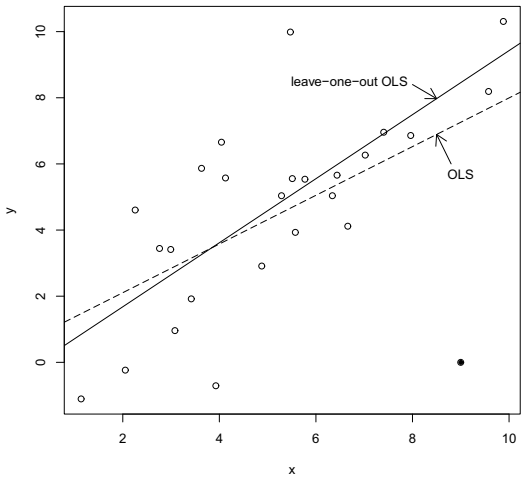

Figure 3.2: Impact of an influential observation on the least-squares estimator

fitted line towards the 26^{th} observation. In fact, the slope coefficient decreases from 0.97 (which is close to the true value of 1.00) to 0.56, which is substantially reduced. Neither y_{26} nor x_{26} are unusual values relative to their marginal distributions, so this outlier would not have been detected from examination of the marginal distributions of the data. The change in the slope coefficient of -0.41 is meaningful and should raise concern to an applied economist.

From (3.37)-(3.38) we know that

$$\begin{aligned}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)} &= (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \tilde{e}_i. \end{aligned} \tag{3.42}$$

By direct calculation of this quantity for each observation i, we can directly discover if a specific observation i is influential for a coefficient estimate of interest.

For a general assessment, we can focus on the predicted values. The difference between the full-sample and leave-one-out predicted values is

$$\begin{aligned}\hat{y}_i - \tilde{y}_i &= \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \\ &= \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \tilde{e}_i \\ &= h_{ii}\tilde{e}_i \end{aligned}$$

which is a simple function of the leverage values h_{ii} and prediction errors $\tilde{e}_i$. Observation i is influential for the predicted value if $|h_{ii}\tilde{e}_i|$ is large, which requires that both h_{ii} and $|\tilde{e}_i|$ are large.

One way to think about this is that a large leverage value h_{ii} gives the potential for observation i to be influential. A large h_{ii} means that observation i is unusual in the sense that the regressor $\boldsymbol{x}_i$ is far from its sample mean. We call an observation with large h_{ii} a **leverage point**. A leverage point is not necessarily influential as the latter also requires that the prediction error $\tilde{e}_i$ is large.

To determine if any individual observations are influential in this sense, several diagnostics have been proposed (some names include DFITS, Cook's Distance, and Welsch Distance). Unfortunately, from a statistical perspective it is difficult to recommend these diagnostics for applications as they are not based on statistical theory. Probably the most relevant measure is the change in the coefficient estimates given in (3.42). The ratio of these changes to the coefficient's standard error is called its DFBETA, and is a postestimation diagnostic available in STATA. While there is no magic threshold, the concern is whether or not an individual observation meaningfully changes an

estimated coefficient of interest. A simple diagnostic for influential observations is to calculate

$$Influence = \max_{1\leq i\leq n} |\hat{y}_i - \tilde{y}_i| = \max_{1\leq i\leq n} |h_{ii}\tilde{e}_i| .$$

This is the largest (absolute) change in the predicted value due to a single observation. If this diagnostic is large relative to the distribution of y_i, it may indicate that that observation is influential.

If an observation is determined to be influential, what should be done? As a common cause of influential observations is data entry error, the influential observations should be examined for evidence that the observation was mis-recorded. Perhaps the observation falls outside of permitted ranges, or some observables are inconsistent (for example, a person is listed as having a job but receives earnings of $0). If it is determined that an observation is incorrectly recorded, then the observation is typically deleted from the sample. This process is often called "cleaning the data". The decisions made in this process involve an fair amount of individual judgement. When this is done it is proper empirical practice to document such choices. (It is useful to keep the source data in its original form, a revised data file after cleaning, and a record describing the revision process. This is especially useful when revising empirical work at a later date.)

It is also possible that an observation is correctly measured, but unusual and influential. In this case it is unclear how to proceed. Some researchers will try to alter the specification to properly model the influential observation. Other researchers will delete the observation from the sample. The motivation for this choice is to prevent the results from being skewed or determined by individual observations, but this practice is viewed skeptically by many researchers who believe it reduces the integrity of reported empirical results.

For an empirical illustration, consider the log wage regression (3.12) for single asian males. This regression, which has 268 observations, has $Influence = 0.29$. This means that the most influential observation, when deleted, changes the predicted (fitted) value of the dependent variable $\log(Wage)$ by 0.29, or equivalently the wage by 29%. This is a meaningful change and suggests further investigation. We examine the influential observation, and find that its leverage h_{ii} is 0.33, which is disturbingly large. (Recall that the leverage values are all positive and sum to k. One twelth of the leverage in this sample of 268 observations is contained in just this single observation!) Examining further, we find that this individual is 65 years old with 8 years education, so that his potential experience is 51 years. This is the highest experience in the subsample – the next highest is 41 years. The large leverage is due to to his unusual characteristics (very low education and very high experience) within this sample. Essentially, regression (3.12) is attempting to estimate the conditional mean at *experience*= 51 with only one observation, so it is not surprising that this observation determines the fit and is thus influential. A reasonable conclusion is the regression function can only be estimated over a smaller range of *experience*. We restrict the sample to individuals with less than 45 years experience, re-estimate, and obtain the following estimates.

$$\widehat{\log(Wage)} = 0.144\ education + 0.043\ experience - 0.095\ experience^2/100 + 0.531. \tag{3.43}$$

For this regression, we calculate that $Influence = 0.11$, which is greatly reduced relative to the regression (3.12). Comparing (3.43) with (3.12), the slope coefficient for education is essentially unchanged, but the coefficients in experience and its square have slightly increased.

By eliminating the influential observation, equation (3.43) can be viewed as a more robust estimate of the conditional mean for most levels of *experience*. Whether to report (3.12) or (3.43) in an application is largely a matter of judgment.

3.18 Normal Regression Model

The normal regression model is the linear regression model under the restriction that the error e_i is independent of $\boldsymbol{x}_i$ and has the distribution $\mathrm{N}\left(0, \sigma^2\right)$. We can write this as

$$e_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(0, \sigma^2\right).$$

This assumption implies

$$y_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2\right).$$

Normal regression is a parametric model, where likelihood methods can be used for estimation, testing, and distribution theory.

The log-likelihood function for the normal regression model is

$$\begin{aligned}\log L(\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^{n} \log\left(\frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right)\right) \\ &= -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2} SSE_n(\boldsymbol{\beta}). \end{aligned} \tag{3.44}$$

The maximum likelihood estimator (MLE) $(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}^2_{\text{mle}})$ maximizes $\log L(\boldsymbol{\beta}, \sigma^2)$. Since the latter is a function of $\boldsymbol{\beta}$ only through the sum of squared errors $SSE_n(\boldsymbol{\beta})$, maximizing the likelihood is identical to minimizing $SSE_n(\boldsymbol{\beta})$. Hence

$$\widehat{\boldsymbol{\beta}}_{\text{mle}} = \widehat{\boldsymbol{\beta}}_{\text{ols}},$$

the MLE for $\boldsymbol{\beta}$ equals the OLS estimator. Due to this equivalence, the least squares estimator $\widehat{\boldsymbol{\beta}}_{\text{ols}}$ is often called the MLE.

We can also find the MLE for σ^2. Plugging $\widehat{\boldsymbol{\beta}}$ into the log-likelihood we obtain

$$\log L\left(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{SSE_n(\widehat{\boldsymbol{\beta}}_{\text{mle}})}{2\sigma^2}.$$

Maximization with respect to σ^2 yields the first-order condition

$$\frac{\partial}{\partial\sigma^2}\log L\left(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}^2\right) = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\left(\hat{\sigma}^2\right)^2} SSE_n(\widehat{\boldsymbol{\beta}}_{\text{mle}}) = 0.$$

Solving for $\hat{\sigma}^2$ yields the MLE for σ^2

$$\hat{\sigma}^2_{\text{mle}} = \frac{SSE_n(\widehat{\boldsymbol{\beta}}_{\text{mle}})}{n} = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

which is the same as the moment estimator (3.28).

Plugging the estimates into (3.44) we obtain the maximized log-likelihood

$$\log L\left(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}^2_{\text{mle}}\right) = -\frac{n}{2}\left(\log\left(2\pi\right) + 1\right) - \frac{n}{2}\log\left(\hat{\sigma}^2_{\text{mle}}\right). \tag{3.45}$$

The log-likelihood (or the negative log-likelihood) is typically reported as a measure of fit.

It may seem surprising that the MLE $\widehat{\boldsymbol{\beta}}_{\text{mle}}$ is numerically equal to the OLS estimator, despite emerging from quite different motivations. It is not completely accidental. The least-squares estimator minimizes a particular sample loss function – the sum of squared error criterion – and most loss functions are equivalent to the likelihood of a specific parametric distribution, in this case the normal regression model. In this sense it is not surprising that the least-squares estimator can be motivated as either the minimizer of a sample loss function or as the maximizer of a likelihood function.

Carl Friedrich Gauss

The mathematician Carl Friedrich Gauss (1777-1855) proposed the normal regression model, and derived the least squares estimator as the maximum likelihood estimator for this model. He claimed to have discovered the method in 1795 at the age of eighteen, but did not publish the result until 1809. Interest in Gauss's approach was reinforced by Laplace's simultaneous discovery of the central limit theorem, which provided a justification for viewing random disturbances as approximately normal.

3.19 CPS Data Set

In this section we describe the data set used in the empirical illustrations.

The Current Population Survey (CPS) is a monthly survey of about 57,000 U.S. households conducted by the Bureau of the Census of the Bureau of Labor Statistics. The CPS is the primary source of information on the labor force characteristics of the U.S. population. The survey covers employment, earnings, educational attainment, income, poverty, health insurance coverage, job experience, voting and registration, computer usage, veteran status, and other variables. Details can be found at www.census.gov/cps and dataferrett.census.gov.

From the March 2009 survey we extracted the individuals with non-allocated variables who were full-time employed (defined as those who had worked at least 36 hours per week for at least 48 weeks the past year), and excluded those in the military. This sample has 50,742 individuals. We extracted 14 variables from the CPS on these individuals and created the data files cps09mar.dta (Stata format) and cps09mar.txt (text format). The variables are:

1. age: years, capped at 85
2. female: 1 if female, 0 otherwise
3. hisp: 1 if Spanish, Hispanic, or Latino, 0 otherwise
4. education
   - 0 Less than 1st grade
   - 4 1st, 2nd, 3rd, or 4th grade
   - 6 5th or 6th grade
   - 8 7th or 8th grade
   - 9 9th grade
   - 10 10th grade
   - 11 11th grade or 12th grade with no high school diploma
   - 12 High school graduate, high school diploma or equivalent
   - 13 Some college but no degree
   - 14 Associate degree in college, including occupation/vocation programs
   - 16 Bachelor's degree or equivalent (BA, AB, BS)
   - 18 Master's degree (MA, MS MENG, MED, MSW, MBA)
   - 20 Professional degree or Doctorate degree (MD, DDS, DVM, LLB, JD, PHD, EDD)
5. earnings: total annual wage and salary earnings

6. hours: number of hours worked per week
7. week: number of weeks worked per year
8. union: 1 for member of a labor union, 0 otherwise
9. uncov: 1 if covered by a union or employee association contract, 0 otherwise
10. region
    1 Northeast
    2 Midwest
    3 South
    4 West
11. Race
    1 White only
    2 Black only
    3 American Indian, Alaskan Native (AI) only
    4 Asian only
    5 Hawaiian/Pacific Islander (HP) only
    6 White-Black
    7 White-AI
    8 White-Asian
    9 White-HP
    10 Black-AI
    11 Black-Asian
    12 Black-HP
    13 AI-Asian
    14 Asian-HP
    15 White-Black-AI
    16 White-Black-Asian
    17 White-AI-Asian
    18 White-Asian-HP
    19 White-Black-AI-Asian
    20 2 or 3 races
    21 4 or 5 races
12. marital
    1 Married - civilian spouse present
    2 Married - Armed Forces spouse present
    3 Married - spouse absent (except separated)
    4 Widowed
    5 Divorced
    6 Separated
    7 Never married

3.20 Programming

Most packages allow both interactive programming (where you enter commands one-by-one) and batch programming (where you run a pre-written sequence of commands from a file). Interactive programming can be useful for exploratory analysis, but eventually all work should be executed in batch mode. This is the best way to control and document your work.

Batch programs are text files where each line executes a single command. For Stata, this file needs to have the filename extension ".do", and for Matlab ".m", while for Gauss and R there are no specific naming requirements.

To execute a program file, you type a command within the program.

Stata: `do chapter3` executes the file *chapter3.do*

Gauss: `run chapter3.prg` executes the file *chapter3.prg*

Matlab: `run chapter3` executes the file *chapter3.m*

R: `source("chapter3.r")` executes the file *chatper3.r*

When writing batch files, it is useful to include comments for documentation and readability.

We illustrate programming files for Stata, Gauss, R, and Matlab, which execute a portion of the empirical illustrations from Sections 3.7 and 3.17.

Stata do File

```
*       Clear memory and load the data
clear
use cps09mar.dta
*       Generate transformations
gen wage=ln(earnings/(hours*week))
gen experience = age - education - 6
gen exp2 = (experience^2)/100
*       Create indicator for subsamples
gen mbf = (race == 2) & (marital <= 2) & (female == 1)
gen sam = (race == 4) & (marital == 7) & (female == 0)
*        Regressions
reg wage education if (mbf == 1) & (experience == 12)
reg wage education experience exp2 if sam == 1
*       Leverage and influence
predict leverage,hat
predict e,residual
gen d=e*leverage/(1-leverage)
summarize d if sam ==1
```

Gauss Program File

```
/* Load the data and create subsamples */
load dat[50742,12]=cps09mar.txt;
experience=dat[.,1]-dat[.,4]-6;
mbf=(dat[.,11].==2).*(dat[.,12].<=2).*(dat[.,2].==1).*(experience.==12);
sam=(dat[.,11].==4).*(dat[.,12].==7).*(dat[.,2].==0);
dat1=selif(dat,mbf);
dat2=selif(dat,sam);
/* First regression */
y=ln(dat1[.,5]./(dat1[.,6].*dat1[.,7]));
x=dat1[.,4]~ones(rows(dat1),1);
beta=invpd(x'x)*(x'y);
beta;
/* Second regression */
y=ln(dat2[.,5]./(dat2[.,6].*dat2[.,7]));
experience=dat2[.,1]-dat2[.,4]-6;
exp2 = (experience.^2)/100;
x=dat2[.,4]~experience~exp2~ones(rows(dat2),1);
beta=invpd(x'x)*(x'y);
beta;
/* Create leverage and influence */
e=y-x*beta;
leverage=sumc((x.*(x*invpd(x'x)))');
d=leverage.*e./(1-leverage);
"Influence " maxc(abs(d));
```

R Program File

```
#      Load the data and create subsamples
dat <- read.table("cps09mar.txt")
experience <- dat[,1]-dat[,4]-6
mbf <- (dat[,11]==2)&(dat[,12]<=2)&(dat[,2]==1)&(experience==12)
sam <- (dat[,11]==4)&(dat[,12]==7)&(dat[,2]==0)
dat1 <- dat[mbf,]
dat2 <- dat[sam,]
#      First regression
y <- as.matrix(log(dat1[,5]/(dat1[,6]*dat1[,7])))
x <- cbind(dat1[,4],matrix(1,nrow(dat1),1))
beta <- solve(t(x)%*%x,t(x)%*%y)
print(beta)
#      Second regression
y <- as.matrix(log(dat2[,5]/(dat2[,6]*dat2[,7])))
experience <- dat2[,1]-dat2[,4]-6
exp2 <- (experience^2)/100
x <- cbind(dat2[,4],experience,exp2,matrix(1,nrow(dat2),1))
beta <- solve(t(x)%*%x,t(x)%*%y)
print(beta)
#      Create leverage and influence
e <- y-x%*%beta
leverage <- rowSums(x*(x%*%solve(t(x)%*%x)))
r <- e/(1-leverage)
d <- leverage*e/(1-leverage)
print(max(abs(d)))
```

Matlab Program File

```
% Load the data and create subsamples
load cps09mar.txt;
dat=cps09mar;
experience=dat(:,1)-dat(:,4)-6;
mbf = (dat(:,11)==2)&(dat(:,12)<=2)&(dat(:,2)==1)&(experience==12);
sam = (dat(:,11)==4)&(dat(:,12)==7)&(dat(:,2)==0);
dat1=dat(mbf,:);
dat2=dat(sam,:);
%      First regression
y=log(dat1(:,5)./(dat1(:,6).*dat1(:,7)));
x=[dat1(:,4),ones(length(dat1),1)];
beta=inv(x'*x)*(x'*y);
display(beta);
%      Second regression
y=log(dat2(:,5)./(dat2(:,6).*dat2(:,7)));
experience=dat2(:,1)-dat2(:,4)-6;
exp2 = (experience.^2)/100;
x=[dat2(:,4),experience,exp2,ones(length(dat2),1)];
beta=inv(x'*x)*(x'*y);
display(beta);
%      Create leverage and influence
e=y-x*beta;
leverage=sum((x.*(x*inv(x'*x)))')';
d=leverage.*e./(1-leverage);
influence=max(abs(d));
display(influence);
```

3.21 Technical Proofs*

Proof of Theorem 3.10.1, equation (3.23): First, $h_{ii} = \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \geq 0$ since it is a quadratic form and $\boldsymbol{X}'\boldsymbol{X} > 0$. Next, since h_{ii} is the i'th diagonal element of the projection matrix $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}$, then

$$h_{ii} = \boldsymbol{s}'\boldsymbol{P}\boldsymbol{s}$$

where

$$\boldsymbol{s} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

is a unit vector with a 1 in the i'th place (and zeros elsewhere).

By the spectral decomposition of the idempotent matrix $\boldsymbol{P}$ (see equation (A.5))

$$\boldsymbol{P} = \boldsymbol{B}' \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{B}$$

where $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_n$. Thus letting $\boldsymbol{b} = \boldsymbol{B}\boldsymbol{s}$ denote the i'th column of $\boldsymbol{B}$, and partitioning $\boldsymbol{b}' = \begin{pmatrix} \boldsymbol{b}_1' & \boldsymbol{b}_2' \end{pmatrix}$ then

$$\begin{aligned} h_{ii} &= \boldsymbol{s}'\boldsymbol{B}' \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{B}\boldsymbol{s} \\ &= \boldsymbol{b}_1' \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{b}_1 \\ &= \boldsymbol{b}_1'\boldsymbol{b}_1 \\ &\leq \boldsymbol{b}'\boldsymbol{b} \\ &= 1 \end{aligned}$$

the final equality since $\boldsymbol{b}$ is the i'th column of $\boldsymbol{B}$ and $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_n$. We have shown that $h_{ii} \leq 1$, establishing (3.23). ■

Proof of Equation (3.37). The Sherman–Morrison formula (A.3) from Appendix A.5 states that for nonsingular $\boldsymbol{A}$ and vector $\boldsymbol{b}$

$$\left(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{b}'\right)^{-1} = \boldsymbol{A}^{-1} + \left(1 - \boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1} \boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{A}^{-1}.$$

This implies

$$\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

and thus

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{(-i)} &= \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i y_i \\ &+ (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i\right) \\ &= \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i y_i + (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - h_{ii} y_i\right) \\ &= \widehat{\boldsymbol{\beta}} - (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \left((1 - h_{ii})\, y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + h_{ii} y_i\right) \\ &= \widehat{\boldsymbol{\beta}} - (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \end{aligned}$$

the third equality making the substitutions $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $h_{ii} = \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i$, and the remainder collecting terms. ■

Exercises

Exercise 3.1 Let y be a random variable with $\mu = \mathbb{E}y$ and $\sigma^2 = \text{var}(y)$. Define

$$g\left(y, \mu, \sigma^2\right) = \begin{pmatrix} y - \mu \\ (y-\mu)^2 - \sigma^2 \end{pmatrix}.$$

Let $(\hat{\mu}, \hat{\sigma}^2)$ be the values such that $\overline{g}_n(\hat{\mu}, \hat{\sigma}^2) = \mathbf{0}$ where $\overline{g}_n(m, s) = n^{-1} \sum_{i=1}^n g\left(y_i, m, s\right)$. Show that $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and variance.

Exercise 3.2 Consider the OLS regression of the $n \times 1$ vector $\boldsymbol{y}$ on the $n \times k$ matrix $\boldsymbol{X}$. Consider an alternative set of regressors $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{C}$, where $\boldsymbol{C}$ is a $k \times k$ non-singular matrix. Thus, each column of $\boldsymbol{Z}$ is a mixture of some of the columns of $\boldsymbol{X}$. Compare the OLS estimates and residuals from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ to the OLS estimates from the regression of $\boldsymbol{y}$ on $\boldsymbol{Z}$.

Exercise 3.3 Using matrix algebra, show $\boldsymbol{X}'\hat{\boldsymbol{e}} = \mathbf{0}$.

Exercise 3.4 Let $\hat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$. Find $\boldsymbol{X}_2'\hat{\boldsymbol{e}}$.

Exercise 3.5 Let $\hat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}$. Find the OLS coefficient from a regression of $\hat{\boldsymbol{e}}$ on $\boldsymbol{X}$.

Exercise 3.6 Let $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Find the OLS coefficient from a regression of $\hat{\boldsymbol{y}}$ on $\boldsymbol{X}$.

Exercise 3.7 Show that if $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$ and $\boldsymbol{M}\boldsymbol{X}_1 = \mathbf{0}$.

Exercise 3.8 Show that $\boldsymbol{M}$ is idempotent: $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

Exercise 3.9 Show that $\text{tr}\,\boldsymbol{M} = n - k$.

Exercise 3.10 Show that if $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ and $\boldsymbol{X}_1'\boldsymbol{X}_2 = 0$ then $\boldsymbol{P} = \boldsymbol{P}_1 + \boldsymbol{P}_2$.

Exercise 3.11 Show that when $\boldsymbol{X}$ contains a constant, $\dfrac{1}{n}\sum_{i=1}^n \hat{y}_i = \bar{y}$.

Exercise 3.12 A dummy variable takes on only the values 0 and 1. It is used for categorical data, such as an individual's gender. Let $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ be vectors of 1's and 0's, with the i'th element of $\boldsymbol{d}_1$ equaling 1 and that of $\boldsymbol{d}_2$ equaling 0 if the person is a man, and the reverse if the person is a woman. Suppose that there are n_1 men and n_2 women in the sample. Consider fitting the following three equations by OLS

$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.46}$$
$$\boldsymbol{y} = \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.47}$$
$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\phi + \boldsymbol{e} \tag{3.48}$$

Can all three equations (3.46), (3.47), and (3.48) be estimated by OLS? Explain if not.

(a) Compare regressions (3.47) and (3.48). Is one more general than the other? Explain the relationship between the parameters in (3.47) and (3.48).

(b) Compute $\boldsymbol{\iota}'\boldsymbol{d}_1$ and $\boldsymbol{\iota}'\boldsymbol{d}_2$, where $\boldsymbol{\iota}$ is an $n \times 1$ is a vector of ones.

(c) Letting $\boldsymbol{\alpha} = (\alpha_1 \ \alpha_2)'$, write equation (3.47) as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + e$. Consider the assumption $\mathbb{E}(\boldsymbol{x}_i e_i) = 0$. Is there any content to this assumption in this setting?

Exercise 3.13 Let $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ be defined as in the previous exercise.

(a) In the OLS regression

$$\boldsymbol{y} = \boldsymbol{d}_1\hat{\gamma}_1 + \boldsymbol{d}_2\hat{\gamma}_2 + \hat{\boldsymbol{u}},$$

show that $\hat{\gamma}_1$ is the sample mean of the dependent variable among the men of the sample ($\overline{y}_1$), and that $\hat{\gamma}_2$ is the sample mean among the women ($\overline{y}_2$).

(b) Let $\boldsymbol{X}$ $(n \times k)$ be an additional matrix of regressors. Describe in words the transformations

$$\boldsymbol{y}^* = \boldsymbol{y} - \boldsymbol{d}_1\overline{y}_1 - \boldsymbol{d}_2\overline{y}_2$$
$$\boldsymbol{X}^* = \boldsymbol{X} - \boldsymbol{d}_1\overline{\boldsymbol{x}}_1' - \boldsymbol{d}_2\overline{\boldsymbol{x}}_2'$$

where $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ are the $k \times 1$ means of the regressors for men and women, respectively.

(c) Compare $\widetilde{\boldsymbol{\beta}}$ from the OLS regresion

$$\boldsymbol{y}^* = \boldsymbol{X}^*\widetilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{e}}$$

with $\widehat{\boldsymbol{\beta}}$ from the OLS regression

$$\boldsymbol{y} = \boldsymbol{d}_1\hat{\alpha}_1 + \boldsymbol{d}_2\hat{\alpha}_2 + \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}}.$$

Exercise 3.14 Let $\widehat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n'\boldsymbol{y}_n$ denote the OLS estimate when $\boldsymbol{y}_n$ is $n \times 1$ and $\boldsymbol{X}_n$ is $n \times k$. A new observation $(y_{n+1}, \boldsymbol{x}_{n+1})$ becomes available. Prove that the OLS estimate computed using this additional observation is

$$\widehat{\boldsymbol{\beta}}_{n+1} = \widehat{\boldsymbol{\beta}}_n + \frac{1}{1 + \boldsymbol{x}_{n+1}'(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_{n+1}}(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_{n+1}\left(y_{n+1} - \boldsymbol{x}_{n+1}'\widehat{\boldsymbol{\beta}}_n\right).$$

Exercise 3.15 Prove that R^2 is the square of the sample correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$.

Exercise 3.16 Consider two least-squares regressions

$$\boldsymbol{y} = \boldsymbol{X}_1\widetilde{\boldsymbol{\beta}}_1 + \tilde{\boldsymbol{e}}$$

and

$$\boldsymbol{y} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}.$$

Let R_1^2 and R_2^2 be the R-squared from the two regressions. Show that $R_2^2 \geq R_1^2$. Is there a case (explain) when there is equality $R_2^2 = R_1^2$?

Exercise 3.17 Show that $\tilde{\sigma}^2 \geq \hat{\sigma}^2$. Is equality possible?

Exercise 3.18 For which observations will $\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}}$?

Exercise 3.19 Use the data set from Section 3.19 and the sub-sample used for equation (3.43) (see Section 3.20) for data construction)

1. Estimate equation (3.43) and compute the equation R^2 and sum of squared errors.
2. Re-estimate the slope on education using the residual regression approach. Regress log(Wage) on experience and its square, regress education on experience and its square, and the residuals on the residuals. Report the estimates from this final regression, along with the equation R^2 and sum of squared errors. Does the slope coefficient equal the value in (3.43)? Explain.
3. Do the R^2 and sum-of-squared errors from parts 1 and 2 equal? Explain.

Exercise 3.20 Estimate equation (3.43) as in part 1 of the previous question. Let $\hat{e}_i$ be the OLS residual, $\hat{y}_i$ the predicted value from the regression, x_{1i} be education and x_{2i} be experience. Numerically calculate the following:

(a) $\sum_{i=1}^{n} \hat{e}_i$

(b) $\sum_{i=1}^{n} x_{1i}\hat{e}_i$

(c) $\sum_{i=1}^{n} x_{2i}\hat{e}_i$

(d) $\sum_{i=1}^{n} x_{1i}^2\hat{e}_i$

(e) $\sum_{i=1}^{n} x_{2i}^2\hat{e}_i$

(f) $\sum_{i=1}^{n} \hat{y}_i\hat{e}_i$

(g) $\sum_{i=1}^{n} \hat{e}_i^2$

Are these calculations consistent with the theoretical properties of OLS? Explain.

Exercise 3.21 Use the data set from Section 3.19.

1. Estimate a log wage regression for the subsample of white male Hispanics. In addition to education, experience, and its square, include a set of binary variables for regions and marital status. For regions, you create dummy variables for Northeast, South and West so that Midwest is the excluded group. For marital status, create variables for married, widowed or divorced, and separated, so that single (never married) is the excluded group.

2. Repeat this estimation using a different econometric package. Compare your results. Do they agree?

Chapter 4

Least Squares Regression

4.1 Introduction

In this chapter we investigate some finite-sample properties of least-squares applied to a random sample in the the linear regression model. In particular, we calculate the finite-sample mean and covariance matrix and propose standard errors for the coefficient estimates.

4.2 Sample Mean

To start with the simplest setting, we first consider the intercept-only model

$$\begin{aligned} y_i &= \mu + e_i \\ \mathbb{E}(e_i) &= 0. \end{aligned}$$

which is equivalent to the regression model with $k = 1$ and $x_i = 1$. In the intercept model, $\mu = \mathbb{E}(y_i)$ is the mean of y_i. (See Exercise 2.15.) The least-squares estimator $\widehat{\mu} = \overline{y}$ equals the sample mean as shown in equation (3.7).

We now calculate the mean and variance of the estimator $\overline{y}$. Since the sample mean is a linear function of the observations, its expectation is simple to calculate

$$\mathbb{E}(\overline{y}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}y_i = \mu.$$

This shows that the expected value of least-squares estimator (the sample mean) equals the projection coefficient (the population mean). An estimator with the property that its expectation equals the parameter it is estimating is called **unbiased**.

Definition 4.2.1 *An estimator $\widehat{\theta}$ for θ is **unbiased** if $\mathbb{E}\widehat{\theta} = \theta$.*

We next calculate the variance of the estimator $\overline{y}$. Making the substitution $y_i = \mu + e_i$ we find

$$\overline{y} - \mu = \frac{1}{n}\sum_{i=1}^{n} e_i.$$

Then

$$\begin{aligned}
\operatorname{var}(\overline{y}) &= \mathbb{E}(\overline{y} - \mu)^2 \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} e_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} e_j\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(e_i e_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
&= \frac{1}{n}\sigma^2.
\end{aligned}$$

The second-to-last inequality is because $\mathbb{E}(e_i e_j) = \sigma^2$ for $i = j$ yet $\mathbb{E}(e_i e_j) = 0$ for $i \neq j$ due to independence.

We have shown that $\operatorname{var}(\overline{y}) = \frac{1}{n}\sigma^2$. This is the familiar formula for the variance of the sample mean.

4.3 Linear Regression Model

We now consider the linear regression model. Throughout the remainder of this chapter we maintain the following.

Assumption 4.3.1 ***Linear Regression Model***
The observations $(y_i, \boldsymbol{x}_i)$ come from a random sample and satisfy the linear regression equation

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{4.1}$$

$$\mathbb{E}(e_i \mid \boldsymbol{x}_i) = 0. \tag{4.2}$$

The variables have finite second moments

$$\mathbb{E}y_i^2 < \infty,$$

$$\mathbb{E}\left\|\boldsymbol{x}_i\right\|^2 < \infty,$$

and an invertible design matrix

$$\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) > 0.$$

We will consider both the general case of heteroskedastic regression, where the conditional variance

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2(\boldsymbol{x}_i) = \sigma_i^2$$

is unrestricted, and the specialized case of homoskedastic regression, where the conditional variance is constant. In the latter case we add the following assumption.

Assumption 4.3.2 ***Homoskedastic Linear Regression Model***
In addition to Assumption 4.3.1,

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2(\boldsymbol{x}_i) = \sigma^2 \tag{4.3}$$

is independent of $\boldsymbol{x}_i$.

4.4 Mean of Least-Squares Estimator

In this section we show that the OLS estimator is unbiased in the linear regression model. This calculation can be done using either summation notation or matrix notation. We will use both.

First take summation notation. Observe that under (4.1)-(4.2)

$$\mathbb{E}\left(y_i \mid \boldsymbol{X}\right) = \mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right) = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{4.4}$$

The first equality states that the conditional expectation of y_i given $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ only depends on $\boldsymbol{x}_i$, since the observations are independent across i. The second equality is the assumption of a linear conditional mean.

Using definition (3.9), the conditioning theorem, the linearity of expectations, (4.4), and properties of the matrix inverse,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_i y_i\right) \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \mathbb{E}\left(\left(\sum_{i=1}^n \boldsymbol{x}_i y_i\right) \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^n \mathbb{E}\left(\boldsymbol{x}_i y_i \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^n \boldsymbol{x}_i \mathbb{E}\left(y_i \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{\beta} \\
&= \boldsymbol{\beta}.
\end{aligned}$$

Now let's show the same result using matrix notation. (4.4) implies

$$\mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \begin{pmatrix} \vdots \\ \mathbb{E}\left(y_i \mid \boldsymbol{X}\right) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \boldsymbol{x}_i'\boldsymbol{\beta} \\ \vdots \end{pmatrix} = \boldsymbol{X}\boldsymbol{\beta}. \tag{4.5}$$

Similarly

$$\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \begin{pmatrix} \vdots \\ \mathbb{E}\left(e_i \mid \boldsymbol{X}\right) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) \\ \vdots \end{pmatrix} = \mathbf{0}. \tag{4.6}$$

Using definition (3.18), the conditioning theorem, the linearity of expectations, (4.5), and the properties of the matrix inverse,

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}.\end{aligned}$$

At the risk of belaboring the derivation, another way to calculate the same result is as follows. Insert $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ into the formula (3.18) for $\widehat{\boldsymbol{\beta}}$ to obtain

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}\right)\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{e}\right) \\ &= \boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}. \qquad (4.7)\end{aligned}$$

This is a useful linear decomposition of the estimator $\widehat{\boldsymbol{\beta}}$ into the true parameter $\boldsymbol{\beta}$ and the stochastic component $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}$. Once again, we can calculate that

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) \\ &= \mathbf{0}.\end{aligned}$$

Regardless of the method, we have shown that $\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$. Applying the law of iterated expectations, we find that

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbb{E}\left(\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)\right) = \boldsymbol{\beta}.$$

We have shown the following theorem.

Theorem 4.4.1 ***Mean of Least-Squares Estimator***
In the linear regression model (Assumption 4.3.1)

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta} \qquad (4.8)$$

and

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}. \qquad (4.9)$$

Equation (4.9) says that the estimator $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$, meaning that the distribution of $\widehat{\boldsymbol{\beta}}$ is centered at $\boldsymbol{\beta}$. Equation (4.8) says that the estimator is conditionally unbiased, which is a stronger result. It says that $\widehat{\boldsymbol{\beta}}$ is unbiased for any realization of the regressor matrix $\boldsymbol{X}$.

4.5 Variance of Least Squares Estimator

In this section we calculate the conditional variance of the OLS estimator.

For any $r \times 1$ random vector $\mathbf{Z}$ define the $r \times r$ covariance matrix

$$\begin{aligned}\operatorname{var}(\mathbf{Z}) &= \mathbb{E}\left(\mathbf{Z} - \mathbb{E}\mathbf{Z}\right)\left(\mathbf{Z} - \mathbb{E}\mathbf{Z}\right)' \\ &= \mathbb{E}\mathbf{Z}\mathbf{Z}' - \left(\mathbb{E}\mathbf{Z}\right)\left(\mathbb{E}\mathbf{Z}\right)'\end{aligned}$$

and for any pair $(\mathbf{Z}, \mathbf{X})$ define the conditional covariance matrix

$$\operatorname{var}(\boldsymbol{Z} \mid \boldsymbol{X}) = \mathbb{E}\left((\boldsymbol{Z} - \mathbb{E}(\boldsymbol{Z} \mid \boldsymbol{X}))(\boldsymbol{Z} - \mathbb{E}(\boldsymbol{Z} \mid \boldsymbol{X}))' \mid \boldsymbol{X}\right).$$

We define

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} \stackrel{def}{=} \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$$

the conditional covariance matrix of the regression coefficient estimates. We now derive its form.

The conditional covariance matrix of the $n \times 1$ regression error $\boldsymbol{e}$ is the $n \times n$ matrix

$$\boldsymbol{D} = \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right).$$

The i'th diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{X}\right) = \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma_i^2$$

while the ij'th off-diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}(e_i e_j \mid \boldsymbol{X}) = \mathbb{E}(e_i \mid \boldsymbol{x}_i)\,\mathbb{E}(e_j \mid \boldsymbol{x}_j) = 0.$$

where the first equality uses independence of the observations (Assumption 1.5.1) and the second is (4.2). Thus $\boldsymbol{D}$ is a diagonal matrix with i'th diagonal element σ_i^2:

$$\boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}. \tag{4.10}$$

In the special case of the linear homoskedastic regression model (4.3), then

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma_i^2 = \sigma^2$$

and we have the simplification

$$\boldsymbol{D} = \boldsymbol{I}_n \sigma^2.$$

In general, however, $\boldsymbol{D}$ need not necessarily take this simplified form.

For any $n \times r$ matrix $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{X})$,

$$\operatorname{var}(\boldsymbol{A}'\boldsymbol{y} \mid \boldsymbol{X}) = \operatorname{var}(\boldsymbol{A}'\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A}. \tag{4.11}$$

In particular, we can write $\widehat{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{y}$ where $\boldsymbol{A} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and thus

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \operatorname{var}(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

It is useful to note that

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \sigma_i^2,$$

a weighted version of $\boldsymbol{X}'\boldsymbol{X}$.

In the special case of the linear homoskedastic regression model, $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$, so $\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{X}\sigma^2$, and the variance matrix simplifies to

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2.$$

Theorem 4.5.1 ***Variance of Least-Squares Estimator***
In the linear regression model (Assumption 4.3.1)

$$\begin{aligned} \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} &= \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \end{aligned} \tag{4.12}$$

where $\boldsymbol{D}$ *is defined in (4.10).*
In the homoskedastic linear regression model (Assumption 4.3.2)

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2.$$

4.6 Gauss-Markov Theorem

Now consider the class of estimators of $\boldsymbol{\beta}$ which are linear functions of the vector $\boldsymbol{y}$, and thus can be written as

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{y}$$

where $\boldsymbol{A}$ is an $n \times k$ function of $\boldsymbol{X}$. As noted before, the least-squares estimator is the special case obtained by setting $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. What is the best choice of $\boldsymbol{A}$? The Gauss-Markov theorem, which we now present, says that the least-squares estimator is the best choice among linear unbiased estimators when the errors are homoskedastic, in the sense that the least-squares estimator has the smallest variance among all unbiased linear estimators.

To see this, since $\mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \boldsymbol{X}\boldsymbol{\beta}$**,** then for any linear estimator $\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{y}$ we have

$$\mathbb{E}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{A}'\mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \boldsymbol{A}'\boldsymbol{X}\boldsymbol{\beta},$$

so $\widetilde{\boldsymbol{\beta}}$ is unbiased if (and only if) $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$. Furthermore, we saw in (4.11) that

$$\operatorname{var}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \operatorname{var}\left(\boldsymbol{A}'\boldsymbol{y} \mid \boldsymbol{X}\right) = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} = \boldsymbol{A}'\boldsymbol{A}\sigma^2$$

the last equality using the homoskedasticity assumption $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$. The "best" unbiased linear estimator is obtained by finding the matrix $\boldsymbol{A}_0$ satisfying $\boldsymbol{A}_0'\boldsymbol{X} = \boldsymbol{I}_k$ such that $\boldsymbol{A}_0'\boldsymbol{A}_0$ is minimized in the positive definite sense, in that for any other matrix $\boldsymbol{A}$ satisfying $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$, then $\boldsymbol{A}'\boldsymbol{A} - \boldsymbol{A}_0'\boldsymbol{A}_0$ is positive semi-definite.

Theorem 4.6.1 ***Gauss-Markov***

1. *In the homoskedastic linear regression model (Assumption 4.3.2), the best (minimum-variance) unbiased linear estimator is the least-squares estimator*
$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$
2. *In the linear regression model (Assumption 4.3.1), the best unbiased linear estimator is*
$$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{y} \tag{4.13}$$

The first part of the Gauss-Markov theorem is a limited efficiency justification for the least-squares estimator. The justification is limited because the class of models is restricted to homoskedastic linear regression and the class of potential estimators is restricted to linear unbiased estimators. This latter restriction is particularly unsatisfactory as the theorem leaves open the possibility that a non-linear or biased estimator could have lower mean squared error than the least-squares estimator.

The second part of the theorem shows that in the (heteroskedastic) linear regression model, within the class of linear unbiased estimators the best estimator is not least-squares but is (4.13). This is called the **Generalized Least Squares** (GLS) estimator. The GLS estimator is infeasible as the matrix $\boldsymbol{D}$ is unknown. This result does not suggest a practical alternative to least-squares. We return to the issue of feasible implementation of GLS in Section 9.2.

We give a proof of the first part of the theorem below, and leave the proof of the second part for Exercise 4.3.

Proof of Theorem 4.6.1.1. Let $\boldsymbol{A}$ be any $n \times k$ function of $\boldsymbol{X}$ such that $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$. The variance of the least-squares estimator is $(\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$ and that of $\boldsymbol{A}'\boldsymbol{y}$ is $\boldsymbol{A}'\boldsymbol{A}\sigma^2$. It is sufficient to show that the difference $\boldsymbol{A}'\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}$ is positive semi-definite. Set $\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Note that $\boldsymbol{X}'\boldsymbol{C} = \boldsymbol{0}$. Then we calculate that

$$\begin{aligned}
\boldsymbol{A}'\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1} &= \left(\boldsymbol{C} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)'\left(\boldsymbol{C} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) - (\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \boldsymbol{C}'\boldsymbol{C} + \boldsymbol{C}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C} \\
&+ (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - (\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \boldsymbol{C}'\boldsymbol{C}.
\end{aligned}$$

The matrix $\boldsymbol{C}'\boldsymbol{C}$ is positive semi-definite (see Appendix A.8) as required.

4.7 Residuals

What are some properties of the residuals $\hat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ and prediction errors $\tilde{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}$, at least in the context of the linear regression model?

Recall from (3.26) that we can write the residuals in vector notation as

$$\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$$

where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the orthogonal projection matrix. Using the properties of conditional expectation

$$\mathbb{E}(\hat{\boldsymbol{e}} \mid \boldsymbol{X}) = \mathbb{E}(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{M}\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{0}$$

and

$$\operatorname{var}(\hat{\boldsymbol{e}} \mid \boldsymbol{X}) = \operatorname{var}(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{M}\operatorname{\mathbf{var}}(\boldsymbol{e} \mid \boldsymbol{X})\boldsymbol{M} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{M} \tag{4.14}$$

where $\boldsymbol{D}$ is defined in (4.10).

We can simplify this expression under the assumption of conditional homoskedasticity

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2.$$

In this case (4.14) simplies to

$$\operatorname{var}(\hat{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}\sigma^2. \tag{4.15}$$

In particular, for a single observation i, we can find the (conditional) variance of $\hat{e}_i$ by taking the i^{th} diagonal element of (4.16). Since the i^{th} diagonal element of $\boldsymbol{M}$ is $1 - h_{ii}$ as defined in (3.21) we obtain

$$\operatorname{var}(\hat{e}_i \mid \boldsymbol{X}) = \mathbb{E}\left(\hat{e}_i^2 \mid \boldsymbol{X}\right) = (1 - h_{ii})\,\sigma^2. \tag{4.16}$$

As this variance is a function of h_{ii} and hence $\boldsymbol{x}_i$, the residuals $\hat{e}_i$ are heteroskedastic even if the errors e_i are homoskedastic.

Similarly, recall from (3.40) that the prediction errors $\tilde{e}_i = (1 - h_{ii})^{-1}\hat{e}_i$ can be written in vector notation as $\tilde{\boldsymbol{e}} = \boldsymbol{M}^*\hat{\boldsymbol{e}}$ where $\boldsymbol{M}^*$ is a diagonal matrix with i^{th} diagonal element $(1 - h_{ii})^{-1}$. Thus $\tilde{\boldsymbol{e}} = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{e}$. We can calculate that

$$\mathbb{E}(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^*\boldsymbol{M}\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{0}$$

and

$$\operatorname{var}(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^*\boldsymbol{M}\operatorname{var}(\boldsymbol{e} \mid \boldsymbol{X})\boldsymbol{M}\boldsymbol{M}^* = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^*$$

which simplifies under homoskedasticity to

$$\begin{aligned}\operatorname{var}(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}) &= \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}\boldsymbol{M}^*\sigma^2 \\ &= \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}^*\sigma^2.\end{aligned}$$

The variance of the i^{th} prediction error is then

$$\begin{aligned}\operatorname{var}(\tilde{e}_i \mid \boldsymbol{X}) &= \mathbb{E}\left(\tilde{e}_i^2 \mid \boldsymbol{X}\right) \\ &= (1 - h_{ii})^{-1}(1 - h_{ii})(1 - h_{ii})^{-1}\sigma^2 \\ &= (1 - h_{ii})^{-1}\sigma^2.\end{aligned}$$

A residual with constant conditional variance can be obtained by rescaling. The **standardized residuals** are

$$\bar{e}_i = (1 - h_{ii})^{-1/2}\hat{e}_i, \tag{4.17}$$

and in vector notation

$$\bar{\boldsymbol{e}} = (\bar{e}_1, ..., \bar{e}_n)' = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{e}.$$

From our above calculations, under homoskedasticity,

$$\operatorname{var}(\bar{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{M}^{*1/2}\sigma^2$$

and

$$\operatorname{var}(\bar{e}_i \mid \boldsymbol{X}) = \mathbb{E}\left(\bar{e}_i^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.18}$$

and thus these standardized residuals have the same bias and variance as the original errors when the latter are homoskedastic.

4.8 Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}e_i^2$ can be a parameter of interest, even in a heteroskedastic regression or a projection model. σ^2 measures the variation in the "unexplained" part of the regression. Its method of moments estimator (MME) is the sample average of the squared residuals:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2$$

and equals the MLE in the normal regression model (3.28).

In the linear regression model we can calculate the mean of $\hat{\sigma}^2$. From (3.26), the properties of projection matrices and the trace operator, observe that

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = \frac{1}{n}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{M}\boldsymbol{e} = \frac{1}{n}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}'\right).$$

Then

$$\begin{aligned}\mathbb{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\right) \\ &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\right) \\ &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\boldsymbol{D}\right). \end{aligned} \tag{4.19}$$

Adding the assumption of conditional homoskedasticity $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, so that $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$, then (4.19) simplifies to

$$\begin{aligned}\mathbb{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\sigma^2\right) \\ &= \sigma^2\left(\frac{n-k}{n}\right),\end{aligned}$$

the final equality by (3.24). This calculation shows that $\hat{\sigma}^2$ is biased towards zero. The order of the bias depends on k/n, the ratio of the number of estimated coefficients to the sample size.

Another way to see this is to use (4.16). Note that

$$\begin{aligned}\mathbb{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\hat{e}_i^2 \mid \boldsymbol{X}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(1 - h_{ii}\right)\sigma^2 \\ &= \left(\frac{n-k}{n}\right)\sigma^2 \end{aligned} \tag{4.20}$$

the last equality using Theorem 3.10.1.

Since the bias takes a scale form, a classic method to obtain an unbiased estimator is by rescaling the estimator. Define

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{e}_i^2. \tag{4.21}$$

By the above calculation,

$$\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.22}$$

so

$$\mathbb{E}\left(s^2\right) = \sigma^2$$

and the estimator s^2 is unbiased for σ^2. Consequently, s^2 is known as the "bias-corrected estimator" for σ^2 and in empirical practice s^2 is the most widely used estimator for σ^2.

Interestingly, this is not the only method to construct an unbiased estimator for σ^2. An estimator constructed with the standardized residuals $\bar{e}_i$ from (4.17) is

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\bar{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(1 - h_{ii}\right)^{-1}\hat{e}_i^2. \tag{4.23}$$

You can show (see Exercise 4.6) that

$$\mathbb{E}\left(\bar{\sigma}^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.24}$$

and thus $\bar{\sigma}^2$ is unbiased for σ^2 (in the homoskedastic linear regression model).

When k/n is small (typically, this occurs when n is large), the estimators $\hat{\sigma}^2$, s^2 and $\bar{\sigma}^2$ are likely to be close. However, if not then s^2 and $\bar{\sigma}^2$ are generally preferred to $\hat{\sigma}^2$. Consequently it is best to use one of the bias-corrected variance estimators in applications.

4.9 Mean-Square Forecast Error

A major purpose of estimated regressions is to predict out-of-sample values. Consider an out-of-sample observation $(y_{n+1}, \boldsymbol{x}_{n+1})$ where $\boldsymbol{x}_{n+1}$ will be observed but not y_{n+1}. Given the coefficient estimate $\widehat{\boldsymbol{\beta}}$ the standard point estimate of $\mathbb{E}\left(y_{n+1} \mid \boldsymbol{x}_{n+1}\right) = \boldsymbol{x}_{n+1}'\boldsymbol{\beta}$ is $\tilde{y}_{n+1} = \boldsymbol{x}_{n+1}'\widehat{\boldsymbol{\beta}}$. The forecast error is the difference between the actual value y_{n+1} and the point forecast, $\tilde{e}_{n+1} = y_{n+1} - \tilde{y}_{n+1}$. The mean-squared forecast error (MSFE) is

$$MSFE_n = \mathbb{E}\tilde{e}_{n+1}^2.$$

In the linear regression model, $\tilde{e}_{n+1} = e_{n+1} - \boldsymbol{x}_{n+1}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$, so

$$\begin{aligned} MSFE_n = \mathbb{E}e_{n+1}^2 - 2\mathbb{E}\left(e_{n+1}\boldsymbol{x}_{n+1}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right) \\ + \mathbb{E}\left(\boldsymbol{x}_{n+1}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{x}_{n+1}\right). \end{aligned} \tag{4.25}$$

The first term in (4.25) is σ^2. The second term in (4.25) is zero since $e_{n+1}\boldsymbol{x}_{n+1}'$ is independent of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and both are mean zero. Using the properties of the trace operator, the third term in (4.25) is

$$\begin{aligned} &\operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\mathbb{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right) \\ &= \operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\mathbb{E}\left(\mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mid \boldsymbol{X}\right)\right)\right) \\ &= \operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\mathbb{E}\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\right) \\ &= \mathbb{E}\operatorname{tr}\left(\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\right) \\ &= \mathbb{E}\left(\boldsymbol{x}_{n+1}'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}_{n+1}\right) \end{aligned} \tag{4.26}$$

where we use the fact that $\boldsymbol{x}_{n+1}$ is independent of $\widehat{\boldsymbol{\beta}}$, the definition $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mid \boldsymbol{X}\right)$ and the fact that $\boldsymbol{x}_{n+1}$ is independent of $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$. Thus

$$MSFE_n = \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_{n+1}'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}_{n+1}\right).$$

Under conditional homoskedasticity, this simplifies to

$$MSFE_n = \sigma^2\left(1 + \mathbb{E}\left(\boldsymbol{x}_{n+1}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_{n+1}\right)\right).$$

A simple estimator for the MSFE is obtained by averaging the squared prediction errors (3.41)

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2$$

where $\tilde{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} = \hat{e}_i(1-h_{ii})^{-1}$. Indeed, we can calculate that

$$\begin{aligned}\mathbb{E}\tilde{\sigma}^2 &= \mathbb{E}\tilde{e}_i^2 \\ &= \mathbb{E}\left(e_i - \boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)\right)^2 \\ &= \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)'\boldsymbol{x}_i\right).\end{aligned}$$

By a similar calculation as in (4.26) we find

$$\mathbb{E}\tilde{\sigma}^2 = \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_i'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}_{(-i)}}\boldsymbol{x}_i\right) = MSFE_{n-1}.$$

his is the MSFE based on a sample of size $n-1$, rather than size n. The difference arises because the in-sample prediction errors $\tilde{e}_i$ for $i \leq n$ are calculated using an effective sample size of $n-1$, while the out-of sample prediction error $\tilde{e}_{n+1}$ is calculated from a sample with the full n observations. Unless n is very small we should expect $MSFE_{n-1}$ (the MSFE based on $n-1$ observations) to be close to $MSFE_n$ (the MSFE based on n observations). Thus $\tilde{\sigma}^2$ is a reasonable estimator for $MSFE_n$.

Theorem 4.9.1 ***MSFE***
In the linear regression model (Assumption 4.3.1)

$$MSFE_n = \mathbb{E}\tilde{e}_{n+1}^2 = \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_{n+1}'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}_{n+1}\right)$$

where $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$. *Furthermore,* $\tilde{\sigma}^2$ *defined in (3.41) is an unbiased estimator of* $MSFE_{n-1}$:

$$\mathbb{E}\tilde{\sigma}^2 = MSFE_{n-1}$$

4.10 Covariance Matrix Estimation Under Homoskedasticity

For inference, we need an estimate of the covariance matrix $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ of the least-squares estimator. In this section we consider the homoskedastic regression model (Assumption 4.3.2).

Under homoskedasticity, the covariance matrix takes the relatively simple form

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2$$

which is known up to the unknown scale σ^2. In Section 4.8 we discussed three estimators of σ^2. The most commonly used choice is s^2, leading to the classic covariance matrix estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^0 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}s^2. \tag{4.27}$$

Since s^2 is conditionally unbiased for σ^2, it is simple to calculate that $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^0$ is conditionally unbiased for $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ under the assumption of homoskedasticity:

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^0 \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\ &= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}.\end{aligned}$$

This estimator was the dominant covariance matrix estimator in applied econometrics for many years, and is still the default method in most regression packages.

If the estimator (4.27) is used, but the regression error is heteroskedastic, it is possible for $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$ to be quite biased for the correct covariance matrix $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$. For example, suppose $k = 1$ and $\sigma_i^2 = x_i^2$ with $\mathbb{E}x_i = 0$. The ratio of the true variance of the least-squares estimator to the expectation of the variance estimator is

$$\frac{\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}}{\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0} \mid \boldsymbol{X}\right)} = \frac{\sum_{i=1}^{n} x_i^4}{\sigma^2 \sum_{i=1}^{n} x_i^2} \simeq \frac{\mathbb{E}x_i^4}{\left(\mathbb{E}x_i^2\right)^2} = \kappa.$$

(Notice that we use the fact that $\sigma_i^2 = x_i^2$ implies $\sigma^2 = \mathbb{E}\sigma_i^2 = \mathbb{E}x_i^2$.) The constant κ is the standardized forth moment (or kurtosis) of the regressor x_i, and can be any number greater than one. For example, if $x_i \sim \mathrm{N}\left(0, \sigma^2\right)$ then $\kappa = 3$, so the true variance $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ is three times larger than the expected homoskedastic estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$. But κ can be much larger. Suppose, for example, that $x_i \sim \chi_1^2 - 1$. In this case $\kappa = 15$, so that the true variance $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ is fifteen times larger than the expected homoskedastic estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$. While this is an extreme and constructed example, the point is that the classic covariance matrix estimator (4.27) may be quite biased when the homoskedasticity assumption fails.

4.11 Covariance Matrix Estimation Under Heteroskedasticity

In the previous section we showed that that the classic covariance matrix estimator can be highly biased if homoskedasticity fails. In this section we show how to contruct covariance matrix estimators which do not require homoskedasticity.

Recall that the general form for the covariance matrix is

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

This depends on the unknown matrix $\boldsymbol{D}$ which we can write as

$$\begin{aligned}\boldsymbol{D} &= \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right) \\ &= \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right) \\ &= \mathbb{E}\left(\boldsymbol{D}_0 \mid \boldsymbol{X}\right)\end{aligned}$$

where $\boldsymbol{D}_0 = \operatorname{diag}\left(e_1^2, ..., e_n^2\right)$. Thus $\boldsymbol{D}_0$ is a conditionally unbiased estimator for $\boldsymbol{D}$. If the squared errors e_i^2 were observable, we could construct the unbiased estimator

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{ideal} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}_0\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' e_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

Indeed,

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{ideal} \mid \boldsymbol{X}\right) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\mathbb{E}\left(e_i^2 \mid \boldsymbol{X}\right)\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\sigma_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\end{aligned}$$

verifying that $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{ideal}$ is unbiased for $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$

Since the errors e_i^2 are unobserved, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{ideal}$ is not a feasible estimator. However, we can replace the errors e_i with the least-squares residuals $\hat{e}_i$. Making this substitution we obtain the estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.28}$$

We know, however, that $\hat{e}_i^2$ is biased towards zero. To estimate the variance σ^2 the unbiased estimator s^2 scales the moment estimator $\hat{\sigma}^2$ by $n/(n-k)$. Making the same adjustment we obtain the estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \left(\frac{n}{n-k}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.29}$$

While the scaling by $n/(n-k)$ is ad hoc, it is recommended over the unscaled estimator (4.28).

Alternatively, we could use the prediction errors $\tilde{e}_i$ or the standardized residuals $\bar{e}_i$, yielding the estimators

$$\begin{aligned}\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\tilde{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\end{aligned} \tag{4.30}$$

and

$$\begin{aligned}\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\bar{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned} \tag{4.31}$$

The four estimators $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, and $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ are collectively called **robust**, **heteroskedasticity-consistent**, or **heteroskedasticity-robust** covariance matrix estimators. The estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ was first developed by Eicker (1963) and introduced to econometrics by White (1980), and is sometimes called the **Eicker-White** or **White** covariance matrix estimator. The scaled estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is the default robust covariance matrix estimator implemented in Stata. The estimator $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ was introduced by Andrews (1991) based on the principle of leave-one-out cross-validation (and is implemented using the vce(hc3) option in Stata). The estimator $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ was introduced by Horn, Horn and Duncan (1975) (and is implemented using the vce(hc2) option in Stata).

Since $(1-h_{ii})^{-2} > (1-h_{ii})^{-1} > 1$ it is straightforward to show that

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} < \overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} < \widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \tag{4.32}$$

(See Exercise 4.7). The inequality $\boldsymbol{A} < \boldsymbol{B}$ when applied to matrices means that the matrix $\boldsymbol{B} - \boldsymbol{A}$ is positive definite.

In general, the bias of the covariance matrix estimators is quite complicated, but they greatly

simplify under the assumption of homoskedasticity (4.3). For example, using (4.16),

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\mathbb{E}\left(\hat{e}_i^2 \mid \boldsymbol{X}\right)\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\left(1-h_{ii}\right)\sigma^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'h_{ii}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\
&< \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\
&= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}.
\end{aligned}$$

This calculation shows that $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ is biased towards zero.

Similarly, (again under homoskedasticity) we can calculate that $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is biased away from zero, specifically

$$\mathbb{E}\left(\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) > \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \tag{4.33}$$

while the estimator $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is unbiased

$$\mathbb{E}\left(\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2. \tag{4.34}$$

(See Exercise 4.8.)

It might seem rather odd to compare the bias of heteroskedasticity-robust estimators under the assumption of homoskedasticity, but it does give us a baseline for comparison.

We have introduced five covariance matrix estimators, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, and $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$. Which should you use? The classic estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$ is typically a poor choice, as it is only valid under the unlikely homoskedasticity restriction. For this reason it is not typically used in contemporary econometric research. Unfortunately, standard regression packages set their default choice as $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$, so users must intentionally select a robust covariance matrix estimator.

Of the four robust estimators, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ and $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ are the most commonly used, and in particular $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is the default robust covariance matrix option in Stata. However, $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ may be the preferred choice based on its improved bias. As $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is simple to implement, this should not be a barrier.

Halbert L. White

Hal White (1950-2012) of the United States was an influential econometrician of recent years. His 1980 paper on heteroskedasticity-consistent covariance matrix estimation for many years has been the most cited paper in economics. His research was central to the movement to view econometric models as approximations, and to the drive for increased mathematical rigor in the discipline. In addition to being a highly prolific and influential scholar, he also co-founded the economic consulting firm Bates White.

4.12 Standard Errors

A variance estimator such as $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is an estimate of the variance of the distribution of $\widehat{\boldsymbol{\beta}}$. A more easily interpretable measure of spread is its square root – the standard deviation. This is so important when discussing the distribution of parameter estimates, we have a special name for estimates of their standard deviation.

Definition 4.12.1 *A **standard error** $s(\widehat{\beta})$ for a real-valued estimator $\widehat{\beta}$ is an estimate of the standard deviation of the distribution of $\widehat{\beta}$.*

When $\boldsymbol{\beta}$ is a vector with estimate $\widehat{\boldsymbol{\beta}}$ and covariance matrix estimate $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, standard errors for individual elements are the square roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$. That is,

$$s(\hat{\beta}_j) = \sqrt{\widehat{\boldsymbol{V}}_{\hat{\beta}_j}} = \sqrt{\left[\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\right]_{jj}}.$$

As we discussed in the previous section, there are multiple possible covariance matrix estimators, so standard errors are not unique. It is therefore important to understand what formula and method is used by an author when studying their work. It is also important to understand that a particular standard error may be relevant under one set of model assumptions, but not under another set of assumptions.

To illustrate, we return to the log wage regression (3.11) of Section 3.7. We calculate that $s^2 = 0.160$. Therefore the homoskedastic covariance matrix estimate is

$$\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}} = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} 0.160 = \begin{pmatrix} 0.002 & -0.031 \\ -0.031 & 0.499 \end{pmatrix}.$$

We also calculate that

$$\sum_{i=1}^{n} (1 - h_{ii})^{-1} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2 = \begin{pmatrix} 763.26 & 48.513 \\ 48.513 & 3.1078 \end{pmatrix}.$$

Therefore the Horn-Horn-Duncan covariance matrix estimate is

$$\begin{aligned} \overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} &= \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \begin{pmatrix} 763.26 & 48.513 \\ 48.513 & 3.1078 \end{pmatrix} \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.001 & -0.015 \\ -0.015 & 0.243 \end{pmatrix}. \end{aligned} \tag{4.35}$$

The standard errors are the square roots of the diagonal elements of these matrices. A conventional format to write the estimated equation with standard errors is

$$\begin{array}{rccc} \widehat{\log(Wage)} = & 0.155 & Education + & 0.698 \\ & (0.031) & & (0.493) \end{array}.$$

Alternatively, standard errors could be calculated using the other formulae. We report the different standard errors in the following table.

	Education	Intercept
Homoskedastic (4.27)	0.045	0.707
White (4.28)	0.029	0.461
Scaled White (4.29)	0.030	0.486
Andrews (4.30)	0.033	0.527
Horn-Horn-Duncan (4.31)	0.031	0.493

The homoskedastic standard errors are noticably different (larger, in this case) than the others, but the four robust standard errors are quite close to one another.

4.13 Computation

We illustrate methods to compute standard errors for equation (3.12) extending the code of Section 3.20.

Stata do File (continued)

```
*       Homoskedastic formula (4.27):
reg wage education experience exp2 if (mnwf == 1)
*       Scaled White formula (4.29):
reg wage education experience exp2 if (mnwf == 1), r
*       Andrews formula (4.30):
reg wage education experience exp2 if (mnwf == 1), vce(hc3)
*       Horn-Horn-Duncan formula (4.31):
reg wage education experience exp2 if (mnwf == 1), vce(hc2)
```

Gauss Program File (continued)

```
n=rows(y);
k=cols(x);
a=n/(n-k);
sig2=(e'e)/(n-k);
u1=x.*e;
u2=x.*(e./(1-leverage));
u3=x.*(e./sqrt(1-leverage));
xx=inv(x'x);
v0=xx*sig2;
v1=xx*(u1'u1)*xx;
v1a=a*xx*(u1'u1)*xx;
v2=xx*(u2'u2)*xx;
v3=xx*(u3'u3)*xx
s0=sqrt(diag(v0));          @ Homoskedastic formula @
s1=sqrt(diag(v1));          @ White formula @
s1a=sqrt(diag(v1a));        @ Scaled White formula @
s2=sqrt(diag(v2));          @ Andrews formula @
s3=sqrt(diag(v3));          @ Horn-Horn-Duncan formula @
```

R Program File (continued)

```
n <- nrow(y)
k <- ncol(x)
a <- n/(n-k)
sig2 <- (t(e) %*% e)/(n-k)
u1 <- x*(e%*%matrix(1,1,k))
u2 <- x*((e/(1-leverage))%*%matrix(1,1,k))
u3 <- x*((e/sqrt(1-leverage))%*%matrix(1,1,k))
v0 <- xx*sig2
xx <- solve(t(x)%*%x)
v1 <- xx %*% (t(u1)%*%u1) %*% xx
v1a <- a * xx %*% (t(u1)%*%u1) %*% xx
v2 <- xx %*% (t(u2)%*%u2) %*% xx
v3 <- xx %*% (t(u3)%*%u3) %*% xx
s0 <- sqrt(diag(v0))          # Homoskedastic formula
s1 <- sqrt(diag(v1))          # White formula
s1a <- sqrt(diag(v1a))        # Scaled White formula
s2 <- sqrt(diag(v2))          # Andrews formula
s3 <- sqrt(diag(v3))          # Horn-Horn-Duncan formula
```

Matlab Program File (continued)

```
[n,k]=size(x);
a=n/(n-k);
sig2=(e'*e)/(n-k);
u1=x.*(e*ones(1,k));
u2=x.*((e./(1-leverage))*ones(1,k));
u3=x.*((e./sqrt(1-leverage))*ones(1,k));
xx=inv(x'*x);
v0=xx*sig2;
v1=xx*(u1'*u1)*xx;
v1a=a*xx*(u1'*u1)*xx;
v2=xx*(u2'*u2)*xx;
v3=xx*(u3'*u3)*xx;
s0=sqrt(diag(v0));          # Homoskedastic formula
s1=sqrt(diag(v1));          # White formula
s1a=sqrt(diag(v1a));        # Scaled White formula
s2=sqrt(diag(v2));          # Andrews formula
s3=sqrt(diag(v3));          # Horn-Horn-Duncan formula
```

4.14 Measures of Fit

As we described in the previous chapter, a commonly reported measure of regression fit is the regression R^2 defined as

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

where $\hat{\sigma}_y^2 = n^{-1}\sum_{i=1}^n (y_i - \overline{y})^2$. R^2 can be viewed as an estimator of the population parameter

$$\rho^2 = \frac{\text{var}\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)}{\text{var}(y_i)} = 1 - \frac{\sigma^2}{\sigma_y^2}.$$

However, $\hat{\sigma}^2$ and $\hat{\sigma}_y^2$ are biased estimators. Theil (1961) proposed replacing these by the unbiased versions s^2 and $\tilde{\sigma}_y^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \overline{y})^2$ yielding what is known as **R-bar-squared** or **adjusted R-squared**:

$$\overline{R}^2 = 1 - \frac{s^2}{\tilde{\sigma}_y^2} = 1 - \frac{(n-1)\sum_{i=1}^n \hat{e}_i^2}{(n-k)\sum_{i=1}^n (y_i - \bar{y})^2}.$$

While $\overline{R}^2$ is an improvement on R^2, a much better improvement is

$$\widetilde{R}^2 = 1 - \frac{\sum_{i=1}^n \tilde{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\tilde{\sigma}^2}{\hat{\sigma}_y^2}$$

where $\tilde{e}_i$ are the prediction errors (3.38) and $\tilde{\sigma}^2$ is the MSPE from (3.41). As described in Section (4.9), $\tilde{\sigma}^2$ is a good estimator of the out-of-sample mean-squared forecast error, so $\widetilde{R}^2$ is a good estimator of the percentage of the forecast variance which is explained by the regression forecast. In this sense, $\widetilde{R}^2$ is a good measure of fit.

One problem with R^2, which is partially corrected by $\overline{R}^2$ and fully corrected by $\widetilde{R}^2$, is that R^2 necessarily increases when regressors are added to a regression model. This occurs because R^2 is a negative function of the sum of squared residuals which cannot increase when a regressor is added. In contrast, $\overline{R}^2$ and $\widetilde{R}^2$ are non-monotonic in the number of regressors. $\widetilde{R}^2$ can even be negative, which occurs when an estimated model predicts worse than a constant-only model.

In the statistical literature the MSPE $\tilde{\sigma}^2$ is known as the **leave-one-out cross validation** criterion, and is popular for model comparison and selection, especially in high-dimensional (nonparametric) contexts. It is equivalent to use $\widetilde{R}^2$ or $\tilde{\sigma}^2$ to compare and select models. Models with high $\widetilde{R}^2$ (or low $\tilde{\sigma}^2$) are better models in terms of expected out of sample squared error. In contrast, R^2 cannot be used for model selection, as it necessarily increases when regressors are added to a regression model. $\overline{R}^2$ is also an inappropriate choice for model selection (it tends to select models with too many parameters), though a justification of this assertion requires a study of the theory of model selection. Unfortunately, $\overline{R}^2$ is routinely used by some economists, possibly as a hold-over from previous generations.

In summary, it is recommended to calculate and report $\widetilde{R}^2$ and/or $\tilde{\sigma}^2$ in regression analysis, and omit R^2 and $\overline{R}^2$.

Henri Theil

Henri Theil (1924-2000) of Holland invented $\overline{R}^2$ and two-stage least squares, both of which are routinely seen in applied econometrics. He also wrote an early influential advanced textbook on econometrics (Theil, 1971).

4.15 Empirical Example

We again return to our wage equation, but use a much larger sample of all individuals with at least 12 years of education. For regressors we include years of education, potential work experience, experience squared, and dummy variable indicators for the following: female, female union member,

male union member, married female[1], married male, formerly married female[2], formerly married male, hispanic, black, American Indian, Asian, and mixed race[3] . The available sample is 46,943 so the parameter estimates are quite precise and reported in Table 4.1. For standard errors we use the unbiased Horn-Horn-Duncan formula.

Table 4.1 displays the parameter estimates in a standard tabular format. The table clearly states the estimation method (OLS), the dependent variable (log(Wage)), and the regressors are clearly labeled. Both parameter estimates and standard errors are reported for all coefficients. In addition to the coefficient estimates, the table also reports the estimated error standard deviation and the sample size. These are useful summary measures of fit which aid readers.

Table 4.1
OLS Estimates of Linear Equation for Log(Wage)

	$\hat{\beta}$	$s(\hat{\beta})$
Education	0.117	0.001
Experience	0.033	0.001
Experience^2/100	-0.056	0.002
Female	-0.098	0.011
Female Union Member	0.023	0.020
Male Union Member	0.095	0.020
Married Female	0.016	0.010
Married Male	0.211	0.010
Formerly Married Female	-0.006	0.012
Formerly Married Male	0.083	0.015
Hispanic	-0.108	0.008
Black	-0.096	0.008
American Indian	-0.137	0.027
Asian	-0.038	0.013
Mixed Race	-0.041	0.021
Intercept	0.909	0.021
$\hat{\sigma}$	0.565	
Sample Size	46,943	

Note: Standard errors are heteroskedasticity-consistent (Horn-Horn-Duncan formula)

As a general rule, it is advisable to always report standard errors along with parameter estimates. This allows readers to assess the precision of the parameter estimates, and as we will discuss in later chapters, form confidence intervals and t-tests for individual coefficients if desired.

The results in Table 4.1 confirm our earlier findings that the return to a year of education is approximately 12%, the return to experience is concave, that single women earn approximately 10% less then single men, and blacks earn about 10% less than whites. In addition, we see that Hispanics earn about 11% less than whites, American Indians 14% less, and Asians and Mixed races about 4% less. We also see there are wage premiums for men who are members of a labor union (about 10%), married (about 21%) or formerly married (about 8%), but no similar premiums are apparant for women.

[1] Defining "married" as marital code 1, 2, or 3.
[2] Defining "formerly married" as marital code 4, 5, or 6.
[3] Race code 6 or higher.

4.16 Multicollinearity

If $\boldsymbol{X}'\boldsymbol{X}$ is singular, then $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\widehat{\boldsymbol{\beta}}$ are not defined. This situation is called **strict multicollinearity**, as the columns of $\boldsymbol{X}$ are linearly dependent, i.e., there is some $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $\boldsymbol{X}\boldsymbol{\alpha} = \mathbf{0}$. Most commonly, this arises when sets of regressors are included which are identically related. For example, if $\boldsymbol{X}$ includes both the logs of two prices and the log of the relative prices, $\log(p_1)$, $\log(p_2)$ and $\log(p_1/p_2)$, for then $\boldsymbol{X}'\boldsymbol{X}$ will necessarily be singular. When this happens, the applied researcher quickly discovers the error as the statistical software will be unable to construct $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Since the error is discovered quickly, this is rarely a *problem* for applied econometric practice.

The more relevant situation is **near multicollinearity**, which is often called "multicollinearity" for brevity. This is the situation when the $\boldsymbol{X}'\boldsymbol{X}$ matrix is *near* singular, when the columns of $\boldsymbol{X}$ are *close* to linearly dependent. This definition is not precise, because we have not said what it means for a matrix to be "near singular". This is one difficulty with the definition and interpretation of multicollinearity.

One potential complication of near singularity of matrices is that the numerical reliability of the calculations may be reduced. In practice this is rarely an important concern, except when the number of regressors is very large.

A more relevant implication of near multicollinearity is that individual coefficient estimates will be imprecise. We can see this most simply in a homoskedastic linear regression model with two regressors

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i,$$

and

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In this case

$$\operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n\left(1-\rho^2\right)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

The correlation ρ indexes collinearity, since as ρ approaches 1 the matrix becomes singular. We can see the effect of collinearity on precision by observing that the variance of a coefficient estimate $\sigma^2\left[n\left(1-\rho^2\right)\right]^{-1}$ approaches infinity as ρ approaches 1. Thus the more "collinear" are the regressors, the worse the precision of the individual coefficient estimates.

What is happening is that when the regressors are highly dependent, it is statistically difficult to disentangle the impact of β_1 from that of β_2. As a consequence, the precision of individual estimates are reduced. The imprecision, however, will be reflected by large standard errors, so there is no distortion in inference.

Some earlier textbooks overemphasized a concern about multicollinearity. A very amusing parody of these texts appeared in Chapter 23.3 of Goldberger's *A Course in Econometrics* (1991), which is reprinted below. To understand his basic point, you should notice how the estimation variance $\sigma^2\left[n\left(1-\rho^2\right)\right]^{-1}$ depends equally and symmetrically on the correlation ρ and the sample size n.

Arthur S. Goldberger

Art Goldberger (1930-2009) was one of the most distinguished members of the Department of Economics at the University of Wisconsin. His PhD thesis developed an early macroeconometric forecasting model (known as the Klein-Goldberger model) but most of his career focused on microeconometric issues. He was the leading pioneer of what has been called the Wisconsin Tradition of empirical work – a combination of formal econometric theory with a careful critical analysis of empirical work. Goldberger wrote a series of highly regarded and influential graduate econometric textbooks, including including *Econometric Theory* (1964), *Topics in Regression Analysis* (1968), and *A Course in Econometrics* (1991).

Micronumerosity

Arthur S. Goldberger

A Course in Econometrics (1991), Chapter 23.3

Econometrics texts devote many pages to the problem of multicollinearity in multiple regression, but they say little about the closely analogous problem of small sample size in estimating a univariate mean. Perhaps that imbalance is attributable to the lack of an exotic polysyllabic name for "small sample size." If so, we can remove that impediment by introducing the term *micronumerosity*.

Suppose an econometrician set out to write a chapter about small sample size in sampling from a univariate population. Judging from what is now written about multicollinearity, the chapter might look like this:

1. *Micronumerosity*

   The extreme case, "exact micronumerosity," arises when $n = 0$, in which case the sample estimate of μ is not unique. (Technically, there is a violation of the rank condition $n > 0$: the matrix 0 is singular.) The extreme case is easy enough to recognize. "Near micronumerosity" is more subtle, and yet very serious. It arises when the rank condition $n > 0$ is barely satisfied. Near micronumerosity is very prevalent in empirical economics.

2. *Consequences of micronumerosity*

   The consequences of micronumerosity are serious. Precision of estimation is reduced. There are two aspects of this reduction: estimates of μ may have large errors, and not only that, but $V_{\bar{y}}$ will be large.

   Investigators will sometimes be led to accept the hypothesis $\mu = 0$ because $\bar{y}/\hat{\sigma}_{\bar{y}}$ is small, even though the true situation may be not that $\mu = 0$ but simply that the sample data have not enabled us to pick μ up.

   The estimate of μ will be very sensitive to sample data, and the addition of a few more observations can sometimes produce drastic shifts in the sample mean.

   The true μ may be sufficiently large for the null hypothesis $\mu = 0$ to be rejected, even though $V_{\bar{y}} = \sigma^2/n$ is large because of micronumerosity. But if the true μ is small (although nonzero) the hypothesis $\mu = 0$ may mistakenly be accepted.

3. *Testing for micronumerosity*

   Tests for the presence of micronumerosity require the judicious use of various fingers. Some researchers prefer a single finger, others use their toes, still others let their thumbs rule.

   A generally reliable guide may be obtained by counting the number of observations. Most of the time in econometric analysis, when n is close to zero, it is also far from infinity.

   Several test procedures develop critical values n^*, such that micronumerosity is a problem only if n is smaller than n^*. But those procedures are questionable.

4. *Remedies for micronumerosity*

   If micronumerosity proves serious in the sense that the estimate of μ has an unsatisfactorily low degree of precision, we are in the statistical position of not being able to make bricks without straw. The remedy lies essentially in the acquisition, if possible, of larger samples from the same population.

   But more data are no remedy for micronumerosity if the additional data are simply "more of the same." So obtaining lots of small samples from the same population will not help.

4.17 Normal Regression Model

In the special case of the normal linear regression model introduced in Section 3.18, we can derive exact sampling distributions for the least-squares estimator, residuals, and variance estimator.

In particular, under the normality assumption $e_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(0, \sigma^2\right)$ then we have the multivariate implication

$$\boldsymbol{e} \mid \boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_n \sigma^2\right).$$

That is, the error vector $\boldsymbol{e}$ is independent of $\boldsymbol{X}$ and is normally distributed. Since linear functions of normals are also normal, this implies that conditional on $\boldsymbol{X}$

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\boldsymbol{e}} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\ \boldsymbol{M} \end{pmatrix} \boldsymbol{e} \sim \mathrm{N}\left(0, \begin{pmatrix} \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} & 0 \\ 0 & \sigma^2\boldsymbol{M} \end{pmatrix}\right)$$

where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. Since uncorrelated normal variables are independent, it follows that $\widehat{\boldsymbol{\beta}}$ is independent of any function of the OLS residuals including the estimated error variance s^2 or $\hat{\sigma}^2$ or prediction errors $\tilde{\boldsymbol{e}}$.

The spectral decomposition (see equation (A.5)) of $\boldsymbol{M}$ yields

$$\boldsymbol{M} = \boldsymbol{H}\begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{H}'$$

where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$. Let $\boldsymbol{u} = \sigma^{-1}\boldsymbol{H}'\boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{H}'\boldsymbol{H}) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Then

$$\begin{aligned}
\frac{n\hat{\sigma}^2}{\sigma^2} &= \frac{(n-k)\,s^2}{\sigma^2} \\
&= \frac{1}{\sigma^2}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \\
&= \frac{1}{\sigma^2}\boldsymbol{e}'\boldsymbol{M}\,\boldsymbol{e} \\
&= \frac{1}{\sigma^2}\boldsymbol{e}'\boldsymbol{H}\begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{H}'\boldsymbol{e} \\
&= \boldsymbol{u}'\begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{u} \\
&\sim \chi^2_{n-k},
\end{aligned}$$

a chi-square distribution with $n-k$ degrees of freedom.

Furthermore, if standard errors are calculated using the homoskedastic formula (4.27)

$$\frac{\widehat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} = \frac{\widehat{\beta}_j - \beta_j}{s\sqrt{\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}}} \sim \frac{\mathrm{N}\left(0, \sigma^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}\right)}{\sqrt{\frac{\sigma^2}{n-k}\chi^2_{n-k}}\sqrt{\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}}} = \frac{\mathrm{N}(0,1)}{\sqrt{\frac{\chi^2_{n-k}}{n-k}}} \sim t_{n-k}$$

a t distribution with $n-k$ degrees of freedom.

Theorem 4.17.1 ***Normal Regression***
In the linear regression model (Assumption 4.3.1) if e_i is independent of $\boldsymbol{x}_i$ and distributed $\mathrm{N}\left(0, \sigma^2\right)$ then

- $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim \mathrm{N}\left(\boldsymbol{0}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$
- $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$
- $\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$

These are the exact finite-sample distributions of the least-squares estimator and variance estimators, and are the basis for traditional inference in linear regression.

While elegant, the difficulty in applying Theorem 4.17.1 is that the normality assumption is too restrictive to be empirically plausible, and therefore inference based on Theorem 4.17.1 has no guarantee of accuracy. We develop an alternative inference theory based on large sample (asymptotic) approximations in the following chapter.

William Gosset

William S. Gosset (1876-1937) of England is most famous for his derivation of the student's t distribution, published in the paper "The probable error of a mean" in 1908. At the time, Gosset worked at Guiness Brewery, which prohibited its employees from publishing in order to prevent the possible loss of trade secrets. To circumvent this barrier, Gosset published under the pseudonym "Student". Consequently, this famous distribution is known as the student's t rather than Gosset's t!

Exercises

Exercise 4.1 Explain the difference between $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'$ and $\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)$.

Exercise 4.2 True or False. If $y_i = x_i\beta + e_i$, $x_i \in \mathbb{R}$, $\mathbb{E}(e_i \mid x_i) = 0$, and $\hat{e}_i$ is the OLS residual from the regression of y_i on x_i, then $\sum_{i=1}^{n} x_i^2\hat{e}_i = 0$.

Exercise 4.3 Prove Theorem 4.6.1.2.

Exercise 4.4 In a linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = 0, \quad \operatorname{var}(\boldsymbol{e} \mid \boldsymbol{X}) = \sigma^2\boldsymbol{\Omega}$$

with $\boldsymbol{\Omega}$ a known function of $\boldsymbol{X}$, the GLS estimator is

$$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}\right),$$

the residual vector is $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$, and an estimate of σ^2 is

$$s^2 = \frac{1}{n-k}\hat{\boldsymbol{e}}'\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{e}}.$$

(a) Find $\mathbb{E}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$.

(b) Find $\operatorname{var}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$.

(c) Prove that $\hat{\boldsymbol{e}} = \boldsymbol{M}_1\boldsymbol{e}$, where $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}$.

(d) Prove that $\boldsymbol{M}_1'\boldsymbol{\Omega}^{-1}\boldsymbol{M}_1 = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}$.

(e) Find $\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right)$.

(f) Is s^2 a reasonable estimator for σ^2?

Exercise 4.5 Let $(y_i, \boldsymbol{x}_i)$ be a random sample with $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$. Consider the **Weighted Least Squares** (WLS) estimator of $\boldsymbol{\beta}$

$$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}\right)$$

where $\boldsymbol{W} = \operatorname{diag}(w_1, ..., w_n)$ and $w_i = x_{ji}^{-2}$, where x_{ji} is one of the $\boldsymbol{x}_i$.

(a) In which contexts would $\widetilde{\boldsymbol{\beta}}$ be a good estimator?

(b) Using your intuition, in which situations would you expect that $\widetilde{\boldsymbol{\beta}}$ would perform better than OLS?

Exercise 4.6 Show (4.24) in the homoskedastic regression model.

Exercise 4.7 Prove (4.32).

Exercise 4.8 Show (4.33) and (4.34) in the homoskedastic regression model.

Exercise 4.9 Let $\mu = \mathbb{E}(y_i)$, $\sigma^2 = \mathbb{E}(y_i - \mu)^2$ and $\mu_3 = \mathbb{E}(y_i - \mu)^3$ and consider the sample mean $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Find $\mathbb{E}(\overline{y} - \mu)^3$ as a function of μ, σ^2, μ_3 and n.

Exercise 4.10 Take the simple regression model $y_i = x_i\beta + e_i$, $x_i \in \mathbb{R}$, $\mathbb{E}(e_i \mid x_i) = 0$. Define $\sigma_i^2 = \mathbb{E}(e_i^2 \mid x_i)$ and $\mu_{3i} = \mathbb{E}(e_i^3 \mid x_i)$ and consider the OLS coefficient $\widehat{\beta}$. Find $\mathbb{E}\left(\left(\widehat{\beta} - \beta\right)^3 \mid \boldsymbol{X}\right)$.

Exercise 4.11 Continue the empirical analysis in Exercise 3.19.

1. Calculate standard errors using the homoskedasticity formula and using the four covariance matrices from Section 4.11.
2. Repeat in your second programming language. Are they identical?

Exercise 4.12 Continue the empirical analysis in Exercise 3.21. Calculate standard errors using the Horn-Horn-Duncan method. Repeat in your second programming language. Are they identical?