Chapter 6

Bivariate Least Squares

In economics we frequently encounter models where the relationship between variables enables us to classify the variables into an independent variables which are not modeled and and a dependent variable that is generated by the model. In order to better understand the underlying structure of such models we begin by restricting our attention to the bivariate case, where there is only one independent variable, and the linear case, where the relationship among the variables is linear. Since the relationship is not precise, we specify an additive error term, which is assumed to be a random variable. An example would the wage equation \( w_t = \alpha + \beta E_t + u_t \), where \( w_t \) is the wage earned by individual \( t \) and \( E_t \) is the education level attained by the individual.

6.1 Introduction

6.1.1 The Bivariate Linear Model

Consider the model

\[
y_i = \alpha + \beta x_i + u_i, \quad i = 1, 2, \ldots, n
\]

where \( y_i \) is the dependent variable, \( x_i \) is the independent or explanatory variable, and \( u_i \) is the unobservable disturbance. This is the bivariate linear regression model. It is linear in the variables (given \( \alpha \) and \( \beta \)), linear in the parameters (given \( x_i \)), and linear in the disturbances.

In matrix form, we can gather all the observations together as

\[
\begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n \\
\end{pmatrix} = \begin{pmatrix}
1 & x_1 \\
1 & x_2 \\
\vdots & \vdots \\
1 & x_n \\
\end{pmatrix} \begin{pmatrix}
\alpha \\
\beta \\
\end{pmatrix} + \begin{pmatrix}
u_1 \\
u_2 \\
\vdots \\
u_n \\
\end{pmatrix},
\]

or, more compactly,

\[
y = X\beta + u,
\]
where \( y \) is an \( n \times 1 \) vector of all \( n \) observations on \( y_i \), \( u \) is an \( n \times 1 \) vector of all \( n \) observations on \( u_i \), \( X \) is an \( n \times 2 \) matrix with the first column a vector of ones and the second column the observations on \( x_i \), and \( \beta=(\alpha \beta)' \).

**Example 6.1.** The consumption example from the first chapter is linear:

\[
C_t = \alpha + \beta D_t + u_t
\]

where \( C_t \) is aggregate consumption and \( D_t \) is aggregate disposable income at time \( t \). Except for the definition of the variables and use of the index \( t \) instead of \( i \) it is the same model. □

**Example 6.2.** Neither of the following two equations is linear:

\[
y_i = (\alpha + \beta x_i) u_i
\]

\[
y_i = \alpha x_i^\beta + u_i.
\]

The first is nonlinear in the variables while the second is nonlinear in the parameters. □

**6.1.2 Assumptions**

For the disturbances, we assume that

(i) \( E(u_i) = 0 \), for all \( i \)

(ii) \( E(u_i^2) = \sigma^2 \), for all \( i \)

(iii) \( E(u_i u_j) = 0 \), for all \( i \neq j \)

For the independent variable, We suppose

(iv) \( x_i \) nonstochastic for all \( i \)

(v) \( x_i \) nonconstant

For purposes of inference in finite samples, we sometime assume

(vi) \( u_i \sim i.i.d. N(0, \sigma^2) \), for all \( i \).

**6.1.3 Line Fitting**

Consider a scatter of points for two unspecified variables \( x \) and \( y \) as shown in Figure 6.1.
We assume that $x$ and $y$ are, in some way, linearly related. If this were exactly true then we can write

$$y_i = \alpha + \beta x_i,$$

for some $\alpha$ and $\beta$. Unfortunately, we find that no single line “matches” all the observations. Since no single line is entirely consistent with the data, we might choose $\alpha$ and $\beta$ that best “fits” the data, in some sense. Define

$$e_i = y_i - (\alpha + \beta x_i), \text{ for } i = 1, 2, \ldots, n \quad (6.4)$$

as the discrepancy between the line chosen and the observations. Our objective then is to choose $\alpha$ and $\beta$ to minimize the discrepancy.

### 6.1.4 Least Squares

A possible criterion for minimum discrepancy is

$$\min_{\alpha, \beta} \left| \sum_i e_i \right|,$$

but this will be zero for any line passing through $(x, y)$. Another possibility is

$$\min_{\alpha, \beta} \sum_i |e_i|,$$

which is called the minimum absolute distance (MAD) or $L_1$ estimator. The MAD estimator has problems since the mathematics (and statistical distributions) are relatively complicated. A closely related choice is

$$\min_{\alpha, \beta} \sum_i e_i^2. \quad (6.5)$$

This yields the least squares or $L_2$ estimator.
6.2 Least Squares Regression

6.2.1 The First-Order Conditions

Let
\[ \psi = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) \]  

(6.6)

Then the minimum values, say, \( \hat{\alpha} \) and \( \hat{\beta} \) must satisfy the following first-order conditions:

\[ 0 = \frac{\partial \psi}{\partial \alpha} = \sum_{i=1}^{n} -2(y_i - \hat{\alpha} - \hat{\beta} x_i) \]

\[ 0 = \frac{\partial \psi}{\partial \beta} = \sum_{i=1}^{n} -2(y_i - \hat{\alpha} - \hat{\beta} x_i) x_i \]

These first-order conditions may be rewritten as the normal equations

\[ \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (\hat{\alpha} + \hat{\beta} x_i) \]  

(6.7)

\[ \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} (\hat{\alpha} x_i + \hat{\beta} x_i^2) \]  

(6.8)

Clearly, (6.7) implies

\[ \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}, \]  

(6.9)

where \( \bar{y} = \sum_{i=1}^{n} y_i / n \) and \( \bar{x} = \sum_{i=1}^{n} x_i / n \).

Substituting (6.9) into (6.8) yields

\[ \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} [(\bar{y} - \hat{\beta} \bar{x}) x_i + \hat{\beta} x_i^2] \]

\[ \quad = \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i (x_i - \bar{x}) \]

and rearranging we obtain

\[ \sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x} = \hat{\beta} \left( \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \]

\[ \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})^2. \]

Provided \( \sum_{i=1}^{n} (x_i - \bar{x})^2 \neq 0 \), which is assured by Assumption (v) above, this may be solved for \( \hat{\beta} \) to obtain

\[ \hat{\beta} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \]  

(6.10)
6.2.2 The Second-Order Conditions

Note that the least squares estimators \( \hat{\beta} \), and in turn, \( \hat{\alpha} \), are unique. Taking the second derivatives of the objective function yields

\[
\frac{\partial^2 \psi}{\partial \alpha^2} = 2n,
\]

\[
\frac{\partial^2 \psi}{\partial \beta^2} = 2 \sum_{i=1}^{n} x_i^2,
\]

and

\[
\frac{\partial^2 \psi}{\partial \alpha \partial \beta} = 2 \sum_{i=1}^{n} x_i.
\]

Thus, the Hessian matrix is

\[
H = \begin{pmatrix}
\frac{2n}{2 \sum_{i=1}^{n} x_i} & \frac{2 \sum_{i=1}^{n} x_i}{2 \sum_{i=1}^{n} x_i^2} \\
\frac{2 \sum_{i=1}^{n} x_i}{2 \sum_{i=1}^{n} x_i^2} & \frac{2 \sum_{i=1}^{n} x_i}{2 \sum_{i=1}^{n} x_i^2}
\end{pmatrix},
\]

which is a positive definite matrix, provided \( \sum_{i=1}^{n} (x_i - \bar{x})^2 \neq 0 \), so \( \hat{\alpha} \) and \( \hat{\beta} \) yield a unique minimum.

6.2.3 Matrix Interpretation

Matrix algebra provides an alternative route to the same solutions. Note that the normal equations (6.7) and (6.8), are linear in \( \hat{\alpha} \) and \( \hat{\beta} \). In matrix form, we have

\[
\begin{pmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} y_i x_i
\end{pmatrix} = \begin{pmatrix}
\sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} x_i^2
\end{pmatrix} \begin{pmatrix}
\hat{\alpha} \\
\hat{\beta}
\end{pmatrix},
\]

which has the solution

\[
\begin{pmatrix}
\hat{\alpha} \\
\hat{\beta}
\end{pmatrix} = \begin{pmatrix}
\sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} x_i^2
\end{pmatrix}^{-1} \begin{pmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} y_i x_i
\end{pmatrix} \begin{pmatrix}
\sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} x_i^2
\end{pmatrix} \begin{pmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} y_i x_i
\end{pmatrix}^{-1}
\]

(6.11)

\[
= \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix}
\sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} x_i & n
\end{pmatrix}^{-1} \begin{pmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} y_i x_i
\end{pmatrix}
\]

\[
= \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix}
\sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} x_i & n
\end{pmatrix}^{-1} \begin{pmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} y_i x_i
\end{pmatrix}
\]

\[
= \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix}
\sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} x_i & n
\end{pmatrix}^{-1} \begin{pmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} y_i x_i + n \sum_{i=1}^{n} x_i y_i
\end{pmatrix}.
\]
Now, $\hat{\beta}$, according to this formula, is

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

$$
= \frac{\sum_{i=1}^{n} x_i y_i \overline{x} - \overline{x} \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i}
$$

$$
= \frac{\sum_{i=1}^{n} x_i y_i - n \overline{x} \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \overline{x}^2}
$$

$$
= \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2},
$$

while

$$
\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}
$$

$$
= \frac{\overline{y} \sum_{i=1}^{n} x_i^2 \overline{x} - \overline{x} \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i}
$$

$$
= \frac{\overline{y} \sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i y_i + \overline{x} \sum_{i=1}^{n} y_i - \overline{x} \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i}
$$

$$
= \overline{y} - \frac{\overline{x}}{\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i}
$$

$$
= \overline{y} - \hat{\beta} \overline{x}.
$$

Note that in the compact notation given following (6.2), we can write the solution to (6.11) using the general solution from the next chapter, namely

$$
\hat{\beta} = (X'X)^{-1}X'Y.
$$

### 6.3 Basic Statistical Properties

#### 6.3.1 Method Of Moments Interpretation

Suppose, for the moment, that $x_i$ is a random variable that is uncorrelated with $u_i$. Let $\mu_x = E(x_i)$. Then

$$
\mu_y = E(y_i) = E(\alpha + \beta x_i + u_i) = \alpha + \beta \mu_x.
$$

(6.12)

Thus, subtracting (6.12) from (6.1) yields

$$
(y_i - \mu_y) = \alpha + \beta x_i + u_i - (\alpha + \beta \mu_x) = \beta(x_i - \mu_x) + u_i.
$$
and

$$E(y_i - \mu_y)(x_i - \mu_x) = \beta E(x_i - \mu_x) + E(x_i - \mu_x)u_i = \beta E(x_i - \mu_x)^2,$$

since $x_i$ and $u_i$ are uncorrelated. Solving for $\beta$, we have

$$\beta = \frac{E[(x_i - \mu_x)(y_i - \mu_y)]}{E[(x_i - \mu_x)^2]},$$

while

$$\alpha = \mu_y - \beta \mu_x.$$

These values are unique provided the variance of the independent variable is nonzero.

These expressions form the basis for estimators by replacing the various population moments with sample moments. Specifically, we have, by analogy

$$\hat{\mu}_x = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

$$\hat{\mu}_y = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y),$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \hat{\mu}_x)^2 + \beta \sum_{i=1}^{n} (x_i - \bar{x})u_i,$$

$$\hat{\alpha} = \hat{\mu}_y - \hat{\beta} \hat{\mu}_x,$$

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

Thus the least-squares estimators are method of moments estimators with sample moments replacing the population moments.

### 6.3.2 Mean of Estimators

Next, note that

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

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\alpha + \beta x_i + u_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\alpha \sum_{i=1}^{n} (x_i - \bar{x}) + \beta \sum_{i=1}^{n} (x_i - \bar{x}) x_i + \sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \beta + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$  \hspace{2cm} (6.13)
6.3. BASIC STATISTICAL PROPERTIES

So, \( E[\hat{\beta}] = \beta \), since \( E[u_i] = 0 \) for all \( i \). Thus, \( \hat{\beta} \) is an unbiased estimator of \( \beta \).

Further

\[
\hat{\alpha} = \bar{y} - \beta \bar{x} = \sum_{i=1}^{n} \frac{y_i}{n} - \bar{x} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) y_i
\]

\[
= \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} \sum_{i=1}^{n}(x_i - \bar{x})^2 \right) y_i
\]

\[
= \alpha \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} \sum_{i=1}^{n}(x_i - \bar{x})^2 \right)
\]

\[
+ \beta \sum_{i=1}^{n} \left( \frac{x_i}{n} - \bar{x} \sum_{i=1}^{n}(x_i - \bar{x})^2 \right) + \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} \sum_{i=1}^{n}(x_i - \bar{x})^2 \right) u_i
\]

\[
= \alpha + \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} \sum_{i=1}^{n}(x_i - \bar{x})^2 \right) u_i
\]  \hspace{1cm} (6.14)

So we have \( E[\hat{\alpha}] = \alpha \), and \( \hat{\alpha} \) is also an unbiased estimator of \( \alpha \).

6.3.3 Variance of Estimators

Now, from (6.13) we have

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

where

\[
w_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\]

So,

\[
\text{Var}(\hat{\beta}) = E[(\hat{\beta} - \beta)^2] \quad \text{(6.15)}
\]

\[
= E[\sum_{i=1}^{n} w_i u_i] = E[(w_1 u_1 + w_2 u_2 + \cdots + w_n u_n)^2]
\]

\[
= E\left[w_1^2 u_1^2 + w_1 u_1 w_2 u_2 + \cdots + w_1 u_1 w_n u_n + w_2 u_2 w_1 u_1 + w_2^2 u_2^2 + \cdots + w_2 u_2 w_n u_n + \cdots + w_n u_n w_1 u_1 + w_n u_n w_2 u_2 + \cdots + w_n^2 u_n^2\right]
\]

\[
= w_1^2 \sigma^2 + w_2^2 \sigma^2 + \cdots + w_n^2 \sigma^2
\]

\[
= \sigma^2 \sum_{i=1}^{n} w_i^2 = \sigma^2 \sum_{i=1}^{n} \left( \frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right)^2
\]

\[
= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \quad \text{(6.16)}
\]
Next, we note from (6.14) that
\[
\hat{\alpha} - \alpha = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{y}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) u_i = \sum_{i=1}^{n} v_i u_i,
\]
where
\[
v_i = \left( \frac{1}{n} - \frac{x_i - \bar{y}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right).
\]
So, in a fashion similar to above, we find that
\[
\text{Var}(\hat{\alpha} - \alpha) = \sigma^2 \frac{1}{n} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{6.17}
\]
and
\[
\text{Cov}(\hat{\alpha}, \hat{\beta}) = E[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)] = \sigma^2 \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{6.18}
\]

### 6.3.4 Estimation of \( \sigma^2 \)

Next, we would like to get an estimate of \( \sigma^2 \). Let
\[
s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2, \tag{6.19}
\]
where
\[
e_i = y_i - \hat{\alpha} - \hat{\beta} x_i
= (y_i - \bar{y}) - (\hat{\alpha} - \bar{y}) - \hat{\beta} (x_i - \bar{x})
= (y_i - \bar{y}) - \hat{\beta} (x_i - \bar{x})
= \beta (x_i - \bar{x}) + (u_i - \bar{u}) - \hat{\beta} (x_i - \bar{x})
= - (\hat{\beta} - \beta) (x_i - \bar{x}) + (u_i - \bar{u})
\]
So,
\[
\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [-(\hat{\beta} - \beta) (x_i - \bar{x}) + (u_i - \bar{u})]^2
= (\hat{\beta} - \beta)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2
- 2(\hat{\beta} - \beta) \sum_{i=1}^{n} (x_i - \bar{x})(u_i - \bar{u}) + \sum_{i=1}^{n} (u_i - \bar{u})^2.
\]
Now, we have
\[
E[(\hat{\beta} - \beta)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2] = \sigma^2,
\]
6.4. STATISTICAL PROPERTIES UNDER NORMALITY

\[
\begin{align*}
\mathbb{E}\left[\sum_{i=1}^{n} (u_i - \bar{u})^2 \right] &= \mathbb{E}\left[\sum_{i=1}^{n} (u_i^2 - \bar{u}^2) \right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} u_i^2 \right] - \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} u_i \right] = (n-1)\sigma^2,
\end{align*}
\]

and

\[
\begin{align*}
\mathbb{E}\left[(\hat{\beta} - \beta) \sum_{i=1}^{n} (x_i - \bar{x})(u_i - \bar{u}) \right] &= \mathbb{E}\left[\sum_{i=1}^{n} w_i u_i \right] \left( \sum_{i=1}^{n} (x_i - \bar{x})u_i - \sum_{i=1}^{n} (x_i - \bar{x})\bar{u} \right) \\
&= \mathbb{E}\left[\sum_{i=1}^{n} w_i u_i \right] \sum_{i=1}^{n} (x_i - \bar{x})u_i = \sum_{i=1}^{n} w_i (x_i - \bar{x})\sigma^2 = \sigma^2.
\end{align*}
\]

Therefore, collecting terms, we have

\[
\begin{align*}
\mathbb{E}\left[\sum e_i^2 \right] &= \sigma^2 - 2\sigma^2 + (n-1)\sigma^2 = (n-2)\sigma^2 \\
\mathbb{E}\left[s^2 \right] &= \sigma^2.
\end{align*}
\]

(6.20)

6.4 Statistical Properties Under Normality

6.4.1 Distribution of \( \hat{\beta} \)

Suppose, under Assumption (vi), that \( u_i \sim i.i.d. N(0, \sigma^2) \). Recall from (6.13) that

\[
\hat{\beta} = \beta + \frac{1}{n} \sum_{i=1}^{n} w_i u_i,
\]

where \( w_i = (x_i - \bar{x})/\sum_{i=1}^{n} (x_i - \bar{x})^2 \), so \( \hat{\beta} \) is also a normal random variable. Specifically,

\[
\hat{\beta} \sim N(\beta, \sigma^2/q),
\]

(6.21)

where \( q = \sum (x_i - \bar{x})^2 \). It follows, for a given \( \beta_0 \),

\[
\hat{\beta} - \beta_0 \sim N(\beta - \beta_0, \sigma^2/q),
\]

and

\[
z_\beta = \frac{\hat{\beta} - \beta_0}{\sqrt{\sigma^2/q}} \sim N\left( \frac{\beta - \beta_0}{\sqrt{\sigma^2/q}}, 1 \right).
\]

This ratio forms the basis for inferences regarding \( \beta \). Under the null hypothesis \( H_0 : \beta = \beta_0 \), then

\[
\frac{\hat{\beta} - \beta_0}{\sqrt{\sigma^2/q}} \sim N(0, 1),
\]
and the probability of values exceeding, say 1.96, is well known to be .025. Under the alternative hypothesis \(H_1: \beta = \beta_1 > \beta_0\) we have

\[
\frac{\hat{\beta} - \beta_0}{\sqrt{\sigma^2/q}} \sim N \left( \frac{\beta_1 - \beta_0}{\sqrt{\sigma^2/q}}, 1 \right),
\]

and we would expect the statistic to be centered to the right of zero and the probability of exceeding 1.96 will be increased. Thus for values exceeding 1.96 we are faced with the choice of the null hypothesis where it is an unlikely value or the alternative hypothesis where it is more likely.

### 6.4.2 Distribution of \(\hat{\alpha}\)

Again, suppose that \(u_i \sim i.i.d. N(0, \sigma^2)\). From (6.14) above

\[
\hat{\alpha} = \alpha + \sum_{i=1}^{n} v_i u_i,
\]

where \(v_i = 1/n - \bar{x}(x_i - \bar{x})/(\sum_{i=1}^{n} (x_i - \bar{x})^2)\), is a normal random variable. Specifically,

\[
\hat{\alpha} \sim N(\alpha, \sigma^2 p/q),
\]

where \(p = \sum_{i=1}^{n} x_i^2/n\). For a given \(\alpha_0\),

\[
\hat{\alpha} - \alpha_0 \sim N(\alpha - \alpha_0, \sigma^2 p/q),
\]

and

\[
z_\alpha = \frac{\hat{\alpha} - \alpha_0}{\sigma^2 p/q} \sim N \left( \frac{\hat{\alpha} - \alpha_0}{\sigma^2 p/q}, 1 \right).
\]

Now, suppose that \(H_0: \alpha = \alpha_0\). Then,

\[
\frac{\hat{\alpha} - \alpha_0}{\sigma^2 p/q} \sim N(0, 1),
\]

while for \(H_0: \alpha = \alpha_1 \neq \alpha_0\) we have

\[
\frac{\hat{\alpha} - \alpha_0}{\sigma^2 p/q} \sim N \left( \frac{\alpha_1 - \alpha_0}{\sigma^2 p/q}, 1 \right),
\]

and we would expect the statistic not to be centered around zero.

It is important to note that

\[
p/q = \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2},
\]

and so \(\sqrt{\sigma^2 p/q}\) grows small as \(n\) grows large. Thus, the noncentrality of the distribution of the statistic \(z_\alpha\) will also grow under the alternative. A similar statement can be made concerning \(q\) and the statistic \(z_\beta\).
6.4. STATISTICAL PROPERTIES UNDER NORMALITY

6.4.3 t-Distribution

In most cases, we do not know the value of $\sigma^2$. A possible alternative is to use $s^2$, whereupon we find
\[
\frac{\hat{\alpha} - \alpha_0}{\sqrt{s^2 p/q}} \sim t_{n-2},
\]
under $H_0: \alpha = \alpha_0$, and
\[
\frac{\hat{\beta} - \beta_0}{\sqrt{s^2 q}} \sim t_{n-2},
\]
under $H_0: \beta = \beta_0$.

The t-distribution is quite similar to the standard normal except being slightly fatter. This reflects the added uncertainty introduced by using $s^2$ rather than $\sigma^2$. However, as $n$ increases and the precision of $s^2$ becomes better, the t-distribution grows closer and closer to the $N(0, 1)$. We lose two degrees of freedom (and so $t_{n-2}$) because of the fact that we estimated two coefficients, namely $\alpha$ and $\beta$.

Just as in the case of $z_\alpha$ and $z_\beta$, we would expect the t-distribution to be off-center if the null hypothesis were not true. Specifically, it would have a non-central t-distribution. This will be covered in more detail in the more general regression model.

6.4.4 Maximum Likelihood

Suppose that $u_i \sim i.i.d. N(0, \sigma^2)$. Then,
\[
y_i \sim i.i.d. N(\alpha + \beta x_i, \sigma^2).
\]
Then, the pdf of $y_i$ is given by
\[
f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - (\alpha + \beta x_i))^2 \right\}.
\]
Since the observations are independent, we can write the joint likelihood function as
\[
f(y_1, y_2, \ldots, y_n) = f(y_1) f(y_2) \cdots f(y_n)
= \frac{1}{(2\pi\sigma^2)^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 \right\}
= L(\alpha, \beta, \sigma^2|y, x).
\]

Now, let $L(\alpha, \beta, \sigma^2|y, x) = \ln L(\alpha, \beta, \sigma^2|y, x)$. We seek to maximize
\[
L(\alpha, \beta, \sigma^2|y, x) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.
\]
Note that for (6.27) to be a maximum with respect to $\alpha$ and $\beta$, we must minimize $\sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$. 
CHAPTER 6. BIVARIATE LEAST SQUARES

The first-order conditions for maximizing (6.27) are

\[ \frac{\partial L}{\partial \alpha} = \frac{1}{\sigma^2} \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta}x_i)] = 0, \]  
(6.28)

\[ \frac{\partial L}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta}x_i) x_i] = 0, \]  
(6.29)

and

\[ \frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2. \]  
(6.30)

Note that the first two conditions imply that

\[ \bar{\alpha} = \bar{y} - \hat{\beta} \bar{x}, \]

and

\[ \hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})}, \]

since these are the same as the normal equations (except for \( \sigma^2 \)). The third condition yields

\[ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n} = \frac{n - 2}{n} s^2. \]  
(6.31)

6.5 An Example

Consider the scatter graph given in Figure 6.1. From this, we construct Table 6.1.

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( x_i - \bar{x} )</th>
<th>( y_i - \bar{y} )</th>
<th>( (x_i - \bar{x})^2 )</th>
<th>( (x_i - \bar{x})(y_i - \bar{y}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>12</td>
<td>-2</td>
<td>5</td>
<td>4</td>
<td>-24</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>-7</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2</td>
<td>-4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>-20</td>
</tr>
</tbody>
</table>

Table 6.1: Summary table.
Thus, we have
\[ \hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{-20}{10} = -2, \]
and
\[ \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 7 - (-2)4 = 15. \]

Now, we calculate the residuals in Table 6.2,

<table>
<thead>
<tr>
<th>$\hat{\beta}x_i$</th>
<th>$\hat{\alpha} + \hat{\beta}x_i$</th>
<th>$e_i$</th>
<th>$e_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-6</td>
<td>9</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>-8</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-10</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-12</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.2: Residual calculation.

and find
\[ s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - 2} = \frac{6}{3} = 2. \]

An estimate of the variance of $\hat{\beta}$, namely $\sigma^2q$, is provided by
\[ s^2/q = s^2 \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{2}{10} = 0.2. \]

Suppose we wish to test $H_0: \beta = 0$ against $H_1: \beta \neq 0$. Then
\[ \frac{\hat{\beta} - 0}{\sqrt{s^2/q}} \sim t_3 \]
under the null hypothesis. The critical values for a symmetric 5% test are $\pm 3.182$ which indicate the 2.5% tails. Since
\[ \frac{\hat{\beta} - 0}{\sqrt{s^2/q}} = \frac{-2}{\sqrt{0.2}} = \frac{-2}{0.45} = -4.2 \]
is clearly in the left-hand 2.5% tail of the $t_3$-distribution, we reject the null hypothesis at the 95% significance level.
Chapter 7

Multivariate Least Squares

The bivariate regression model studied in the previous chapter enables us to study the relationship between two variables. One is taken as the dependent variable and the other the independent or explanatory variable. We simplified by assuming that the relationship was linear in the parameters given the variables. In much of economic analysis, however, the models considered involve more than two variables. Accordingly, in this chapter we begin to consider models that have more than one independent or explanatory variables. Again, we simplify by assuming that the relationship is linear in the parameters given the variables and the error term is additive.

7.1 Introduction

7.1.1 Multiple Regression Model

The general $k$-variable linear model can be written as

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \quad i = 1, 2, \ldots, n$$  \hfill (7.1)

where $x_{ij}$ denotes the $i$-th observation on the $j$-th explanatory variable. For the bivariate model studied in the previous chapter, $k = 2$ and $x_{i1} = 1$ yields $\beta_1$ as the intercept and $x_{i2} = x_i$ yields $\beta_2$ as the slope coefficient. In general, for any model with an intercept term the corresponding variable (non-variable) will be unity. The variables $x_{ij}$ are known as regressors or covariates.

Using matrix notation, we can equivalently write all observations on this model as

$$
\begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{pmatrix} = 
\begin{pmatrix}
  x_{11} & x_{12} & \cdots & x_{1k} \\
  x_{21} & x_{22} & \cdots & x_{2k} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nk}
\end{pmatrix} 
\begin{pmatrix}
  \beta_1 \\
  \beta_2 \\
  \vdots \\
  \beta_k
\end{pmatrix} + 
\begin{pmatrix}
  u_1 \\
  u_2 \\
  \vdots \\
  u_n
\end{pmatrix},
$$

86
or, more compactly, as
\[ y = X\beta + u. \quad (7.2) \]

where
\[
\begin{align*}
\mathbf{y} &\sim \mathbb{N}(\mu, \Sigma) \\
\mathbf{u} &\sim \mathbb{N}(0, \sigma^2 I)
\end{align*}
\]

\[
\begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{pmatrix}
= 
\begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1k} \\
x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk}
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_k
\end{pmatrix}
\]
and
\[
\begin{pmatrix}
u_1 \\
u_2 \\
\vdots \\
u_n
\end{pmatrix}
= 
\begin{pmatrix}
u_1 \\
u_2 \\
\vdots \\
u_n
\end{pmatrix}
\]

### 7.1.2 Assumptions

For the disturbances, we assume
(i) \( \mathbb{E}(u_i) = 0 \) for all \( i \)
(ii) \( \mathbb{E}(u_i^2) = 0 \) for all \( i \)
(iii) \( \mathbb{E}(u_i u_\ell) = 0 \), for all \( i \neq \ell \).

For the independent or explanatory variables, we assume
(iv) \( x_{ij} \) non-stochastic for all \( i, j \)
(v) \( (x_{i1}, x_{i2}, \ldots, x_{ik}) \) linearly independent for all \( i \).

For the purposes of inference in finite samples, we sometimes also assume
(vi) \( u_i \sim i.i.d. \mathbb{N}(0, \sigma^2) \) for all \( i \).

The assumptions on the disturbances are the same as for the bivariate model. Likewise, the assumption that the explanatory variables are all non-stochastic is the same and covers one of the variables being unity for the intercept term. The linear independence assumption is introduced to make sure that no term is redundant since if one of the variables were a constant linear combination of other variables for all \( i \), then we could eliminate that variable and use a model with one less explanatory variable.

### 7.1.3 Matrix Form Assumptions

The assumptions can be rewritten in terms of the matrix form of the model. Specifically we have, for the disturbances,
(i) \( \mathbb{E}(u) = 0 \)
and

\[(ii),(iii) \quad \text{Cov}(u) = E(uu') = \sigma^2 I_n,\]

where $I_n$ is an $n \times n$ identity matrix. And for the explanatory variables, we have

\[(iv) \quad X \text{ is nonstochastic.}\]

\[(v) \quad X \text{ has full column rank (the columns are linearly independent).}\]

And for inferences in finite samples, we have

\[(vi) \quad u \sim N(0, \sigma^2 I_n).\]

### 7.1.4 Plane Fitting

It is useful to consider the approach taken in choosing an estimator for $\beta$ for the tri-variate model, where the relationships may still be visualized geometrically. Accordingly, we have $k = 3$ and $x_{1i} = 1$, so

\[y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i \quad i = 1, 2, \ldots, n.\]

Now the triples $(y_i, x_{i2}, x_{i3})$ define a cloud of $n$ points in the three-dimensional space of $y, x_2, \text{and } x_3$ and for some choice of, say, $\beta_1, \beta_2$, and $\beta_3$

\[\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} \quad i = 1, 2, \ldots, n.\]

defines a plane in that space. We seek to choose $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ so that the points on the plane corresponding to $x_{i2}$ and $x_{i3}$, namely $\hat{y}_i$, will be close to $y_i$. That is, we will “fit” a plane to the cloud of observations.

As in the two-dimensional case, we choose to measure closeness in the vertical distance. Specifically, define

\[e_i = y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}) \quad i = 1, 2, \ldots, n,\]

and we choose $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ to minimize the sum of these squared residuals. The results of this minimization define the least-squares estimator.

More generally, for the the $k$-variate case, we choose $\hat{\beta}_1$ through $\hat{\beta}_k$ so that the points lying on the hyper-plane in the $k$-dimensional space of the $y$’s and $x$’s define by

\[\hat{y}_i = \hat{\beta}_{i1} + \hat{\beta}_{i2} x_{i2} + \ldots + \hat{\beta}_{ik} x_{ik} \quad i = 1, 2, \ldots, n.\]

are close to $y_i$. Specifically, we define

\[e_i = y_i - (\hat{\beta}_{i1} + \hat{\beta}_{i2} x_{i2} + \ldots + \hat{\beta}_{ik} x_{ik}) \quad i = 1, 2, \ldots, n,\]
and our measure of distance is given by

$$
\psi(\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[ y_i - \left( \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \right) \right]^2.
$$

(7.3)

The least-squares estimator for this general case is then yielded, as before, by

$$
\hat{\beta} = \arg\min_{\beta} \psi(\beta).
$$

(7.4)

7.2 Least Squares Regression

7.2.1 The OLS Estimator

A necessary condition for the minimization of this function is that the first partial derivatives be zero at the solution. These first-order conditions are

$$
0 = \frac{\partial \psi}{\partial \beta_1} = 2 \sum_{i=1}^{n} \left[ y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{i1} + \cdots + \hat{\beta}_k x_{ik}) \right] x_{i1},
$$

$$
0 = \frac{\partial \psi}{\partial \beta_2} = 2 \sum_{i=1}^{n} \left[ y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}) \right] x_{i2},
$$

$$
\vdots
$$

$$
0 = \frac{\partial \psi}{\partial \beta_k} = 2 \sum_{i=1}^{n} \left[ y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}) \right] x_{ik}.
$$

where \((\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)\) are solutions. Rearranging, we have the normal equations:

$$
\hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{1i} x_{ik} = \sum_{i=1}^{n} x_{1i} y_i
$$

$$
\hat{\beta}_1 \sum_{i=1}^{n} x_{2i} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{2i} x_{ik} = \sum_{i=1}^{n} x_{2i} y_i
$$

$$
\vdots
$$

$$
\hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} x_{ik} = \sum_{i=1}^{n} x_{ik} y_i
$$
which can be rewritten using matrix notation as

\[
\begin{pmatrix}
\sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1} x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1} x_{ik} \\
\sum_{i=1}^{n} x_{i2} x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2} x_{ik} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} x_{ik} x_{i1} & \sum_{i=1}^{n} x_{ik} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2
\end{pmatrix}
\begin{pmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2 \\
\vdots \\
\hat{\beta}_k
\end{pmatrix}
= \begin{pmatrix}
\sum_{i=1}^{n} x_{i1} y_i \\
\sum_{i=1}^{n} x_{i2} y_i \\
\vdots \\
\sum_{i=1}^{n} x_{ik} y_i
\end{pmatrix}.
\]  

(7.5)

Analogous to the bivariate model, we find

\[
X'X = \begin{pmatrix}
\sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1} x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1} x_{ik} \\
\sum_{i=1}^{n} x_{i2} x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2} x_{ik} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} x_{ik} x_{i1} & \sum_{i=1}^{n} x_{ik} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2
\end{pmatrix},
\]  

(7.6)

and

\[
X'y = \begin{pmatrix}
\sum_{i=1}^{n} x_{i1} y_i \\
\sum_{i=1}^{n} x_{i2} y_i \\
\vdots \\
\sum_{i=1}^{n} x_{ik} y_i
\end{pmatrix}.
\]  

(7.7)

Substituting, we can write the normal equations (7.5) as

\[
X'X\hat{\beta} = X'y.
\]  

(7.8)

where \(\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)\). Therefore, we have the unique solution

\[
\hat{\beta} = (X'X)^{-1}X'y,
\]  

(7.9)

as long as \(|X'X| \neq 0\), which is assured by Assumption (v). This estimator is known as the ordinary least squares (OLS) estimator. It is ordinary in the sense that it works well under the standard assumptions. When we attempt to deal with relaxation of these assumptions more complicated estimators are suggested.

### 7.2.2 Some Algebraic Results

Define the fitted value for each \(i\) as \(\hat{y}_i = x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + \cdots + x_{ik}\hat{\beta}_k = x'_i\hat{\beta}\) whereupon

\[
\hat{y} = X\hat{\beta}.
\]  

(7.10)

Next define the OLS residual for each \(i\) as \(e_i = y_i - \hat{y}_i\) so

\[
e = y - \hat{y}.
\]  

(7.11)

Then,

\[
X'e = X'(y - \hat{y}) = X'y - X'X(X'X)^{-1}X'y = 0,
\]  

(7.12)
and we say that the residuals are orthogonal to the regressors. Also, we find

\[ \hat{y}'\hat{y} = (X\hat{\beta}'(X\hat{\beta} + e) \]
\[ = \hat{\beta}'X'X\hat{\beta} + \hat{\beta}'X'e \]
\[ = \hat{\beta}'X'X\hat{\beta} \quad \text{by (7.12)} \]
\[ = \hat{y}'\hat{y}. \quad (7.13) \]

Now, suppose that the first coefficient is the intercept. Then, the first column of \( X \) and hence the first row of \( X' \) are all ones. This means that

\[ 0 = X'e = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{nn} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} e_i \\ \sum_{i=1}^{n} x_{i2} e_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik} e_i \end{bmatrix}. \]

So, \( \sum_{i=1}^{n} e_i = 0 \), which means that

\[ \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i + e_i = \sum_{i=1}^{n} \hat{y}_i. \quad (7.14) \]

Finally, we note that

\[ e = y - X\hat{\beta} \]
\[ = y - X(X'X)^{-1}X'y \]
\[ = (I_n - X(X'X)^{-1}X')y \]
\[ = My \]
\[ = M(X\beta + u) \]
\[ = [I_n - X(X'X)^{-1}X'](X\beta) + Mu \]
\[ = Mu. \quad (7.15) \]

We see that the OLS residuals are a linear transformation of the underlying disturbances. The matrix \( M = I_n - X(X'X)^{-1}X' \) which is called the influence function, plays an important role in the sequel. Note that it is both symmetric and idempotent, the latter meaning \( M = M \cdot M \).

### 7.2.3 The \( R^2 \) Statistic

Define the following:

\[ \text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad (7.16) \]
\[ \text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad (7.17) \]
\[ \text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \quad (7.18) \]

Note that SSE is the variation of actuals around the fitted plane and is called the unexplained or error sum-of-squares. SSR, or residual sum-of-squares, is the variation of the fitted values around the sample mean and SST, or total sum-of-squares, is the variation of the actual around the sample mean.

The three sums-of-squares are closely related. Consider

\[
\text{SST} - \text{SSE} = \sum_{i=1}^{n} [(y_i - \bar{y})^2 - (y_i - \hat{y}_i)^2] \\
= \sum_{i=1}^{n} (y_i^2 - 2y_i\bar{y} + \bar{y}^2) - \sum_{i=1}^{n} (y_i^2 - 2y_i\hat{y} + \hat{y}^2) \\
= \sum_{i=1}^{n} \bar{y}^2 - 2\bar{y} \sum_{i=1}^{n} y_i + 2 \sum_{i=1}^{n} y_i\hat{y} - \sum_{i=1}^{n} \hat{y}_i^2 \\
= \sum_{i=1}^{n} \bar{y}^2 - 2\bar{y} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \hat{y}_i^2 \quad \text{by (7.13)} \\
= \sum_{i=1}^{n} \hat{y}_i^2 - 2\bar{y} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \hat{y}_i^2 \quad \text{by (7.14)} \\
= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \text{SSR}.
\]

Thus we see the residual sum-of-squares SSR is total sum of squares SST less the unexplained SSE, hence the name. Alternatively, we have \( \text{SST} = \text{SSE} + \text{SSR} \).

We now define

\[
R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SST}}{\text{SST}} - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}},
\]

as one less the percentage of the total variation that is not explained by the model or the percent of of total variation explained by the model.

This statistic can also be interpreted as a squared correlation coefficient. Consider the sample second moments,

\[
\hat{\text{Var}}(y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n} \text{SST},
\]
\[
\hat{\text{Var}}(\hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \frac{1}{n} \text{SSR},
\]
and

\[
\widehat{\text{Cov}}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y})
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} y_i \hat{y}_i - \bar{y} \hat{y}_i - \bar{y} \bar{y}_i + \bar{y}^2
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i^2 - 2 \bar{y} \hat{y}_i + \bar{y}^2
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2
\]

\[
= \frac{1}{n} \text{SSR}.
\]

Then the sample correlation between \( y_i \) and \( \hat{y}_i \) can be written as,

\[
r_{y, \hat{y}} = \frac{\widehat{\text{Cov}}(y, \hat{y})}{\sqrt{\text{Var}(y) \text{Var}(\hat{y})}}
\]

\[
= \frac{\frac{1}{n} \text{SSR}}{\sqrt{\frac{1}{n} \text{SST} \frac{1}{n} \text{SSR}}}
\]

\[
= \frac{\sqrt{\text{SSR}}}{\sqrt{\text{SST}}}
\]

Squaring, we have the \( R^2 \) as a squared correlation between \( y \) and \( \hat{y} \),

\[
r_{y, \hat{y}}^2 = \frac{\text{SSR}}{\text{SST}} = R^2.
\]

This statistic is variously called the coefficient of determination, the multiple correlation statistic and the “\( R \)”-squared statistic. It is a measure of goodness-of-fit and always lies between 0, in the event that there is no linear relationship between \( y_i \) and the \( x \)'s, and 1, in the event that there is a perfect relationship. It is worth noting that it is typically large in the case of time-series data due to shared trends and much smaller with cross-sectional data. While an invaluable tool, it is subject to a great deal of misuse in making selections between model alternatives. The topic of model selection will receive extended treatment in Chapter 15.
7.3 Basic Statistical Results

7.3.1 Mean and Covariance of $\hat{\beta}$

From the previous section, under assumption (v), we have

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$= (X'X)^{-1}X'(X\beta + u)$$
$$= (X'X)^{-1}X'\beta + (X'X)^{-1}X'u$$
$$= \beta + (X'X)^{-1}X'u.$$  \hspace{1cm} (7.20)

Since $X$ is nonstochastic by assumption (iv), we have, also using assumption (i),

$$E(\hat{\beta}) = \beta + E[(X'X)^{-1}X'u]$$
$$= \beta + (X'X)^{-1}X'E(u)$$
$$= \beta.$$  \hspace{1cm} (7.21)

Thus, the OLS estimator is unbiased. Also, using assumptions (ii) and (iii),

$$\text{Cov}(\hat{\beta}) = E(\hat{\beta} - \beta)[\hat{\beta} - \beta']$$
$$= E[(X'X)^{-1}X'u(X'X)^{-1}]$$
$$= (X'X)^{-1}X'E(\mu')X(X'X)^{-1}$$
$$= (X'X)^{-1}(X'(\sigma^2I_n)X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}. \hspace{1cm} (7.22)$$

7.3.2 Best Linear Unbiased Estimator (BLUE)

The OLS estimator is linear in $y$ and it is an unbiased estimator, as we saw above. Let

$$\tilde{\beta} = \tilde{A}y,$$

where $\tilde{A}$ is a $k \times n$ matrix that is nonstochastic, be any other unbiased estimator. Define

$$A = \tilde{A} - (X'X)^{-1}X',$$

then

$$\tilde{\beta} = [A + (X'X)^{-1}X']y$$
$$= [A + (X'X)^{-1}X'][X\beta + u]$$
$$= AX\beta + \beta + [A + (X'X)^{-1}X']u.$$  

Now, $\tilde{\beta}$ is an unbiased estimator, so

$$E(\tilde{\beta}) = AX\beta + \beta + [A + (X'X)^{-1}X']E(u)$$
$$= AX\beta + \beta = \beta,$$
for any true value of $\beta$, which implies that $AX = 0$. Thus,

$$\tilde{\beta} = \beta + [A + (X'X)^{-1}X']u,$$

and

$$\text{Cov}(\tilde{\beta}) = E[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta')]
= E[(A + (X'X)^{-1}X')uu'[A + (X'X)^{-1}X']]
= [A + (X'X)^{-1}X']E(uu')[A + (X'X)^{-1}X']'
= \sigma^2[AA' + (X'X)^{-1}]\quad\text{since } AX = 0
= \sigma^2 AA' + \sigma^2(X'X)^{-1}.$$

This shows that the covariance matrix of any other linear unbiased estimator exceeds the covariance matrix of the OLS estimator by a positive semi-definite matrix $\sigma^2AA'$. Hence, OLS is said to be best linear unbiased estimator (BLUE). This result is known as the Gauss-Markov theorem. Note that we have used all of the Assumptions (i)-(v) to get to this point.

### 7.3.3 Consistency

Typically, the elements of $X'X$ are unbounded (they go to infinity) as $n$ gets very large. For example, the 1,1 element is $n$ and the $j,j$ element is $\sum_{i=1}^n x_{ij}^2$. Therefore, we typically have

$$\lim_{n \to \infty} (X'X)^{-1} = 0,$$

and the variances of $\tilde{\beta}$ converge to zero. This means that the distribution collapses about its expected value, namely $\beta$. So, by convergence in quadratic mean,

$$\text{plim}_{n \to \infty} \tilde{\beta} = \beta,$$

and OLS estimation is consistent. A more formal proof of this property will be given in the chapter on stochastic regressors.

### 7.3.4 Estimation of $\sigma^2$

Recall that $e = Mu$, then,

$$e'e = (Mu)'Mu = u'M'Mu = u'Mu,$$

since $M$ is symmetric and idempotent. Now,

$$e'e = \text{tr}(e'e) = \text{tr}(u'Mu) = \text{tr}(Mu'u'),$$
since $e'e$ is a scalar, and $\text{tr}AB = \text{tr}BA$, for any matrices $A$ and $B$, provided both multiplications are conformable. Thus,

$$E[e'e] = E[\text{tr}(Mu'u')] = \text{tr}(ME[u'u']) \quad \text{since } E \text{ and } \text{tr} \text{ are linear operators}$$

$$= \text{tr}(Ms^2I_n)$$

$$= s^2\text{tr}M.$$ 

But, $M = I_n - X(X'X)^{-1}X'$, so

$$\text{tr}M = \text{tr}(I_n - X(X'X)^{-1}X')$$

$$= n - \text{tr}((X'X)^{-1}X'X)$$

$$= n - k.$$ 

Now, define

$$s^2 = \frac{e'e}{n-k},$$

then it follows that

$$E(s^2) = \frac{E(e'e)}{n-k} = \frac{\sigma^2(n-k)}{n-k} = \sigma^2,$$

so $s^2$ is an unbiased estimator of $\sigma^2$. 

We can also establish, under consistency of $\hat{\beta}$, that $s^2$ is a consistent estimator of $\sigma^2$. That is,

$$\underset{n \to \infty}{\text{plim}} s^2 = \sigma^2.$$ (7.27)

This finding will be given more extensive attention in the chapter on stochastic regressors.

### 7.4 Statistical Properties Under Normality

#### 7.4.1 Distribution Of $\hat{\beta}$

Suppose that we introduce assumption (vi), so the $u_i$’s are jointly normal:

$$u \sim N(0, \sigma^2I_n).$$ (7.28)

Recall that

$$\hat{\beta} = \beta + (X'X)^{-1}X'u,$$

so $\hat{\beta}$ is linear in $u$ since $(X'X)^{-1}X'$, which is nonstochastic, may be treated as a constant matrix. It follows that $\hat{\beta}$ is also jointly normally distributed. We already know the mean and covariance matrices so we have

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}).$$ (7.29)
as the joint distribution of \( \hat{\beta} \).

The normality of the estimator will prove useful in performing inferences in finite samples as discussed in the next chapter. The individual elements of \( \hat{\beta} \) will be marginally normal, so

\[
\hat{\beta}_j \sim N(\beta_j, \sigma^2 d_{jj})
\]

for \( j = 1, 2, ..., k \), where \( d_{jj} = [(X'X)^{-1}]_{jj} \). We may now apply the standard normal transformation to obtain

\[
z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 d_{jj}}} \sim N(0, 1).
\]

(7.30)

which will form the basis for testing hypotheses on a single coefficient. Under \( H_0 : \beta_j = \beta^0_j \) we have

\[
\frac{\hat{\beta}_j - \beta^0_j}{\sqrt{\sigma^2 d_{jj}}} \sim N(0, 1),
\]

while for \( H_1 : \beta_j = \beta^1_j \neq \beta^0_j \) we have

\[
\frac{\hat{\beta}_j - \beta^0_j}{\sqrt{\sigma^2 d_{jj}}} \sim N\left(\frac{\beta^1_j - \beta^0_j}{\sqrt{\sigma^2 d_{jj}}}, 1\right),
\]

and we would expect the statistic not to be centered around zero.

### 7.4.2 t-Distribution

In most cases, we do not know the value of \( \sigma^2 \). A possible alternative is to use \( s^2 \), whereupon we find

\[
\frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 d_{jj}}} \sim t_{n-2}.
\]

(7.31)

Under \( H_0 : \beta_j = \beta^0_j \), then

\[
\frac{\hat{\beta}_j - \beta^0_j}{\sqrt{s^2 d_{jj}}} \sim t_{n-2}
\]

while for \( H_1 : \beta_j = \beta^1_j \neq \beta^0_j \)

\[
\frac{\hat{\beta}_j - \beta^0_j}{\sqrt{s^2 d_{jj}}} = \frac{\hat{\beta}_j - \beta^1_j}{\sqrt{s^2 d_{jj}}} + \frac{\beta^1_j - \beta^0_j}{\sqrt{s^2 d_{jj}}}
\]

will have the distribution shifted to the right or left depending on whether \( \beta^1_j - \beta^0_j \) is positive or negative. Specifically, it would have a non-central t-distribution. This will be covered in more detail in the next chapter.
7.4.3 Maximum Likelihood Estimation

Now,

\[ y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i = x_i' \beta + u_i, \]

is linear in \( u_i \), so \( y_i \) is also normal given \( x_i' = (x_{i1}, x_{i2}, \ldots, x_{ik}) \):

\[ y_i \sim N( x_i' \beta, \sigma^2 ). \]

Thus, the density for \( y_i \) is given by

\[ f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} [y_i - x_i' \beta]^2 \right\}. \]

Since the \( u_i \)'s and hence the \( y_i \)'s are independent, the joint likelihood function is

\[ f(y_1, y_2, \ldots, y_n) = f(y_1) \cdots f(y_n) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - x_i' \beta]^2 \right\} = L(\beta, \sigma^2|y, X). \]

We may maximize this function with respect to \( \beta \) and \( \sigma^2 \) directly or equivalently, and more conveniently, the log-likelihood function

\[ \mathcal{L}(\beta, \sigma^2|y, X) = \ln L(\beta, \sigma^2|y, X) \quad (7.32) \]

\[ = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - x_i' \beta]^2 \]

which is a monotonic transformation. Notice that maximizing \( \mathcal{L} \) with respect to \( \beta \) is equivalent to minimizing the sum of squares \( \psi(\beta) = \sum_{i=1}^{n} [y_i - x_i' \beta]^2 \), so

\[ \hat{\beta}_{MLE} = (X'X)^{-1}X'y, \quad (7.33) \]

which is the OLS estimator from above. It is easily shown that

\[ \hat{\sigma}^2_{MLE} = \frac{e'e}{n} = \frac{n-k}{n} s^2. \quad (7.34) \]

The distribution of \( s^2 \) and hence \( \hat{\sigma}^2_{MLE} \) will be developed and utilized in the next chapter.

7.4.4 Efficiency of \( \hat{\beta} \) and \( s^2 \)

Since \( \hat{\beta} \) is the MLE and unbiased, we find then it is the minimum variance unbiased estimator (BUE). At the same time, \( s^2 \) is not the MLE, so it is not necessarily BUE. On the other hand, \( \hat{\sigma}^2_{MLE} \) is biased, so it is not BUE either. The bias disappears asymptotically, however, so they will both be equivalent in large samples and be asymptotically BUE.
7.5  An Example

We now examine a simple trivariate example taken from Wonacott and Wonacott. The details of the calculations are instructive and will be exhibited in full. The model considered is

\[ Y_t = \beta_1 + \beta_2 X_{t,2} + \beta_3 X_{t,3} + u_t, \]

(7.35)

where \( Y_t \) is wheat yield, \( X_{t,2} \) is the amount of fertilizer applied and \( X_{t,3} \) is the annual rainfall. The data are given in Table 8.1.

<table>
<thead>
<tr>
<th>Wheat Yield (Bushels/Acre)</th>
<th>Fertilizer (Pounds/Acre)</th>
<th>Rainfall (Inches/Year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>100</td>
<td>36</td>
</tr>
<tr>
<td>45</td>
<td>200</td>
<td>33</td>
</tr>
<tr>
<td>50</td>
<td>300</td>
<td>37</td>
</tr>
<tr>
<td>65</td>
<td>400</td>
<td>37</td>
</tr>
<tr>
<td>70</td>
<td>500</td>
<td>34</td>
</tr>
<tr>
<td>70</td>
<td>600</td>
<td>32</td>
</tr>
<tr>
<td>80</td>
<td>700</td>
<td>36</td>
</tr>
</tbody>
</table>

Table 8.1: Wheat yield data.

For computational convenience, we rescale the data, and calculate the elements of the moment matrix

\[
\begin{array}{cccccccc}
Y_t & X_{t2} & X_{t3} & X_{t2}Y_t & X_{t3}Y_t & X_{t2}^2 & X_{t2}X_{t3} & X_{t3}^2 \\
4.0 & 1 & 3.6 & 4 & 14.40 & 1 & 3.6 & 12.96 \\
4.5 & 2 & 3.3 & 9 & 14.85 & 4 & 6.6 & 10.89 \\
5.0 & 3 & 3.7 & 15 & 18.50 & 9 & 11.1 & 13.69 \\
6.5 & 4 & 3.7 & 26 & 24.05 & 16 & 14.8 & 13.69 \\
7.0 & 5 & 3.4 & 35 & 23.80 & 25 & 17.0 & 11.56 \\
7.0 & 6 & 3.2 & 42 & 22.40 & 36 & 19.2 & 10.24 \\
8.0 & 7 & 3.6 & 56 & 28.80 & 49 & 25.2 & 12.96 \\
42 & 28 & 24.5 & 187 & 146.80 & 140 & 97.5 & 85.99 \\
\end{array}
\]

Table 8.2: Moment matrix calculations.
Thus,

\[
X'X = \begin{pmatrix}
\sum x_t^2 & \sum x_t x_{t2} & \sum x_t x_{t3} \\
\sum x_t x_{t2} & \sum x_t^2 x_{t2} & \sum x_t^2 x_{t3} \\
\sum x_t x_{t3} & \sum x_t^2 x_{t3} & \sum x_t^3 \\
\end{pmatrix}
\] (7.36)

\[
= \begin{pmatrix}
7 & 28 & 24.5 \\
28 & 14 & 97.5 \\
24.5 & 97.5 & 85.99 \\
\end{pmatrix}
\] (7.37)

\[
X'y = \begin{pmatrix}
\sum y_t \\
\sum x_t y_t \\
\sum x_t^2 y_t \\
\end{pmatrix}
= \begin{pmatrix}
42.0 \\
187.0 \\
146.8 \\
\end{pmatrix}
\] (7.38)

and

\[
det(X'X) = 84,270.2 + 66,885.0 + 66,885.0 \\
- 84,035.0 - 67,416.16 - 66,543.75 \\
= 45.29
\]

\[
cof(X'X) = \begin{pmatrix}
2532.35 & -18.97 & -700 \\
-18.97 & 1.68 & 3.5 \\
-700 & 3.5 & 196.0 \\
\end{pmatrix}
\]

\[
(X'X)^{-1} = \frac{1}{\det(X'X)} \text{cof}(X'X)
\]

\[
= \begin{pmatrix}
55.9141 & -0.4189 & -15.4560 \\
-0.4189 & 0.0371 & 0.0773 \\
-15.4560 & 0.0773 & 4.3277 \\
\end{pmatrix}
\]

We can now calculate the least squares estimator

\[
\hat{\beta} = (X'X)^{-1}X'y
\]

\[
= \begin{pmatrix}
55.9141 & -0.4189 & -15.4560 \\
-0.4189 & 0.0371 & 0.0773 \\
-15.4560 & 0.0773 & 4.3277 \\
\end{pmatrix}
\begin{pmatrix}
42.0 \\
187.0 \\
146.8 \\
\end{pmatrix}
\]

\[
= \begin{pmatrix}
1.132 \\
0.689 \\
0.603 \\
\end{pmatrix}
\]

In order to conduct inference and evaluate the quality of these estimates, we need to obtain the relevant sums-of-squares. Accordingly we obtain
It follows that

\[ s^2 = \frac{\sum e_t^2}{(n - 3)} = \frac{0.5232}{4} = 0.1308, \]

\[ R^2 = 1 - \frac{0.5232}{13.5} = 0.9612, \]

since \( SSE = \sum e_t^2 = 0.5232 \) and \( SST = \sum (Y_t - \bar{Y})^2 = 13.50 \).

Recall

\[ E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\prime] = \text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}, \quad (7.39) \]

which we estimate using

\[ \hat{\text{Cov}}(\hat{\beta}) = s^2(X'X)^{-1}. \quad (7.40) \]

Thus,

\[ \hat{\text{Var}}(\hat{\beta}_1) = s^2d_{11} = 7.3133 \]
\[ \hat{\text{Var}}(\hat{\beta}_2) = s^2d_{22} = 0.0049 \]
\[ \hat{\text{Var}}(\hat{\beta}_3) = s^2d_{33} = 0.5660 \]

For \( H_0: \beta_3 = 0 \) vs \( H_1: \beta_3 \neq 0 \), we have

\[ \frac{\hat{\beta}_3 - \beta_3^0}{\sqrt{\hat{\text{Var}}(\hat{\beta}_3)}} = \frac{0.603}{0.75236} = 0.80119 \sim t_4. \quad (7.41) \]

Now, a 95% acceptance region for a \( t_4 \) distribution is \(-2.776 \leq t_4 \leq 2.776\). Thus, we fail to reject the null hypothesis that rainfall has no effect on yield once we control for fertilizer.

For \( H_0: \beta_2 = 0 \) vs \( H_1: \beta_2 \neq 0 \), we have

\[ \frac{\hat{\beta}_2 - \beta_2^0}{\sqrt{\hat{\text{Var}}(\hat{\beta}_2)}} = \frac{0.6893}{0.0697} = 9.8965 \sim t_4. \quad (7.42) \]

and we reject the null hypothesis at the 95% confidence level. In fact, we reject at the 99.9% confidence level, where the acceptance region is \(-7.173 \leq t_4 \leq 7.173\). Clearly, rainfall is important in determining yield whether or not fertilizer is applied.
Chapter 8

Confidence Intervals and Hypothesis Tests

In the previous chapter we established a number of properties of the least-squares estimator for the linear regression model. These estimates provide point estimates and distributions of the coefficients in question. Ultimately, we are interested in making statements regarding the reasonableness of our preferred values for certain coefficients. This entails making a probability statement regarding the most likely values of the estimators relative to the preferred values. We can formulate these statements as hypothesis tests, confidence intervals, or p-values, which will be developed below. First we review and extend the previous results.

8.1 Preliminaries

8.1.1 Model and Assumptions

The model is a $k$-variable linear model:

$$ y = X\beta + u, $$

where $y$ and $u$ are both $n \times 1$ vectors, $X$ is a $n \times k$ matrix and $\beta$ is a $k \times 1$ vector. We make the following assumptions about the disturbances:

(i) $E[u] = 0$

and

(ii),(iii) $\text{Cov}(u) = E[uu'] = \sigma^2 I_n$,

where $I_n$ is an $n \times n$ identity matrix. The nonstochastic assumptions are

(iv) $X$ is nonstochastic.
(v) $X$ has full column rank (the columns are linearly independent).
For inferences, we assume that $u$ are normally distributed. That is,
(vi) $u \sim N(0, \sigma^2 I_n)$.

### 8.1.2 Ordinary Least Squares Estimation

For some estimate $\hat{\beta}$ of $\beta$, define

$$ e = y - X\hat{\beta} $$

and

$$ \psi(\beta) = e'e $$

Choosing $\hat{\beta}$ to minimize $\psi(\beta)$ yields the ordinary least squares (OLS) estimator

$$ \hat{\beta} = (X'X)^{-1}X'y, $$

Substitution yields

$$ \hat{\beta} = (X'X)^{-1}X'(X\beta + u) $$

$$ = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u $$

$$ = \beta + (X'X)^{-1}X'u. $$

This structure of the estimator being represented as the sum of the target plus an expression which resembles the estimator except it replaces $y$ with $u$ will prove to be very general.

### 8.1.3 Properties of $\hat{\beta}$

Since $X$ is nonstochastic,

$$ E[\hat{\beta}] = \beta + E[(X'X)^{-1}X'u] $$

$$ = \beta + (X'X)^{-1}X'E[u] $$

$$ = \beta. $$

Thus, the OLS estimator is unbiased. Also,

$$ \text{Cov}(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] $$

$$ = E[(X'X)^{-1}X'uu'X(X'X)^{-1}] $$

$$ = (X'X)^{-1}X'E[uu']X(X'X)^{-1} $$

$$ = (X'X)^{-1}X'\sigma^2 I_nX(X'X)^{-1} $$

$$ = \sigma^2(X'X)^{-1}. $$

The elements of $X'X$ are unbounded as $n$ gets very large. Therefore,

$$ \lim_{n \to \infty} (X'X)^{-1} = 0, $$
and the variances of $\hat{\beta}$ converge to zero. This means that the distribution collapses about its expected value, namely $\beta$. So,

$$\lim_{n \to \infty} \hat{\beta} = \beta,$$

and OLS estimation is consistent.

The OLS estimates $\hat{\beta}$ are the best linear unbiased (BLUE) in that they have minimum variance in the class of unbiased estimators of $\beta$ that are also linear in $y$.

Suppose that $u$ is normal, then the linear transformation

$$\hat{\beta} - \beta = (X'X)^{-1}X'u$$

is also normal.

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

or

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

Moreover, the $\hat{\beta}$ are maximum likelihood and hence minimum variance in the class of unbiased estimators (BUE).

### 8.1.4 Properties of $e$

Now, the OLS residuals are

$$e = y - X\hat{\beta}$$

$$= y - X(X'X)^{-1}X'y$$

$$= [I_n - X(X'X)^{-1}X']y$$

$$= My \quad M = I_n - X(X'X)^{-1}X'$$

$$= M(X\beta + u)$$

$$= [I_n - X(X'X)^{-1}X'](X\beta) + Mu$$

$$= Mu.$$ since $MX = 0$. Thus, the OLS residuals are a linear transformation of the underlying disturbances. Also,

$$X'e = X'Mu$$

$$= 0,$$

again, since $MX = 0$, and the OLS residuals are orthogonal or linearly unrelated to $X$. When $u$ are normal, then the linear transformation $e = Mu$ is also normal. Specifically,

$$e \sim N(0, \sigma^2M)$$

since

$$E[e] = E[Mu] = ME[u] = 0,$$
and

\[ E[ee'] = E[Muu'M'] = M (E[uu']) M' = M (\sigma^2 I) M' = \sigma^2 M, \]

since \( MM' = M \). Note that this is not a full rank distribution since as we see below \( \text{rank}(M) = n - k \).

### 8.2 Tests Based on the \( \chi^2 \) Distribution

#### 8.2.1 The \( \chi^2 \) Distribution

Recall that for \( z_1, z_2, \ldots, z_m \) i.i.d. \( N(0,1) \) random variables, then

\[ \sum_{i=1}^{m} z_i^2 \sim \chi^2_m. \]

#### 8.2.2 Distribution of \((n - k)s^2/\sigma^2\)

By definition,

\[ u_i = y_i - x_i'\beta \sim N(0, \sigma^2), \]

so

\[ \frac{u_i}{\sigma} \sim N(0, 1) \]

and

\[ \sum_{i=1}^{n} \left( \frac{u_i}{\sigma} \right)^2 = \sum_{i=1}^{n} \frac{u_i^2}{\sigma^2} \sim \chi^2_n. \]

Now,

\[ e_i = y_i - x_i'\hat{\beta} \]

is an estimate of \( u_i \) and we might expect that

\[ \sum_{i=1}^{n} \frac{e_i^2}{\sigma^2} \sim \chi^2_n. \]

However, this would be wrong as only \( n - k \) of the observations are independent since \( e \) satisfies the \( k \) equations \( X'e = 0 \).

The properties of \( e = Mu \) follow from the properties of \( M = I_n - X(X'X)^{-1}X' \), which is symmetric idempotent and positive semi-definite and hence has some very special properties. Principle of these is that we can write the decomposition \( M = QD_{n-k}Q' \) where \( D_{n-k} \) is a diagonal matrix of eigenvalues with its first
n - k diagonals unity and the remainder zero, and Q is a matrix of corresponding orthonormal eigenvectors such that \( Q'Q = I_n \) or \( Q' = Q^{-1} \). It follows, that \( \text{rank}(M) = \text{tr}(M) = n - k \).

Let \( v = Q'u \)

then \( v \sim N(0, \sigma^2 I_n) \) and \( u = Qv \). Substitution yields

\[
\frac{1}{\sigma^2} e'e = \frac{1}{\sigma^2} u'Mu
= \frac{1}{\sigma^2} u'QD_{n-k}Q'u
= \frac{1}{\sigma^2} v'QD_{n-k}Q'Qv
= \frac{1}{\sigma^2} v'D_{n-k}v
= \frac{1}{\sigma^2} \sum_{i=1}^{n-k} v_i^2
= \sum_{i=1}^{n-k} \left( \frac{v_i}{\sigma} \right)^2 \sim \chi^2_{n-k}
\]

Thus,

\[
\sum_{i=1}^{n} \frac{e_i^2}{\sigma^2} \sim \chi^2_{n-k}
\]

and

\[
(n - k) \frac{\sum_{i=1}^{n-k} e_i^2}{\sigma^2} = (n - k) \frac{s^2}{\sigma^2} \sim \chi^2_{n-k}.
\]

Not only so, but \( e = Mu \) and \( \hat{\beta} - \beta = (X'X)^{-1}X'u \) are obviously jointly normal and

\[
E[e(\hat{\beta} - \beta)'] = E[Muu'X(X'X)^{-1}]
= M\sigma^2 I_n X(X'X)^{-1}
= \sigma^2 MX(X'X)^{-1} = 0
\]

so they are uncorrelated and hence stochastically independent. But this means \( s^2 \) is independent of \( \hat{\beta} \), since it is a function only of \( e \).

### 8.2.3 An Hypothesis Test

Occasionally, a model will suggest a specific value of the variance which can be subjected to a hypothesis test. Suppose that

\[
H_0: \sigma^2 = \sigma_0^2, \quad H_1: \sigma^2 = \sigma_1^2 \neq \sigma_0^2.
\]

Then we know that

\[
(n - k) \frac{s^2}{\sigma_0^2} \sim \chi^2_{n-k}.
\]
under the null hypothesis. Under the alternative, we have
\[
(n - k) \frac{s^2}{\sigma_0^2} = (n - k) \frac{s^2 \sigma_1^2}{\sigma_1^2 \sigma_0^2}
\]
\[
= (n - k) \frac{s^2}{\sigma_1^2} + (n - k) \left( \frac{\sigma_1^2}{\sigma_0^2} - 1 \right) \frac{s^2}{\sigma_1^2}
\]
with the first term having a \(\chi^2\) distribution and the second being a shift term.
We see that the distribution will be shifted to the right if \(\sigma_1^2 > \sigma_0^2\) and to left if \(\sigma_1^2 < \sigma_0^2\). The size of the shift increases with the sample size and the difference between the null and alternative.

For example, suppose that \(s^2 = 4.0\) and \(\sigma_0^2 = 1\) for \(n - k = 14\), then
\[
(n - k) \frac{s^2}{\sigma_0^2} = 14 \frac{4.0}{1} = 56.
\]
Choose \(\alpha = 0.05\), say, then the critical values corresponding to 2.5% tails are 5.63 and 26.12 for \(n - k = 14\). Since 56 > 26.12 we are in the right-hand tail and we reject the null hypothesis.

### 8.2.4 A Confidence Interval

Now, let \(a\) and \(b\) be numbers such that
\[
Pr(b \leq \chi^2_{n-k} \leq a) = 1 - \alpha = 0.95,
\]
say. Then \(a\) and \(b\) can be obtained from a table. Thus,
\[
Pr\left( b \leq \frac{(n - k)s^2}{\sigma^2} \leq a \right) = 0.95
\]
\[
Pr\left( \frac{1}{b} \geq \frac{\sigma^2}{(n - k)s^2} \geq \frac{1}{a} \right) = 0.95
\]
\[
Pr\left( \frac{(n - k)s^2}{b} \geq \sigma^2 \geq \frac{(n - k)s^2}{a} \right) = 0.95
\]
establishes a 95% confidence interval for \(\sigma^2\).

For example, for \(n - k = 14\), we have
\[
Pr\left( 5.64 \leq \frac{14s^2}{\sigma^2} \leq 26.12 \right) = 0.95
\]
\[
Pr\left( \frac{14s^2}{5.64} \geq \sigma^2 \geq \frac{14s^2}{26.12} \right) = 0.95
\]
and if \(s^2 = 4.0\), then
\[
Pr\left( \frac{56}{5.63} \geq \sigma^2 \geq \frac{56}{26.12} \right) = 0.95
\]
or

\[ \Pr (10 \geq \sigma^2 \geq 2.1) = 0.95 \]

is the confidence interval.

This confidence interval comprises the set of values for \( \sigma^2 \) that would not be rejected if used as a null hypothesis when \( \alpha \) is chosen as the size. Conversely, any null value that falls outside the interval would lead to rejection then used as the basis of a hypothesis test with that \( \alpha \). In the example above, for \( \alpha = .05 \), the null hypothesis \( \sigma_0^2 = 1 \) is not included in this interval and would be rejected.

### 8.3 Tests Based on the \( t \) Distribution

#### 8.3.1 The \( t \) Distribution

Suppose that \( z \) is a \( N(0, 1) \) random variable and that \( w \sim \chi^2_m \) independent of \( z \), then

\[ \frac{z}{\sqrt{\frac{w}{m}}} \sim t_m. \]

#### 8.3.2 The Distribution of \( \left( \hat{\beta}_j - \beta_j \right) / \left( s^2 d_{jj} \right)^{1/2} \)

We have seen that

\[ \hat{\beta}_j \sim N(\beta_j, \sigma^2 d_{jj}), \]

where \( d_{jj} \) is the \( (j,j) \) element of the matrix \((X'X)^{-1}\). Then,

\[ z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 d_{jj}}} \sim N(0, 1), \]

while

\[ w = (n - k) \frac{s^2}{\sigma^2} \sim \chi^2_{n-k}. \]

Since \( \hat{\beta} \) and \( s^2 \) are independent, we have

\[ \frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{(n-k)s^2}{\sigma^2}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 d_{jj}}} \sim t_{n-k}. \]

#### 8.3.3 Testing a Simple Hypothesis

Suppose that

\[ H_0: \beta_j = \beta_j^0, \quad H_1: \beta_j = \beta_j^1 \neq \beta_j^0. \]

We know that

\[ \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\frac{s^2}{d_{jj}}}} \sim t_{n-k}. \]

under the null hypothesis.
Under the alternative hypothesis, we find the numerator
\[
\begin{align*}
    z &= \frac{\hat{\beta}_j - \beta^0_j}{\sqrt{\sigma^2 d_{jj}}} \\
    &= \frac{\hat{\beta}_j - \beta^1_j}{\sqrt{\sigma^2 d_{jj}}} + \frac{\beta^1_j - \beta^0_j}{\sqrt{\sigma^2 d_{jj}}} \\
    &\sim N\left( \frac{\beta^1_j - \beta^0_j}{\sqrt{\sigma^2 d_{jj}}}, 1 \right)
\end{align*}
\]
has a noncentral normal, so the statistic will have a non-central t-distribution
with non-centrality \((\beta^1_j - \beta^0_j)/\sqrt{\sigma^2 d_{jj}}\). The size of the noncentrality can depend
on a number of factors other than the difference between the null and alternative
values, which will be studied in Chapter 10.

Now, choose \(\alpha = 0.05\), say, then critical values corresponding to 2.5% tails
of a t distribution with \((n - k)\) degrees of freedom are taken from the tables as
\(\pm a\), say, so if
\[
\left| \frac{\hat{\beta}_j - \beta^0_j}{\sqrt{\sigma^2 d_{jj}}} \right| > a,
\]
we reject the null hypothesis. Otherwise we fail to reject the null hypothesis.
Here again we are making a choice between a value which is unlikely under the
null or likely under the alternative.

### 8.3.4 Confidence Interval and P-value for \(\beta_j\)

We consider a two-sided interval. First, for \(n - k\) degrees of freedom, we obtain
\(a\) such that
\[
\Pr(-a \leq t_{n-k} \leq a) = 1 - \alpha = 0.95,
\]
say, from a table. Then,
\[
\begin{align*}
    \Pr \left( -a \leq \frac{\hat{\beta}_j - \beta^1_j}{\sqrt{\sigma^2 d_{jj}}} \leq a \right) &= 0.95 \\
    \Pr \left( \hat{\beta}_j + a\sqrt{\sigma^2 d_{jj}} \geq \beta_j \geq \hat{\beta}_j - a\sqrt{\sigma^2 d_{jj}} \right) &= 0.95
\end{align*}
\]
and \(\hat{\beta}_j \pm a\sqrt{\sigma^2 d_{jj}}\) defines a 95% confidence interval for \(\beta_j\).

It is important to emphasize that the interval is random and \(\beta_j\) is not.
Specifically, the center and width of the interval are random since \(\hat{\beta}_j\) and \(\hat{s}^2\) are
both random variables. The probability statement above says that the random
interval has a 95% chance of including the true (non-random) \(\beta_j\).

As with the chi-squared interval, this interval comprises the set of non-
rejectable null hypotheses for \(\beta_j\), when \(\alpha\) is chosen as the size. Conversely, any
null value that falls outside the interval would lead to rejection when used as
the basis of a hypothesis test with that \(\alpha\). This is a much more informative
procedure than the hypothesis test, which gives a yes-no answer for a particular \( \alpha \) and null value.

The other side of the inferential coin to the confidence interval is the p-value. The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. Suppose that \( \tau \) is the observed value of the statistic and \( t \) is a random variable that has the specified \( t_{n-k} \)-distribution, then

\[
p\text{-value} = \Pr(|t| \geq |\tau|) = 2 \Pr[t \geq |\tau|]
\]

since the \( t \)-distribution is symmetric. For a two-sided hypothesis test, if \( \alpha \) is smaller than the p-value then we would not reject with the realization \( \tau \) of the statistic. Thus it yields the set of \( \alpha \) for which the given null hypothesis will not be rejected. Again, this is much more informative than a single hypothesis test.

### 8.3.5 Testing a Linear Restriction on \( \beta \)

We frequently encounter situations where we want to test a linear restriction such as \( H_0: c'\beta = a \) where \( c \) and \( a \) are known weights. For example, \( \beta_1 + \beta_2 = 1 \) or, as a special case, the usual zero restriction \( \beta_j = 0 \). Consider the linear combination

\[
c'\hat{\beta}.
\]

It follows that

\[
c'\hat{\beta} \sim N(c'\beta, \sigma^2 c'(X'X)^{-1}c),
\]

and

\[
\frac{c'(\hat{\beta} - \beta)}{\sqrt{\sigma^2 c'(X'X)^{-1}c}} \sim N(1, 0).
\]

As before, we use \( s^2 \) instead of \( \sigma^2 \), whereupon we find

\[
\frac{c'(\hat{\beta} - \beta)}{\sqrt{s^2 c'(X'X)^{-1}c}} \sim t_{n-k}.
\]

We can perform inferences and calculate confidence intervals and p-values as before. In the next section we will consider testing more than one linear restriction at the same time.

### 8.4 Tests Based on the F Distribution

#### 8.4.1 The F Distribution

Suppose that

\[
v \sim \chi^2_l \quad \text{and} \quad v \sim \chi^2_m
\]
8.4. Tests Based on the F Distribution

If $v$ and $w$ are independent, then
\[
\frac{v/l}{w/m} \sim F_{l,m}.
\]

8.4.2 Testing a Set of Linear Restrictions on $\beta$

Suppose we are interested in testing a set of $q$ linear restrictions. Examples would be $\beta_1 + \beta_2 + ... + \beta_k = 1$ and $\beta_3 = 2\beta_2$. More generally, we consider
\[
H_0: \quad R\beta = r \quad H_1: \quad R\beta \neq r
\]
where $r$ is a $q \times 1$ known vector and $R$ is a $q \times k$ known matrix. Due to the multivariate normality of $\hat{\beta}$, then under the null hypothesis, we have
\[
R\hat{\beta} - r \sim N(0, \sigma^2 R(X'X)^{-1}R')
\]
and hence
\[
(R\hat{\beta} - r)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r) \sim \chi^2_q.
\]

Recall that $\hat{\beta}$ and $\hat{s}^2$ are independent so $(n - k)\frac{\hat{s}^2}{\sigma^2} \sim \chi^2_{n-k}$ is independent of the quadratic form in (8.53). Thus, under the null hypothesis,
\[
F = \frac{(R\hat{\beta} - r)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{(n - k)\frac{\hat{s}^2}{\sigma^2}/(n - k)} \sim F_{q,n-k}
\]
and after some simplification
\[
F = (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/(s^2q) \sim F_{q,n-k}. \tag{8.1}
\]

Under the alternative hypothesis $R\beta - r = d \neq 0$, then the numerator is distributed $\chi^2_q(\delta)$ where $\delta = d'[\sigma^2 R(X'X)^{-1}R']^{-1}d$. Accordingly, $F$ is distributed $F_{q,n-k}(\delta)$, which diverges at the rate $n$, and we expect large positive values of the statistic with high probability. This is why we typically only consult the RHS tail values of the distribution to establish critical values. Values of the statistic exceeding these critical values are rare events under the null but typical under the alternative, so we reject when the realization exceeds the critical value.

8.4.3 A Sums of Squares Approach

An alternative to the quadratic form test just proposed can be based on sums of squares from a restricted and unrestricted regression. The unrestricted regression is the usual least squares estimator $\hat{\beta} = (X'X)^{-1}X'y$ with corresponding unrestricted residuals
\[
e_u = e = y - X\hat{\beta}
\]
and unrestricted sum of squares
\[
SSE_u = e_u'e_u.
\]
The restricted regression is obtained as
\[
\tilde{\beta} = \arg\min_{\beta} (y - X\beta)'(y - X\beta) \quad s.t. \quad R\beta - r = 0.
\]
Define the corresponding Lagrangian function
\[
\varphi(\beta, \lambda) = [(y - X\beta)'(y - X\beta)] - \lambda' (R\beta - r)
\]
where \( \lambda \) is a \( k \times 1 \) vector of Lagrange multipliers. The first-order conditions for optimizing this function, evaluated at the solution \( \tilde{\beta} \) are
\[
\frac{\partial \varphi(\tilde{\beta}, \lambda)}{\partial \beta} = -2X'(y - X\tilde{\beta}) - R'\lambda = 0
\]
\[
\frac{\partial \varphi(\tilde{\beta}, \lambda)}{\partial \lambda} = R\beta - r = 0.
\]
Multiplying the first equation by \( R(X'X)^{-1} \) and solving for \( \lambda \) yields
\[
\lambda = 2[R(X'X)^{-1}R']^{-1}R(\tilde{\beta} - \hat{\beta})
\]
\[
= 2[R(X'X)^{-1}R']^{-1}(R\tilde{\beta} - r)
\]
since \( R\tilde{\beta} = r \). Substituting for \( \lambda \) in the first equation and solving for \( \tilde{\beta} \) yields
\[
\tilde{\beta} = \beta - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\tilde{\beta} - r).
\] (8.2)

Now define the corresponding restricted residuals
\[
e_r = y - X\tilde{\beta}
\]
\[
= y - X\tilde{\beta} - X(\beta - \tilde{\beta})
\]
\[
= e_u - X(\beta - \tilde{\beta}).
\]
Thus the restricted sum of squares is
\[
SSE_r = e_r'e_r
\]
\[
= e_u'e_u - 2e_u'X(\beta - \tilde{\beta}) + (\beta - \tilde{\beta})'X'X(\beta - \tilde{\beta})
\]
\[
= e_u'e_u + (\beta - \tilde{\beta})'X'X(\beta - \tilde{\beta})
\]
since \( X'e_u = X'e = 0 \). Rearranging and substituting for \( (\beta - \tilde{\beta}) \) from (8.2) yields
\[
e_r'e_r - e_u'e_u = (\beta - \tilde{\beta})'X'X(\beta - \tilde{\beta})
\]
\[
= (R\tilde{\beta} - r)'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'X(\beta - \tilde{\beta})
\]
\[
= (R\tilde{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\tilde{\beta} - r).
\]
8.4. TESTS BASED ON THE F DISTRIBUTION

Substituting into (8.1), we find the F-statistic can be alternatively and equivalently written

\[
F = \frac{(e'_r e_r - e'_u e_u) / (s^2 q)}{SSE_u / (n - k)} \sim F_{q, n - k}.
\]

This is by far the most frequently see form of the F-statistic and is directly related to the likelihood ratio statistic, as is shown in the Appendix to the Chapter.

8.4.4 Testing a Set of Zero Restrictions on \( \beta \)

Although superficially simpler, the sums of squares form of the \( F \)-statistic requires obtaining the restricted least squares estimates which, in general, is no simpler than calculating the quadratic form alternative. In some cases, however, where the restricted estimates are easy to obtain, it has substantial appeal. The leading such case is when the null hypothesis eliminates a set of variables from the regression.

Suppose the “unrestricted” model can be written as

\[
y = X_1 \beta_1 + X_2 \beta_2 + u
\]

and

\[
H_0: \beta_2 = 0 \quad H_1: \beta_2 \neq 0
\]

where \( \beta_1 \) is \((k_1 \times 1)\), \( \beta_2 \) is \((k_2 \times 1)\), and \( k_1 + k_2 = k \) so \( q = k_2 \). For \( \hat{\beta}' = (\hat{\beta}_2', \hat{\beta}_2') \), define the unrestricted residuals

\[
e_u = y - X_1 \hat{\beta}_1 - X_2 \hat{\beta}_2
\]

and

\[
SSE_u = e'_u e_u
\]

from the OLS regression of \( y \) on \( X_1 \) and \( X_2 \), namely \( \hat{\beta} = (X'X)^{-1}X'y \).

Under the null hypothesis the “restricted” model can be written

\[
y = X_1 \beta_1 + u
\]

Next, define the restricted residuals

\[
e_r = y - X_1 \tilde{\beta}_1
\]

and

\[
SSE_r = e'_r e_r
\]

from the OLS regression of \( y \) on \( X_1 \) only, namely \( \tilde{\beta}_1 = (X_1'X_1)^{-1}X_1'y \).

Thus specific for the zero restriction case, the \( F \)-statistic becomes

\[
F = \frac{(SSE_r - SSE_u) / k_2}{SSE_u / (n - (k_1 + k_2))} \sim F_{k_1, n - (k_1 + k_2)}
\]
We can consult the tables to find the critical point, \( c_{0.05} \), corresponding to \( \alpha = 0.05 \), say. Then, if \( F > c_{0.05} \), we reject the null hypothesis at the 5% level. Note that for \( k_2 = 1 \), that is, one restriction, the critical point for the \( F \)-distribution is the square of the critical point for the \( t_{n-(k_1+1)} \) distribution. So, in this simple case, the test based on the \( F \) statistic is equivalent to the two-sided test based on the \( t \)-ratio for the \( \beta_2 \) coefficient.

### 8.5 An Example

We now undertake an extended example which examines the possibility that differences in wages may or may not be explained by gender discrimination. We have a series of observations (527) on wages by individual for a firm. We also have a number of other variables which might be potential explanatory variables in a wage equation.

We begin with a bivariate regression equation with level of wages explained by education attainment in years. The results are:

\[
  w_i = -1.54 + 0.804E_i + e_i 
\]

with \( SSE = 10574.16 \) and \( R^2 = 0.166 \). The estimated standard errors for each coefficient estimate is in parentheses below the estimate, which is a standard way to present the results. The coefficient on \( E_i \) is highly significant and indicates that an additional year of schooling adds 80 cents to the wage. A plot of the residual against \( E_i \), however reveals a problem.
8.5. AN EXAMPLE

It is clear that the spread of the residuals is larger at higher values of education and wages. This is not surprising since most researchers find that income and wage data are best approximated by a log-normal distribution.

Accordingly we turn to a specification with the log of wages as the dependent variable. The results are

$$\ln(w_i) = 0.985 + 0.0822E_i + e_i$$

with \( SSE = 118.13 \) and \( R^2 = 0.157 \). Note that neither of the latter two statistics are comparable since the dependent variable is different. The coefficient on \( E_i \) is also highly significant here but has a different interpretation as the percentage increase in the wage for a year additional schooling. A plot of the residuals for this form is
This plot has largely corrected the problem of spread increasing with education and appears approximately normal which is desirable.

Using this regression as a base we begin to examine the possibility that wages are paid differentially on the basis of gender. Accordingly, we add $G_i$, which represents gender, to the regression. Now $G_i$ is a “dummy” or indicator variable that takes on a value of 1 for females and 0 for males. This specification allows a different intercept for male and female employees. The results are

$$\ln(w_i) = 1.114 + 0.0807E_i - 0.240 G_i + e_i$$

with $SSE = 110.60$ and $R^2 = 0.211$. These results indicate that, on average, a female receives 24 percent less in wages, correcting for education. Testing the null hypothesis of no discrimination or the coefficient on gender being zero we obtain the statistic $(\hat{\beta}_G - \beta_G)/\sqrt{s.e.(\hat{\beta}_G)} = -0.240/0.0402 = -5.97$. This statistic should have a $t$-value which is very closely approximated by the standard normal. Looking at the normal tables we see that this realization is far far out in the left-hand tail of the distribution and hence very unlikely. The probability of such an unlikely value is effectively zero. We would reject the null for any positive choice of $\alpha$. There seems to be convincing evidence of discrimination. A plot of the residuals for males and females is presented below.
Males are represented by “o” and females by “x” with the two regression lines reflecting the shifted intercept.

Some analysts might caution us not to be so bold in our conclusions. They would argue that there may be other factors important to determining the wage that differ by gender and might provide an explanation for the differences. Foremost would be the suggestion that we also consider $X_i$ or experience as an explanatory variable. It is likely important in explaining the wage and due to family interruptions likely to differ for males and females. Accordingly, we add $X_i$ and $X_i^2$ as explanatory variables. The square is added because research has shown that the impact of experience on wages is nonlinear - first increasing the wage and then ultimately decreasing it in later years. The results are

$$\ln(w_i) = 0.498 + 0.0944 E_i + 0.0446 X_i - 0.00074 X_i^2 - 0.267 G_i + e_i$$

with $SSE = 92.11$ and $R^2 = 0.343$. Surprisingly, this did not detract from the finding that women received a significantly lower wage. In fact, the statistic for testing the null of no discrimination is marginally higher at $-7.23$. Note that both of the experience coefficients are also highly significant.

Our cautious analysts might suggest there is more work to be done. In particular, it might be argued that what we are witnessing is discrimination by race rather than gender. They would argue that minorities are over-represented in the females employee group as a justification. Accordingly, we add two more
“dummy” variables: $B_i$ for black and $H_i$ for hispanic. The results are

$$\ln(w_i) = 0.539 + 0.0927E_i + 0.0446X_i - 0.00074X_i^2$$

$$- 0.101B_i - 0.094H_i - 0.268 G_i + e_i$$

with $SSE = 91.37$ and $R^2 = 0.348$. Again, adding the variables did not change the result with respect to gender discrimination. The statistic for testing that null is $-7.28$ and just as significant – hugely. It is interesting that neither of the other dummy variables has a significant coefficient at the 5 percent level when conducting a two-sided test with values of $-1.83$ for testing the null that $\beta_B = 0$ and $-1.08$ for testing the null that $\beta_H = 0$. If we conduct a one-sided test with negative as the alternative, then the black coefficient is significant at the 5 percent level. We can conduct an $F$-test for testing the null that both coefficients are zero. This compares the last two regressions and yields the statistic

$$\frac{(92.11 - 91.37)/2}{91.37/520} = \frac{0.37}{0.175} = 2.10$$

which should have an $F_{2,520}$ distribution under the null. This has a p-value of 0.125 and is hence not significant for any of the preferred choices for $\alpha$.

The bottom line is that there is strong evidence of gender discrimination controlling for the likely other covariates.

### 8.A Appendix

Since we have the maximum likelihood estimator we can directly apply the results of Chapter 5. There we introduced the trinity of tests for a vector restriction. They were the likelihood ratio, Wald and Lagrange multiplier tests. It is instructive to present these tests in some detail for the present case. As above, we seek to test the null hypothesis $H_0 : R\hat{\beta} = r$ where $r$ is a $q \times 1$ known vector and $R$ is a $q \times k$ known matrix full row rank.

First consider the Wald statistic from above. Now, from above, $R\hat{\beta} - r \sim N(0, \sigma^2 R(X'X)^{-1}R')$ so using the maximum likelihood estimate $\hat{\sigma}^2$ for $\sigma^2$, we have the Wald statistic as the feasible quadratic form

$$W = \frac{(R\hat{\beta} - r)'[\hat{\sigma}^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{q \cdot \frac{n}{n-k}}$$

Except for being divided by $q$ and use of $s^2$ rather than $\hat{\sigma}^2$, we see that the $F$-statistic is just a Wald-type test. The critical points for this statistic will just be the critical points of the appropriate $F$-distribution, multiplied by $\frac{q}{n-k}$. The $F$-statistic is really just a Wald-type test.
Next, we consider the likelihood ratio statistic. From Chapter 7, the log-likelihood function is given by

\[
L_u = -\frac{n}{2} \ln 2\pi \hat{\sigma}^2 \frac{1}{2\sigma^2} (y - X\hat{\beta})'(y - X\beta).
\]

Suppose \(\hat{\beta}\) denotes the unrestricted maximum likelihood estimator used in the Wald-type test and \(\beta\) is the restricted maximum likelihood estimator, then \(R\beta - r = 0\). Let \(e_u = y - X\hat{\beta} = e\) and \(e_r = y - X\beta\) denote the unrestricted and restricted residuals, then \(\hat{\sigma}^2 = \hat{e}_u'^e_u / n\) and \(\hat{\sigma}^2 = e_r'^e_r / n\) are the unrestricted and restricted maximum likelihood estimators of \(\sigma^2\).

Now substitution of \(\hat{\sigma}^2\) into the unrestricted log-likelihood function yields the concentrated unrestricted log-likelihood function

\[
L_u = -\frac{n}{2} \ln 2\pi \hat{\sigma}^2 \frac{1}{2\hat{\sigma}^2} (y - X\hat{\beta})'(y - X\hat{\beta})
= -\frac{n}{2} \ln 2\pi \hat{\sigma}^2 \frac{1}{2\hat{\sigma}^2} \hat{u}'\hat{u}
= -\frac{n}{2} \ln 2\pi \frac{n}{2} \ln \hat{\sigma}^2 - \frac{n}{2}.
\]

Similarly the concentrated restricted log-likelihood is

\[
L_r = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \hat{\sigma}^2 - \frac{n}{2}.
\]

Thus we have the likelihood ratio statistic and its limiting distribution

\[
LR = 2(L_u - L_r)
= 2\left(-\frac{n}{2} \ln \hat{\sigma}^2 + \frac{n}{2} \ln \sigma^2\right)
= n \cdot \ln \frac{\hat{\sigma}^2}{\sigma^2} \rightarrow_d \chi^2.
\]

The likelihood ratio statistic is obviously a monotonic function of the ratio of restricted to unrestricted sums-of-squares. Moreover,

\[
\frac{\hat{\sigma}^2}{\sigma^2} = \frac{e_r'^e_r / n}{e_u'^e_u / n} = \frac{SSE_r}{SSE_u}
= \frac{SSE_r - SSE_u}{SSE_u} + 1
= \frac{SSE_r - SSE_u}{SSE_u} + 1
= \frac{(SSE_r - SSE_u) / k_2}{SSE_u / (n - (k_1 + k_2))} \times \frac{(n - (k_1 + k_2))}{k_2} + 1
= F \times \frac{(n - (k_1 + k_2))}{k_2} + 1.
\]

We see that the likelihood ratio is a simple nonstochastic monotonic transformation of the \(F\)-statistic for the linear model. Accordingly, the finite-sample
critical values for the likelihood ratio are the same nonstochastic transformation of the $F$-statistic critical values. Both the likelihood ratio and Wald-type test will yield exactly the same inferences in finite samples as the $F$-statistic.

The Lagrange multiplier test examines the unrestricted scores (first derivatives of the log-likelihood function) evaluated at the restricted estimates. If the restriction is correct then the average of the scores will converge in probability to zero and appropriately normalized be asymptotically normal. For this model we have

$$\frac{\partial \ln L(\tilde{\beta})}{\partial \beta} = \frac{1}{\sigma^2} X'(y - X\tilde{\beta}).$$

For simplicity, we consider the case where the restrictions have been rewritten in zero form as discussed above, so $\tilde{\beta} = (\tilde{\beta}_1', 0)'$.

For simplicity, we consider the case from Section 8.4.3 where the restrictions require the final $q = k_2$ coefficients to be zero. Now, by definition, the restricted estimates yield

$$0 = \frac{\partial \ln L(\tilde{\beta})}{\partial \beta_1} = \frac{1}{\sigma^2} X'_1(y - X\tilde{\beta}),$$

so for the LM statistic, we only consider the final $q$, possibly nonzero, elements of the score

$$\frac{\partial \ln L(\tilde{\beta})}{\partial \beta_2} = \frac{1}{\sigma^2} X'_2(y - X\tilde{\beta}) = \frac{1}{\sigma^2} X'_2\tilde{u} = \frac{1}{\sigma^2} X'_2M_1y = \frac{1}{\sigma^2} X'_2M_1u$$

where $M_1 = I_n - X_1(X'_1X_1)^{-1}X'_1$.

By normality of $u$, we have

$$\frac{\partial \ln L(\tilde{\beta})}{\partial \beta_2} \sim N(0, \frac{1}{\sigma^2} X'_2M_1X_2).$$

The corresponding quadratic form,

$$\frac{\partial \ln L(\tilde{\beta})}{\partial \beta_2} \left(\frac{1}{\sigma^2} X'_2M_1X_2\right)^{-1} \frac{\partial \ln L(\tilde{\beta})}{\partial \beta_2} = \frac{u'M_1X_2(X'_2M_1X_2)^{-1}X'_2M_1u}{\sigma^2}$$

will have a $\chi^2_q$ distribution. The LM statistic is the feasible version of this
8.5. AN EXAMPLE

quadratic form

\[
LM = \frac{u'\mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2' \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2' \mathbf{M}_1 \mathbf{u}}{\hat{\sigma}^2} \\
= \frac{u'\mathbf{M}_1 \mathbf{u} - u' [\mathbf{M}_1 - \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2' \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2' \mathbf{M}_1] \mathbf{u}}{\hat{\sigma}^2} \\
= \frac{\mathbf{e}_r' \mathbf{e}_r - \mathbf{e}_u' \mathbf{e}_u}{\hat{\sigma}^2} = \frac{\mathbf{e}_r' \mathbf{e}_r - \mathbf{e}_r' \mathbf{e}_u}{\mathbf{e}_u' \mathbf{e}_u / n} = \frac{(\mathbf{e}_r' \mathbf{e}_r - \mathbf{e}_u' \mathbf{e}_u) / k_2}{\mathbf{e}_u' \mathbf{e}_u / (n - k)} = \frac{n \cdot k_2}{n - k}
\]

And we see that the \( LM \) statistic is, like the \( LR \), a monotonically increasing nonstochastic function of \( F \)-statistic. So the finite-sample critical values for the \( LM \) statistic are the same nonstochastic transformation of the \( F \)-statistic critical values. Not only so, but the \( LM \) is exactly the same as the \( W \).
Chapter 9
Prediction in Linear Models

In the first chapter, one of the three statistical tasks we outlined for econometrics was to forecast the behavior of the variable of interest using an estimated model of the process that generates the variable. For example, we might be interested in predicting the GPA of an incoming college student using a model based on high school grades and test scores. Or we might seek to predict the response of GDP levels to various choices of levels of the policy instruments such as the money supply, government expenditures, and tax levels. In either case we have a set of explanatory variables, whose values we know, and seek to predict the corresponding values of the dependent variable based on an estimated relationship. Estimation and inference concerning this relationship were studied in the previous two chapters. In this chapter, we will examine how to use these results to accomplish prediction when the underlying model is the linear regression model.

9.1 Notation and Review

It is useful in this chapter to introduce a more compact notation for a single observation on the linear relationship. Specifically, from before we have

\[ y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \quad i = 1, 2, \ldots, n \]

which can be rewritten as

\[ y_i = x_i' \beta + u_i \quad i = 1, 2, \ldots, n \]

where \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ik}) \) is the \( 1 \times k \) vector of explanatory variables for observation \( i \). For the bivariate model studied in the previous chapter, \( k = 2 \) and \( x_{i1} = 1 \) yields \( \beta_1 \) as the intercept and \( x_{i2} = x_i \) yields \( \beta_2 \) as the slope coefficient.

In terms of the single observation representation, we have the assumptions from Chapter 7. For the disturbances we assume
(i) \( E(u_i) = 0 \) for all \( i \)
(ii) \( E(u_i^2) = 0 \) for all \( i \)
(iii) \( E(u_i u_\ell) = 0 \), for all \( i \neq \ell \).

For the independent or explanatory variables, we assume
(iv) \( x_{ij} \) non-stochastic for all \( i,j \)
(v) \( (x_{i1}, x_{i2}, ..., x_{ik}) \) linearly independent for all \( i \).

For the purposes of inference in finite samples, we also assume
(vi) \( u_i \sim \text{i.i.d. } N(0, \sigma^2) \) for all \( i \).

Using matrix notation, we can equivalently write all observations on this model as or, more compactly, as
\[
y = X\beta + u. \tag{9.1}
\]

where
\[
\begin{align*}
\mathbf{y} \sim n \times 1 & = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \\
\mathbf{X} \sim n \times k & = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \\
\beta \sim k \times 1 & = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \\
\mathbf{u} \sim n \times 1 & = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.
\end{align*}
\]

Note that \( x'_i \) is the \( i \)th row of \( \mathbf{X} \).

The assumptions can be rewritten in terms of the matrix form of the model. Specifically we have, for the disturbances,
(i) \( E(u) = 0 \)
and
(ii),(iii) \( \text{Cov}(u) = E uu' = \sigma^2 \mathbf{I}_n \),

where \( \mathbf{I}_n \) is an \( n \times n \) identity matrix. And for the explanatory variables, we have
(iv) \( \mathbf{X} \) is nonstochastic.
(v) \( \mathbf{X} \) has full column rank (the columns are linearly independent).

And for inferences in finite samples, we have
(vi) \( u \sim N(0, \sigma^2 I_n) \).

In Chapter 7, we introduced the least squares estimator of the coefficient vector \( \beta \). Specifically, we obtained

\[
\hat{\beta} = (X'X)^{-1}X'y \tag{9.2}
\]
\[
= \beta + (X'X)^{-1}X'u \tag{9.3}
\]
as long as \(|X'X| \neq 0\), which is assured by Assumption (v). Since \( X \) is non-stochastic by assumption (iv), we have, also using assumption (i),

\[ E(\hat{\beta}) = \beta. \]

Thus, the OLS estimator is unbiased. Also, using assumptions (ii) and (iii),

\[ \text{Cov}(\hat{\beta}) = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \sigma^2(X'X)^{-1}. \]

OLS was shown to be the best unbiased estimator within the class of estimators that are linear in \( y \) (BLUE). Note that we have used all of the Assumptions (i)-(v) to get to this point.

Suppose that we introduce assumption (vi), so the \( u_i \)'s are jointly normal:

\[ u \sim N(0, \sigma^2 I_n). \tag{9.4} \]

Since \( \hat{\beta} \) is linear in \( u \) and \( (X'X)^{-1}X' \) is nonstochastic, it follows that \( \hat{\beta} \) is also jointly normally distributed. We already know the mean and covariance matrices so we have

\[ \hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}). \tag{9.5} \]
as the joint distribution of \( \hat{\beta} \).

The normality of the estimator proved useful in performing inferences in finite samples, as discussed in the previous chapter. The individual elements of \( \hat{\beta} \) will be marginally normal, so

\[ \hat{\beta}_j \sim N(\beta_j, \sigma^2 d_{jj}) \]

for \( j = 1, 2, ..., k \) and \( d_{jj} = [(X'X)^{-1}]_{jj} \). We may now apply the standard normal transformation to obtain

\[ \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 d_{jj}}} \sim N(0, 1). \tag{9.6} \]

which was the basis for testing hypotheses on single coefficients.

The problem with using the ratio in the previous paragraph for inference is that \( \sigma^2 \) is unknown. An estimator is provided by

\[ s^2 = \frac{e'e}{n-k}. \tag{9.7} \]
which was shown to be unbiased. Moreover, under the normality assumption (vi), we found that
\[
(n - k) \frac{s^2}{\sigma^2} \sim \chi^2_{n-k}
\]
and is stochastically independent of \(\hat{\beta}\). Combining this result with the infeasible standard normal ratio yielded the feasible statistic
\[
\frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2((X'X)^{-1})_{jj}}} \sim t_{n-k}
\]
This ratio was the basis for conducting hypothesis testing, confidence intervals, and prob-values in the previous chapter.

9.2 Prediction Properties

Suppose that we wish to predict the value of the dependent variable, possibly outside the sample, given only the values of the independent or explanatory variables. Let the subscript * denote the prediction period, then the value of the dependent variable generated by the model is
\[
y_* = \beta_1 x_{*1} + \beta_2 x_{*2} + \cdots + \beta_k x_{*k} + u_*
\]
where \(x_*' = (x_{*1}, x_{*2}, \cdots, x_{*k})\). Note that the objective of our inferences \(y_*\), which we label *predictand*, is now a random variable. This is fundamentally different from the estimation problem where our objective is a real number. Since the disturbance term \(u_*\) is outside the sample period and has yet to be drawn, it is *a priori* unknowable. Consistent with previous chapters, we assume that \(\mathbf{x}_*\) is nonstochastic and known and that \(u_*\) has zero mean, variance \(\sigma^2\), and is uncorrelated with the sample disturbances \(\mathbf{u}\).

We consider prediction strategies or *predictors*, denoted \(\hat{y}_*\), that are based on the sample information and given values of the independent variables. The performance of the predictor is based on the properties of the prediction error
\[
y_* - \hat{y}_* = u_* + x_*' \beta - \hat{\beta}_*
\]
which is the difference between the predictor and predictand. The expectation of this error is the prediction bias
\[
B_* = E[y_* - \hat{y}_*] = x_*' \beta - E[\hat{\beta}_*]
\]
and the expectation of its square is the *mean squared prediction error* (MSPE)
\[
M_* = E[(y_* - \hat{y}_*)^2].
\]
Rather obviously, we prefer predictors which yield zero values of \(B_*\) and small values of \(M_*\).
Some additional structure can be given to the mean squared prediction error. Define $\mu_\ast = x_\ast' \beta$, then we find

$$M_\ast = \mathbb{E}[(y_\ast - \mu_\ast) + (\mu_\ast - \hat{y}_\ast)^2] = \mathbb{E}[y_\ast^2 + 2u_\ast^2 + 2u_\ast(\mu_\ast - \hat{y}_\ast) + (\mu_\ast - \hat{y}_\ast)^2] = \sigma^2 + \mathbb{E}[(x_\ast' \beta - \hat{y}_\ast)^2]$$

since $\hat{y}_\ast$ is based on sample information and hence uncorrelated with $u_\ast$. We see that there are two components to the mean square prediction error. The first, $\sigma^2$, arises because it is a prediction problem and $u_\ast$ has yet to be drawn and is unknowable. The second depends on the precision with which we estimate $x_\ast' \beta$, the expectation of the target given the explanatory variables $x_\ast$. This naturally suggests $\hat{y}_\ast = x_\ast' \hat{\beta}$ as a predictor, whose behavior we consider in the next two sections.

### 9.3 Bivariate Least Squares Prediction

We now consider the prediction problem in the bivariate regression model. Although all the results will be special cases of the more general results in the next section, particular insight is gained by examining this simpler problem first. From Chapter 6, the model is

$$y_i = \alpha + \beta x_i + u_i \quad i = 1, 2, \ldots, n$$

so the corresponding predictand is

$$y_\ast = \alpha + \beta x_\ast + u_\ast.$$

From the previous section, the MSPE is minimized by the infeasible predictor $\hat{y}_\ast = \alpha + \beta x_\ast$. Since $\alpha$ and $\beta$ are unknown we use the least squares estimators to obtain

$$\hat{y}_\ast = \hat{\alpha} + \hat{\beta} x_\ast$$

as a feasible predictor. Note that this predictor is a single point on the real line and is hence termed a point predictor.

The prediction error for this simpler problem is given by

$$y_\ast - \hat{y}_\ast = (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) x_\ast + u_\ast$$

Clearly since $u_\ast$ has mean zero and the coefficient estimators are both unbiased

$$B_\ast = \mathbb{E}[y_\ast - \hat{y}_\ast] = 0$$

and the least squares predictor is unbiased. This is a measure of the location of the predictor relative to the unrealized predictand. On average, we have the correct level.
9.3. BIVARIATE LEAST SQUARES PREDICTION

We are much more interested in the precision of the predictor, which is measured by the MSPE. Specifically, we examine

\[
M_* = E[(y_* - \widehat{y}_*)^2]
\]

\[
= E[((\alpha - \widehat{\alpha}) + (\beta - \widehat{\beta})x_* + u_*)^2]
\]

\[
= E[(\alpha - \widehat{\alpha})^2] + 2E[(\alpha - \widehat{\alpha})(\beta - \widehat{\beta})x_*] + E[(\beta - \widehat{\beta})^2x_*^2] + E[u_*^2].
\]

Using the results for variances and covariances from Chapter 6, we find

\[
M_* = \sigma^2 \left( \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} - 2\sigma^2 \frac{\sum_{i=1}^{n} x_i x_*}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + \sigma^2 \left( \frac{x_*^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + \sigma^2
\]

\[
= \sigma^2 \left[ 1 + \frac{1}{n} \sum_{i=1}^{n} (x_i - x_*)^2 \right]
\]

\[
= \sigma^2 \left[ 1 + \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]
\]

\[
= \sigma^2 \left[ 1 + \frac{1}{n} \left( \sum_{i=1}^{n} (x_i - \bar{x}) + (\bar{x} - x_*) \right)^2 \right]
\]

\[
= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(\bar{x} - x_*)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]
\]

\[
= \sigma^2 m_*
\]

where \(m_* = 1 + \frac{1}{n} + (\bar{x} - x_*)^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2 = M_*/\sigma^2\). Notice that this measure is quadratic in \((\bar{x} - x_*)\) and achieves a minimum at \(x_* = \bar{x}\). This means that the predictors performance will degrade if we attempt to predict in situations where the conditioning variables are far from their historical averages. Also notice that the second and third terms will become very small as \(n\) grows large. So the impact of using estimated coefficients becomes negligible in large samples.

The basic limitation of the point predictor \(\widehat{y}_*\) is that it is a single point on the real line. The probability of the value predicted being realized is zero. Accordingly, we will move to interval predictors which have some probability of including he yet to be drawn realization. This requires a distributional assumption, so we add Assumption (vi) or normality to the properties of the \(u_i\) and \(u_*\). Since the prediction error is linear \((\widehat{\alpha}, \widehat{\beta})\) and hence \(u_i\) and also \(u_*\) it will also be normal. Specifically,

\[
(y_* - \widehat{y}_*) \sim N(0, \sigma^2 m_*)
\]

and

\[
\frac{(y_* - \widehat{y}_*)}{\sqrt{\sigma^2 m_*}} \sim N(0, 1).
\]

From the previous chapter we know \((n - 2)s^2/\sigma^2 \sim \chi^2_{n-2}\) and independent of \((\widehat{\alpha}, \widehat{\beta})\) and now \(u_*\), so

\[
z_* = \frac{(y_* - \widehat{y}_*)/\sqrt{\sigma^2 m_*}}{\sqrt{((n - 2)s^2/\sigma^2)/(n - 2)}} = \frac{(y_* - \widehat{y}_*)}{\sqrt{s^2 m_*}} \sim t_{n-2}.
\]
This result may now be utilized to form a prediction interval. Choosing a significance level, say \( \alpha = 0.05 \), and a corresponding critical value from the \( t \)-tables for \( n - 2 \) degrees of freedom, say \( t_{0.05} \), we know that

\[
0.95 = \Pr[t_{0.05} \leq \frac{(y_\star - \hat{y}_\star)}{\sqrt{s^2 m_*}} \leq t_{0.05}]
= \Pr \left[ \hat{y}_\star - t_{0.025} \sqrt{s^2 m_*} \leq y_\star \leq \hat{y}_\star + t_{0.025} \sqrt{s^2 m_*} \right].
\]

Thus \( \hat{y}_\star \pm t_{0.025} \sqrt{s^2 m_*} \) forms a 95% prediction interval for the yet to be realized \( y_\star \). Notice that like the confidence interval for a coefficient estimate the interval is a random variable with both the center and width of the interval random. Unlike the confidence interval, the target is also a random variable. We have a 95% chance of our random interval including the random predictand. Since the width depends directly on \( m_* \), the size of the prediction interval will be smallest for \( x_\star = \bar{x} \).

### 9.4 Multivariate Least Squares Prediction

We now consider the more general problem where the predictor is

\[
\hat{y}_\star = x'_\star \hat{\beta} = x'_\star (X'X)^{-1} X'y = x'_\star (X'X)^{-1} X'(X/\beta + u) = x_\star \beta + x'_\star (X'X)^{-1} X'u. \tag{9.10}
\]

Note that this predictor is linear in \( y \) and hence \( u \). Now, it is easy to see that

\[
E(\hat{y}_\star | x_\star) = x'_\star \beta, \tag{9.11}
\]

so

\[
E[(y_\star - \hat{y}_\star) | x_\star] = B_* = 0. \tag{9.12}
\]

whereupon \( \hat{y}_\star \) is an unbiased predictor of \( y_\star \).

In terms of expected squared error, we find that the variance or expected squared error of the predictor around its mean is

\[
\text{Var}(\hat{y}_\star) = E(\hat{y}_\star - x'_\star \beta)^2 = \sigma^2 x'_\star (X'X)^{-1} x_\star, \tag{9.13}
\]

and that the mean squared prediction error (MSPE) or expected squared error of the predictor around its stochastic target \( y_\star \) is

\[
M_* = E[(y_\star - \hat{y}_\star)^2] = \sigma^2 [1 + x'_\star (X'X)^{-1} x_\star] = \sigma^2 m_* . \tag{9.14}
\]

Note that the typical diagonal element of \( X'X \) is a sum of squares and grows large and hence \( (X'X)^{-1} \) grows small with the sample size \( n \). Thus, in large samples, \( \sigma^2 \), which reflects the variation introduced by the a priori unknowable
9.4. MULTIVARIATE LEAST SQUARES PREDICTION

\( u_s \), dominates the MSPE and the variability arising from the estimation of \( \beta \) becomes relatively negligible.

Notice that the MSPE is quadratic in \( x_s \). As with the bivariate model it is of interest what values of \( x_s \) yield smaller values of \( M_s \). Specifically, we seek to

\[
\min_{x_s} x_s' (X'X)^{-1} x_s
\]

which, following some math, yields the same solution as the bivariate case, namely

\[ x_s = \bar{x}. \]

Thus we see that, in general, the predictions perform more accurately when the conditioning information is typical of the sample on which the estimates/predictions are based. And the imprecision of the predictor goes up exponentially as we depart from the mean of the conditioning variables.

It is this precision and its dependence on the conditioning variables which leads us to entertain interval rather than point predictors. If we assume, following Assumption (vi), that \( u_i \sim i.i.d. N(0, \sigma^2) \), including for the prediction observation, then

\[
(y_s - \hat{y}_s) = x_s' (\beta - \hat{\beta}) + u_s
\]

\[
= -x_s' (X'X)^{-1} X' u + u_s
\]

\[
\sim N(0, \sigma^2 m_s)
\]

so

\[
\frac{(y_s - \hat{y}_s)}{\sqrt{\sigma^2 m_s}} \sim N(0, 1).
\]

And for general \( k \)-variate case

\[
(n - k)s^2 / \sigma^2 \sim \chi^2_{n-k}
\]

independent of \( \hat{\beta}, u_s, \) and \( (y_s - \hat{y}_s) \), whereupon

\[
\frac{(y_s - \hat{y}_s)}{\sqrt{s^2 m_s}} \sim t_{n-k}.
\]

Proceeding as above, we find the 95% interval

\[
0.95 = \Pr \left[ \hat{y}_s - t_{0.025} \sqrt{s^2 m_s} \leq y_s \leq \hat{y}_s + t_{0.025} \sqrt{s^2 m_s} \right]
\]

where \( t_{0.025} \) is a critical value taken from the \( t \)-tables with \( n - k \) degrees of freedom.

The predictor \( \hat{y}_s \) and its standard error \( \sqrt{s^2 m_s} \) can be easily calculated from standard least squares statistics produced by an auxiliary regression. Let \( x_i' = (1, x_{i2}', \ldots, x_{ik}') \) denote the conditioning values of the explanatory variables, \( \beta' = \)
\((\beta_1, \beta_2')\), and recall \(\mu_* = x_*'\beta\) as the target of our predictor. Then form the regression

\[
y_i - \mu_* = x_i'\beta + u_i - x_*'\beta = (x_i - x_*)'\beta + u_i = (x_{2i} - x_2)'/\beta_2 + u_i
\]

since the first element of \(x_i\) and \(x_*\) are both unity. Thus an equivalent regression is

\[
y_i = \mu_* + (x_{2i} - x_2)'/\beta_2 + u_i = (1, (x_{2i} - x_2)')\begin{pmatrix} \mu_* \\ \beta_2 \end{pmatrix} + u_i.
\]

Performing least squares on this regression yields \(\hat{y}_* = \hat{\mu}_*\) as the intercept estimator and the usual standard error estimator for the intercept will yield \(\sqrt{s^2/m_*}\). These may be used to find an interval predictor as discussed in the last paragraph.

If we restrict our attention to unbiased predictors, which have \(E(\hat{y}_*|x_*) = x_*'\beta\), then the least square predictor has efficiency properties. Consider any predictor, say \(\tilde{y}\), that is unbiased then for any \(x_2\)

\[
E(\tilde{y}|x_2) = x_*'\tilde{\beta}
\]

which implies that

\[
\tilde{y}_* = x_*'\tilde{\beta}
\]

for some unbiased estimator \(\tilde{\beta}\). If \(\tilde{y}_*\) is linear in \(y\) then so is \(\tilde{\beta}\) and we find that \(\tilde{y}_*\) has smaller variance as a consequence of the Gauss-Markov theorem and in fact is the best (minimum variance) linear unbiased predictor (BLUP) of \(y_*\). Note that this efficiency applies whether or not the disturbances are normally distributed. When we add the normality assumption this result strengthens and \(\tilde{y}_*\) is the best unbiased predictor (BUP).

### 9.5 An Example

We will now apply the results obtained in this chapter to the extended example presented in the previous chapter. We will generate a series of point and interval predictions for both the full sample of 527 observations and the first 25 observations. This will enable us to see the impact of more observations on the prediction intervals. We will also generate predictions for both the simplest regression considered and the most complicated. This will enable us to see the effect adding regressors on the predictions.

First we look at predictions for the simplest regression, which was a bivariate regression of the log-wage on education. The regression results for a sample size of 25 are:

\[
\ln(w_i) = 0.959 + 0.095 E_i + e_i
\]
with $s^2 = 0.268$ and $R^2 = 0.156$. A plot of the in-sample predictions and a 50 percent prediction interval yields Figure 9.1.

![Figure 9.1: Bivariate Regression, $n = 25$](image)

Note the quadratic nature of the prediction interval and how it is smallest at the mean value of the independent variable.

The complete regression over the smaller sample size yields

$$
\ln(w_i) = 0.705 + 0.109 E_i + 0.0157 X_i - 0.00018 X_i^2
$$

$$
- 0.169 B_i - 0.068 H_i - 0.408 G_i + e_i
$$

with $s^2 = 0.278$ and $R^2 = 0.313$. Note that the $R^2$ is much larger, which reflects the added regressors. For prediction purposes, we set the explanatory variables other than education to their means and generate point and interval predictions for education levels 1 to 25. A plot for the in-sample predictions for this case is presented in Figure 9.2. Note how the confidence intervals for this case are smaller, which reflects the smaller $R^2$.

We now turn to the predictions for the full sample. The regression results for the bivariate regression are (from last chapter)

$$
\ln(w_i) = 0.985 + 0.0822 E_i + e_i
$$

with $s^2 = 0.225$ and $R^2 = 0.157$. Note that both $s^2$ and $R^2$ are roughly the same as for the smaller sample size, which is to be expected. A plot of the point
and 50 percent interval predictions for this case yields Figure 9.3. Note how the upper and lower components of the interval are low almost linear. This reflects the fact that the variance of the predictor \( \sigma^2 m_\ast = \sigma^2 [1 + x_\ast' (X'X)^{-1} x_\ast] \approx \sigma^2 \) since the second term is \( O(1/n) \). The estimation error disappears from the prediction error in large samples.

Finally, we look at predictions for the complete regression over the full sample. The regression results are (from last chapter)

\[
\ln(w_i) = 0.539 + 0.0927E_i + 0.0446X_i - 0.00074X_i^2 - 0.101B_i - 0.094H_i - 0.268G_i + e_i
\]

with \( s^2 = 0.176 \) and \( R^2 = 0.348 \). Again, both \( s^2 \) and \( R^2 \) are roughly the same as for the smaller sample size. The differences in the size of \( s^2 \) are probably due to sampling error. A plot of the point and interval predictions for this case gives Figure 9.4. The notable differences from the previous plots are the smaller intervals and the almost linear upper and lower bound lines for the intervals. These reflect the impact of additional variables and additional observations respectively. Both of these phenomena are predicted by the theory.
9.5. AN EXAMPLE

Figure 9.3: Bivariate Regression, \( n = 527 \)

Figure 9.4: Complete Regression, \( n = 527 \)