Chapter 3

Classical Least Squares Theory

In the field of economics, numerous hypotheses and theories have been proposed in order to describe the behavior of economic agents and the relationships between economic variables. Although these propositions may be theoretically appealing and logically correct, they may not be practically relevant unless they are supported by real world data. A theory with supporting empirical evidence is of course more convincing. Therefore, empirical analysis has become an indispensable ingredient of contemporary economic research. By *econometrics* we mean the collection of statistical and mathematical methods that utilize data to analyze the relationships between economic variables.

A leading approach in econometrics is the *regression analysis*. For this analysis one must first specify a regression model that characterizes the relationship of economic variables; the simplest and most commonly used specification is the linear model. The linear regression analysis then involves estimating unknown parameters of this specification, testing various economic and econometric hypotheses, and drawing inferences from the testing results. This chapter is concerned with one of the most important estimation methods in linear regression, namely, the method of *ordinary least squares* (OLS). We will analyze the OLS estimators of parameters and their properties. Testing methods based on the OLS estimation results will also be presented. We will not discuss asymptotic properties of the OLS estimators until Chapter 6. Readers can find related topics in other econometrics textbooks, e.g., Davidson and MacKinnon (1993), Goldberger (1991), Greene (2000), Harvey (1990), Intriligator et al. (1996), Johnston (1984), Judge et al. (1988), Maddala (1992), Ruud (2000), and Theil (1971), among many others.

3.1 The Method of Ordinary Least Squares

Suppose that there is a variable, y, whose behavior over time (or across individual units) is of interest to us. A theory may suggest that the behavior of y can be well characterized by some function f of the variables $x_1, \ldots, x_k$. Then, $f(x_1, \ldots, x_k)$ may be viewed as a "systematic" component of y provided that no other variables can further account for the behavior of the residual $y - f(x_1, \ldots, x_k)$. In the context of linear regression, the function f is specified as a linear function. The unknown linear weights (parameters) of the linear specification can then be determined using the OLS method.

3.1.1 Simple Linear Regression

In simple linear regression, only one variable x is designated to describe the behavior of the variable y. The linear specification is

$$\alpha + \beta x,$$

where α and β are unknown parameters. We can then write

$$y = \alpha + \beta x + e(\alpha, \beta),$$

where $e(\alpha, \beta) = y - \alpha - \beta x$ denotes the error resulted from this specification. Different parameter values result in different errors. In what follows, y will be referred to as the *dependent variable* (*regressand*) and x an *explanatory variable* (*regressor*). Note that both the regressand and regressor may be a function of some other variables. For example, when $x = z^2$,

$$y = \alpha + \beta z^2 + e(\alpha, \beta).$$

This specification is not linear in the variable z but is linear in x (hence linear in parameters). When $y = \log w$ and $x = \log z$, we have

$$\log w = \alpha + \beta(\log z) + e(\alpha, \beta),$$

which is still linear in parameters. Such specifications can all be analyzed in the context of linear regression.

Suppose that we have T observations of the variables y and x. Given the linear specification above, our objective is to find suitable α and β such that the resulting linear function "best" fits the data (y_t, x_t), $t = 1, \ldots, T$. Here, the generic subscript t is used for both cross-section and time-series data. The OLS method suggests to find a straight line

whose sum of squared errors is as small as possible. This amounts to finding α and β that minimize the following OLS criterion function:

$$Q(\alpha,\beta) := \frac{1}{T}\sum_{t=1}^{T} e_t(\alpha,\beta)^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2.$$

The solutions can be easily obtained by solving the first order conditions of this minimization problem.

The first order conditions are:

$$\frac{\partial Q(\alpha,\beta)}{\partial \alpha} = -\frac{2}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t) = 0,$$

$$\frac{\partial Q(\alpha,\beta)}{\partial \beta} = -\frac{2}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)x_t = 0.$$

Solving for α and β we have the following solutions:

$$\hat{\beta}_T = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2},$$

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

where $\bar{y} = \sum_{t=1}^{T} y_t/T$ and $\bar{x} = \sum_{t=1}^{T} x_t/T$. As $\hat{\alpha}_T$ and $\hat{\beta}_T$ are obtained by minimizing the OLS criterion function, they are known as the OLS estimators of α and β, respectively. The subscript T of $\hat{\alpha}_T$ and $\hat{\beta}_T$ signifies that these solutions are obtained from a sample of T observations. Note that if x_t is a constant c for every t, then $\bar{x} = c$, and hence $\hat{\beta}_T$ cannot be computed.

The function $\hat{y} = \hat{\alpha}_T + \hat{\beta}_T x$ is the estimated *regression line* with the intercept $\hat{\alpha}_T$ and slope $\hat{\beta}_T$. We also say that this line is obtained by regressing y on (the constant one and) the regressor x. The regression line so computed gives the "best" fit of data, in the sense that any other linear function of x would yield a larger sum of squared errors. For a given x_t, the OLS fitted value is a point on the regression line:

$$\hat{y}_t = \hat{\alpha}_T + \hat{\beta}_T x_t.$$

The difference between y_t and $\hat{y}_t$ is the t^{th} OLS *residual*:

$$\hat{e}_t := y_t - \hat{y}_t,$$

which corresponds to the error of the specification as

$$\hat{e}_t = e_t(\hat{\alpha}_T, \hat{\beta}_T).$$

Note that regressing y on x and regressing x on y lead to different regression lines in general, except when all (y_t, x_t) lie on the same line; see Exercise 3.9.

Remark: Different criterion functions would result in other estimators. For example, the so-called *least absolute deviation* estimator can be obtained by minimizing the average of the sum of absolute errors:

$$\frac{1}{T}\sum_{t=1}^{T}|y_t - \alpha - \beta x_t|,$$

which in turn determines a different regression line. We refer to Manski (1991) for a comprehensive discussion of this topic.

3.1.2 Multiple Linear Regression

More generally, we may specify a linear function with k explanatory variables to describe the behavior of y:

$$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

so that

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e(\beta_1, \ldots, \beta_k),$$

where $e(\beta_1, \ldots, \beta_k)$ again denotes the error of this specification. Given a sample of T observations, this specification can also be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}(\boldsymbol{\beta}), \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_1 \;\; \beta_2 \;\; \cdots \;\; \beta_k)'$ is the vector of unknown parameters, $\boldsymbol{y}$ and $\boldsymbol{X}$ contain all the observations of the dependent and explanatory variables, i.e.,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tk} \end{bmatrix},$$

where each column of $\boldsymbol{X}$ contains T observations of an explanatory variable, and $\boldsymbol{e}(\boldsymbol{\beta})$ is the vector of errors. It is typical to set the first explanatory variable as the constant one so that the first column of $\boldsymbol{X}$ is the $T \times 1$ vector of ones, $\boldsymbol{\ell}$. For convenience, we also write $\boldsymbol{e}(\boldsymbol{\beta})$ as $\boldsymbol{e}$ and its element $e_t(\boldsymbol{\beta})$ as e_t.

Our objective now is to find a k-dimensional regression hyperplane that "best" fits the data $(\boldsymbol{y}, \boldsymbol{X})$. In the light of Section 3.1.1, we would like to minimize, with respect to $\boldsymbol{\beta}$, the average of the sum of squared errors:

$$Q(\boldsymbol{\beta}) := \frac{1}{T}\boldsymbol{e}(\boldsymbol{\beta})'\boldsymbol{e}(\boldsymbol{\beta}) = \frac{1}{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{3.2}$$

This is a well-defined problem provided that the basic *identification requirement* below holds for the specification (3.1).

[ID-1] The $T \times k$ data matrix $\boldsymbol{X}$ is of full column rank k.

Under [ID-1], the number of regressors, k, must be no greater than the number of observations, T. This is so because if $k > T$, the rank of $\boldsymbol{X}$ must be less than or equal to T, and hence $\boldsymbol{X}$ cannot have full column rank. Moreover, [ID-1] requires that any linear specification does not contain any "redundant" regressor; that is, any column vector of $\boldsymbol{X}$ cannot be written as a linear combination of other column vectors. For example, $\boldsymbol{X}$ contains a column of ones and a column of x_t in simple linear regression. These two columns would be linearly dependent if $x_t = c$ for every t. Thus, [ID-1] requires that x_t in simple linear regression is not a constant.

The first order condition of the OLS minimization problem is

$$\nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} (\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta})/T = \boldsymbol{0}.$$

By the matrix differentiation results in Section 1.2, we have

$$\nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/T = \boldsymbol{0}.$$

Equivalently, we can write

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}. \tag{3.3}$$

These k equations, also known as the *normal equations*, contain exactly k unknowns. Given [ID-1], $\boldsymbol{X}$ is of full column rank so that $\boldsymbol{X}'\boldsymbol{X}$ is positive definite and hence invertible by Lemma 1.13. It follows that the unique solution to the first order condition is

$$\hat{\boldsymbol{\beta}}_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{3.4}$$

Moreover, the second order condition is also satisfied because

$$\nabla^2_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = 2(\boldsymbol{X}'\boldsymbol{X})/T$$

is a positive definite matrix under [ID-1]. Thus, $\hat{\boldsymbol{\beta}}_T$ is the unique minimizer of the OLS criterion function and hence known as the OLS estimator of $\boldsymbol{\beta}$. This result is formally stated below.

Theorem 3.1 *Given the specification* (3.1)*, suppose that* [ID-1] *holds. Then, the OLS estimator* $\hat{\boldsymbol{\beta}}_T$ *given by* (3.4) *uniquely minimizes the OLS criterion function* (3.2)*.*

If $\boldsymbol{X}$ is not of full column rank, its column vectors are linearly dependent and therefore satisfy an exact linear relationship. This is the problem of *exact multicollinearity*. In this case, $\boldsymbol{X}'\boldsymbol{X}$ is not invertible so that there exist infinitely many solutions to the normal equations $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$. As such, the OLS estimator $\hat{\boldsymbol{\beta}}_T$ cannot be uniquely determined. See Exercise 3.4 for a geometric interpretation of this result. Exact multicollinearity usually arises from inappropriate model specifications. For example, including both total income, total wage income, and total non-wage income as regressors results in exact multicollinearity because total income is, by definition, the sum of wage and non-wage income; see also Section 3.5.2 for another example. In what follows, the identification requirement for the linear specification (3.1) is always assumed.

Remarks:

1. Theorem 3.1 does not depend on the "true" relationship between $\boldsymbol{y}$ and $\boldsymbol{X}$. Thus, whether (3.1) agrees with true relationship between $\boldsymbol{y}$ and $\boldsymbol{X}$ is irrelevant to the existence and uniqueness of the OLS estimator.

2. It is easy to verify that the magnitudes of the coefficient estimates $\hat{\beta}_i$, $i = 1, \ldots, k$, are affected by the measurement units of dependent and explanatory variables; see Exercise 3.7. As such, a larger coefficient estimate does not necessarily imply that the associated regressor is more important in explaining the behavior of $\boldsymbol{y}$. In fact, the coefficient estimates are not directly comparable in general; cf. Exercise 3.5.

Once the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is obtained, we can plug it into the original linear specification and obtain the vector of OLS fitted values:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_T.$$

The vector of OLS residuals is then

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{e}(\hat{\boldsymbol{\beta}}_T).$$

From the normal equations (3.3) we can deduce the following algebraic results. First, the OLS residual vector must satisfy the normal equations:

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_T) = \boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0},$$

so that $\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}$. When $\boldsymbol{X}$ contains a column of constants (i.e., a column of $\boldsymbol{X}$ is proportional to $\boldsymbol{\ell}$, the vector of ones), $\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}$ implies

$$\boldsymbol{\ell}'\hat{\boldsymbol{e}} = \sum_{t=1}^{T} \hat{e}_t = 0.$$

That is, the sum of OLS residuals must be zero. Second,

$$\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = \hat{\boldsymbol{\beta}}_T'\boldsymbol{X}'\hat{\boldsymbol{e}} = 0.$$

These results are summarized below.

Theorem 3.2 *Given the specification* (3.1)*, suppose that* [ID-1] *holds. Then, the vector of OLS fitted values* $\hat{\boldsymbol{y}}$ *and the vector of OLS residuals* $\hat{\boldsymbol{e}}$ *have the following properties.*

(a) $\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}$*; in particular, if* $\boldsymbol{X}$ *contains a column of constants,* $\sum_{t=1}^{T} \hat{e}_t = 0$*.*

(b) $\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = 0$*.*

Note that when $\boldsymbol{\ell}'\hat{\boldsymbol{e}} = \boldsymbol{\ell}'(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0$, we have

$$\frac{1}{T}\sum_{t=1}^{T} y_t = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t.$$

That is, the sample average of the data y_t is the same as the sample average of the fitted values $\hat{y}_t$ when $\boldsymbol{X}$ contains a column of constants.

3.1.3 Geometric Interpretations

The OLS estimation result has nice geometric interpretations. These interpretations have nothing to do with the stochastic properties to be discussed in Section 3.2, and they are valid as long as the OLS estimator exists.

In what follows, we write $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ which is an orthogonal projection matrix that projects vectors onto span($\boldsymbol{X}$) by Lemma 1.14. The vector of OLS fitted values can be written as

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

Hence, $\hat{\boldsymbol{y}}$ is the orthogonal projection of $\boldsymbol{y}$ onto span($\boldsymbol{X}$). The OLS residual vector is

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y},$$

which is the orthogonal projection of $\boldsymbol{y}$ onto span($\boldsymbol{X}$)$^\perp$ and hence is orthogonal to $\hat{\boldsymbol{y}}$ and $\boldsymbol{X}$; cf. Theorem 3.2. Consequently, $\hat{\boldsymbol{y}}$ is the "best approximation" of $\boldsymbol{y}$, given the information

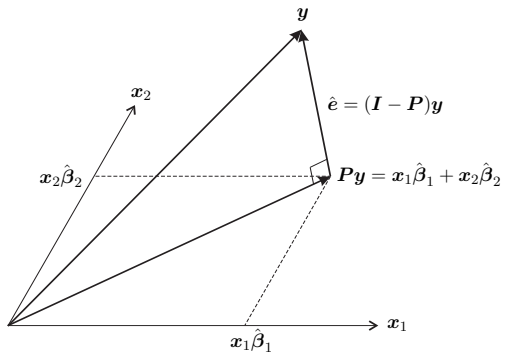

Figure 3.1: The orthogonal projection of $\boldsymbol{y}$ onto span($\boldsymbol{x}_1$,$\boldsymbol{x}_2$)

contained in $\boldsymbol{X}$, as shown in Lemma 1.10. Figure 3.1 illustrates a simple case where there are only two explanatory variables in the specification.

The following results are useful in many applications.

Theorem 3.3 (Frisch-Waugh-Lovell) *Given the specification*

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e},$$

where $\boldsymbol{X}_1$ is of full column rank k_1 and $\boldsymbol{X}_2$ is of full column rank k_2, let $\hat{\boldsymbol{\beta}}_T = (\hat{\boldsymbol{\beta}}_{1,T}' \; \hat{\boldsymbol{\beta}}_{2,T}')'$ denote the corresponding OLS estimators. Then,

$$\hat{\boldsymbol{\beta}}_{1,T} = [\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y},$$
$$\hat{\boldsymbol{\beta}}_{2,T} = [\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y},$$

where $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ and $\boldsymbol{P}_2 = \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$.

Proof: These results can be directly verified from (3.4) using the matrix inversion formula in Section 1.4. Alternatively, write

$$\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,T} + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{2,T} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y},$$

where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ with $\boldsymbol{X} = [\boldsymbol{X}_1 \; \boldsymbol{X}_2]$. Pre-multiplying both sides by $\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)$, we have

$$\begin{aligned}&\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y} \\ &\quad = \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,T} + \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{2,T} + \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.\end{aligned}$$

The second term on the right-hand side vanishes because $(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_2 = \boldsymbol{0}$. For the third term, we know span($\boldsymbol{X}_2$) $\subseteq$ span($\boldsymbol{X}$), so that span($\boldsymbol{X}$)$^\perp \subseteq$ span($\boldsymbol{X}_2$)$^\perp$. As each column

vector of $\boldsymbol{I}-\boldsymbol{P}$ is in $\text{span}(\boldsymbol{X})^{\perp}$, $\boldsymbol{I}-\boldsymbol{P}$ is not affected if it is projected onto $\text{span}(\boldsymbol{X}_2)^{\perp}$. That is,

$$(\boldsymbol{I}-\boldsymbol{P}_2)(\boldsymbol{I}-\boldsymbol{P})=\boldsymbol{I}-\boldsymbol{P}.$$

Similarly, $\boldsymbol{X}_1$ is in $\text{span}(\boldsymbol{X})$, and hence $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X}_1=\boldsymbol{0}$. It follows that

$$\boldsymbol{X}_1'(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{y}=\boldsymbol{X}_1'(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,T},$$

from which we obtain the expression for $\hat{\boldsymbol{\beta}}_{1,T}$. The proof for $\hat{\boldsymbol{\beta}}_{2,T}$ is similar. □

Theorem 3.3 shows that $\hat{\boldsymbol{\beta}}_{1,T}$ can be computed from regressing $(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{y}$ on $(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{X}_1$, where $(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{y}$ and $(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{X}_1$ are the residual vectors of the "purging" regressions of $\boldsymbol{y}$ on $\boldsymbol{X}_2$ and $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$, respectively. Similarly, $\hat{\boldsymbol{\beta}}_{2,T}$ can be obtained by regressing $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}$ on $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2$, where $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}$ and $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2$ are the residual vectors of the regressions of $\boldsymbol{y}$ on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, respectively.

From Theorem 3.3 we can deduce the following results. Consider the regression of $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}$ on $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2$. By Theorem 3.3 we have

$$(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}=(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{2,T}+\text{ residual vector}, \tag{3.5}$$

where the residual vector is

$$(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}=(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}.$$

Thus, the residual vector of (3.5) is identical to the residual vector of regressing $\boldsymbol{y}$ on $\boldsymbol{X}=[\boldsymbol{X}_1\ \boldsymbol{X}_2]$. Note that $(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{I}-\boldsymbol{P})=\boldsymbol{I}-\boldsymbol{P}$ implies $\boldsymbol{P}_1=\boldsymbol{P}_1\boldsymbol{P}$. That is, the orthogonal projection of $\boldsymbol{y}$ directly on $\text{span}(\boldsymbol{X}_1)$ is equivalent to performing iterated projections of $\boldsymbol{y}$ on $\text{span}(\boldsymbol{X})$ and then on $\text{span}(\boldsymbol{X}_1)$. The orthogonal projection part of (3.5) now can be expressed as

$$(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{2,T}=(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{P}\boldsymbol{y}=(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{y}.$$

These relationships are illustrated in Figure 3.2. Similarly, we have

$$(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{y}=(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,T}+\text{ residual vector},$$

where the residual vector is also $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}$, and the orthogonal projection part of this regression is $(\boldsymbol{P}-\boldsymbol{P}_2)\boldsymbol{y}$. See also Davidson and MacKinnon (1993) for more details.

Intuitively, Theorem 3.3 suggests that $\hat{\boldsymbol{\beta}}_{1,T}$ in effect describes how $\boldsymbol{X}_1$ characterizes $\boldsymbol{y}$, after the effect of $\boldsymbol{X}_2$ is excluded. Thus, $\hat{\boldsymbol{\beta}}_{1,T}$ is different from the OLS estimator of regressing $\boldsymbol{y}$ on $\boldsymbol{X}_1$ because the effect of $\boldsymbol{X}_2$ is not controlled in the latter. These two

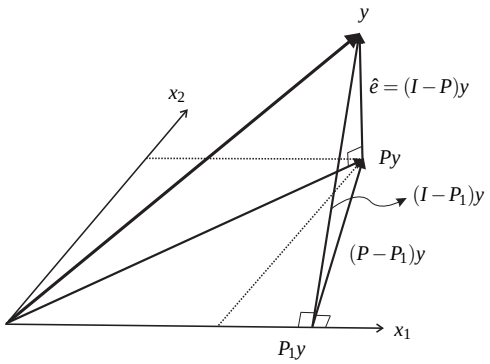

Figure 3.2: An illustration of the Frisch-Waugh-Lovell Theorem

estimators would be the same if $\boldsymbol{P}_2\boldsymbol{X}_1 = \boldsymbol{0}$, i.e., $\boldsymbol{X}_1$ is orthogonal to $\boldsymbol{X}_2$. Also, $\hat{\boldsymbol{\beta}}_{2,T}$ describes how $\boldsymbol{X}_2$ characterizes $\boldsymbol{y}$, after the effect of $\boldsymbol{X}_1$ is excluded, and it is different from the OLS estimator from regressing $\boldsymbol{y}$ on $\boldsymbol{X}_2$, unless $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal to each other.

As an application, consider the specification with $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$, where $\boldsymbol{X}_1$ contains the constant term and a time trend variable t, and $\boldsymbol{X}_2$ includes the other $k-2$ explanatory variables. This specification is useful when the variables of interest exhibit a trending behavior. Then, the OLS estimators of the coefficients of $\boldsymbol{X}_2$ are the same as those obtained from regressing (detrended) $\boldsymbol{y}$ on detrended $\boldsymbol{X}_2$, where detrended $\boldsymbol{y}$ and $\boldsymbol{X}_2$ are the residuals of regressing $\boldsymbol{y}$ and $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, respectively. See Section 3.5.2 and Exercise 3.11 for other applications.

3.1.4 Measures of Goodness of Fit

We have learned that from previous sections that, when the explanatory variables in a linear specification are given, the OLS method yields the best fit of data. In practice, one may consdier a linear specfication with different sets of regressors and try to choose a particular one from them. It is therefore of interest to compare the performance across different specifications. In this section we discuss how to measure the *goodness of fit* of a specification. A natural goodness-of-fit measure is of course the sum of squared errors $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$. Unfortunately, this measure is not invariant with respect to measurement units of the dependent variable and hence is not appropriate for model comparison. Instead, we consider the following "relative" measures of goodness of fit.

Recall from Theorem 3.2(b) that $\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = 0$. Then,

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

This equation can be written in terms of sum of squares:

$$\underbrace{\sum_{t=1}^{T} y_t^2}_{\text{TSS}} = \underbrace{\sum_{t=1}^{T} \hat{y}_t^2}_{\text{RSS}} + \underbrace{\sum_{t=1}^{T} \hat{e}_t^2}_{\text{ESS}},$$

where TSS stands for *total sum of squares* and is a measure of total squared variations of y_t, RSS stands for *regression sum of squares* and is a measures of squared variations of fitted values, and ESS stands for *error sum of squares* and is a measure of squared variation of residuals. The non-centered *coefficient of determination* (or non-centered R^2) is defined as the proportion of TSS that can be explained by the regression hyperplane:

$$R^2 = \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{ESS}}{\text{TSS}}. \tag{3.6}$$

Clearly, $0 \leq R^2 \leq 1$, and the larger the R^2, the better the model fits the data. In particular, a model has a perfect fit if $R^2 = 1$, and it does not account for any variation of $\boldsymbol{y}$ if $R^2 = 0$. It is also easy to verify that this measure does not depend on the measurement units of the dependent and explanatory variables; see Exercise 3.7.

As $\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}'\boldsymbol{y}$, we can also write

$$R^2 = \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}} = \frac{(\hat{\boldsymbol{y}}'\boldsymbol{y})^2}{(\boldsymbol{y}'\boldsymbol{y})(\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}})}.$$

It follows from the discussion of inner product and Euclidean norm in Section 1.2 that the right-hand side is just $\cos^2\theta$, where θ is the angle between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$. Thus, R^2 can be interpreted as a measure of the linear association between these two vectors. A perfect fit is equivalent to the fact that $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are collinear, so that $\boldsymbol{y}$ must be in $\text{span}(\boldsymbol{X})$. When $R^2 = 0$, $\boldsymbol{y}$ is orthogonal to $\hat{\boldsymbol{y}}$ so that $\boldsymbol{y}$ is in $\text{span}(\boldsymbol{X})^\perp$.

It can be verified that when a constant is added to all observations of the dependent variable, the resulting coefficient of determination also changes. This is clearly a drawback because a sensible measure of fit should not be affected by the location of the dependent variable. Another drawback of the coefficient of determination is that it is non-decreasing in the number of variables in the specification. That is, adding more variables to a linear specification will *not* reduce its R^2. To see this, consider a specification with k_1 regressors and a more complex one containing the same k_1 regressors and additional k_2 regressors. In this case, the former specification is "nested" in the latter, in the sense that the former can be obtained from the latter by setting the coefficients of those additional regressors to zero. Since the OLS method searches for the best fit of data without any constraint, the more complex model cannot have a worse fit than the specifications nested in it. See also Exercise 3.8.

A measure that is invariant with respect to constant addition is the centered coefficient of determination (or centered R^2). When a specification contains a constant term,

$$\underbrace{\sum_{t=1}^{T}(y_t-\bar{y})^2}_{\text{Centered TSS}}=\underbrace{\sum_{t=1}^{T}(\hat{y}_t-\bar{\hat{y}})^2}_{\text{Centered RSS}}+\underbrace{\sum_{t=1}^{T}\hat{e}_t^2}_{\text{ESS}},$$

where $\bar{\hat{y}}=\bar{y}=\sum_{t=1}^{T}y_t/T$. Analogous to (3.6), the centered R^2 is defined as

$$\text{Centered } R^2=\frac{\text{Centered RSS}}{\text{Centered TSS}}=1-\frac{\text{ESS}}{\text{Centered TSS}}. \tag{3.7}$$

Centered R^2 also takes on values between 0 and 1 and is non-decreasing in the number of variables in the specification. In contrast with non-centered R^2, this measure *excludes* the effect of the constant term and hence is invariant with respect to constant addition.

When a specfication contains a constant term, we have

$$\sum_{t=1}^{T}(y_t-\bar{y})(\hat{y}_t-\bar{y})=\sum_{t=1}^{T}(\hat{y}_t-\bar{y}+\hat{e}_t)(\hat{y}_t-\bar{y})=\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2,$$

because $\sum_{t=1}^{T}\hat{y}_t\hat{e}_t=\sum_{t=1}^{t}\hat{e}_t=0$ by Theorem 3.2. It follows that

$$R^2=\frac{\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2}{\sum_{t=1}^{T}(y_t-\bar{y})^2}=\frac{[\sum_{t=1}^{T}(y_t-\bar{y})(\hat{y}_t-\bar{y})]^2}{[\sum_{t=1}^{T}(y_t-\bar{y})^2][\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2]}.$$

That is, the centered R^2 is also the squared sample correlation coefficient of y_t and $\hat{y}_t$, also known as the *squared multiple correlation coefficient*. If a specification does *not* contain a constant term, the centered R^2 may be negative; see Exercise 3.10.

Both centered and non-centered R^2 are still non-decreasing in the number of regressors. This property implies that a more complex model would be preferred if R^2 is the only criterion for choosing a specification. A modified measure is the adjusted R^2, $\bar{R}^2$, which is the centered R^2 adjusted for the degrees of freedom:

$$\bar{R}^2=1-\frac{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(T-k)}{(\boldsymbol{y}'\boldsymbol{y}-T\bar{y}^2)/(T-1)}.$$

This measure can also be expressed in different forms:

$$\bar{R}^2=1-\frac{T-1}{T-k}(1-R^2)=R^2-\frac{k-1}{T-k}(1-R^2).$$

That is, $\bar{R}^2$ is the centered R^2 with a penalty term depending on model complexity and explanatory ability. Observe that when k increases, $(k-1)/(T-k)$ increases but $1-R^2$ decreases. Whether the penalty term is larger or smaller depends on the trade-off between

these two terms. Thus, $\bar{R}^2$ need not be increasing with the number of explanatory variables. Clearly, $\bar{R}^2 < R^2$ except for $k = 1$ or $R^2 = 1$. It can also be verified that $\bar{R}^2 < 0$ when $R^2 < (k-1)/(T-1)$.

Remark: As different dependent variables have different TSS, the associated specifications are therefore not comparable in terms of their R^2. For example, R^2 of the specifications with y and $\log y$ as dependent variables are not comparable.

3.2 Statistical Properties of the OLS Estimators

Readers should have noticed that the previous results, which are either algebraic or geometric, hold regardless of the random nature of data. To derive the statistical properties of the OLS estimator, some probabilistic conditions must be imposed.

3.2.1 Classical Conditions

The following conditions on data are usually known as the *classical conditions*.

[A1] $\boldsymbol{X}$ is non-stochastic.

[A2] $\boldsymbol{y}$ is a random vector such that

- **(i)** $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}_o$ for some $\boldsymbol{\beta}_o$;
- **(ii)** $\mathrm{var}(\boldsymbol{y}) = \sigma_o^2 \boldsymbol{I}_T$ for some $\sigma_o^2 > 0$.

[A3] $\boldsymbol{y}$ is a random vector such that $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}_o, \sigma_o^2 \boldsymbol{I}_T)$ for some $\boldsymbol{\beta}_o$ and $\sigma_o^2 > 0$.

Condition [A1] is not crucial, but, as we will see below, it is quite convenient for subsequent analysis. Concerning [A2](i), we first note that $\mathbb{E}(\boldsymbol{y})$ is the "averaging" behavior of y and may be interpreted as a systematic component of y. [A2](i) is thus a condition ensuring that the postulated linear function $\boldsymbol{X}\boldsymbol{\beta}$ is a specification of this systematic component, correct up to unknown parameters. Condition [A2](ii) regulates that the variance-covariance matrix of $\boldsymbol{y}$ depends only on one parameter σ_o^2; such a matrix is also known as a *scalar covariance matrix*. Under [A2](ii), y_t, $t = 1, \ldots, T$, have the constant variance σ_o^2 and are pairwise uncorrelated (but not necessarily independent). Although conditions [A2] and [A3] impose the same structures on the mean and variance of $\boldsymbol{y}$, the latter is much stronger because it also specifies the distribution of $\boldsymbol{y}$. We have seen in Section 2.3 that uncorrelated normal random variables are also independent. Therefore, y_t, $t = 1, \ldots, T$, are i.i.d. (independently and identically distributed) normal random variables under [A3]. The linear specification (3.1) with [A1] and [A2] is known as the *classical linear model*, and (3.1)

with [A1] and [A3] is also known as the *classical normal linear model*. The limitations of these conditions will be discussed in Section 3.6.

In addition to $\hat{\boldsymbol{\beta}}_T$, the new unknown parameter $\mathrm{var}(y_t) = \sigma_o^2$ in [A2](ii) and [A3] should be estimated as well. The OLS estimator for σ_o^2 is

$$\hat{\sigma}_T^2 = \frac{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{T-k} = \frac{1}{T-k}\sum_{t=1}^{T}\hat{e}_t^2, \tag{3.8}$$

where k is the number of regressors. While $\hat{\boldsymbol{\beta}}_T$ is a *linear estimator* in the sense that it is a linear transformation of $\boldsymbol{y}$, $\hat{\sigma}_T^2$ is not. In the sections below we will derive the properties of the OLS estimators $\hat{\boldsymbol{\beta}}_T$ and $\hat{\sigma}_T^2$ under these classical conditions.

3.2.2 Without the Normality Condition

Under the imposed classical conditions, the OLS estimators have the following statistical properties.

Theorem 3.4 *Consider the linear specification* (3.1).

(a) *Given* [A1] *and* [A2](i), $\hat{\boldsymbol{\beta}}_T$ *is unbiased for* $\boldsymbol{\beta}_o$.

(b) *Given* [A1] *and* [A2], $\hat{\sigma}_T^2$ *is unbiased for* σ_o^2.

(c) *Given* [A1] *and* [A2], $\mathrm{var}(\hat{\boldsymbol{\beta}}_T) = \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Proof: Given [A1] and [A2](i), $\hat{\boldsymbol{\beta}}_T$ is unbiased because

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_T) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}(\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}_o = \boldsymbol{\beta}_o.$$

To prove (b), recall that $(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$ so that the OLS residual vector can be written as

$$\hat{\boldsymbol{e}} = (\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y} = (\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o).$$

Then, $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'(\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)$ which is a scalar, and

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}) &= \mathbb{E}\big[\mathrm{trace}\big((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'(\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\big)\big] \\ &= \mathbb{E}[\mathrm{trace}\big((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'(\boldsymbol{I}_T - \boldsymbol{P})\big)\big].\end{aligned}$$

By interchanging the trace and expectation operators, we have from [A2](ii) that

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}) &= \mathrm{trace}\big(\mathbb{E}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'(\boldsymbol{I}_T - \boldsymbol{P})]\big) \\ &= \mathrm{trace}\big(\mathbb{E}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'](\boldsymbol{I}_T - \boldsymbol{P})\big) \\ &= \mathrm{trace}\big(\sigma_o^2\boldsymbol{I}_T(\boldsymbol{I}_T - \boldsymbol{P})\big) \\ &= \sigma_o^2\,\mathrm{trace}(\boldsymbol{I}_T - \boldsymbol{P}).\end{aligned}$$

By Lemmas 1.12 and 1.14, $\mathrm{trace}(\boldsymbol{I}_T - \boldsymbol{P}) = \mathrm{rank}(\boldsymbol{I}_T - \boldsymbol{P}) = T - k$. Consequently,

$$\mathbb{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}) = \sigma_o^2(T-k),$$

so that

$$\mathbb{E}(\hat{\sigma}_T^2) = \mathbb{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}})/(T-k) = \sigma_o^2.$$

This proves the unbiasedness of $\hat{\sigma}_T^2$. Given that $\hat{\boldsymbol{\beta}}_T$ is a linear transformation of $\boldsymbol{y}$, we have from Lemma 2.4 that

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}_T) &= \mathrm{var}\big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\big) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma_o^2\boldsymbol{I}_T)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

This establishes (c). □

It can be seen that the unbiasedness of $\hat{\boldsymbol{\beta}}_T$ does not depend on [A2](ii), the variance property of $\boldsymbol{y}$. It is also clear that when $\hat{\sigma}_T^2$ is unbiased, the estimator

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_T) = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

is also unbiased for $\mathrm{var}(\hat{\boldsymbol{\beta}}_T)$. The result below, known as the *Gauss-Markov theorem*, indicates that when [A1] and [A2] hold, $\hat{\boldsymbol{\beta}}_T$ is not only unbiased but also the best (most efficient) among all linear unbiased estimators for $\boldsymbol{\beta}_o$.

Theorem 3.5 (Gauss-Markov) *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A2] *hold. Then the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}_o$.*

Proof: Consider an arbitrary linear estimator $\check{\boldsymbol{\beta}}_T = \boldsymbol{A}\boldsymbol{y}$, where $\boldsymbol{A}$ is non-stochastic. Writing $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}$, $\check{\boldsymbol{\beta}}_T = \hat{\boldsymbol{\beta}}_T + \boldsymbol{C}\boldsymbol{y}$. Then,

$$\mathrm{var}(\check{\boldsymbol{\beta}}_T) = \mathrm{var}(\hat{\boldsymbol{\beta}}_T) + \mathrm{var}(\boldsymbol{C}\boldsymbol{y}) + 2\ \mathrm{cov}(\hat{\boldsymbol{\beta}}_T, \boldsymbol{C}\boldsymbol{y}).$$

By [A1] and [A2](i),

$$\mathbb{E}(\check{\boldsymbol{\beta}}_T) = \boldsymbol{\beta}_o + \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}_o.$$

Since $\boldsymbol{\beta}_o$ is arbitrary, this estimator would be unbiased if, and only if, $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{0}$. This property further implies that

$$\begin{aligned}\mathrm{cov}(\hat{\boldsymbol{\beta}}_T, \boldsymbol{C}\boldsymbol{y}) &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\boldsymbol{y}'\boldsymbol{C}'] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\boldsymbol{y}']\boldsymbol{C}' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma_o^2\boldsymbol{I}_T)\boldsymbol{C}' \\ &= \boldsymbol{0}.\end{aligned}$$

Thus,

$$\mathrm{var}(\check{\boldsymbol{\beta}}_T) = \mathrm{var}(\hat{\boldsymbol{\beta}}_T) + \mathrm{var}(\boldsymbol{C}\boldsymbol{y}) = \mathrm{var}(\hat{\boldsymbol{\beta}}_T) + \sigma_o^2\boldsymbol{C}\boldsymbol{C}',$$

where $\sigma_o^2\boldsymbol{C}\boldsymbol{C}'$ is clearly a positive semi-definite matrix. This shows that for any linear unbiased estimator $\check{\boldsymbol{\beta}}_T$, $\mathrm{var}(\check{\boldsymbol{\beta}}_T) - \mathrm{var}(\hat{\boldsymbol{\beta}}_T)$ is positive semi-definite, so that $\hat{\boldsymbol{\beta}}_T$ is more efficient. □

Example 3.6 Given the data $[\boldsymbol{y}\ \boldsymbol{X}]$, where $\boldsymbol{X}$ is a nonstochastic matrix and can be partitioned as $[\boldsymbol{X}_1\ \boldsymbol{X}_2]$. Suppose that $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}_1\boldsymbol{b}_1$ for some $\boldsymbol{b}_1$ and $\mathrm{var}(\boldsymbol{y}) = \sigma_o^2\boldsymbol{I}_T$ for some $\sigma_o^2 > 0$. Consider first the specification that contains only $\boldsymbol{X}_1$ but not $\boldsymbol{X}_2$:

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

Let $\hat{\boldsymbol{b}}_{1,T}$ denote the resulting OLS estimator. It is clear that $\hat{\boldsymbol{b}}_{1,T}$ is still a linear estimator and unbiased for $\boldsymbol{b}_1$ by Theorem 3.4(a). Moreover, it is the BLUE for $\boldsymbol{b}_1$ by Theorem 3.5 with the variance-covariance matrix

$$\mathrm{var}(\hat{\boldsymbol{b}}_{1,T}) = \sigma_o^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1},$$

by Theorem 3.4(c).

Consider now the linear specification that involves both $\boldsymbol{X}_1$ and irrelevant regressors $\boldsymbol{X}_2$:

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

This specification would be a correct specification if some of the parameters ($\boldsymbol{\beta}_2$) are restricted to zero. Let $\hat{\boldsymbol{\beta}}_T = (\hat{\boldsymbol{\beta}}_{1,T}'\ \hat{\boldsymbol{\beta}}_{2,T}')'$ be the OLS estimator of $\boldsymbol{\beta}$. Using Theorem 3.3, we find

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}_{1,T}) &= \mathbb{E}\left([\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{y}\right) = \boldsymbol{b}_1, \\
\mathbb{E}(\hat{\boldsymbol{\beta}}_{2,T}) &= \mathbb{E}\left([\boldsymbol{X}_2'(\boldsymbol{I}_T - \boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'(\boldsymbol{I}_T - \boldsymbol{P}_1)\boldsymbol{y}\right) = \boldsymbol{0},
\end{aligned}$$

where $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ and $\boldsymbol{P}_2 = \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$. This shows that $\hat{\boldsymbol{\beta}}_T$ is unbiased for $(\boldsymbol{b}_1'\ \boldsymbol{0}')'$. Also,

$$\begin{aligned}
\mathrm{var}(\hat{\boldsymbol{\beta}}_{1,T}) &= \mathrm{var}([\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{y}) \\
&= \sigma_o^2[\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}.
\end{aligned}$$

Given that $\boldsymbol{P}_2$ is a positive semi-definite matrix,

$$\boldsymbol{X}_1'\boldsymbol{X}_1 - \boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1 = \boldsymbol{X}_1'\boldsymbol{P}_2\boldsymbol{X}_1,$$

must also be positive semi-definite. It follows from Lemma 1.9 that

$$[\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1} - (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$$

is a positive semi-definite matrix. This shows that $\hat{\boldsymbol{b}}_{1,T}$ is more efficient than $\hat{\boldsymbol{\beta}}_{1,T}$, as it ought to be. When $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{0}$, i.e., the columns of $\boldsymbol{X}_1$ are orthogonal to the columns of $\boldsymbol{X}_2$, we immediately have $(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1 = \boldsymbol{X}_1$, so that $\hat{\boldsymbol{\beta}}_{1,T} = \hat{\boldsymbol{b}}_{1,T}$. In this case, estimating a more complex specification does not result in efficiency loss. □

Remark: This example shows that, given the specification $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, the OLS estimator of $\boldsymbol{\beta}_1$ is not the most efficient when $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}_1\boldsymbol{b}_1$. This result thus suggests that when the restrictions on parameters are not taken into account, the resulting OLS estimator would suffer from efficiency loss.

3.2.3 With the Normality Condition

We have learned that the normality condition [A3] is much stronger than [A2]. With this stronger condition, more can be said about the OLS estimators.

Theorem 3.7 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold.*

(a) $\hat{\boldsymbol{\beta}}_T \sim \mathcal{N}\big(\boldsymbol{\beta}_o, \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\big)$.

(b) $(T-k)\hat{\sigma}_T^2/\sigma_o^2 \sim \chi^2(T-k)$.

(c) $\hat{\sigma}_T^2$ *has mean* σ_o^2 *and variance* $2\sigma_o^4/(T-k)$.

Proof: As $\hat{\boldsymbol{\beta}}_T$ is a linear transformation of $\boldsymbol{y}$, it is also normally distributed as

$$\hat{\boldsymbol{\beta}}_T \sim \mathcal{N}\big(\boldsymbol{\beta}_o, \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\big),$$

by Lemma 2.6, where its mean and variance-covariance matrix are as in Theorem 3.4(a) and (c). To prove the assertion (b), we again write $\hat{\boldsymbol{e}} = (\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)$ and deduce

$$(T-k)\hat{\sigma}_T^2/\sigma_o^2 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/\sigma_o^2 = \boldsymbol{y}^{*\prime}(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y}^*,$$

where $\boldsymbol{y}^* = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)/\sigma_o$. Let $\boldsymbol{C}$ be the orthogonal matrix that diagonalizes the symmetric and idempotent matrix $\boldsymbol{I}_T - \boldsymbol{P}$. Then, $\boldsymbol{C}'(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{C} = \boldsymbol{\Lambda}$. Since $\operatorname{rank}(\boldsymbol{I}_T - \boldsymbol{P}) = T - k$, $\boldsymbol{\Lambda}$ contains $T-k$ eigenvalues equal to one and k eigenvalues equal to zero by Lemma 1.11. Without loss of generality we can write

$$\boldsymbol{y}^{*\prime}(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y}^* = \boldsymbol{y}^{*\prime}\boldsymbol{C}[\boldsymbol{C}'(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{C}]\boldsymbol{C}'\boldsymbol{y}^* = \boldsymbol{\eta}'\begin{bmatrix} \boldsymbol{I}_{T-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{\eta},$$

where $\boldsymbol{\eta} = \boldsymbol{C}'\boldsymbol{y}^*$. Again by Lemma 2.6, $\boldsymbol{y}^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_T)$ under [A3]. Hence, $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_T)$, so that η_i are independent, standard normal random variables. Consequently,

$$\boldsymbol{y}^{*\prime}(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y}^* = \sum_{i=1}^{T-k} \eta_i^2 \sim \chi^2(T-k).$$

This proves (b). Noting that the mean of $\chi^2(T-k)$ is $T-k$ and variance is $2(T-k)$, the assertion (c) is just a direct consequence of (b). □

Suppose that we believe that [A3] is true and specify the log-likelihood function of $\boldsymbol{y}$ as:

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The first order conditions of maximizing this log-likelihood are

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sigma^2}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}, \\
\nabla_{\sigma^2} \log L(\boldsymbol{\beta}, \sigma^2) &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0,
\end{aligned}$$

and their solutions are the MLEs $\tilde{\boldsymbol{\beta}}_T$ and $\tilde{\sigma}_T^2$. The first k equations above are equivalent to the OLS normal equations (3.3). It follows that the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is also the MLE $\tilde{\boldsymbol{\beta}}_T$. Plugging $\hat{\boldsymbol{\beta}}_T$ into the first order conditions we can solve for σ^2 and obtain

$$\tilde{\sigma}_T^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_T)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_T)}{T} = \frac{\hat{e}'\hat{e}}{T}, \tag{3.9}$$

which is different from the OLS variance estimator (3.8).

The conclusion below is stronger than the Gauss-Markov theorem (Theorem 3.5).

Theorem 3.8 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold. Then the OLS estimators* $\hat{\boldsymbol{\beta}}_T$ *and* $\hat{\sigma}_T^2$ *are the best unbiased estimators for* $\boldsymbol{\beta}_o$ *and* σ_o^2*, respectively.*

Proof: The score vector is

$$\boldsymbol{s}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{bmatrix},$$

and the Hessian matrix of the log-likelihood function is

$$\boldsymbol{H}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} -\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} & -\frac{1}{\sigma^4}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ -\frac{1}{\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{X} & \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{bmatrix}.$$

It is easily verified that when [A3] is true, $\mathbb{E}[\boldsymbol{s}(\boldsymbol{\beta}_o, \sigma_o^2)] = \mathbf{0}$ and

$$\mathbb{E}[\boldsymbol{H}(\boldsymbol{\beta}_o, \sigma_o^2)] = \begin{bmatrix} -\frac{1}{\sigma_o^2}\boldsymbol{X}'\boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & -\frac{T}{2\sigma_o^4} \end{bmatrix}.$$

The information matrix equality (Lemma 2.9) ensures that the negative of $\mathbb{E}[\boldsymbol{H}(\boldsymbol{\beta}_o, \sigma_o^2)]$ equals the information matrix. The inverse of the information matrix is then

$$\begin{bmatrix} \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma_o^4}{T} \end{bmatrix},$$

which is the Cramér-Rao lower bound by Lemma 2.10. Clearly, $\text{var}(\hat{\boldsymbol{\beta}}_T)$ achieves this lower bound so that $\hat{\boldsymbol{\beta}}_T$ must be the best unbiased estimator for $\boldsymbol{\beta}_o$. Although the variance of $\hat{\sigma}_T^2$ is greater than the lower bound, it can be shown that $\hat{\sigma}_T^2$ is still the best unbiased estimator for σ_o^2; see, e.g., Rao (1973, p. 319) for a proof. □

Remark: Comparing to the Gauss-Markov theorem, Theorem 3.8 gives a stronger result at the expense of a stronger condition (the normality condition [A3]). The OLS estimators now are the best (most efficient) in a much larger class of estimators, namely, the class of unbiased estimators. Note also that Theorem 3.8 covers $\hat{\sigma}_T^2$, whereas the Gauss-Markov theorem does not.

3.3 Hypotheses Testing

After a specification is estimated, it is often desirable to test various economic and econometric hypotheses. Given the classical conditions [A1] and [A3], we consider the linear hypothesis

$$\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}, \tag{3.10}$$

where $\boldsymbol{R}$ is a $q \times k$ non-stochastic matrix with rank $q < k$, and $\boldsymbol{r}$ is a vector of pre-specified, hypothetical values.

3.3.1 Tests for Linear Hypotheses

If the null hypothesis (3.10) is true, it is reasonable to expect that $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ is "close" to the hypothetical value $\boldsymbol{r}$; otherwise, they should be quite different. Here, the closeness between $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ and $\boldsymbol{r}$ must be justified by the null distribution of the test statistics.

If there is only a single hypothesis, the null hypothesis (3.10) is such that $\boldsymbol{R}$ is a row vector ($q = 1$) and $\boldsymbol{r}$ is a scalar. Note that a single hypothesis may involve two or more

parameters. Consider the following statistic:

$$\frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\sigma_o[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}}.$$

By Theorem 3.7(a), $\hat{\boldsymbol{\beta}}_T \sim \mathcal{N}(\boldsymbol{\beta}_o, \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$, and hence

$$\boldsymbol{R}\hat{\boldsymbol{\beta}}_T \sim \mathcal{N}(\boldsymbol{R}\boldsymbol{\beta}_o, \sigma_o^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}').$$

Under the null hypothesis, we have

$$\frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\sigma_o[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}} = \frac{\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)}{\sigma_o[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}} \sim \mathcal{N}(0,1). \tag{3.11}$$

Although the left-hand side has a known distribution, it cannot be used as a test statistic because σ_o is unknown. Replacing σ_o by its OLS estimator $\hat{\sigma}_T$ yields an operational statistic:

$$\tau = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\hat{\sigma}_T[R(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}}. \tag{3.12}$$

The null distribution of τ is given in the result below.

Theorem 3.9 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold. Then under the null hypothesis* (3.10) *with* $\boldsymbol{R}$ *a* $1 \times k$ *vector,*

$$\tau \sim t(T-k),$$

where τ *is given by* (3.12)*.*

Proof: We first write the statistic τ as

$$\tau = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\sigma_o[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}} \Bigg/ \sqrt{\frac{(T-k)\hat{\sigma}_T^2/\sigma_o^2}{T-k}},$$

where the numerator is distributed as $\mathcal{N}(0,1)$ by (3.11), and $(T-k)\hat{\sigma}_T^2/\sigma_o^2$ is distributed as $\chi^2(T-k)$ by Theorem 3.7(b). Hence, the square of the denominator is a central χ^2 random variable divided by its degrees of freedom $T-k$. The assertion follows if we can show that the numerator and denominator are independent. Note that the random components of the numerator and denominator are, respectively, $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$, where $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{e}}$ are two normally distributed random vectors with the covariance matrix

$$\begin{aligned}
\operatorname{cov}(\hat{\boldsymbol{e}}, \hat{\boldsymbol{\beta}}_T) &= \mathbb{E}[(\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\
&= (\boldsymbol{I}_T - \boldsymbol{P})\,\mathbb{E}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\boldsymbol{y}']\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma_o^2(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \boldsymbol{0}.
\end{aligned}$$

Since uncorrelated normal random vectors are also independent, $\hat{\boldsymbol{\beta}}_T$ is independent of $\hat{\boldsymbol{e}}$. By Lemma 2.1, we conclude that $\hat{\boldsymbol{\beta}}_T$ is also independent of $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$. □

As the null distribution of the statistic τ is $t(T-k)$ by Theorem 3.9, τ is known as the t statistic. When the alternative hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_o \neq \boldsymbol{r}$, this is a two-sided test; when the alternative hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_o > \boldsymbol{r}$ (or $\boldsymbol{R}\boldsymbol{\beta}_o < \boldsymbol{r}$), this is a one-sided test. For each test, we first choose a small significance level α and then determine the critical region C_α. For the two-sided t test, we can find the values $\pm t_{\alpha/2}(T-k)$ from the table of t distributions such that

$$\begin{aligned}\alpha &= \mathbb{P}\{\tau < -t_{\alpha/2}(T-k) \text{ or } \tau > t_{\alpha/2}(T-k)\} \\ &= 1 - \mathbb{P}\{-t_{\alpha/2}(T-k) \leq \tau \leq t_{\alpha/2}(T-k)\}.\end{aligned}$$

The critical region is then

$$C_\alpha = (-\infty, -t_{\alpha/2}(T-k)) \cup (t_{\alpha/2}(T-k), \infty),$$

and $\pm t_{\alpha/2}(T-k)$ are the critical values at the significance level α. For the alternative hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o > \boldsymbol{r}$, the critical region is $(t_\alpha(T-k), \infty)$, where $t_\alpha(T-k)$ is the critical value such that

$$\alpha = \mathbb{P}\{\tau > t_\alpha(T-k)\}.$$

Similarly, for the alternative $\boldsymbol{R}\boldsymbol{\beta}_o < \boldsymbol{r}$, the critical region is $(-\infty, -t_\alpha(T-k))$.

The null hypothesis is rejected at the significance level α when τ falls in the critical region. As α is small, the event $\{\tau \in C_\alpha\}$ is unlikely under the null hypothesis. When τ does take an extreme value relative to the critical values, it is an evidence against the null hypothesis. The decision of rejecting the null hypothesis could be wrong, but the probability of the type I error will not exceed α. When τ takes a "reasonable" value in the sense that it falls in the complement of the critical region, the null hypothesis is not rejected.

Example 3.10 To test a single coefficient equal to zero: $\beta_i = 0$, we choose $\boldsymbol{R}$ as the transpose of the i^{th} Cartesian unit vector:

$$\boldsymbol{R} = [\,0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0\,].$$

Let m^{ii} be the i^{th} diagonal element of $\boldsymbol{M}^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Then, $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' = m^{ii}$. The t statistic for this hypothesis, also known as the t *ratio*, is

$$\tau = \frac{\hat{\beta}_{i,T}}{\hat{\sigma}_T\sqrt{m^{ii}}} \sim t(T-k).$$

When a t ratio rejects the null hypothesis, it is said that the corresponding estimated coefficient is significantly different from zero; econometrics and statistics packages usually report t ratios along with the coefficient estimates. □

Example 3.11 To test the single hypothesis $\beta_i + \beta_j = 0$, we set $\boldsymbol{R}$ as

$$\boldsymbol{R} = [\, 0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0 \,].$$

Hence, $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' = m^{ii} + 2m^{ij} + m^{jj}$, where m^{ij} is the $(i,j)^{\text{th}}$ element of $\boldsymbol{M}^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. The t statistic is

$$\tau = \frac{\hat{\beta}_{i,T} + \hat{\beta}_{j,T}}{\hat{\sigma}_T (m^{ii} + 2m^{ij} + m^{jj})^{1/2}} \sim t(T-k). \quad \square$$

Several hypotheses can also be tested jointly. Consider the null hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$, where $\boldsymbol{R}$ is now a $q \times k$ matrix ($q \geq 2$) and $\boldsymbol{r}$ is a vector. This hypothesis involves q single hypotheses. Similar to (3.11), we have under the null hypothesis that

$$[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q).$$

Therefore,

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o^2 \sim \chi^2(q). \tag{3.13}$$

Again, we can replace σ_o^2 by its OLS estimator $\hat{\sigma}_T^2$ to obtain an operational statistic:

$$\varphi = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})}{\hat{\sigma}_T^2 q}. \tag{3.14}$$

The next result gives the null distribution of φ.

Theorem 3.12 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold. Then under the null hypothesis* (3.10) *with* $\boldsymbol{R}$ *a* $q \times k$ *matrix with rank* $q < k$*, we have*

$$\varphi \sim F(q, T-k),$$

where φ *is given by* (3.14)*.*

Proof: Note that

$$\varphi = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/(\sigma_o^2 q)}{(T-k)\frac{\hat{\sigma}_T^2}{\sigma_o^2}\Big/(T-k)}.$$

In view of (3.13) and the proof of Theorem 3.9, the numerator and denominator terms are two independent χ^2 random variables, each divided by its degrees of freedom. The assertion follows from the definition of F random variable. □

The statistic φ is known as the F statistic. We reject the null hypothesis at the significance level α when φ is too large relative to the critical value $F_\alpha(q, T-k)$ from the table of F distributions, where $F_\alpha(q, T-k)$ is such that

$$\alpha = \mathbb{P}\{\varphi > F_\alpha(q, T-k)\}.$$

If there is only a single hypothesis, the F statistic is just the square of the corresponding t statistic. When φ rejects the null hypothesis, it simply suggests that there is evidence against at least one single hypothesis. The inference of a joint test is, however, not necessary the same as the inference of individual tests; see also Section 3.4.

Example 3.13 Joint null hypothesis: H_o: $\beta_1 = b_1$ and $\beta_2 = b_2$. The F statistic is

$$\varphi = \frac{1}{2\hat{\sigma}_T^2} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix}' \begin{bmatrix} m^{11} & m^{12} \\ m^{21} & m^{22} \end{bmatrix}^{-1} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix} \sim F(2, T-k),$$

where m^{ij} is as defined in Example 3.11. □

Remark: For the null hypothesis of s coefficients being zero, if the corresponding F statistic $\varphi > 1$ ($\varphi < 1$), dropping these s regressors will reduce (increase) $\bar{R}^2$; see Exercise 3.12.

3.3.2 Power of the Tests

Recall that the power of a test is the probability of rejecting the null hypothesis when the null hypothesis is indeed false. In this section, we consider the hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r} + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ characterizes the deviation from the null hypothesis, and analyze the power performance of the t and F tests.

Theorem 3.14 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold. Then under the hypothesis that* $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r} + \boldsymbol{\delta}$*, where* $\boldsymbol{R}$ *is a* $q \times k$ *matrix with rank* $q < k$*, we have*

$$\varphi \sim F(q, T-k;\, \boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}, 0),$$

where φ *is given by* (3.14)*,* $\boldsymbol{D} = \sigma_o^2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']$*, and* $\boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}$ *is the non-centrality parameter of the numerator term.*

Proof: When $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r} + \boldsymbol{\delta}$,

$$\begin{aligned} &[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o \\ &\quad = [\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}[\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) + \boldsymbol{\delta}]/\sigma_o. \end{aligned}$$

Given [A3],

$$[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)/\sigma_o. \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q),$$

and hence

$$[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o \sim \mathcal{N}(\boldsymbol{D}^{-1/2}\boldsymbol{\delta}, \boldsymbol{I}_q).$$

It follows from Lemma 2.7 that

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o^2 \sim \chi^2(q;\, \boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}),$$

which is the non-central χ^2 distribution with q degrees of freedom and the non-centrality parameter $\boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}$. This is in contrast with (3.13) which has a central χ^2 distribution under the null hypothesis. As $(T-k)\hat{\sigma}_T^2/\sigma_o^2$ is still distributed as $\chi^2(T-k)$ by Theorem 3.7(b), the assertion follows because the numerator and denominator of φ are independent. □

Clearly, when the null hypothesis is correct, we have $\boldsymbol{\delta} = \boldsymbol{0}$, so that $\varphi \sim F(q, T-k)$. Theorem 3.14 thus includes Theorem 3.12 as a special case. In particular, for testing a single hypothesis, we have

$$\tau \sim t(T-k;\, \boldsymbol{D}^{-1/2}\boldsymbol{\delta}),$$

which reduces to $t(T-k)$ when $\boldsymbol{\delta} = 0$, as in Theorem 3.9.

Theorem 3.14 implies that when $\boldsymbol{R}\boldsymbol{\beta}_o$ deviates farther from the hypothetical value $\boldsymbol{r}$, the non-centrality parameter $\boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}$ increases, and so does the power. We illustrate this point using the following two examples, where the power are computed using the GAUSS program. For the null distribution $F(2, 20)$, the critical value at 5% level is 3.49. Then for $F(2, 20; \nu_1, 0)$ with the non-centrality parameter $\nu_1 = 1, 3, 5$, the probabilities that φ exceeds 3.49 are approximately 12.1%, 28.2%, and 44.3%, respectively. For the null distribution $F(5, 60)$, the critical value at 5% level is 2.37. Then for $F(5, 60; \nu_1, 0)$ with $\nu_1 = 1, 3, 5$, the probabilities that φ exceeds 2.37 are approximately 9.4%, 20.5%, and 33.2%, respectively. In both cases, the power increases with the non-centrality parameter.

3.3.3 An Alternative Approach

Given the specification (3.1), we may take the constraint $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$ into account and consider the *constrained* OLS estimation that finds the saddle point of the Lagrangian:

$$\min_{\boldsymbol{\beta},\boldsymbol{\lambda}} \frac{1}{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r})'\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is the $q \times 1$ vector of Lagrangian multipliers. It is straightforward to show that the solutions are

$$\begin{aligned} \ddot{\boldsymbol{\lambda}}_T &= 2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}), \\ \ddot{\boldsymbol{\beta}}_T &= \hat{\boldsymbol{\beta}}_T - (\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T/2, \end{aligned} \tag{3.15}$$

which will be referred to as the constrained OLS estimators.

Given $\ddot{\boldsymbol{\beta}}_T$, the vector of constrained OLS residuals is

$$\ddot{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\ddot{\boldsymbol{\beta}}_T = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_T + \boldsymbol{X}(\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T) = \hat{\boldsymbol{e}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T).$$

It follows from (3.15) that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T &= (\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T/2 \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}). \end{aligned}$$

The inner product of $\ddot{\boldsymbol{e}}$ is then

$$\begin{aligned} \ddot{\boldsymbol{e}}'\ddot{\boldsymbol{e}} &= \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} + (\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T)'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T) \\ &= \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} + (\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}). \end{aligned}$$

Note that the second term on the right-hand side is nothing but the numerator of the F statistic (3.14). The F statistic now can be written as

$$\varphi = \frac{\ddot{\boldsymbol{e}}'\ddot{\boldsymbol{e}} - \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{q\hat{\sigma}_T^2} = \frac{(\mathrm{ESS_c} - \mathrm{ESS_u})/q}{\mathrm{ESS_u}/(T-k)}, \tag{3.16}$$

where $\mathrm{ESS_c} = \ddot{\boldsymbol{e}}'\ddot{\boldsymbol{e}}$ and $\mathrm{ESS_u} = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$ denote, respectively, the ESS resulted from constrained and unconstrained estimations. Dividing the numerator and denominator of (3.16) by centered TSS ($\boldsymbol{y}'\boldsymbol{y} - T\bar{y}^2$) yields another equivalent expression for φ:

$$\varphi = \frac{(R_{\mathrm{u}}^2 - R_{\mathrm{c}}^2)/q}{(1 - R_{\mathrm{u}}^2)/(T-k)}, \tag{3.17}$$

where R_c^2 and R_u^2 are, respectively, the centered coefficient of determination of constrained and unconstrained estimations. As the numerator of (3.17), $R_{\mathrm{u}}^2 - R_{\mathrm{c}}^2$, can be interpreted as the loss of fit due to the imposed constraint, the F test is in effect a loss-of-fit test. The null hypothesis is rejected when the constrained specification fits data much worse.

Example 3.15 Consider the specification: $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + e_t$. Given the hypothesis (constraint) $\beta_2 = \beta_3$, the resulting constrained specification is

$$y_t = \beta_1 + \beta_2(x_{t2} + x_{t3}) + e_t.$$

By estimating these two specifications separately, we obtain $\mathrm{ESS_u}$ and $\mathrm{ESS_c}$, from which the F statistic can be easily computed. □

Example 3.16 Test the null hypothesis that all the coefficients (except the constant term) equal zero. The resulting constrained specification is $y_t = \beta_1 + e_t$, so that $R_{\mathrm{c}}^2 = 0$. Then, (3.17) becomes

$$\varphi = \frac{R_{\mathrm{u}}^2/(k-1)}{(1-R_{\mathrm{u}}^2)/(T-k)} \sim F(k-1, T-k),$$

which requires only estimation of the unconstrained specification. This test statistic is also routinely reported by most of econometrics and statistics packages and known as the "regression F test." □

3.4 Confidence Regions

In addition to point estimators for parameters, we may also be interested in finding confidence intervals for parameters. A confidence interval for $\beta_{i,o}$ with the confidence coefficient $(1-\alpha)$ is the interval $(\underline{g}_\alpha, \overline{g}_\alpha)$ that satisfies

$$\mathbb{P}\{\, \underline{g}_\alpha \leq \beta_{i,o} \leq \overline{g}_\alpha\} = 1-\alpha.$$

That is, we are $(1-\alpha)\times 100$ percent sure that such an interval would include the true parameter $\beta_{i,o}$.

From Theorem 3.9, we know

$$\mathbb{P}\left\{-t_{\alpha/2}(T-k) \leq \frac{\hat{\beta}_{i,T}-\beta_{i,o}}{\hat{\sigma}_T\sqrt{m^{ii}}} \leq t_{\alpha/2}(T-k)\right\} = 1-\alpha,$$

where m^{ii} is the i^{th} diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, and $t_{\alpha/2}(T-k)$ is the critical value of the (two-sided) t test at the significance level α. Equivalently, we have

$$\mathbb{P}\left\{\hat{\beta}_{i,T} - t_{\alpha/2}(T-k)\hat{\sigma}_T\sqrt{m^{ii}} \leq \beta_{i,o} \leq \hat{\beta}_{i,T} + t_{\alpha/2}(T-k)\hat{\sigma}_T\sqrt{m^{ii}}\right\} = 1-\alpha.$$

This shows that the confidence interval for $\beta_{i,o}$ can be constructed by setting

$$\underline{g}_\alpha = \hat{\beta}_{i,T} - t_{\alpha/2}(T-k)\hat{\sigma}_T\sqrt{m^{ii}},$$
$$\overline{g}_\alpha = \hat{\beta}_{i,T} + t_{\alpha/2}(T-k)\hat{\sigma}_T\sqrt{m^{ii}}.$$

It should be clear that the greater the confidence coefficient (i.e., α smaller), the larger is the magnitude of the critical values $\pm t_{\alpha/2}(T-k)$ and hence the resulting confidence interval.

The confidence region for $\boldsymbol{R}\boldsymbol{\beta}_o$ with the confidence coefficient $(1-\alpha)$ satisfies

$$\mathbb{P}\{(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)'\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)/(q\hat{\sigma}_T^2) \leq F_\alpha(q, T-k)\}$$
$$= 1-\alpha,$$

where $F_\alpha(q, T-k)$ is the critical value of the F test at the significance level α.

Example 3.17 The confidence region for $(\beta_{1,o} = b_1, \beta_{2,o} = b_2)$. Suppose $T - k = 30$ and $\alpha = 0.05$, then $F_{0.05}(2, 30) = 3.32$. In view of Example 3.13,

$$\mathbb{P}\left\{ \frac{1}{2\hat{\sigma}_T^2} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix}' \begin{bmatrix} m^{11} & m^{12} \\ m^{21} & m^{22} \end{bmatrix}^{-1} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix} \leq 3.32 \right\} = 0.95,$$

which results in an ellipse with the center $(\hat{\beta}_{1,T}, \hat{\beta}_{2,T})$. □

Remark: A point $(\beta_{1,o}, \beta_{2,o})$ may be outside the joint confidence ellipse but inside the confidence box formed by individual confidence intervals. Hence, each t ratio may show that the corresponding coefficient is insignificantly different from zero, while the F test indicates that both coefficients are not jointly insignificant. It is also possible that (β_1, β_2) is outside the confidence box but inside the joint confidence ellipse. That is, each t ratio may show that the corresponding coefficient is significantly different from zero, while the F test indicates that both coefficients are jointly insignificant. See also an illustrative example in Goldberger (1991, Chap. 19).

3.5 Multicollinearity

In Section 3.1.2 we have seen that a linear specification suffers from the problem of exact multicollinearity if the basic identifiability requirement (i.e., $\boldsymbol{X}$ is of full column rank) is not satisfied. In this case, the OLS estimator cannot be computed as (3.4). This problem may be avoided by modifying the postulated specifications.

3.5.1 Near Multicollinearity

In practice, it is more common that explanatory variables are related to some extent but do not satisfy an exact linear relationship. This is usually referred to as the problem of *near multicollinearity*. But as long as there is no exact multicollinearity, parameters can still be estimated by the OLS method, and the resulting estimator remains the BLUE under [A1] and [A2].

Nevertheless, there are still complaints about near multicollinearity in empirical studies. In some applications, parameter estimates are very sensitive to small changes in data. It is also possible that individual t ratios are all insignificant, but the regression F statistic is highly significant. These symptoms are usually attributed to near multicollinearity. This is not entirely correct, however. Write $\boldsymbol{X} = [\boldsymbol{x}_i \ \boldsymbol{X}_i]$, where $\boldsymbol{X}_i$ is the submatrix of $\boldsymbol{X}$ excluding the i^{th} column $\boldsymbol{x}_i$. By the result of Theorem 3.3, the variance of $\hat{\beta}_{i,T}$ can be expressed as

$$\operatorname{var}(\hat{\beta}_{i,T}) = \operatorname{var}\big([\boldsymbol{x}_i'(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{x}_i]^{-1}\boldsymbol{x}_i'(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{y}\big) = \sigma_o^2[\boldsymbol{x}_i'(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{x}_i]^{-1},$$

where $\boldsymbol{P}_i = \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'$. It can also be verified that

$$\text{var}(\hat{\beta}_{i,T}) = \frac{\sigma_o^2}{\sum_{t=1}^T (x_{ti} - \bar{x}_i)^2[1 - R^2(i)]},$$

where $R^2(i)$ is the centered coefficient of determination from the auxiliary regression of $\boldsymbol{x}_i$ on $\boldsymbol{X}_i$. When $\boldsymbol{x}_i$ is closely related to other explanatory variables, $R^2(i)$ is high so that $\text{var}(\hat{\beta}_{i,T})$ would be large. This explains why $\hat{\beta}_{i,T}$ are sensitive to data changes and why corresponding t ratios are likely to be insignificant. Near multicollinearity is not a necessary condition for these problems, however. Large $\text{var}(\hat{\beta}_{i,T})$ may also arise due to small variations of x_{ti} and/or large σ_o^2.

Even when a large value of $\text{var}(\hat{\beta}_{i,T})$ is indeed resulted from high $R^2(i)$, there is nothing wrong statistically. It is often claimed that "severe multicollinearity can make an important variable look insignificant." As Goldberger (1991) correctly pointed out, this statement simply confuses statistical significance with economic importance. These large variances merely reflect the fact that parameters cannot be precisely estimated from the given data set.

Near multicollinearity is in fact a problem related to data and model specification. If it does cause problems in estimation and hypothesis testing, one may try to break the approximate linear relationship by, e.g., adding more observations to the data set (if plausible) or dropping some variables from the current specification. More sophisticated statistical methods, such as the ridge estimator and principal component regressions, may also be used; we omit the details of these methods.

3.5.2 Digression: Dummy Variables

A linear specification may include some qualitative variables to indicate the presence or absence of certain attributes of the dependent variable. These qualitative variables are typically represented by *dummy variables* which classify data into different categories.

For example, let y_t denote the annual salary of college teacher t and x_t the years of teaching experience of t. Consider the dummy variable: $D_t = 1$ if t is a male and $D_t = 0$ if t is a female. Then, the specification

$$y_t = \alpha_0 + \alpha_1 D_t + \beta x_t + e_t$$

yields two regression lines with different intercepts. The "male" regression line has the intercept $\alpha_0 + \alpha_1$, and the "female" regression line has the intercept α_0. We may test the hypothesis $\alpha_1 = 0$ to see if there is a difference between the starting salaries of male and female teachers.

This specification can be expanded to incorporate an interaction term between D and x:

$$y_t = \alpha_0 + \alpha_1 D_t + \beta_0 x_t + \beta_1 (D_t x_t) + e_t,$$

which yields two regression lines with different intercepts and slopes. The slope of the "male" regression line is mow $\beta_0 + \beta_1$, whereas the slope of the "female" regression line is β_0. By testing $\beta_1 = 0$, we can check whether teaching experience is treated the same in determining salaries for male and female teachers.

In the analysis of quarterly data, it is also common to include the *seasonal dummy variables* D_{1t}, D_{2t} and D_{3t}, where for $i = 1, 2, 3$, $D_{it} = 1$ if t is the observation of the i^{th} quarter and $D_{it} = 0$ otherwise. Similar to the previous example, the following specification,

$$y_t = \alpha_0 + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \beta x_t + e_t,$$

yields four regression lines. The regression line for the data of the i^{th} quarter has the intercept $\alpha_0 + \alpha_i$, $i = 1, 2, 3$, and the regression line for the fourth quarter has the intercept α_0. Including seasonal dummies allows us to classify the levels of y_t into four seasonal patterns. Various interesting hypotheses can be tested based on this specification. For example, one may test the hypotheses that $\alpha_1 = \alpha_2$ and $\alpha_1 = \alpha_2 = \alpha_3 = 0$. By the Frisch-Waugh-Lovell theorem we know that the OLS estimate of β can also be obtained from regressing y_t^* on x_t^*, where y_t^* and x_t^* are the residuals of regressing, respectively, y_t and x_t on seasonal dummies. Although some seasonal effects of y_t and x_t may be eliminated by the regressions on seasonal dummies, their residuals y_t^* on x_t^* are not the so-called "seasonally adjusted data" which are usually computed using different methods or algorithms.

Remark: The preceding examples show that, when a specification contains a constant term, the number of dummy variables is always one *less* than the number of categories that dummy variables intend to classify. Otherwise, the specification would have exact multicollinearity; this is known as the "dummy variable trap."

3.6 Limitations of the Classical Conditions

The previous estimation and testing results are based on the classical conditions. As these conditions may be violated in practice, it is important to understand their limitations.

Condition [A1] postulates that explanatory variables are non-stochastic. Although this condition is quite convenient and facilitates our analysis, it is not practical. When the dependent variable and regressors are economic variables, it does not make too much sense to treat only the dependent variable as a random variable. This condition may also be

violated when a lagged dependent variable is included as a regressor, as in many time-series analysis. Hence, it would be more reasonable to allow regressors to be random as well.

In [A2](i), the linear specification $\boldsymbol{X}\boldsymbol{\beta}$ is assumed to be correct up to some unknown parameters. It is possible that the systematic component $\mathbb{E}(\boldsymbol{y})$ is in fact a non-linear function of $\boldsymbol{X}$. If so, the estimated regression hyperplane could be very misleading. For example, an economic relation may change from one regime to another at some time point so that $\mathbb{E}(\boldsymbol{y})$ is better characterized by a piecewise liner function. This is known as the problem of *structural change*; see e.g., Exercise 3.14. Even when $\mathbb{E}(\boldsymbol{y})$ is a linear function, the specified $\boldsymbol{X}$ may include some irrelevant variables or omit some important variables. Example 3.6 shows that in the former case, the OLS estimator $\hat{\boldsymbol{\beta}}_T$ remains unbiased but is less efficient. In the latter case, it can be shown that $\hat{\boldsymbol{\beta}}_T$ is biased but with a smaller variance-covariance matrix; see Exercise 3.6.

Condition [A2](ii) may also easily break down in many applications. For example, when y_t is the consumption of the t^{th} household, it is likely that y_t has smaller variation for low-income families than for high-income families. When y_t denotes the GDP growth rate of the t^{th} year, it is also likely that y_t are correlated over time. In both cases, the variance-covariance matrix of $\boldsymbol{y}$ cannot be expressed as $\sigma_o^2\boldsymbol{I}_T$. A consequence of the failure of [A2](ii) is that the OLS estimator for $\text{var}(\hat{\boldsymbol{\beta}}_T)$, $\hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, is biased, which in turn renders the tests discussed in Section 3.3 invalid.

Condition [A3] may fail when y_t have non-normal distributions. Although the BLUE property of the OLS estimator does not depend on normality, [A3] is crucial for deriving the distribution results in Section 3.3. When [A3] is not satisfied, the usual t and F tests do not have the desired t and F distributions, and their exact distributions are typically unknown. This causes serious problems for hypothesis testing.

Our discussion thus far suggests that the classical conditions are quite restrictive. In subsequent chapters, we will try to relax these conditions and discuss more generally applicable methods. These methods play an important role in contemporary empirical studies.

Exercises

3.1 Construct a linear regression model for each equation below:

$$y = \alpha\, x^{\beta}, \qquad y = \alpha\, e^{\beta x}, \qquad y = \frac{x}{\alpha x - \beta}, \qquad y = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

3.2 Use the general formula (3.4) to find the OLS estimators from the specifications below:

$$y_t = \alpha + \beta x_t + e,\ t = 1, \ldots, T,$$
$$y_t = \alpha + \beta(x_t - \bar{x}) + e,\ t = 1, \ldots, T,$$
$$y_t = \beta x_t + e,\ t = 1, \ldots, T.$$

Compare the resulting regression lines.

3.3 Given the specification $y_t = \alpha + \beta x_t + e,\ t = 1, \ldots, T$, assume that the classical conditions hold. Let $\hat{\alpha}_T$ and $\hat{\beta}_T$ be the OLS estimators for α and β, respectively.

(a) Apply the general formula of Theorem 3.4(c) to show that

$$\begin{aligned}
\operatorname{var}(\hat{\alpha}_T) &= \sigma_o^2 \frac{\sum_{t=1}^T x_t^2}{T\sum_{t=1}^T (x_t - \bar{x})^2}, \\
\operatorname{var}(\hat{\beta}_T) &= \sigma_o^2 \frac{1}{\sum_{t=1}^T (x_t - \bar{x})^2}, \\
\operatorname{cov}(\hat{\alpha}_T, \hat{\beta}_T) &= -\sigma_o^2 \frac{\bar{x}}{\sum_{t=1}^T (x_t - \bar{x})^2}.
\end{aligned}$$

What kind of data can make the variances of the OLS estimators smaller?

(b) Suppose that a prediction $\hat{y}_{T+1} = \hat{\alpha}_T + \hat{\beta}_T x_{T+1}$ is made based on the new observation x_{T+1}. Show that

$$\begin{aligned}
\mathbb{E}(\hat{y}_{T+1} - y_{T+1}) &= 0, \\
\operatorname{var}(\hat{y}_{T+1} - y_{T+1}) &= \sigma_o^2 \left(1 + \frac{1}{T} + \frac{(x_{T+1} - \bar{x})^2}{\sum_{t=1}^T (x_t - \bar{x})^2} \right).
\end{aligned}$$

What kind of x_{T+1} can make the variance of prediction error smaller?

3.4 Given the specification (3.1), suppose that $\boldsymbol{X}$ is not of full column rank. Does there exist a unique $\hat{\boldsymbol{y}} \in \operatorname{span}(\boldsymbol{X})$ that minimizes $(\boldsymbol{y} - \hat{\boldsymbol{y}})'(\boldsymbol{y} - \hat{\boldsymbol{y}})$? If yes, is there a unique $\hat{\boldsymbol{\beta}}_T$ such that $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_T$? Why or why not?

3.5 Given the estimated model

$$y_t = \hat{\beta}_{1,T} + \hat{\beta}_{2,T} x_{t2} + \cdots + \hat{\beta}_{k,T} x_{tk} + \hat{e}_t,$$

consider the *standardized* regression:

$$y_t^* = \hat{\beta}_{2,T}^* x_{t2}^* + \cdots + \hat{\beta}_{k,T}^* x_{tk}^* + \hat{e}_t^*,$$

where $\hat{\beta}^*_{i,T}$ are known as the *beta coefficients*, and

$$y_t^* = \frac{y_t - \bar{y}}{s_y}, \qquad x_{ti}^* = \frac{x_{ti} - \bar{x}_i}{s_{x_i}}, \qquad \hat{e}_t^* = \frac{\hat{e}_t}{s_y},$$

with $s_y^2 = (T-1)^{-1}\sum_{t=1}^T (y_t - \bar{y})^2$ is the sample variance of y_t and for each i, $s_{x_i}^2 = (T-1)^{-1}\sum_{t=1}^T (x_{ti} - \bar{x}_i)^2$ is the sample variance of x_{ti}. What is the relationship between $\hat{\beta}^*_{i,T}$ and $\hat{\beta}_{i,T}$? Give an interpretation of the beta coefficients.

3.6 Given the following specification

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

where $\boldsymbol{X}_1$ $(T \times k_1)$ is a non-stochastic matrix, let $\hat{\boldsymbol{b}}_{1,T}$ denote the resulting OLS estimator. Suppose that $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{X}_2\boldsymbol{b}_2$ for some $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$, where $\boldsymbol{X}_2$ $(T \times k_2)$ is also a non-stochastic matrix, $\boldsymbol{b}_2 \neq \boldsymbol{0}$, and $\mathrm{var}(y) = \sigma_o^2\boldsymbol{I}$.

(a) Is $\hat{\boldsymbol{b}}_{1,T}$ unbiased?

(b) Is $\hat{\sigma}_T^2$ unbiased?

(c) What is $\mathrm{var}(\hat{\boldsymbol{b}}_{1,T})$?

(d) Let $\hat{\boldsymbol{\beta}}_T = (\hat{\boldsymbol{\beta}}_{1,T}' \; \hat{\boldsymbol{\beta}}_{2,T}')'$ denote the OLS estimator obtained from estimating the specification: $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$. Compare $\mathrm{var}(\hat{\boldsymbol{\beta}}_{1,T})$ and $\mathrm{var}(\hat{\boldsymbol{b}}_{1,T})$.

(e) Does your result in (d) change when $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{0}$?

3.7 Given the specification (3.1), will the changes below affect the resulting OLS estimator $\hat{\boldsymbol{\beta}}_T$, t ratios, and R^2?

(a) $\boldsymbol{y}^* = 1000 \times \boldsymbol{y}$ and $\boldsymbol{X}$ are used as the dependent and explanatory variables.

(b) $\boldsymbol{y}$ and $\boldsymbol{X}^* = 1000 \times \boldsymbol{X}$ are used as the dependent and explanatory variables.

(c) $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$ are used as the dependent and explanatory variables.

3.8 Let R_k^2 denote the centered R^2 obtained from the model with k explanatory variables.

(a) Show that

$$R_k^2 = \sum_{i=1}^k \hat{\beta}_{i,T}\,\frac{\sum_{t=1}^T (x_{ti} - \bar{x}_i)y_t}{\sum_{t=1}^T (y_t - \bar{y})^2},$$

where $\hat{\beta}_{i,T}$ is the i^{th} element of $\hat{\boldsymbol{\beta}}_T$, $\bar{x}_i = \sum_{t=1}^T x_{ti}/T$, and $\bar{y} = \sum_{t=1}^T y_t/T$.

(b) Show that $R_k^2 \geq R_{k-1}^2$.

3.9 Consider the following two regression lines: $\hat{y} = \hat{\alpha} + \hat{\beta}x$ and $\hat{x} = \hat{\gamma} + \hat{\delta}y$. At which point do these two lines intersect? Using the result in Exercise 3.8 to show that these two regression lines coincide if and only if the centered R^2s for both regressions are one.

3.10 Given the specification (3.1), suppose that $\boldsymbol{X}$ does *not* contain the constant term. Show that the centered R^2 need not be bounded between zero and one if it is computed as (3.7).

3.11 Rearrange the matrix $\boldsymbol{X}$ as $[\boldsymbol{x}_i \ \boldsymbol{X}_i]$, where $\boldsymbol{x}_i$ is the i^{th} column of $\boldsymbol{X}$. Let $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ denote the residual vectors of regressing $\boldsymbol{y}$ on $\boldsymbol{X}_i$ and $\boldsymbol{x}_i$ on $\boldsymbol{X}_i$, respectively. Define the *partial correlation coefficient* of $\boldsymbol{y}$ and $\boldsymbol{x}_i$ as

$$r_i = \frac{\boldsymbol{u}_i'\boldsymbol{v}_i}{(\boldsymbol{u}_i'\boldsymbol{u}_i)^{1/2}(\boldsymbol{v}_i'\boldsymbol{v}_i)^{1/2}}.$$

Let R_i^2 and R^2 be obtained from the regressions of $\boldsymbol{y}$ on $\boldsymbol{X}_i$ and $\boldsymbol{y}$ on $\boldsymbol{X}$, respectively.

(a) Apply the Frisch-Waugh-Lovell Theorem to show

$$\boldsymbol{I} - \boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{P}_i) - \frac{(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{I} - \boldsymbol{P}_i)}{\boldsymbol{x}_i'(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{x}_i},$$

where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{P}_i = \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'$. This result can also be derived using the matrix inversion formula; see, e.g., Greene (1993. p. 27).

(b) Show that $(1 - R^2)/(1 - R_i^2) = 1 - r_i^2$, and use this result to verify

$$R^2 - R_i^2 = r_i^2(1 - R_i^2).$$

What does this result tell you?

(c) Let τ_i denote the t ratio of $\hat{\beta}_{i,T}$, the i^{th} element of $\hat{\boldsymbol{\beta}}_T$ obtained from regressing $\boldsymbol{y}$ on $\boldsymbol{X}$. First show that $\tau_i^2 = (T - k)r_i^2/(1 - r_i^2)$, and use this result to verify

$$r_i^2 = \tau_i^2/(\tau_i^2 + T - k).$$

(d) Combine the results in (b) and (c) to show

$$R^2 - R_i^2 = \tau_i^2(1 - R^2)/(T - k).$$

What does this result tell you?

3.12 Suppose that a linear model with k explanatory variables has been estimated.

(a) Show that $\hat{\sigma}_T^2 = \text{Centered TSS}(1 - \bar{R}^2)/(T - 1)$. What does this result tell you?

(b) Suppose that we want to test the hypothesis that s coefficients are zero. Show that the F statistic can be written as

$$\varphi = \frac{(T-k+s)\hat{\sigma}_{\mathrm{c}}^2 - (T-k)\hat{\sigma}_{\mathrm{u}}^2}{s\hat{\sigma}_{\mathrm{u}}^2},$$

where $\hat{\sigma}_{\mathrm{c}}^2$ and $\hat{\sigma}_{\mathrm{u}}^2$ are the variance estimates of the constrained and unconstrained models, respectively. Let $a = (T-k)/s$. Show that

$$\frac{\hat{\sigma}_{\mathrm{c}}^2}{\hat{\sigma}_{\mathrm{u}}^2} = \frac{a+\varphi}{a+1}.$$

(c) Based on the results in (a) and (b), what can you say when $\varphi > 1$ and $\varphi < 1$?

3.13 For the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, an alternative expression of $k-m$ linear restrictions on $\boldsymbol{\beta}$ can be expressed as $\boldsymbol{\beta} = \boldsymbol{S}\boldsymbol{\theta} + \boldsymbol{d}$, where $\boldsymbol{\theta}$ is a m-dimensional vector of unknown parameters, $\boldsymbol{S}$ is a $k \times m$ matrix of pre-specified constants with full column rank, and $\boldsymbol{d}$ is a vector of pre-specified constants.

(a) By incorporating this restriction into the specification, find the OLS estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

(b) The *constrained least squares estimator* of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_d = \boldsymbol{S}\hat{\theta} + \boldsymbol{d}$. Show that

$$\hat{\boldsymbol{\beta}}_d = \boldsymbol{Q}_S\hat{\boldsymbol{\beta}} + (\boldsymbol{I} - \boldsymbol{Q}_S)\boldsymbol{d},$$

where $\boldsymbol{Q}_S = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{X}'\boldsymbol{X}$. Is this decomposition orthogonal?

(c) Show that

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}_d = \boldsymbol{P}_{XS}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{P}_{XS})\boldsymbol{X}\boldsymbol{d},$$

where $\boldsymbol{P}_{XS} = \boldsymbol{X}\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{X}'$. Use a graph to illustrate this result.

3.14 (The Chow Test) Consider the model of a one-time structural change at a known change point:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{X}_2 & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_o \\ \boldsymbol{\delta}_o \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{bmatrix},$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are $T_1 \times 1$ and $T_2 \times 1$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $T_1 \times k$ and $T_2 \times k$, respectively. The null hypothesis is $\boldsymbol{\delta}_o = \boldsymbol{0}$. How would you test this hypothesis based on the constrained and unconstrained models?

References

Davidson, Russell and James G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York, NY: Oxford University Press.

Goldberger, Arthur S. (1991). *A Course in Econometrics*, Cambridge, MA: Harvard University Press.

Greene, William H. (2000). *Econometric Analysis*, 4th ed., Upper Saddle River, NJ: Prentice Hall.

Harvey, Andrew C. (1990). *The Econometric Analysis of Time Series*, Second edition., Cambridge, MA: MIT Press.

Intriligator, Michael D., Ronald G. Bodkin, and Cheng Hsiao (1996). *Econometric Models, Techniques, and Applications*, Second edition, Upper Saddle River, NJ: Prentice Hall.

Johnston, J. (1984). *Econometric Methods*, Third edition, New York, NY: McGraw-Hill.

Judge, Georgge G., R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee (1988). *Introduction to the Theory and Practice of Econometrics*, Second edition, New York, NY: Wiley.

Maddala, G. S. (1992). *Introduction to Econometrics*, Second edition, New York, NY: Macmillan.

Manski, Charles F. (1991). Regression, *Journal of Economic Literature*, **29**, 34–50.

Rao, C. Radhakrishna (1973). *Linear Statistical Inference and Its Applications*, Second edition, New York, NY: Wiley.

Ruud, Paul A. (2000). *An Introduction to Classical Econometric Theory*, New York, NY: Oxford University Press.

Theil, Henri (1971). *Principles of Econometrics*, New York, NY: Wiley.