Ch6. Multiple Regression: Estimation

1 The model

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \cdots, n. \] \hspace{1cm} (1)

The assumptions for \( \epsilon_i \) and \( y_i \) are analogous to those for simple linear regression, namely

1. \( E(\epsilon_i) = 0 \) for all \( i = 1, 2, \cdots, n \), or, equivalently, \( E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \).
2. \( \text{var}(\epsilon_i) = \sigma^2 \) for all \( i = 1, 2, \cdots, n \), or, equivalently, \( \text{var}(y_i) = \sigma^2 \).
3. \( \text{cov}(\epsilon_i, \epsilon_j) = 0 \) for all \( i \neq j \), or, equivalently, \( \text{cov}(y_i, y_j) = 0 \).

In matrix form, the model can be written as

\[
\begin{pmatrix}
  y_1 \\ y_2 \\ \vdots \\ y_n
\end{pmatrix} =
\begin{pmatrix}
  1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk}
\end{pmatrix}
\begin{pmatrix}
  \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k
\end{pmatrix} +
\begin{pmatrix}
  \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n
\end{pmatrix},
\]

or

\[ y = X\beta + \epsilon. \]

The assumptions on \( \epsilon \) or \( y \) can be expressed as

1. \( E(\epsilon) = 0 \) or \( E(y) = X\beta \).
2. \( \text{cov}(\epsilon) = \sigma^2 I \) or \( \text{cov}(y) = \sigma^2 I \).

The matrix \( X \) is \( n \times (k+1) \) and is called the design matrix. In this chapter, we assume that \( n > k + 1 \) and \( \text{rank}(X) = k + 1 \).

2 Estimation of \( \beta \) and \( \sigma^2 \)

2.1 Least squares estimator for \( \beta \)

The least squares approach is to seek the estimators of \( \beta \) which minimize the sums of squares of deviations of the \( n \) observed \( y \)'s from their predicted values \( \hat{y} \), i.e., minimize

\[ \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \]
Theorem 2.1 If \( y = X\beta + \epsilon \), where \( X \) is \( n \times (k+1) \) of rank \( k+1 < n \), then the least squares estimator of \( \beta \) is
\[
\hat{\beta} = (X'X)^{-1}X'y.
\]
Proof: Exercise.

2.2 Properties of the least squares estimator \( \hat{\beta} \)

Theorem 2.2 If \( E(y) = X\beta \), then \( \hat{\beta} \) is an unbiased estimator for \( \beta \).

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

Theorem 2.3 If \( \text{cov}(y) = \sigma^2 I \), the covariance matrix for \( \beta \) is given by \( \sigma^2(X'X)^{-1} \).

Proof: Exercise.

Theorem 2.4 (Gauss-Markov Theorem) If \( E(y) = X\beta \) and \( \text{cov}(y) = \sigma^2 I \), the least squares estimators \( \hat{\beta}_j \), \( j = 0, 1, \cdots, k \), have minimum variance among all linear unbiased estimators, i.e., the least squares estimators \( \hat{\beta}_j \), \( j = 0, 1, \cdots, k \) are best linear unbiased estimators (BLUE).

Proof: We consider a linear estimator \( Ay \) of \( \beta \) and seek the matrix \( A \) for which \( Ay \) is a minimum variance unbiased estimator of \( \beta \). Since \( Ay \) is to be unbiased for \( \beta \), we have
\[
E(Ay) = AE(y) = AX\beta = \beta,
\]
which gives the unbiasedness condition
\[
AX = I
\]
since the relationship \( AX\beta = \beta \) must hold for any possible value of \( \beta \).

The covariance matrix for \( Ay \) is
\[
\text{cov}(Ay) = A(\sigma^2 I)A' = \sigma^2 AA'.
\]
The variance of the \( \beta_j \)'s are on the diagonal of \( \sigma^2 AA' \), and therefore we need to choose \( A \) (subject to \( AX = I \)) so that the diagonal elements of \( AA' \) are minimized. Since
\[
AA' = [A - (X'X)^{-1}X'] [A - (X'X)^{-1}X']' \iddots [A - (X'X)^{-1}X'] [A - (X'X)^{-1}X']' + (X'X)^{-1}
\]

2
Note in the last equality, $AX = I$ is used. Since $[A - (X'X)^{-1}X']^T[A - (X'X)^{-1}X']'$ is positive semidefinite, the diagonal elements are great than or equal to zero. These diagonal elements can be made equal to zero by choosing $A = (X'X)^{-1}X'$. (This value of $A$ also satisfies the unbiasedness condition $AX = I$). The resulting minimum variance estimator of $\beta$ is

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

which is equal to the least square estimator $\hat{\beta}$.

**Remark:** The remarkable feature of the Gauss-Markov theorem is its distributional generality. The result holds for any distribution of $y$; normality is not required. The only assumptions used in the proof are $E(y) = X\beta$ and $\text{cov}(y) = \sigma^2I$. If these assumptions do not hold, $\hat{\beta}$ may be biased or each $\hat{\beta}_j$ may have a larger variance than that of some other estimator.

**Corollary 2.1** If $E(y) = X\beta$ and $\text{cov}(y) = \sigma^2I$, the best linear unbiased estimator of $a'\beta$ is $a'\hat{\beta}$, where $\beta = (X'X)^{-1}X'y$.

A fourth property of $\hat{\beta}$ is the following: the predicted value $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \cdots + \hat{\beta}_kx_k = \hat{\beta}'x$ is invariant to simple linear changes of scale on the $x$'s, where $x = (1, x_1, x_2, \cdots, x_k)'$.

**Theorem 2.5** If $x = (1, x_1, \cdots, x_k)'$ and $z = (1, c_1x_1, \cdots, c_kx_k)$, then $\hat{y} = \hat{\beta}'x = \hat{\beta}'z$, where $\hat{\beta}'z$ is the least squares estimator from the regression of $y$ on $z$.

**Proof:** We can write $Z = XD$, where $D = \text{diag}(1, c_1, c_2, \cdots, c_k)$. Substituting $Z = XD$ into $\hat{\beta}_z$, we have

$$\hat{\beta}_z = [((XD)'(XD))^{-1}(XD)'y = D^{-1}(X'X)^{-1}X'y = D^{-1}\hat{\beta}.$$ 

Hence,

$$\hat{\beta}_z'z = (D^{-1}\hat{\beta})'Dx = \hat{\beta}'x.$$ 

### 2.3 An estimator for $\sigma^2$

By assumption 1, $E(y_i) = x_i'\beta$, and by assumption 2, $\sigma^2 = E[y_i - E(y_i)]^2$, we have

$$\sigma^2 = E(y_i - x_i'\beta)^2.$$ 

Hence, $\sigma^2$ can be estimated by

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

$$= \frac{1}{n - k - 1} (y - X\hat{\beta})'(y - X\hat{\beta})$$

$$= \frac{\text{SSE}}{n - k - 1}.$$ 

3
With the denominator \( n - k - 1 \), \( s^2 \) is an unbiased estimator of \( \sigma^2 \).

**Theorem 2.6** If \( E(y) = X\beta \) and \( \text{cov}(y) = \sigma^2 I \), then
\[
E(s^2) = \sigma^2.
\]

**Proof:** Exercise.

**Corollary 2.2** An unbiased estimator of \( \text{cov}(\hat{\beta}) \) is given by
\[
\hat{\text{cov}}(\hat{\beta}) = s^2(X'X)^{-1}.
\]

**Theorem 2.7** If \( E(y) = X\beta \), \( \text{cov}(y) = \sigma^2 I \), and \( E(\epsilon^4) = 3\sigma^4 \) for the linear model \( y = X\beta + \epsilon \), then \( s^2 \) is the best (minimum variance) quadratic unbiased estimator of \( \sigma^2 \).


### 3 The model in centered form

In matrix form, the centered model for the linear multiple regression becomes
\[
y = (j, X_c) \begin{pmatrix} \alpha \\ \beta_1 \end{pmatrix} + \epsilon,
\]
where \( j \) is a vector of 1’s, \( \beta_1 = (\beta_1, \beta_2, \cdots, \beta_k)' \),
\[
X_c = (I - \frac{1}{n}J)X_1 = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix}.
\]

The matrix \( I - \frac{1}{n}J \) is sometimes called the centering matrix.

The corresponding least squares estimator becomes
\[
\begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{pmatrix} = [(j, X_c)'(j, X_c)]^{-1}(j, X_c)'y
\]
\[
= \begin{pmatrix} n & 0 \\ 0 & X_c'X_c \end{pmatrix}^{-1} \begin{pmatrix} n\bar{y} \\ X_c'y \end{pmatrix}
\]
\[
= \begin{pmatrix} \bar{y} \\ (X_c'X_c)^{-1}X_c'y \end{pmatrix}.
\]
4 Normal model

4.1 Assumptions

Normality assumption:
\[ y \sim N_n(X\beta, \sigma^2 I) \text{ or } \epsilon \sim N_n(0, \sigma^2 I). \]

Under normality, \( \text{cov}(y) = \text{cov}(\epsilon) = \sigma^2 I \) implies that the \( y \)'s are independent as well as uncorrelated.

4.2 Maximum likelihood estimators for \( \beta \) and \( \sigma^2 \)

**Theorem 4.1** If \( y \sim N_n(X\beta, \sigma^2 I) \), where \( X \) is \( n \times (k+1) \) of rank \( k+1 \leq n \), the maximum likelihood estimators of \( \beta \) and \( \sigma^2 \) are

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

**Proof:** Exercise.

The maximum likelihood estimator \( \hat{\beta} \) is the same as the least squares estimator \( \hat{\beta} \). The estimator \( \hat{\sigma}^2 \) is biased since the denominator is \( n \) rather than \( n - k - 1 \). We often use the unbiased estimator \( s^2 \) to estimate \( \sigma^2 \).

4.3 Properties of \( \hat{\beta} \) and \( \hat{\sigma}^2 \)

**Theorem 4.2** Suppose \( y \sim N_n(X\beta, \sigma^2 I) \), where \( X \) is \( n \times (k+1) \) of rank \( k+1 < n \) and \( \beta = (\beta_0, \cdots, \beta_k)' \). Then the maximum likelihood estimators \( \hat{\beta} \) and \( \hat{\sigma}^2 \) have the following distributional properties:

(i) \( \hat{\beta} \sim N_{k+1}(\beta, \sigma^2 (X'X)^{-1}) \).

(ii) \( n\hat{\sigma}^2/\sigma^2 \) is distributed as \( \chi^2(n - k - 1) \), or equivalently, \( (n - k - 1)s^2/\sigma^2 \) is distributed as \( \chi^2(n - k - 1) \).

(iii) \( \hat{\beta} \) and \( \hat{\sigma}^2 \) (or \( s^2 \)) are independent.

**Proof:** (i) Since \( y \) is normal, \( \hat{\beta} = (X'X)^{-1}X'y \) is a linear function of \( y \), \( E(\hat{\beta}) = \beta \) and \( \text{cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1} \), \( \hat{\beta} \sim N_{k+1}(\beta, \sigma^2 (X'X)^{-1}) \).

(ii) \( n\hat{\sigma}^2/\sigma^2 = \frac{y'}{\sigma} (I - X(X'X)X') \frac{y}{\sigma} \)

and \( I - X(X'X)X' \) is idempotent, hence \( n\hat{\sigma}^2/\sigma^2 \) is distributed as \( \chi^2(n - k - 1) \).

(iii) Since \( \hat{\beta} = (X'X)^{-1}X'y \) and \( \hat{\sigma}^2 = y'(I - X(X'X)X')y \),

and that \( (X'X)^{-1}(I - X(X'X)X') = \mathbf{O} \), we have \( \hat{\beta} \) and \( \hat{\sigma}^2 \) (or \( s^2 \)) are independent.
Theorem 4.3 If \( y \) is \( N_n(\mathbf{X}\beta, \sigma^2 I) \), then \( \hat{\beta} \) and \( \hat{\sigma}^2 \) are jointly sufficient for \( \beta \) and \( \sigma^2 \).

PROOF: Using the Neyman factorization theorem. For details, see Rencher (2000, p.144).

Since \( \hat{\beta} \) and \( \hat{\sigma}^2 \) are jointly sufficient for \( \beta \) and \( \sigma^2 \), no other estimators can improve on the information they extract from the sample to estimate \( \beta \) and \( \sigma^2 \). Thus, it is not surprising that \( \hat{\beta} \) and \( \hat{s}^2 \) are minimum variance unbiased estimators.

Theorem 4.4 If \( y \) is \( N_n(\mathbf{X}\beta, \sigma^2 I) \), then \( \hat{\beta} \) and \( s^2 \) have minimum variance among all unbiased estimators.


5 \( R^2 \) in fixed-\( x \) regression

The proportion of the total sum of squares due to regression is measured by

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

where \( SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \), \( SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}' \mathbf{X}'y - n\bar{y}^2 \), and

\[
SST = SSR + SSE,
\]

where \( SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \).

The \( R^2 \) is called the coefficient of determination or the squared multiple correlation. The positive square root \( R \) is called the multiple correlation coefficient. If the \( x \)'s were random, \( R \) would estimate a population multiple correlation.

We list some properties of \( R^2 \) and \( R \).

1. The range of \( R^2 \) is \( 0 \leq R^2 \leq 1 \). If all the \( \hat{\beta}_j \)'s were zero, except for \( \hat{\beta}_0 \), \( R^2 \) would be zero. (This event has probability zero for continuous data.) If all the \( y \)-values fell on the fitted surface, that is, if \( y_i = \hat{y}_i \), \( i = 1, 2, \cdots, n \), then \( R^2 \) would be 1.

2. \( R = r_{yy} \); that is, the multiple correlation is equal to the simple correlation between the observed \( y_i \)'s and the fitted \( \hat{y}_i \)'s.

3. Adding a variable \( x \) to the model generally increases (but does not decrease) the value of \( R^2 \).

4. \( R^2 \) cannot be partitioned into \( k \) components, each of which is uniquely attributable to an \( x_j \), unless the \( x \)'s are mutually orthogonal, that is, unless \( \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m) = 0 \) for \( j \neq m \).
5. $R^2$ is invariant to full-rank linear transformations on the x’s and to a scale change on y (but not invariant to a joint linear transformation including y and the x’s).

Adjusted $R^2$

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

$R^2$ can also be expressed in terms of sample variances and covariances:

$$R^2 = \frac{\hat{\beta}'X'X\hat{\beta}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{s'yxS_x^{-1}(n-1)S_xS_y^{-1}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{s'yxS_x^{-1}s_y}{s_y^2}$$

Note that $\hat{\beta}_1 = (n-1)(X'X)^{-1}X'y = (\frac{X'X_n}{n-1})^{-1}X'y = S_x^{-1}s_y$. This form of $R^2$ will facilitate a comparison with $R^2$ for the random-x case.

Geometrically, $R$ is the cosine of the angle $\theta$ between y and $\hat{y}$ corrected for their means. The mean of $\hat{y}$ is $\bar{y}$, the same as the mean of y. Thus, the centered form of y and $\hat{y}$ are $y - \bar{y}$ and $\hat{y} - \bar{y}$.

$$\cos \theta = \frac{(y - \bar{y})(\hat{y} - \bar{y})}{\sqrt{[(y - \bar{y})(\hat{y} - \bar{y})][(y - \bar{y})(\hat{y} - \bar{y})]}}$$

Note that

$$(y - \bar{y})(\hat{y} - \bar{y}) = [(\hat{y} - \bar{y}) + (y - \hat{y})][(\hat{y} - \bar{y}) + (y - \hat{y})]$$

$$= (\hat{y} - \bar{y})(\hat{y} - \bar{y}) + (y - \hat{y})(\hat{y} - \bar{y})$$

$$= (\hat{y} - \bar{y})(\hat{y} - \bar{y}) + 0.$$  

Hence,

$$\cos \theta = \frac{\sqrt{(y - \bar{y}\hat{y} - \bar{y})^2}}{(y - \bar{y})(\hat{y} - \bar{y})} = \sqrt{\frac{SSR}{SST}} = R.$$  

6 Generalized least squares: $cov(y) = \sigma^2V$

The model

$$y = X\beta + \epsilon, \quad E(y) = X\beta, \quad cov(y) = \Sigma = \sigma^2V,$$

**Theorem 6.1** Let $y = X\beta + \epsilon, E(y) = X\beta,$ and $cov(y) = cov(\epsilon) = \sigma^2V,$ where $X$ is a full-rank matrix and $V$ is a known positive definite matrix. For this model, we obtain the following results:

(i) The best linear unbiased estimator (BLUE) of $\beta$ is

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y.$$
(ii) The covariance matrix for $\hat{\beta}$ is

$$\text{cov}(\beta) = \sigma^2 (X'V^{-1}X)^{-1}.$$

(iii) An unbiased estimator of $\sigma^2$ is

$$s^2 = \frac{(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta})}{n - k - 1}.$$

**Proof:** (i) Since $V$ is positive definite, there exists an $n \times n$ nonsingular matrix $P$ such that $V = PP'$. Multiplying $y = X\beta + \epsilon$ by $P^{-1}$, we obtain

$$P^{-1}y = P^{-1}X\beta + P^{-1}\epsilon.$$

Applying the least square approach to this transformed model, we get

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

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

Note that since $X$ is full rank, $X'V^{-1}X$ is positive definite. The estimator $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$ is usually called the **generalized least squares** estimator.

(ii) and (iii) are left as exercises.

**Theorem 6.2** If $y$ is $N_n(X\beta, \sigma^2V)$, where $X$ is $n \times (k + 1)$ of rank $k + 1$ and $V$ is a known positive definite matrix, then the maximum likelihood estimators for $\beta$ and $\sigma^2$ are

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

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}).$$

**Proof:** Exercise.