ECONOMETRICS (I)

MEI-YUAN CHEN

Department of Finance
National Chung Hsing University

July 17, 2003

Contents

1 Introduction

Economists have proposed numerous theories to characterize the relationships between economic variables; whether these theories are supported by real world data is an empirical issue. By *econometrics* we mean the application of statistical and mathematical methods to the analysis of economic data, with a purpose of verifying or refuting economic theories. One of the most commonly used econometric techniques is *regression* analysis.

In the nineteenth century, Sir Francis Galton (1822–1911) studied the relationship between the heights of children and their parents. He observed that although tall parents tended to have tall children and short parents tended to have short children, there was a tendency for children's heights to converge toward the average. He termed this as a "regression toward mediocrity". Contemporary regression analysis is concerned with describing and evaluating the relationship between a dependent variable and one or more explanatory variables. This involves formulating an econometric model, estimating its unknown parameters, and drawing statistical inference about the estimated results.

2 Reviews of Statistics

2.1 Random Variables

A *random variable* is a variable whose values are determined by an experiment of chance (i.e., governed by a probability distribution). We use capital letter to denote a random variable and lower case to denote its value.

1. Discrete random variable X.
   - Probability: $\mathrm{P}\{X = x\}$.
   - Probability distribution: $\mathrm{P}\{X \leq a\} = \sum_{\{i:\, x_i \leq a\}} \mathrm{P}\{X = x_i\}$.
2. Continuous random variable X.
   - Probability density function (p.d.f.): $f(x)$.
   - Cumulative distribution function (c.d.f.) $F(a) = \mathrm{P}\{X \leq a\} = \int_{-\infty}^{a} f(x)\, dx$.

The behavior of a random variable is completely determined by its probability density function. *Moments* are numerical measures summarizing certain behavior of a random variable, e.g., *expected value* and *variance*.

1. Expected value: $\mathrm{E}(X) = \mu$.
   - If X is discrete, $\mathrm{E}(X) = \sum_i x_i \mathrm{P}\{X = x_i\}$.
   - If X is continuous, $\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx$.
   - If c is nonstochastic, $\mathrm{E}(c) = c$, and $\mathrm{E}(cX) = c\,\mathrm{E}(X)$.
2. Variance: $\mathrm{var}(X) = \mathrm{E}(X-\mu)^2 = \mathrm{E}(X^2) - \mu^2 = \sigma^2$.
   - If X is discrete, $\mathrm{var}(X) = \sum_i (x_i - \mu)^2\, \mathrm{P}\{X = x_i\}$.
   - If X is continuous, $\mathrm{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$.
   - If c is nonstochastic, $\mathrm{var}(c) = 0$, and $\mathrm{var}(cX) = c^2\,\mathrm{var}(X)$.

The behavior of two (or more) random variables is determined by their joint probability density function.

1. Joint p.d.f. $f_{XY}(x,y) = \mathrm{P}\{X = x, Y = y\}$.
2. Joint c.d.f. $F_{XY}(a,b) = \mathrm{P}\{X \le a, Y \le b\} = \int_{-\infty}^{b} \int_{-\infty}^{a} f_{XY}(x,y)\, dx\, dy$.
3. Marginal p.d.f. $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy$; $f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx$.
4. Conditional p.d.f. $f(x|y) = f_{XY}(x,y)/f_Y(y)$; $f(y|x) = f_{XY}(x,y)/f_X(x)$.
5. If $f_{XY}(x,y) = f_X(x) f_Y(y)$, then X and Y are said to be independent.

The linear association between two random variables are characterized by their *covariance* (or *correlation*).

1. $\mathrm{cov}(X,Y) = \mathrm{E}((X-\mu_X)(Y-\mu_Y)) = \mathrm{E}(XY) - \mu_X\mu_Y = \sigma_{XY}$.
2. $\mathrm{corr}(X,Y) = \sigma_{XY}/(\sigma_X\sigma_Y) = \rho_{XY}$ and $-1 \le \rho_{XY} \le 1$. Note that $\mathrm{corr}(X,Y)$ is nothing but the covariance between Z_X and Z_Y, where $Z_X = [X - \mathrm{E}(X)]/\sqrt{\mathrm{var}(X)}$ and $Z_Y = [Y - \mathrm{E}(Y)]/\sqrt{\mathrm{var}(Y)}$ are Z-scores of X and Y.

3. If X and Y are independent, then $\mathrm{cov}(X, Y) = 0$; the converse is not true.

4. $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$; $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X, Y)$.

Some frequently used random variables are:

- Normal random variable $X \sim N(\mu, \sigma^2)$. $(X - \mu)/\sigma \sim N(0, 1)$.
- If $X_1, \ldots X_m$ are independent $N(0, 1)$, then $Z = \sum_{i=1}^{m} X_i^2 \sim \chi_m^2$.
- If $X \sim N(0, 1)$ and $Y \sim \chi_m^2$ are independent, then $W = X/\sqrt{Y/m} \sim t_m$.
- If $X \sim \chi_n^2$ and $Y \sim \chi_m^2$ are independent, then $U = (X/n)/(Y/m) \sim F_{n,m}$.

2.2 Estimation

Typically, we do not know the population characteristics θ (e.g., mean and variance) because we not know the probabilistic structure governing the random variable. Hence, we collect data to estimate these unknown parameters.

1. Point estimation: An *estimator* is a function (a rule) of sample data; an *estimate* is its particular value. Given a sample $x_1, \ldots, x_n$, an estimator of θ can be represented as $\hat{\theta} = g(x_1, \ldots, x_n)$.

   Examples: Given a sample $(x_1, y_1), \ldots, (x_n, y_n)$:

   - An estimator of mean: the sample average $\bar{x} = \sum_{i=1}^{n} x_i/n$.
   - An estimator of variance: the sample variance $\sum_{i=1}^{n}(x_i - \bar{x})^2/(n - 1)$.
   - An estimator of covariance: the sample covariance $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/(n - 1)$.
   - An estimator of correlation: the sample correlation
   $$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^{1/2}[\sum_{i=1}^{n}(y_i - \bar{y})^2]^{1/2}}.$$

2. Criteria to evaluate an estimator:

   - Unbiasedness: $\mathrm{E}(\hat{\theta}) = \theta$.
   - Efficiency: If $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators, then $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if $\mathrm{var}(\hat{\theta}_1) < \mathrm{var}(\hat{\theta}_2)$.

- Mean Square Error: $\mathrm{E}(\hat{\theta} - \theta)^2$. This criterion allows us to compare biased estimators.

3. Interval estimation: Instead of providing a particular estimate of an unknown parameter, it may be desirable to provide a range of values which may contain the true parameter. To do this, we first specify a *confidence coefficient* γ, which is a probability, say, 0.95. Then construct two functions $g_1(x_1, \ldots, x_n)$ and $g_2(x_1, \ldots x_n)$ such that

$$\mathrm{P}\{g_1(x_1, \cdots, x_n) \leq \theta \leq g_2(x_1, \cdots, x_n)\} = \gamma.$$

The interval (g_1, g_2) is called the *confidence interval*. In words, we are 95% sure that this interval would contain the parameter θ.

Example: x_i are drawn from independent $N(\mu, 1)$. Let $\gamma = 0.95$. Consider an estimator $\bar{x} = \sum_{i=1}^{n} x_i/n$ for μ. It can be verified that $\bar{x} \sim N(\mu, 1/n)$ so that $\sqrt{n}(\bar{x} - \mu) \sim N(0, 1)$. From the table of the standard normal random variable,

$$\mathrm{P}\{-1.96 < \sqrt{n}(\bar{x} - \mu) < 1.96\} = 0.95.$$

Hence, the 95% confidence interval of μ is $(\bar{x} - 1.96/\sqrt{n} < \mu < \bar{x} + 1.96/\sqrt{n})$.

2.3 Hypothesis Testing

Theory (or prior belief) may suggest that the true parameter θ equals a particular value a. Hence, we may be interested in testing the *null* hypothesis $H_0 : \theta = a$ against the *alternative* hypothesis $H_a : \theta \neq a$ (or $\theta > a$).

1. Test statistic T: it typically involves the difference between the estimate and the hypothesized value, e.g., $T = \sqrt{n}(\bar{x} - a)$ is used to test $H_0 : \mu = a$. A test statistic is a random variable, hence has a distribution from which we can check its probability, e.g., $T \sim N(0, 1)$ under the null hypothesis. A large value of T is considered to be improbable, hence suggests rejection of the null hypothesis.

2. Significance level α: a probability that we would tolerate when we incorrectly reject the null hypothesis. (This probability is also known as the type I error.) Given α,

a *critical value* c_α is such that $\mathrm{P}\{|T| > c_\alpha\} = \alpha$. We reject H_0 if the observed value $T = t$ is such that $|t| > c_\alpha$, and we say that $T = t$ is significant at the level α.

3. Power: the probability of rejecting the null hypothesis when it is indeed false. The type II error is the probability of incorrectly accepting the null hypothesis. Hence, the power of a test is (1 − type II error).

4. p-value: given an observed test statistic $T = t$, the probability of observing more extreme T (i.e., $T \geq t$ and $T \leq -t$). That is, the p-value is the "α" at which $T = t$ is just significant.

3 Random Sampling Model, Projection, and Regression

3.1 Random Sampling

Suppose an eonometrician has the observational data

$$\{\boldsymbol{w}_i, i = 1, \ldots, n\} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n\},$$

where each $\boldsymbol{w}_i$ is a vector of numerical values which represent the characteristics of individuals. Typically, the data can be written as

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_n \end{bmatrix} &= \begin{bmatrix} (y_1, \boldsymbol{x}_1) \\ (y_2, \boldsymbol{x}_2) \\ \vdots \\ (y_n, \boldsymbol{x}_n) \end{bmatrix} \quad y_i \in R, \boldsymbol{x}_i \in R^k \\
&= \begin{bmatrix} y_1 & x_{11} & x_{12} & \cdots & x_{1k} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \cdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \\
&= (\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k).
\end{aligned}$$

If this data is *cross-sectional* (data $\mathbf{w}_i, i = 1, \ldots, n$ were observed at a certain time and i represents "individual"), it is reasonable to assume they are mutually independent ("spatial" data are an exception). Furthermore, if the data are symmetrically gathered (e.g., randomly), it is also reasonable to model each observation as a "random draw" from the same probability distribution. Thus, the data are *independent and identical distributed,* or i.i.d. We call this a *random sample.*

3.2 Regression

In regression, we want to find the *central tendency* of the conditional distribution of y given $x = x_i$. A standard measure of central tendency is the *mean.* The conditional analog is the *conditional mean.* Let $f(y, x)$ denote the joint density of (y, x), then the conditional density

$$f(y|x = x_i) = \frac{f(y, x = x_i)}{f_x(x = x_i)}$$

exists, where $f_x(x = x_i) = \int_{-\infty}^{\infty} f(y, x = x_i)dy$ is the marginal density of x at x_i. The conditional mean is defined as the function

$$m(x_i) = E(y|x = x_i) = \int_{-\infty}^{\infty} yf(y|x = x_i)dy.$$

Note that this definition requires the existence of densities. The conditional mean $m(x_i) = E(y|x = x_i)$ is a function, meaning that when x equals x_i, then the expected value of y is $m(x_i)$. Clearly, it is a random variable since it is a function of random variable x_i.

The regression error e_i is defined to be the difference between y_i given at $x = x_i$ and its conditional mean:

$$e = (y|x = x_i) - m(x_i).$$

By construction, this yields the formula

$$(y|x = x_i) = m(x_i) + e. \tag{1}$$

For the joint observed data $(x_i, y_i), i = 1, \ldots, n$, the considered regression can be expressed as

$$y_i = m(x_i) + e_i, \qquad i = 1, \ldots, n.$$

It is worth emphasizing that no assumptions have been imposed to develop (1), other than that (y, x) have a joint distribution and $E|y| < \infty$.

Proposition 3.1 *Properties of the regression errors e_i*

1. $E(e_i|x_i) = 0$.

2. $E(e_i) = 0$.

3. $E[h(x_i)e_i] = \mathbf{0}$ *for all function* $h(\cdot)$.

4. $E(x_ie_i) = \mathbf{0}$.

Proof:

1. By the definition of e_i and the linearity of conditional expectation,

$$\begin{aligned} E(e_i|x_i) &= E[(y_i - m(x_i))|x_i] \\ &= E(y_i|x_i) - E[m(x_i)|x_i] \\ &= m(x_i) - m(x_i), \qquad \text{as}\, E[m(x_i)|x_i] = m(x_i) \\ &= 0. \qquad\qquad \square \end{aligned}$$

2. By the law of iterated expectations and the first result

$$\begin{aligned} E(e_i) &= E[E(e_i|x_i)] \\ &= E[0] = 0. \qquad\qquad \square \end{aligned}$$

3. By essentially the same argument,

$$\begin{aligned} E[h(x_i)e_i] &= E\{E[h(x_i)e_i|x_i]\} \\ &= E\{h(x_i)E[e_i|x_i]\} \\ &= E\{h(x_i) \times 0\} = \mathbf{0}. \qquad\qquad \square \end{aligned}$$

4. Follows from the third result setting $h(x_i) = x_i$. $\qquad\qquad \square$

The final result implies that e_i and x_i are *uncorrelated.* It is important to understand that despite being uncorrelated, in general e_i need not be independent of x_i.

Generally, the following equations

$$\begin{aligned} y_i &= m(x_i) + e_i \\ E(e_i|x_i) &= 0, \forall i, \end{aligned}$$

are often stated jointly as the regression framework. It is important to understand that this is a framework, not a model, because no restrictions have been placed on the joint distribution of the data. These equations hold true by definition. A regression model imposes further restrictions on the joint distribution; most typically, restrictions on the permissible class of regression function $m(x)$.

3.3 Linear Models

While $m(\boldsymbol{x})$ in general can take any shape, a parametric family $\{m(\boldsymbol{x}, \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{R}^k\}$ is typically picked to simplify estimation and interpretation. Sometimes, the form of $m(\boldsymbol{x}, \boldsymbol{\beta})$ is given by an economic theory or model. Most often, however, we consider a linear form for convenience and data coherence.

A linear model for $m(\boldsymbol{x})$ is written as

$$m(\boldsymbol{x}_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is the parameter vector. In matrix notation,

$$m(\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta},$$

where $\boldsymbol{x}_i = (1, x_{2i}, \ldots, x_{ki})'$. Then the linear regression model becomes

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \qquad (2) \\ E(e_i|\boldsymbol{x}_i) &= 0 \end{aligned}$$

This is a model because $m(\cdot)$ has been restricted to the linear form.

While linearity is substantively restricted, it has still a great deal of flexibility. For example, if $\boldsymbol{x}_i$ is real-valued and

$$m(\boldsymbol{x}_i) = \beta_1 + x_i\beta_2 + x_i^2\beta_3 + \cdots + x_i^{k-1}\beta_k$$

is a polynomial, then a linear regression model still holds, by the redefinition of $\boldsymbol{x}_i$ as $(1, x_i, x_i^2, \ldots, x_i^{k-1})'$. The linear conditional mean model is illustrated in the following figure.

3.4 Linear Projection

The linear regression model (2) implies $E(\boldsymbol{x}_i e_i) = \mathbf{0}$ as

$$\begin{aligned} E(\boldsymbol{x}_i e_i) &= E[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})] \\ &= E\{E[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})|\boldsymbol{x}_i]\} \\ &= E\{\boldsymbol{x}_i[E(y_i|\boldsymbol{x}_i) - \boldsymbol{x}_i'\boldsymbol{\beta}]\} \\ &= E(\boldsymbol{x}_i \times 0) = \mathbf{0}. \end{aligned}$$

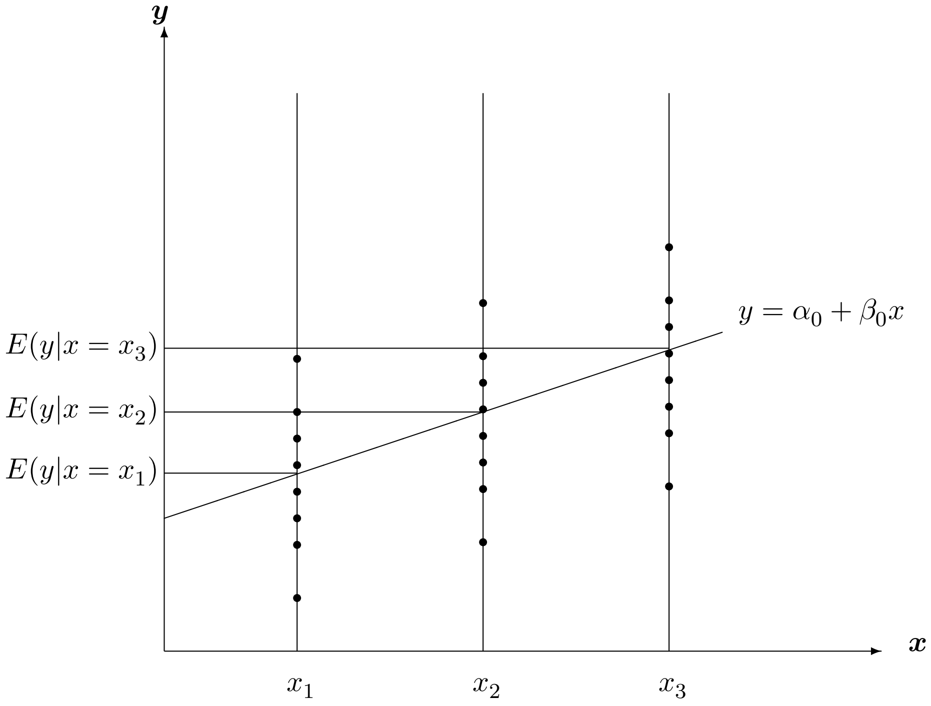

Figure 1: An illustration of the Linear Conditional Mean.

This condition is sufficient for many asymptotic results. It is interesting to observe that in linear models, there is always a vector $\boldsymbol{\beta}$ such that this equation holds. This vector $\boldsymbol{\beta}$ may be called the *linear projection coefficient* or *linear predictor*.

Proposition 3.2 *For any random variables* $(y_i, \boldsymbol{x}_i)$, *let*

$$\beta = [E(\boldsymbol{x}_i\boldsymbol{x}_i')]^{-1}E(\boldsymbol{x}_i y_i) \tag{3}$$

and

$$e_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Then

$$E(\boldsymbol{x}_i e_i) = \mathbf{0}.$$

Proof:

$$\begin{aligned} E(\boldsymbol{x}_i e_i) &= E[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})] \\ &= E\{\boldsymbol{x}_i[y_i - \boldsymbol{x}_i'E(\boldsymbol{x}_i\boldsymbol{x}_i')^{-1}E(\boldsymbol{x}_i y_i)]\} \end{aligned}$$

$$
\begin{aligned}
&= E\{E\{\boldsymbol{x}_i[y_i - \boldsymbol{x}_i' E(\boldsymbol{x}_i\boldsymbol{x}_i')^{-1}E(\boldsymbol{x}_i y_i)]|\boldsymbol{x}_i\}\} \\
&= E\{\boldsymbol{x}_i E(y_i|\boldsymbol{x}_i) - \boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{x}_i\boldsymbol{x}_i')^{-1}\boldsymbol{x}_i E(y_i|\boldsymbol{x}_i)\} \\
&= \mathbf{0}. \qquad\qquad \Box
\end{aligned}
$$

If $\boldsymbol{\beta}$ is defined as in (3), then $E(\boldsymbol{x}_i e_i) = \mathbf{0}$ holds by construction. It does not necessarily follow that $E(e_i|\boldsymbol{x}_i) = 0$. This only holds if the true conditional mean of y_i is $\boldsymbol{x}_i'\boldsymbol{\beta}$, i.e., $m(\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, which is substantive restriction. Thus the linear regression assumption that $E(e_i|\boldsymbol{x}_i) = 0$ is more restrictive than the linear projection construction. It turns out that for most issues in statistical inferences, the projection assumption is sufficient. Therefore, the more general assumption $E(\boldsymbol{x}_i e_i) = \mathbf{0}$ is adopted.

For econometric practice, however, it is typical desirable for $\boldsymbol{x}_i'\boldsymbol{\beta}$ to represent the conditional mean of y_i, rather than a simple linear projection. So while it is not necessary for inference on $\boldsymbol{\beta}$, it may be necessary for inference on an economic relationship of interest.

3.5 Assumptions on the Regression Errors

While the regression motivation leads naturally to the model (2), at times it is more convenient to adopt assumptions which are either more restrictive or less restrictive. The standard types of models considered by econometricians and their strength and weakness are discussed as the follows. All the models are based on the decomposition

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i. \tag{4}$$

In addition, all models normalized the error so that $E(e_i) = 0$ and presume a finite variance $E(e_i^2) = \sigma^2 < \infty$.

Definition 3.1 *The Linear Projection Model is (4) plus*

$$E(\boldsymbol{x}_i e_i) = \mathbf{0}.$$

The advantage of he linear projection model is that it is true by construction, and many inferential results hold under this broad condition. The disadvantage is that the coefficient vector β may not have useful economic interpretations without additional structure.

Definition 3.2 *The Linear Regression Model is (4) plus*

$$E(e_i|\boldsymbol{x}_i) = 0.$$

This model leads naturally from the derivation of the conditional mean function. The primary advantage is that the parameter β is easily interpretable.

Definition 3.3 *The Homoskedastic Regression Model is the Linear Regression Model plus*

$$E(e_i^2|\boldsymbol{x}_i) = \sigma^2. \tag{5}$$

This model adds the auxiliary assumption (5) that the regression is *conditionally homoskedastic.* This assumption greatly simplifies many theoretical arguments and calculations, and it therefore very useful in illustrative arguments. Many formulae simplify under this assumption, and as a result, alternative estimators and techniques are utilized. The danger in this assumption is that these simplifications result in incorrect answers and inferences if indeed the homoskedasticity assumption is false.

Another meaningful justification for assumption (5) is that while it may not be precisely true in the data, it may be approximately true, and in some applications the cost of imposing homoskedasticity on the estimates may be less than the cost of using the more general techniques appropriate for the linear regression model.

Definition 3.4 *The Classical Regression Model is (4) plus that e_i is independent of $\boldsymbol{x}_i$. Usually, $\boldsymbol{x}_i$ is assumed to be nonstochastic.*

This model is more restrictive than the homoskedastic regression model, and is a common starting point in classical econometrics textbooks.

Definition 3.5 *The Normal Regression Model is (4) plus that e_i is independent of $\boldsymbol{x}_i$ and distributed as $N(0, \sigma^2)$.*

The above five models are strictly nested, with the first (the linear projection model) the less restrictive, and the last (the normal regression model) the most restrictive.

The conditional variance function is

$$\operatorname{var}(y_i|\boldsymbol{x} = \boldsymbol{x}_i) = E(e_i^2|\boldsymbol{x} = \boldsymbol{x}_i) = \sigma^2(\boldsymbol{x}_i)$$

which is (potentially) a function of $\boldsymbol{x}_i$. Just as the conditional mean function may take any form, so may the conditional variance function (other than the restriction that it is non-negative). Given the random variable x_i, the conditional variance is $\sigma_i^2 = \sigma^2(\boldsymbol{x}_i)$.

In the general case where $\sigma^2(\boldsymbol{x}_i)$ is not a constant function, so σ_i^2 is different across i, we say that the error e_i is *heteroskedastic*. On the other hand, when the function $\sigma^2(\boldsymbol{x}_i)$ is a constant so that the conditional variance σ_i^2 all equal the same constant value σ^2, we say that the error e_i is *homoskedastic*.

4 Classical Linear Regression Models

In classical analysis, the tools of regression has been applied to study how response variable y is affected by the independent variables $\boldsymbol{x}$ of an experiment in the Lab. Usually, the values of $\boldsymbol{x}$ are designed by scientists so that they are controllable. Therefore, $\boldsymbol{x}$ are also called the *controlled variables* or *designed variables.* Thus, the explanatory variables are assumed to be "nonstochastic" in the classical regression analysis.

4.1 Simple Linear Regression

It is typical and convenient to describe an economic relationship using a linear model. Hence, given a set of economic data, one would like to find a linear equation (straight line) that best fits the data.

We know that a good estimator is the one has smallest mean squared errors. In regression analysis, we are trying to estimate the dependent variable y with a set of explanatory variables x. That is, we want to find an estimator $\hat{y} = f(x)$ to estimate y. Then the mean squared errors of $\hat{y}$ is $\text{mse}(\hat{y}) = E[(\hat{y}-y)^2]$. As we know that the arithmetic average of sample observations is nothing but the expectation value evaluated with the sample relative frequencies as its probabilities. Therefore, the sample counter part of the $\text{mse}(\hat{y})$ is the arithmetic average, $\sum_{i=1}^n (y_i - f(x_i))^2/n$. In linear regression content, the estimator $f(x)$ is restricted to a linear function form, i.e., $f(x) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. The discussion of simple linear regression focuses on $k = 2$. Thus, we want to find an estimator which makes $\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2/n$ as small as possible. This is the ordinary least square estimator we are going to discuss.

Given the linear conditional mean $\text{E}(y|x) = \alpha_0 + \beta_0 x$ is assumed (usually it is unknown) from an economic or financial theory, a linear regression model is specified as $y = \alpha + \beta x + u$. Under the "believe" of the "representativity" on obtained sample observations $\{x_i, y_i\}_{i=1}^n$, the relations $y_i = \alpha_0 + \beta_0 x_i + e_i, i = 1, \ldots, n$ are believed and the regression model is appropriate for sample observations, i.e., $y_i = \alpha + \beta x_i + u_i, i = 1, \ldots, n$. The relations $y_i = \alpha_0 + \beta_0 x_i + e_i, i = 1, \ldots, n$ is sometimes called the "identification" of relation between y and x, denoted as ID1 hereafter. That is

ID 1: $y_i = \alpha_0 + \beta_0 x_i + e_i, i = 1, \ldots, n.$

The OLS estimators of α and β are obtained by minimizing the average of squared errors:

$$f(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

The first order conditions are

$$\begin{aligned} \frac{\partial}{\partial\alpha} f(\alpha,\beta) &= -2\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha}_n - \hat{\beta}_n x_i) = 0, && (6) \\ \frac{\partial}{\partial\alpha} f(\alpha,\beta) &= -2\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha}_n - \hat{\beta}_n x_i)x_i = 0, && (7) \end{aligned}$$

which are also called "normal equations". From (6) we obtain

$$\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_n \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y}_n - \hat{\beta}_n \bar{x}_n, \tag{8}$$

and by plugging this $\hat{\alpha}_n$ into (7) we get

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

so that

$$\hat{\beta}_n \left(\frac{1}{n}\sum_{i=1}^{n} x_i(x_i - \bar{x}_n)\right) = \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \bar{y}_n). \tag{9}$$

It follows from (8) and (9) that the OLS estimators of α and β are

$$\begin{aligned} \hat{\beta}_n &= \frac{\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}, && (10) \\ \hat{\alpha}_n &= \bar{y}_n - \hat{\beta}_n \bar{x}_n. && (11) \end{aligned}$$

Note that $\hat{\beta}_n$ exists uniquely if $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ is not equal to zero "deterministically". It is obvious $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = 0$ if all x_is are constant and also could be zero when x_i is stochastic. Therefore, we have to impose the following assumption to have $\hat{\beta}_n$ uniquely and deteriministically,

A1: $x_i, i = 1, \ldots, n$ are not all constant and nonstochastic.

The equation $\hat{y} = \hat{\alpha}_n + \hat{\beta}_n x$ is the *regression line.* The values $\hat{y}_i$ are called *fitted* values, and $e_i = y_i - \hat{y}_i$ are called *residuals.* Note that by normal equations (6), $\sum_{i=1}^{n} e_i = 0$ so that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$ and $\bar{y}_n = \bar{\hat{y}}$. Also, note that by (7), $\sum_{i=1}^{n} x_i e_i = 0$ so that $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.

The OLS estimators have the following properties under some appropriate assumptions.

1. Given **A1** OLS estimators are *linear* estimators in y_i, i.e., $\hat{\beta}_n = \sum_{i=1}^n k_i y_i$ and $\hat{\alpha}_n = \sum_{i=1}^n h_i y_i$.

$$\begin{aligned}\hat{\beta}_n &= \frac{\sum_{i=1}^n (x_i - \bar{x}_n) y_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\ &= \sum_{i=1}^n \frac{x_i - \bar{x}_n}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} y_i \\ &= \sum_{i=1}^n k_i y_i.\end{aligned}$$

and

$$\begin{aligned}\hat{\alpha}_n &= \bar{y}_n - \hat{\beta}_n \bar{x}_n \\ &= \sum_{i=1}^n y_i/n - \sum_{i=1}^n k_i y_i \bar{x}_n \\ &= \sum_{i=1}^n \left(\frac{1}{n} - k_i \bar{x}_n\right) y_i \\ &= \sum_{i=1}^n h_i y_i.\end{aligned}$$

Note that $\sum_{i=1}^n k_i = 0$ and $\sum_{i=1}^n k_i^2 = 1/\sum_{i=1}^n (x_i - \bar{x}_n)^2$.

2. Given **ID** 1: $y_i = \alpha_0 + \beta_0 x_i + e_i, i = 1, \ldots, n$ and **A1**, the OLS estimators are conditional *unbiased*.

First observe that, denote $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)'$,

$$\begin{aligned}\hat{\beta}_n &= \frac{\sum_{i=1}^n (x_i - \bar{x}_n) y_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(\alpha_0 + \beta_0 x_i + e_i)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \qquad \text{under ID1} \\ &= \alpha_0 \frac{\sum_{i=1}^n (x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} + \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x}_n) x_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} + \frac{\sum_{i=1}^n (x_i - \bar{x}_n) e_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\ &= \beta_0 + \frac{\sum_{i=1}^n (x_i - \bar{x}_n) e_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\ &= \beta_0 + \sum_{i=1}^n k_i \, e_i.\end{aligned}$$

To prove the unbiasedness, take expectation to both sides in above equation,

$$\begin{aligned}
E(\hat{\beta}_n|\boldsymbol{X}) &= E\left(\beta_0 + \frac{\sum_{i=1}^n (x_i - \bar{x}_n)e_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}|\boldsymbol{X}\right) \\
&= \beta_0 + E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}_n)e_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}|\boldsymbol{X}\right) \\
&= \beta_0 + \frac{\sum_{i=1}^n (x_i - \bar{x}_n)E(e_i|\boldsymbol{X})}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\
&= \beta_0.
\end{aligned}$$

As, by iterated expectation,

$$E(\hat{\beta}_n) = E\{E[\hat{\beta}_n|\boldsymbol{X}]\} = E\{\beta_0\} = \beta_0.$$

That is, $\hat{\beta}_n$ is also unconditional unbiased.

Besides, it can be seen that

$$\begin{aligned}
E(\hat{\alpha}_n|\boldsymbol{X}) &= E(\bar{y}_n - \hat{\beta}_n\bar{x}_n|\boldsymbol{X}) \\
&= E(\sum_{i=1}^n y_i/n - \hat{\beta}_n\bar{x}_n|\boldsymbol{X}) \\
&= E(\sum_{i=1}^n (\alpha_0 + \beta_0 x_i + \epsilon_i)/n - \hat{\beta}_n\bar{x}_n|\boldsymbol{X}) \\
&= E(\alpha_0 + (\beta_0 - \hat{\beta}_n)\bar{x}_n + \sum_{i=1}^n e_i/n|\boldsymbol{X}) \\
&= \alpha_0 + E(\beta_0 - \hat{\beta}_n)\bar{x}_n + \sum_{i=1}^n E(e_i|\boldsymbol{X})/n \\
&= \alpha_0,
\end{aligned}$$

since $\hat{\beta}_n$ is conditionally unbiased for β_0 so that $E(\beta_0 - \hat{\beta}_n|\boldsymbol{X}) = 0$. Alternatively, as $\mathrm{E}(\bar{y}_n|\boldsymbol{X}) = \alpha_0 + \beta_0\bar{x}_n$, it follows that

$$\mathrm{E}(\hat{\alpha}_n|\boldsymbol{X}) = \mathrm{E}(\bar{y}_n - \hat{\beta}_n\bar{x}_n|\boldsymbol{X}) = \alpha_0.$$

Note that

$$\hat{\alpha}_n = \alpha_0 + (\beta_0 - \hat{\beta}_n)\bar{x}_n + \sum_{i=1}^n e_i/n$$

$$\begin{aligned}
&= \alpha_0 - \sum_{i=1}^{n} k_i\, \bar{x}_n e_i + \sum_{i=1}^{n} e_i/n \\
&= \alpha_0 + \sum_{i=1}^{n} [1/n - k_i \bar{x}_n] e_i \\
&= \alpha_0 + \sum_{i=1}^{n} h_i e_i.
\end{aligned}$$

3. Under the homoskedastic liner model, i.e., **ID** 1 plus $\mathrm{E}(e_i) = 0$ and $\mathrm{var}(e_i) = \sigma_0^2$ It can be shown that

$$\begin{aligned}
\sigma^2_{\hat{\alpha}_n} &:= \mathrm{var}(\hat{\alpha}_n) \;=\; \sigma_0^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} \right), \\
\sigma^2_{\hat{\beta}_n} &:= \mathrm{var}(\hat{\beta}_n) \;=\; \frac{\sigma_0^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}, \\
\sigma_{\hat{\alpha}_n \hat{\beta}_n} &:= \mathrm{cov}(\hat{\alpha}_n, \hat{\beta}_n) \;=\; \sigma_0^2 \frac{-\bar{x}_n}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}.
\end{aligned}$$

First, we observe that

$$\begin{aligned}
E(y_i) &= E(\alpha_0 + \beta_0 x_i + \epsilon_i) = \alpha_0 + \beta_0 x_i, \\
\mathrm{var}(y_i) &= E[(y_i - E(y_i))^2] \\
&= E(\alpha_0 + \beta_0 x_i + \epsilon_i - \alpha_0 - \beta_0 x_i)^2] \\
&= E(\epsilon_i^2) = \sigma_0^2 \\
\mathrm{cov}(y_i, y_j) &= E[(y_i - E(y_i))(y_j - E(y_j))] \\
&= E(\epsilon_i \epsilon_j) = 0. \qquad \text{by [A.3]}
\end{aligned}$$

Then, we prove above results as follows:

$$\begin{aligned}
\sigma^2_{\hat{\beta}_n} &= \mathrm{var}(\sum_{i=1}^{n} k_i y_i) \\
&= \sum_{i=1}^{n} k_i^2 \mathrm{var}(y_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} k_i k_j \mathrm{cov}(y_i, y_j) \\
&= \sigma_0^2 \sum_{i=1}^{n} k_i^2 = \frac{\sigma_0^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}.
\end{aligned}$$

Besides,

$$\begin{aligned}
\sigma^2_{\hat{\alpha}_n} &= \mathrm{var}(\sum_{i=1}^n h_i y_i) \\
&= \sum_{i=1}^n h_i^2 \mathrm{var}(y_i) + 2\sum_{i=1}^n \sum_{j=i+1}^n h_i h_j \mathrm{cov}(y_i, y_j) \\
&= \sigma_0^2 \sum_{i=1}^n h_i^2 \\
&= \sigma_0^2 \sum_{i=1}^n (1/n - k_i \bar{x}_n)^2 \\
&= \sigma_0^2 \sum_{i=1}^n (1/n^2 - 2k_i \bar{x}_n/n + k_i^2 \bar{x}_n^2) \\
&= \sigma_0^2 [1/n - 2\bar{x}_n \sum_{i=1}^n k_i + \bar{x}_n^2 \sum_{i=1}^n k_i^2] \\
&= \sigma_0^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right).
\end{aligned}$$

Finally,

$$\begin{aligned}
\sigma_{\hat{\alpha}_n \hat{\beta}_n} &= \mathrm{cov}(\sum_{i=1}^n k_i y_i, \sum_{i=1}^n h_i y_i) \\
&= E\{[\sum_{i=1}^n k_i y_i - E(\sum_{i=1}^n k_i y_i)][\sum_{i=1}^n h_i y_i - E(\sum_{i=1}^n h_i y_i)]\} \\
&= E\{[\sum_{i=1}^n k_i (y_i - E(y_i))][\sum_{i=1}^n h_i (y_i - E(y_i))]\} \\
&= \sum_{i=1}^n h_i k_i \mathrm{var}(y_i) + 2\sum_{i=1}^n \sum_{j=i+1}^n h_i k_j \mathrm{cov}(y_i, y_j) \\
&= \sigma_0^2 \sum_{i=1}^n h_i k_i \\
&= \sigma_0^2 \sum_{i=1}^n (1/n - k_i \bar{x}_n) k_i \\
&= \sigma_0^2/n \sum_{i=1}^n k_i - \sigma_0^2 \bar{x}_n \sum_{i=1}^n k_i^2
\end{aligned}$$

$$= \frac{-\bar{x}_n \sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

4. (Gauss-Markov Theorem) result says that, given **ID** 1, **A2** $\hat{\alpha}_n$ and $\hat{\beta}_n$ have the smallest variance (the most efficient) among all linear and unbiased estimators of α_0 and β_0, i.e., they are the Best Linear Unbiased Estimators (BLUE).

   proof: Let $\tilde{\beta}_n$ be any other linear estimator in y_i than $\hat{\beta}_n$ so that it can be written as

$$\begin{aligned}
\tilde{\beta}_n &= \sum_{i=1}^n (k_i + c_i)\, y_i \\
&= \sum_{i=1}^n (k_i + c_i)(\alpha_0 + \beta_0 x_i + \epsilon_i) \\
&= \alpha_0 \sum_{i=1}^n (k_i + c_i) + \beta_0 \sum_{i=1}^n (k_i + c_i)\, x_i + \sum_{i=1}^n (k_i + c_i)\, \epsilon_i,
\end{aligned}$$

   given $c_i \neq 0, i = 1, \ldots, n$. By unbiasedness of $\tilde{\beta}_n$, $\sum_{i=1}^n (k_i + c_i) = 0$ and $\sum_{i=1}^n (k_i + c_i)\, x_i = 1$. Thus, given $\sum_{i=1}^n k_i = 0$ and $\sum_{i=1}^n k_i\, x_i = 1$,

$$\begin{aligned}
\sum_{i=1}^n c_i &= 0 \\
\sum_{i=1}^n x_i\, c_i &= 0.
\end{aligned}$$

   As

$$\begin{aligned}
\text{var}(\tilde{\beta}_n) &= \text{var}\left(\sum_{i=1}^n (k_i + c_i)\, y_i\right) \\
&= \sum_{i=1}^n (k_i + c_i)^2 \text{var}(y_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n k_i k_j \text{cov}(y_i, y_j) \\
&= \sum_{i=1}^n (k_i + c_i)^2 \text{var}(y_i) \\
&= \sum_{i=1}^n k_i^2 \sigma_0^2 + \sum_{i=1}^n c_i^2 \sigma_0^2 + 2 \sum_{i=1}^n k_i\, c_i \sigma_0^2.
\end{aligned}$$

and

$$\sum_{i=1}^{n} k_i\, c_i = \frac{\sum(x_i - \bar{x}_n)c_i}{\sum(x_i - \bar{x}_n)^2} = \frac{\sum x_i c_i - \bar{x}_n \sum c_i}{\sum(x_i - \bar{x}_n)^2} = 0,$$

we have

$$\mathrm{var}(\tilde{\beta}_n) = \frac{\sigma_0^2}{\sum(x_i - \bar{x}_n)^2} + \sigma_0^2 \sum_{i=1}^{n} c_i^2 \geq \mathrm{var}(\hat{\beta}_n).$$

Therefore, $\hat{\beta}_n$ has the smallest variance among linear and unbiased estimators.

5. $\hat{\sigma}_n^2 = \sum_{i=1}^n e_i^2/(n-2)$ is unbiased for σ_0^2. As

$$\begin{aligned}
e_i &= y_i - \hat{y}_i = y_i - \hat{\alpha}_n - \hat{\beta}_n x_i \\
&= (\alpha_0 - \beta_0 x_i + \epsilon_i) - (\bar{y}_n - \hat{\beta}_n \bar{x}_n) - \hat{\beta}_n x_i \\
&= (\alpha_0 - \beta_0 x_i + \epsilon_i) - (\sum_{i=1}(\alpha_0 + \beta_0 + \epsilon_i)/n - \hat{\beta}_n \bar{x}_n) - \hat{\beta}_n x_i \\
&= \beta_0 x_i + \epsilon_i - \beta_0 \bar{x}_n - \bar{\epsilon}_n - \hat{\beta}_n \bar{x}_n - \hat{\beta}_n x_i \\
&= -(\hat{\beta}_n - \beta_0)(x_i - \bar{x}_n) + (\epsilon_i - \bar{\epsilon}_n),
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon}_n) + (\hat{\beta}_n - \beta_0)^2 \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \\
&\quad -2(\hat{\beta}_n - \beta_0)\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon}_n)(x_i - \bar{x}_n).
\end{aligned}$$

Observe that

$$\begin{aligned}
E[\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon}_n)^2] &= E(\sum_{i=1}^{n}\epsilon_i^2 - n\bar{\epsilon}_n^2) \\
&= \sum_{i=1}^{n}\mathrm{var}(\epsilon_i) - n\mathrm{var}(\bar{\epsilon}_n) \\
&= n\sigma_0^2 - n(\sigma_0^2/n) = (n-1)\sigma_0^2.
\end{aligned}$$

And,

$$\begin{aligned}
E[(\hat{\beta}_n-\beta_0)^2\sum_{i=1}^n(x_i-\bar{x}_n)^2] &= \sum_{i=1}^n(x_i-\bar{x}_n)^2E(\hat{\beta}_n-\beta_0)^2\\
&= \sum_{i=1}^n(x_i-\bar{x}_n)^2[\sigma_0^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2]\\
&= \sigma_0^2.
\end{aligned}$$

Finally, as $\hat{\beta}_n=\beta_0+\sum_{i=1}^n k_i\epsilon_i$,

$$\begin{aligned}
E[(\hat{\beta}_n-\beta_0)\epsilon_i] &= E[(\sum_{i=1}^n k_i\epsilon_i)\epsilon_i]\\
&= k_iE(\epsilon_i^2)=k_i\sigma_0^2,
\end{aligned}$$

and

$$\begin{aligned}
E[(\hat{\beta}_n-\beta_0)\bar{\epsilon}_n] &= E[(\sum_{i=1}^n k_i\epsilon_i)(\sum_{i=1}^n \epsilon_i/n)]\\
&= \frac{1}{n}\sum_{i=1}^n k_i\sigma_0^2=0.
\end{aligned}$$

This implies

$$\begin{aligned}
&E[-2(\hat{\beta}_n-\beta_0)\sum_{i=1}^n(\epsilon_i-\bar{\epsilon}_n)(x_i-\bar{x}_n)]\\
&= -2[\sum_{i=1}^n(x_i-\bar{x}_n)E[(\hat{\beta}_n-\beta_0)\epsilon_i]]+2\sum_{i=1}^n(x_i-\bar{x}_n)E[(\hat{\beta}_n-\beta_0)\bar{\epsilon}_n]\\
&= -2\sum_{i=1}^n(x_i-\bar{x}_n)k_i\sigma_0^2\\
&= -2\sigma_0^2.
\end{aligned}$$

Thus,

$$E(\sum_{i=1}^n e_i^2)=(n-1)\sigma_0^2+\sigma_0^2-2\sigma_0^2=(n-2)\sigma_0^2.$$

We have proved that $E(\hat{\sigma}_n^2)=\sigma_0^2$.

6. As σ_0^2 is unknown, $\mathrm{var}(\hat{\alpha}_n)$ and $\mathrm{var}(\hat{\beta}_n)$ can be estimated by

$$\begin{aligned} s_{\hat{\alpha}_n}^2 &:= \widehat{\mathrm{var}(\hat{\alpha}_n)} = \hat{\sigma}_n^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right), \\ s_{\hat{\beta}_n}^2 &:= \widehat{\mathrm{var}(\hat{\beta}_n)} = \frac{\hat{\sigma}_n^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}, \\ s_{\hat{\alpha}_n \hat{\beta}_n} &:= \widehat{\mathrm{cov}(\hat{\alpha}_n, \hat{\beta}_n)} = \sigma_0^2 \frac{-\bar{x}_n}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}. \end{aligned}$$

Since $\hat{\sigma}_n^2$ is unbiased for σ_0^2, $s_{\hat{\alpha}_n}^2$, $s_{\hat{\beta}_n}^2$ and $s_{\hat{\alpha}_n \hat{\beta}_n}$ are all unbiased for $\sigma_{\hat{\alpha}_n}^2$, $\sigma_{\hat{\beta}_n}^2$, and $\sigma_{\hat{\alpha}_n \hat{\beta}_n}$, respectively.

Note that, the identification plus the assumptions mentioned previously are usually called the classical assumptions:

A1 $y_i = \alpha_0 + \beta_0 x_i + \epsilon_i$, $i = 1, \ldots, n$.

A2 x_i are nonstochastic and nonconstant.

A3 $E(\epsilon_i) = 0$.

A4 $E(\epsilon_i^2) = \sigma_0^2$, and $E(\epsilon_i \epsilon_j) = 0$ for all $i \neq j$.

A5 ϵ_i are i.i.d. $N(0, \sigma_0^2)$.

4.1.1 Hypothesis Testing

To perform hypothesis testing, we now assume assumption 5 (ϵ_i are i.i.d. $N(0, \sigma_0^2)$) holds. As assumption 5 implies assumptions 3 and 4, previous results remain valid. From assumption 5 we have

1. y_i are independent $N(\alpha_0 + \beta_0 x_i, \sigma_0^2)$.
2. $\hat{\alpha}_n \sim N(\alpha_0, \sigma_{\hat{\alpha}_n}^2)$ and $\hat{\beta}_n \sim N(\beta_0, \sigma_{\hat{\beta}_n}^2)$.
3. $\dfrac{\hat{\alpha}_n - \alpha_0}{\sigma_{\hat{\alpha}_n}} \sim N(0, 1)$ and $\dfrac{\hat{\beta}_n - \beta_0}{\sigma_{\hat{\beta}_n}} \sim N(0, 1)$.

4. $\sum_{i=1}^{n} e_t^2/\sigma_0^2 = (n-2)\hat{\sigma}_n^2/\sigma_0^2 \sim \chi^2_{n-2}$. Also, $\hat{\alpha}_n$ and $\hat{\beta}_n$ are independent of $\hat{\sigma}_n^2$.

proof: As

$$e_i = -(\hat{\beta}_n - \beta_0)(x_i - \bar{x}_n) + (\epsilon_i - \bar{\epsilon}_n),$$

we have

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon}_n)^2 + (\hat{\beta}_n - \beta_0)^2 \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \\ &\quad -2(\hat{\beta}_n - \beta_0)\sum_{i=1}^{n}(x_i - \bar{x}_n)(\epsilon_i - \bar{\epsilon}_n). \end{aligned} \tag{12}$$

For the first term in (12), we know

$$\begin{aligned}&\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon}_n)^2 \\ &= \sum_{i=1}^{n}[(\epsilon_i - E(\epsilon_i)) - (\bar{\epsilon}_n - E(\epsilon_i))]^2 \\ &= \sum_{i=1}^{n}[\epsilon_i - E(\epsilon_i)]^2 + \sum_{i=1}^{n}[\bar{\epsilon}_n - E(\epsilon_i)]^2 \\ &\quad -2\sum_{i=1}^{n}[(\epsilon_i - E(\epsilon_i))(\bar{\epsilon}_n - E(\epsilon_i))] \\ &= \sum_{i=1}^{n}[\epsilon_i - E(\epsilon_i)]^2 + \sum_{i=1}^{n}[\bar{\epsilon}_n - E(\epsilon_i)]^2 \\ &\quad -2(\bar{\epsilon}_n - E(\epsilon_i)\left(\sum_{i=1}^{n}\epsilon_i - nE(\epsilon_i)\right) \\ &= \sum_{i=1}^{n}[\epsilon_i - E(\epsilon_i)]^2 + \sum_{i=1}^{n}[\bar{\epsilon}_n - E(\epsilon_i)]^2 \\ &\quad -2(\bar{\epsilon}_n - E(\epsilon_i)\left(n\bar{\epsilon}_n - nE(\epsilon_i)\right) \\ &= \sum_{i=1}^{n}[\epsilon_i - E(\epsilon_i)]^2 + \sum_{i=1}^{n}[\bar{\epsilon}_n - E(\epsilon_i)]^2 \\ &\quad -2n(\bar{\epsilon}_n - E(\epsilon_i)^2,\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon}_n)^2/\sigma_0^2 \\
&= \sum_{i=1}^{n}\left[\frac{\epsilon_i - E(\epsilon_i)}{\sigma}\right]^2 + \sum_{i=1}^{n}\left[\frac{\bar{\epsilon}_n - E(\epsilon_i)}{\sigma_0}\right]^2 \\
&\quad -2\left[\frac{(\bar{\epsilon}_n - E(\epsilon_i)}{\sigma_0/\sqrt{n}}\right]^2 \\
&\sim \chi^2(n) + \chi^2(1) - 2\chi^2(2) = \chi^2(n-1).
\end{aligned}$$

Next, for the second term in (12),

$$\begin{aligned}
&(\hat{\beta}_n - \beta_0)^2 \sum_{i=1}^{n}(x_i - \bar{x}_n)^2/\sigma_0^2 \\
&= \left(\frac{\hat{\beta}_n - \beta_0}{\sqrt{\sigma_0^2/\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}\right) \\
&\sim \chi^2(1).
\end{aligned}$$

Finally, for the last term in (12), as $\hat{\beta}_n - \beta_0 = \sum_{i=1}^{n} k_i\epsilon_i$

$$\begin{aligned}
&2(\hat{\beta}_n - \beta_0)\sum_{i=1}^{n}(x_i - \bar{x}_n)(\epsilon_i - \bar{\epsilon}_n)/\sigma_0^2 \\
&= 2(\hat{\beta}_n - \beta_0)\sum_{i=1}^{n}(x_i - \bar{x}_n)\epsilon_i/\sigma_0^2 \\
&= 2(\hat{\beta}_n - \beta_0)/\sigma_0^2\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\sum_{i=1}^{n}\frac{x_i - \bar{x}_n}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\epsilon_i \\
&= 2(\hat{\beta}_n - \beta_0)/\sigma_0^2\sum_{i=1}^{n}(x_i - \bar{x}_n)^2(\hat{\beta}_n - \beta_0) \\
&= 2\frac{(\hat{\beta}_n - \beta_0)^2}{\frac{\sigma_0^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}} \\
&= 2\left(\frac{\hat{\beta}_n - \beta_0}{\sqrt{\frac{\sigma_0^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}}\right)^2 \\
&\quad \sim 2N(0,1)^2 2\chi^2(1).
\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{(n-2)\hat{\sigma}_n^2}{\sigma_0^2} &= \sum_{i=1}^n e_i^2/\sigma_0^2 \\ &\sim \chi^2(n-1)+\chi^2(1)-2\chi^2(1)=\chi^2(n-2).\end{aligned}$$

5. $\dfrac{\hat{\alpha}_n-\alpha_0}{s_{\hat{\alpha}_n}} \sim t_{n-2}$ and $\dfrac{\hat{\beta}_n-\beta_0}{s_{\hat{\beta}_n}} \sim t_{n-2}$.

   proof:

$$\begin{aligned}\frac{\hat{\alpha}_n-\alpha_0}{s_{\hat{\alpha}_n}} &= \frac{\hat{\alpha}_n-\alpha_0}{\sqrt{\hat{\sigma}_n^2[1/n+\bar{x}_n^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2]}} \\ &= \frac{\frac{\hat{\alpha}_n-\alpha_0}{\sqrt{\sigma_0^2[1/n+\bar{x}_n^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2]}}}{[(n-2)\hat{\sigma}_n^2/\sigma_0^2]/(n-2)} \\ &\sim \frac{N(0,1)}{\frac{\sqrt{\chi^2(n-2)}}{n-2}} = t(n-2).\end{aligned}$$

   Similarly,

$$\begin{aligned}\frac{\hat{\beta}_n-\beta_0}{s_{\hat{\beta}_n}} &= \frac{\hat{\beta}_n-\beta_0}{\sqrt{\hat{\sigma}_n^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2}} \\ &= \frac{\frac{\hat{\beta}_n-\beta_0}{\sqrt{\sigma_0^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2}}}{(n-2)\hat{\sigma}_n^2/\sigma_0^2]/(n-2)} \\ &\sim \frac{N(0,1)}{\frac{\sqrt{\chi^2(n-2)}}{n-2}} = t(n-2).\end{aligned}$$

To test the null hypothesis $H_0:\alpha_0=a$ or $H_0:\beta_0=b$ we can use t-tests.

1. One-sided test:

   - $H_a:\beta_0>b$. Under the null hypothesis,

     $$\tau_{\hat{\beta}_n}=(\hat{\beta}_n-b)/s_{\hat{\beta}_n}\sim t_{n-2}.$$

     Given the significance level γ and degrees of freedom $n-2$, the critical value $c_{\gamma,n-2}$ can be found in the t-table, and we reject H_0 if $\tau_{\hat{\beta}_n}>c_{\gamma,n-2}$. Similarly, we can test against $H_a:\alpha_0>a$ by checking whether $\tau_{\hat{\alpha}_n}=(\hat{\alpha}_n-a)/s_{\hat{\alpha}_n}>c_{\gamma,n-2}$.

- $H_a : \beta_0 < b$. Reject H_0 if $\tau_{\hat{\beta}_n} < -c_{\gamma,n-2}$.

2. Two-sided test: For $H_a : \beta_0 \neq b$, reject H_0 if $\tau_{\hat{\beta}_n} > c_{\gamma/2,n-2}$ or $\tau_{\hat{\beta}_n} < -c_{\gamma/2,n-2}$. For $H_a : \alpha_0 \neq a$, reject H_0 if $\tau_{\hat{\alpha}_n} > c_{\gamma/2,n-2}$ or $\tau_{\hat{\alpha}_n} < -c_{\gamma/2,n-2}$.

3. The $(1-\gamma)$ confidence intervals for β_0 and α_0 are

$$(\hat{\beta}_n - s_{\hat{\beta}_n} c_{\gamma/2,n-2},\ \hat{\beta}_n + s_{\hat{\beta}_n} c_{\gamma/2,n-2}),$$
$$(\hat{\alpha}_n - s_{\hat{\alpha}_n} c_{\gamma/2,n-2},\ \hat{\alpha}_n + s_{\hat{\alpha}_n} c_{\gamma/2,n-2}).$$

4.1.2 Prediction

Based on the regression line estimated with n observations, we can predict $\hat{y}_{n+1} = \hat{\alpha} + \hat{\beta} x_{n+1}$, provided that the new information x_{n+1} is available. Observe that the prediction error has mean zero

$$\mathrm{E}(\hat{y}_{n+1} - y_{n+1}) = \mathrm{E}[(\hat{\alpha}_n + \hat{\beta}_n x_{n+1}) - (\alpha_0 + \beta_0 x_{n+1} + \epsilon_{n+1})] = 0$$

and variance

$$\begin{aligned} \mathrm{E}(\hat{y}_{n+1} - y_{n+1})^2 &= E[(\hat{\alpha}_n - \alpha_0) + (\hat{\beta}_n - \beta_0)x_{n+1} - \epsilon_{n+1}]^2 \\ &= \mathrm{var}(\hat{\alpha}_n) + \mathrm{var}(\hat{\beta}_n)x_{n+1}^2 + \sigma_0^2 + 2x_{n+1}\mathrm{cov}(\hat{\alpha}_n, \hat{\beta}_n) \\ &= \sigma_0^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right). \end{aligned}$$

Hence, we have better prediction if x_{n+1} is close to $\bar{x}$.

4.2 Multiple Linear Regression

More generally, we may postulate a linear model with k explanatory variables to represent the identification equation: of y:

$$y = \beta_{10} x_1 + \beta_{20} x_2 + \cdots + \beta_{k0} x_k + e.$$

Given a sample of T observations, this specification can also be expressed as the identification condition:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}, \tag{13}$$

where $\boldsymbol{\beta}_0 = (\beta_{10}\ \beta_{20}\ \cdots\ \beta_{k0})'$ is the vector of unknown parameters, and $\boldsymbol{y}$ and $\boldsymbol{X}$ contain all the observations of the dependent and explanatory variables, i.e.,

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

where each column vector of $\boldsymbol{X}$ contains T observations for an explanatory variable. The basic "identifiability" requirement of this specification is that the number of regressors, k, is strictly less than the number of observations, T, such that the matrix $\boldsymbol{X}$ is of full column rank k. That is, the model does not contain any "redundant" regressor. It is also typical to set the first explanatory variable as the constant one so that the first column vector of $\boldsymbol{X}$ is a $T \times 1$ vector of ones, ℓ. To summary, the identification condition is **ID** 1: $y_t = \beta_{10} + \beta_{20}x_{t2} + \beta_{30}x_{t3} + \cdots + \beta_{k0}x_{tk} + e_t, t = 1, \ldots, T$.

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 4.1, we must minimize the average of the sum of squared errors:

$$Q(\boldsymbol{\beta}) := \frac{1}{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{14}$$

The first order conditions for the OLS minimization problem, also known as the *normal equations*, are:

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} (\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta})/T \\ &= -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/T \\ &\overset{\text{set}}{=} \mathbf{0}, \end{aligned}$$

the last equality can also be written as

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}_T \overset{\text{set}}{=} \boldsymbol{X}'\boldsymbol{y}$$

which is known as the *normal equation.* To have a unique solution for the system equation for $\boldsymbol{\beta}$, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ has to exist. This is first assumption has to be satisfied to have solution for $\boldsymbol{\beta}$ uniquely.

[A2] The $T \times k$ data matrix $\boldsymbol{X}$ is full column rank.

Given that $\boldsymbol{X}$ is of full column rank, $\boldsymbol{X}'\boldsymbol{X}$ is p.d. and hence invertible. The solution to the normal equations can then be expressed as

$$\hat{\boldsymbol{\beta}}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{15}$$

It is easy to see that the second order condition is also satisfied because

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

is p.d. Hence, $\hat{\beta}_T$ is the minimizer of the OLS criterion function and known as the OLS estimator for β. As the matrix inverse is unique, the OLS estimator is also unique.

The vector of OLS fitted values is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}_T,$$

and the vector of OLS residuals is

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}.$$

By the normal equations, $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$ so that $\hat{\mathbf{y}}'\hat{\mathbf{e}} = 0$. When the first regressor is the constant one, $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$ implies that $\ell'\hat{\mathbf{e}} = \sum_{t=1}^T \hat{e}_t = 0$. It follows that $\sum_{t=1}^T y_t = \sum_{t=1}^T \hat{y}_t$, and the sample average of the data y_t is the same as the sample average of the fitted values $\hat{y}_t$.

If $\mathbf{X}$ is not of full column rank and then its column vectors satisfies an exact linear relationship, this is also known as the problem of *exact multicollinearity*. In this case, without loss of generality we can write

$$\mathbf{x}_1 = \gamma_2\mathbf{x}_2 + \cdots + \gamma_k\mathbf{x}_k,$$

where $\mathbf{x}_i$ is the ith column of $\mathbf{X}$ and $\gamma_2, \ldots, \gamma_k$ are not all zero. Then, for any number $a \neq 0$,

$$\beta_1\mathbf{x}_1 = (1-a)\beta_1\mathbf{x}_1 + a\beta_1(\gamma_2\mathbf{x}_2 + \ldots + \gamma_k\mathbf{x}_k).$$

The linear specification (13) is thus observationally equivalent to

$$\mathbf{X}\beta^* := (1-a)\beta_1\mathbf{x}_1 + (\beta_2 + a\beta_1\gamma_2)\mathbf{x}_2 + \cdots + (\beta_k + a\beta_1\gamma_k)\mathbf{x}_k,$$

where the elements of β^* vary with a and therefore could be anything. That is, the parameter vector β is *not* identified when exact multicollinearity is present. Practically, when $\mathbf{X}$ is not of full column rank, $\mathbf{X}'\mathbf{X}$ is not invertible, and there are infinitely many solutions to the normal equations $\mathbf{X}'\mathbf{X}\beta \overset{\text{set}}{=} \mathbf{X}'\mathbf{y}$. Consequently, the OLS estimator $\hat{\beta}_T$ cannot be computed as (15). 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. .

It is also easy to verify that the magnitude of the coefficient estimates $\hat{\beta}_{iT}$ are affected by the measurement units of variables. Thus, a larger coefficient estimate does not necessarily imply that the associated explanatory variable is more important in explaining the behavior of $\mathbf{y}$. In fact, the coefficient estimates are not comparable in general.

Remark: The OLS estimators are derived without resorting to the knowledge of the "true" relationship between $\mathbf{y}$ and $\mathbf{X}$. That is, whether $\mathbf{y}$ is indeed generated according to our linear specification is irrelevant to the computation of the OLS estimator; it does affect the properties of the OLS estimator, however.

4.3 Geometric Interpretations

We know that the OLS estimation result has nice geometric interpretations. The vector of OLS fitted values can be written as

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}_X\mathbf{y},$$

here, and in what follows, $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is an orthogonal projection matrix. Hence, $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto span($\mathbf{X}$). The OLS residual vector is thus

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

which is the orthogonal projection of $\mathbf{y}$ onto span$(\mathbf{X})^{\perp}$ and orthogonal to $\hat{\mathbf{y}}$ and $\mathbf{X}$. Consequently, $\hat{\mathbf{y}}$ is the "best approximation" of $\mathbf{y}$, given the information contained in $\mathbf{X}$. Figure 2 illustrates a simple case where the model contains only two explanatory variables.

Let $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$, where $\mathbf{X}_1$ is $T \times k_1$ and $\mathbf{X}_2$ is $T \times k_2$, and $k_1 + k_2 = k$. We can write

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \text{random error},$$

and $\hat{\beta}_T = (\hat{\beta}_{1T}' \ \ \hat{\beta}_{2T}')'$. Let $\mathbf{P}_{X_1} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ and $\mathbf{P}_{X_2} = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$ denote the orthogonal projection matrices on span($\mathbf{X}_1$) and span($\mathbf{X}_2$), respectively. We have the following result.

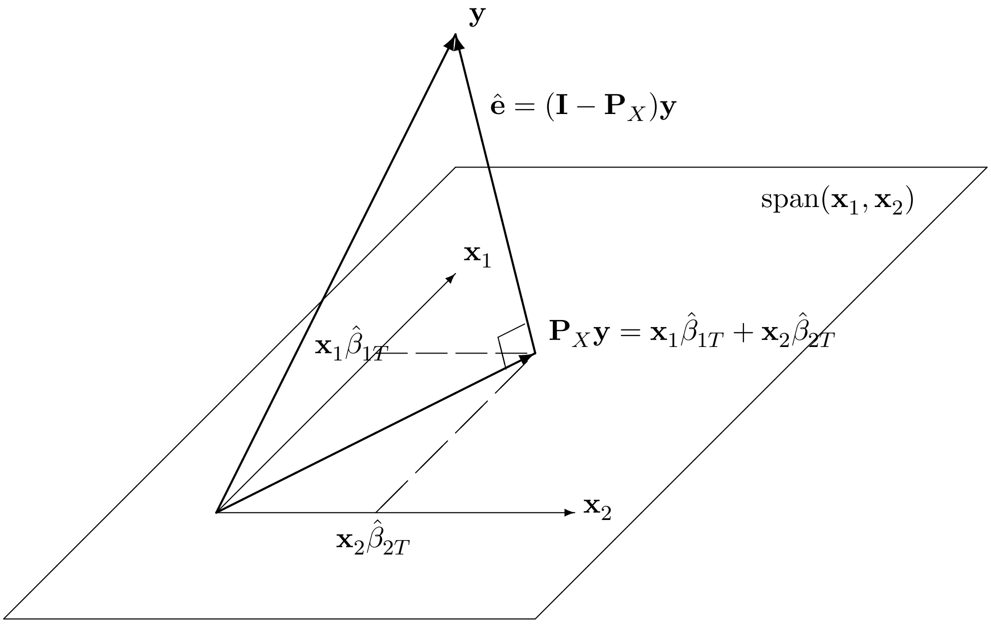
$\mathbf{y}$
$\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{P}_X)\mathbf{y}$
$\text{span}(\mathbf{x}_1, \mathbf{x}_2)$
$\mathbf{x}_1$
$\mathbf{x}_1\hat{\beta}_{1T}$
$\mathbf{P}_X\mathbf{y} = \mathbf{x}_1\hat{\beta}_{1T} + \mathbf{x}_2\hat{\beta}_{2T}$
$\mathbf{x}_2$
$\mathbf{x}_2\hat{\beta}_{2T}$

Figure 2: The orthogonal projection of $\mathbf{y}$ onto $\text{span}(\mathbf{x}_1, \mathbf{x}_2)$.

Theorem 4.1 (Frisch-Waugh-Lovell) *Given a vector* $\mathbf{y}$*,* $(\mathbf{I} - \mathbf{P}_{X_2})\mathbf{y}$ *and* $(\mathbf{I} - \mathbf{P}_{X_1})\mathbf{y}$ *can be uniquely decomposed into two orthogonal components:*

$$\begin{aligned}(\mathbf{I} - \mathbf{P}_{X_2})\mathbf{y} &= (\mathbf{I} - \mathbf{P}_{X_2})\mathbf{X}_1\hat{\beta}_{1T} + (\mathbf{I} - \mathbf{P}_X)\mathbf{y},\\ (\mathbf{I} - \mathbf{P}_{X_1})\mathbf{y} &= (\mathbf{I} - \mathbf{P}_{X_1})\mathbf{X}_{X_2}\hat{\beta}_{2T} + (\mathbf{I} - \mathbf{P}_X)\mathbf{y}.\end{aligned}$$

Proof: As $\mathbf{I} - \mathbf{P}_{X_2}$ is in $\text{span}(\mathbf{X}_2)^\perp$ and $\mathbf{I} - \mathbf{P}_X$ is in $\text{span}(\mathbf{X})^\perp \subseteq \text{span}(\mathbf{X}_2)^\perp$, we have $(\mathbf{I} - \mathbf{P}_{X_2})(\mathbf{I} - \mathbf{P}_X) = \mathbf{I} - \mathbf{P}_X$. Hence,

$$\begin{aligned}(\mathbf{I} - \mathbf{P}_{X_2})\mathbf{y} &= (\mathbf{I} - \mathbf{P}_{X_2})\mathbf{P}_X\mathbf{y} + (\mathbf{I} - \mathbf{P}_{X_2})(\mathbf{I} - \mathbf{P}_X)\mathbf{y}\\ &= (\mathbf{I} - \mathbf{P}_{X_2})\mathbf{X}_1\hat{\beta}_{1T} + (\mathbf{I} - \mathbf{P}_{X_2})\mathbf{X}_2\hat{\beta}_{2T} + (\mathbf{I} - \mathbf{P}_X)\mathbf{y}\\ &= (\mathbf{I} - \mathbf{P}_{X_2})\mathbf{X}_1\hat{\beta}_{1T} + (\mathbf{I} - \mathbf{P}_X)\mathbf{y},\end{aligned}$$

and these two components are orthogonal because

$$\mathbf{y}'\mathbf{P}_X(\mathbf{I} - \mathbf{P}_{X_2})(\mathbf{I} - \mathbf{P}_X)\mathbf{y} = \mathbf{y}'\mathbf{P}_X(\mathbf{I} - \mathbf{P}_X)\mathbf{y} = 0.$$

The second assertion follows similarly. □

An implication of Theorem 4.1 is that $(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{X}_1\hat{\beta}_{1T} = (\mathbf{I}-\mathbf{P}_{X_2})\mathbf{P}_X\mathbf{y}$ is the orthogonal projection of $(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{y}$ onto $\text{span}((\mathbf{I}-\mathbf{P}_{X_2})\mathbf{X}_1)$. Thus, we can write

$$\hat{\beta}_{1T} = [\mathbf{X}_1'(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{X}_1]^{-1}\mathbf{X}_1'(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{y},$$

as can be directly verified from (15) using the matrix inversion formula. That is, $\hat{\beta}_{1T}$ can also be obtained by regressing $(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{y}$ on $(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{X}_1$, where $(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{y}$ and $(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{X}_1$ are, respectively, the residual vectors from two "purging" regressions: $\mathbf{y}$ on $\mathbf{X}_2$ and $\mathbf{X}_1$ on $\mathbf{X}_2$. Moreover, the residual vector from regressing $(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{y}$ on $(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{X}_1$ is the same as the residual vector from regressing $\mathbf{y}$ on $\mathbf{X}$. Similarly,

$$\hat{\beta}_{2T} = [\mathbf{X}_2'(\mathbf{I}-\mathbf{P}_{X_1})\mathbf{X}_2]^{-1}\mathbf{X}_2'(\mathbf{I}-\mathbf{P}_{X_1})\mathbf{y}$$

can be obtained by regressing $(\mathbf{I}-\mathbf{P}_{X_1})\mathbf{y}$ on $(\mathbf{I}-\mathbf{P}_{X_1})\mathbf{X}_2$. Note that $\hat{\beta}_{1T}$ is not the same as the OLS estimator from regressing $\mathbf{y}$ on $\mathbf{X}_1$, and that $\hat{\beta}_{2T}$ is not the same as the OLS estimator from regressing $\mathbf{y}$ on $\mathbf{X}_2$, except when $\mathbf{X}_1$ is orthogonal to $\mathbf{X}_2$.

From Theorem 4.1 we can re-write $(\mathbf{I}-\mathbf{P}_{X_2})\mathbf{y} = (\mathbf{I}-\mathbf{P}_{X_2})\mathbf{P}_X\mathbf{y} + (\mathbf{I}-\mathbf{P}_X)\mathbf{y}$ as

$$\mathbf{P}_{X_2}\mathbf{y} = \mathbf{P}_{X_2}\mathbf{P}_X\mathbf{y}.$$

Thus, a second implication of Theorem 4.1 is that projecting $\mathbf{y}$ directly on $\text{span}(\mathbf{X}_2)$ is equivalent to performing iterated projections of $\mathbf{y}$ on $\text{span}(\mathbf{X})$ then on $\text{span}(\mathbf{X}_2)$. Similarly, we have $\mathbf{P}_{X_1}\mathbf{y} = \mathbf{P}_{X_1}\mathbf{P}_X\mathbf{y}$. For an illustration of Theorem 4.1 see Figure 3; see also Davidson & MacKinnon (1993) for more details.

As an application, consider the model with $\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2]$, where $\mathbf{X}_1$ contains the constant term and a time trend variable t, and $\mathbf{X}_2$ includes $k-2$ other explanatory variables. Then, the OLS estimates of the coefficients of $\mathbf{X}_2$ are the same as those obtained by regressing (detrended) $\mathbf{y}$ on detrended $\mathbf{X}_2$, where detrended $\mathbf{y}$ and $\mathbf{X}_2$ are obtained by regressing $\mathbf{y}$ and $\mathbf{X}_2$ on $\mathbf{X}_1$, respectively.

4.4 Measures of Goodness of Fit

We have learned that for a *given* linear specification, the OLS method yields the best fit of data. In practice, one may postulate different linear models with different regressors and try to choose a particular one among them. It is therefore of interest to compare the

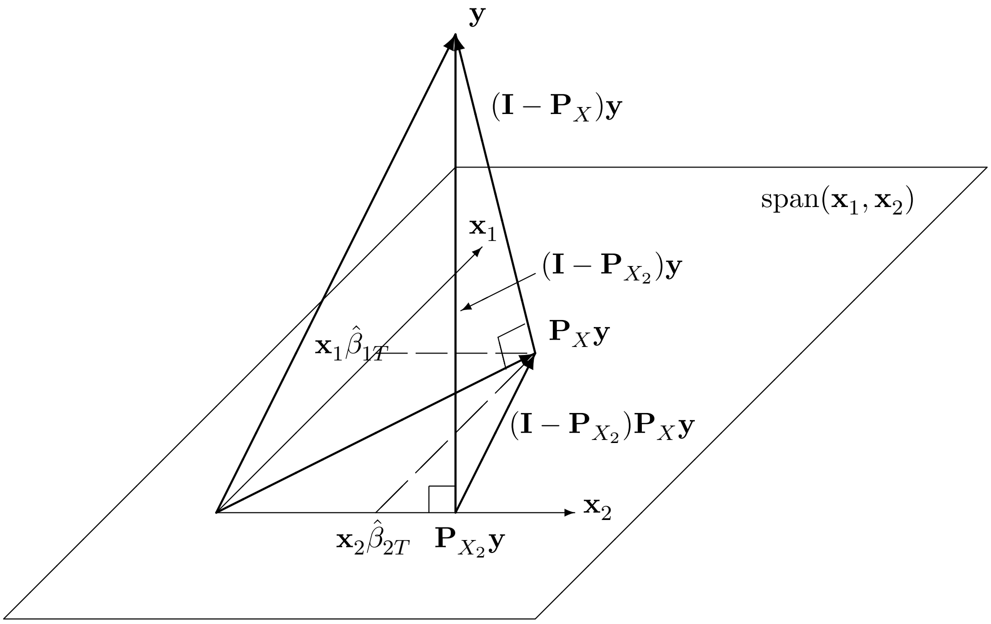
$\mathbf{y}$
$(\mathbf{I} - \mathbf{P}_X)\mathbf{y}$
span$(\mathbf{x}_1, \mathbf{x}_2)$
$\mathbf{x}_1$
$(\mathbf{I} - \mathbf{P}_{X_2})\mathbf{y}$
$\mathbf{P}_X\mathbf{y}$
$\mathbf{x}_1\hat{\beta}_{1T}$
$(\mathbf{I} - \mathbf{P}_{X_2})\mathbf{P}_X\mathbf{y}$
$\mathbf{x}_2$
$\mathbf{x}_2\hat{\beta}_{2T}$
$\mathbf{P}_{X_2}\mathbf{y}$

Figure 3: An illustration of the Frisch-Waugh-Lovell Theorem.

performance across models. In this section we discuss how to measure the *goodness of fit* of models.

A natural goodness-of-fit measure is the regression variance $\hat{\sigma}_T^2 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(T-k)$. This measure, however, is *not* invariant with respect to measurement units of the dependent variable. Instead, the following "relative" measures of goodness of fit are adopted in the linear regression analysis. Recall that

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

where TSS, RSS, and ESS denote total, regression, and error sum of squares, respectively. The non-centered *coefficient of determination* (or non-centered R^2) is defined to be the proportion of TSS that can be explained by the regression hyperplane:

$$R^2 = \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{ESS}}{\text{TSS}}. \tag{16}$$

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 $R^2 = 0$. Note that R^2 is non-decreasing in the number of variables in the model. That is, adding more variables to a model will *not* reduce its R^2. As $\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}'\boldsymbol{y}$, we can also write

$$R^2 = \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}} = \frac{(\hat{\boldsymbol{y}}'\boldsymbol{y})^2}{(\boldsymbol{y}'\boldsymbol{y})(\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}})} = \cos^2\theta,$$

where θ is the angle between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$. That is, R^2 is a measure of the linear association between these two vectors.

It is also easily verified that, when the model 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 (16), the centered coefficient of determination (or centered R^2) is defined as

$$\text{Centered } R^2 = \frac{\text{Centered RSS}}{\text{Centered TSS}} = 1 - \frac{\text{ESS}}{\text{Centered TSS}}. \tag{17}$$

This measure also takes on values between 0 and 1 and is non-decreasing in the number of variables in the model. In contrast with the non-centered R^2, this measure *excludes* the effect of the constant term in the model, and is hence invariant with respect to constant addition. If the model does *not* contain a constant term, the centered R^2 may be negative. As

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

we immediately get

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

That is, the centered R^2 is also the squared sample correlation coefficient of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$.

If R^2 is the only criterion to determine model adequacy, one would tend to select a model with more explanatory variables. The adjusted R^2, $\bar{R}^2$, is the centered R^2 adjusted for the degrees of freedom:

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

It can also be shown that

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

That is, $\bar{R}^2$ is the centered R^2 with a penalty term depending on model complexity and explanatory ability. Clearly, $\bar{R}^2 < R^2$ except for $k = 1$ or $R^2 = 1$. Note also that $\bar{R}^2$ need not be increasing with the number of explanatory variables; in fact, $\bar{R}^2 < 0$ when $R^2 < (k-1)/(T-1)$.

Remark: Models for different dependent variables are not comparable in terms of their R^2 because their total variations (i.e., TSS) are different. For example, R^2 of models for y and $\log y$ are not comparable.

5 Properties of the OLS Estimators

5.1 Bias

Proposition 5.1 *If* $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}$, *then* $\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$.

Proof: Since $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}$,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_T &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} \\
&= \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}.
\end{aligned}$$

To have $\hat{\boldsymbol{\beta}}_T$ to be unbiased for $\boldsymbol{\beta}_0$, the following assumption has to be imposed:
A3: $E(\boldsymbol{e}|\boldsymbol{X}) = \boldsymbol{0}$.

Proposition 5.2 *Given* **ID1**, **A2** *and* **A3**, $E(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0|\boldsymbol{X}) = \boldsymbol{0}$ *and* $E(\hat{\boldsymbol{\beta}}_T) = \boldsymbol{\beta}_0$.

Proof: By the previous result,

$$\begin{aligned}
E[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)|\boldsymbol{X}] &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}|\boldsymbol{X}] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{e}|\boldsymbol{X}) = \boldsymbol{0},
\end{aligned}$$

and $E(\hat{\boldsymbol{\beta}}_T|\boldsymbol{X}) = \boldsymbol{\beta}_0$. And then applying the law of iterated expectations,

$$E(\hat{\boldsymbol{\beta}}_T) = E[E(\hat{\boldsymbol{\beta}}_T|\boldsymbol{X})] = E(\boldsymbol{\beta}_0) = \boldsymbol{\beta}_0. \square$$

Thus $\hat{\boldsymbol{\beta}}_T$ is *unbiased* for $\boldsymbol{\beta}_0$. Indeed, it is *conditionally unbiased*, conditional upon $\boldsymbol{X}$, which is a stronger result.

5.2 Variance-Covariance Matrix of Regression Error

The conditional variance-covariance matrix of the regression error vector $\boldsymbol{e}$ is

$$\begin{aligned} \boldsymbol{D} &= \operatorname{var}(\boldsymbol{e}|\boldsymbol{X}) = E(\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}) \\ &= \begin{bmatrix} E(e_1^2|\boldsymbol{x}_1) & E(e_1e_2|\boldsymbol{x}_1) & E(e_1e_3|\boldsymbol{x}_1) & \cdots & E(e_1e_T|\boldsymbol{x}_1) \\ E(e_2e_1|\boldsymbol{x}_2) & E(e_2^2|\boldsymbol{x}_2) & E(e_2e_3|\boldsymbol{x}_2) & \cdots & E(e_2e_T|\boldsymbol{x}_2) \\ E(e_3e_1|\boldsymbol{x}_3) & E(e_3e_2|\boldsymbol{x}_3) & E(e_3^2|\boldsymbol{x}_3) & \cdots & E(e_3e_T|\boldsymbol{x}_3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ E(e_Te_1|\boldsymbol{x}_T) & E(e_Te_2|\boldsymbol{x}_T) & E(e_Te_3|\boldsymbol{x}_T) & \cdots & E(e_T^2|\boldsymbol{x}_T) \end{bmatrix} \end{aligned}$$

when the data are random sample then $(\boldsymbol{x}_t, e_t)$ is independent of $(\boldsymbol{x}_s, e_s)$ for $t \neq s$, thus

$$\begin{aligned} E(e_t^2|\boldsymbol{X}) &= E(e_t^2|\boldsymbol{x}_t) = \sigma_t^2 \\ E(e_te_s|\boldsymbol{X}) &= E(e_te_s|\boldsymbol{x}_t) = E(e_t|\boldsymbol{x}_t)\, E(e_s|x_t) = 0. \end{aligned}$$

Thus in general

$$\boldsymbol{D} = \operatorname{var}(\boldsymbol{e}|\boldsymbol{X}) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_T^2 \end{bmatrix} \tag{18}$$

when the data are random. Under the homoskedasticity restriction (5), $E(e_t^2|\boldsymbol{x}_t) = \sigma_0^2$ for all t, then

$$\boldsymbol{D} = \begin{bmatrix} \sigma_0^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_0^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_0^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_0^2 \end{bmatrix} = \sigma_0^2 I_T, \tag{19}$$

which is the classical assumption for the linear regression models. That is
A4: $\operatorname{var}(\boldsymbol{e}|\boldsymbol{X}) = \sigma_0^2 I_T$.

5.3 Variance-Covariance Matrix of OLS Estimator

The conditional variance-covariance matrix for $\hat{\boldsymbol{\beta}}_T$ is

$$V_T = E\left[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)' \,|\boldsymbol{X}\right]$$

Since $\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$,

$$\begin{aligned} V_T &= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}\boldsymbol{e}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}\right] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \end{aligned}$$

where $\boldsymbol{D}$ is defined in (18). It may be helpful to observe that

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'\sigma_t^2.$$

In the special case of (5), then $\sigma_t^2 = \sigma_0^2$, $\boldsymbol{D} = \sigma_0^2 I_T$ and $\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{X}\sigma_0^2$. Thus, V_T simplifies to

$$\begin{aligned} V_T &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{aligned}$$

Theorem 5.1 *In the linear regression model,*

$$V_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{20}$$

If (5), $E(e_t^2|\boldsymbol{x}_t) = \sigma_0^2$, holds,

$$V_T = \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{21}$$

The expression $V_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is often called a “sandwich formula”, because the central variance matrix $\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}$ is “sandwiched” between the moment matrices $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

5.4 Gauss-Markov Theorem

Theorem 5.2 *(Gauss-Markov) In the linear regression model, $\hat{\beta}_T$ is the best linear unbiased estimator for β_0.*

Proof: Consider an arbitrary linear estimator $\check{\boldsymbol{\beta}}_T = \boldsymbol{A}\boldsymbol{y} = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}]\boldsymbol{y}$, where $\boldsymbol{C}$ is an arbitrary non-zero matrix. $\check{\boldsymbol{\beta}}_T$ is unbiased if and only if $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{0}$ since

$$\begin{aligned}
\check{\boldsymbol{\beta}}_T &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}]\boldsymbol{y} \\
&= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}](\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}) \\
&= \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} + \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{C}\boldsymbol{e}.
\end{aligned}$$

and

$$\begin{aligned}
E[\check{\boldsymbol{\beta}}_T|\boldsymbol{X}] &= \boldsymbol{\beta}_0 + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}|\boldsymbol{X}] + E[\boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}_0|\boldsymbol{X}] + E[\boldsymbol{C}\boldsymbol{e}|\boldsymbol{X}] \\
&= \boldsymbol{\beta}_0 + \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}_0.
\end{aligned}$$

It follows that when $\check{\boldsymbol{\beta}}_T$ is unbiased

$$\begin{aligned}
\text{var}(\check{\boldsymbol{\beta}}_T|\boldsymbol{X}) &= E[(\check{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)(\check{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)'|\boldsymbol{X}] \\
&= E\{[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} + \boldsymbol{C}\boldsymbol{e}][(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} + \boldsymbol{C}\boldsymbol{e}]'|\boldsymbol{X}\} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma_0^2\boldsymbol{I}_T\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}' \\
&\quad + \boldsymbol{C}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{C}\sigma_0^2\boldsymbol{I}_T\boldsymbol{C}' \\
&= \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \sigma_0^2\boldsymbol{C}\boldsymbol{C}',
\end{aligned}$$

where the first term on the right-hand side is $\text{var}(\hat{\boldsymbol{\beta}}_T)$ and the second term is clearly p.s.d. Thus, for any linear unbiased estimator $\check{\boldsymbol{\beta}}_T$, $\text{var}(\check{\boldsymbol{\beta}}_T) - \text{var}(\hat{\boldsymbol{\beta}}_T)$ is a p.s.d. matrix. □

5.5 OLS Estimation of Error Variance

Under the restriction of (5), $E(e_t^2|\boldsymbol{x}_t) = \sigma_0^2$ is another parameter under estimation. The OLS estimator for σ_0^2 is

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

where k is the number of regressors. It is clear that $\hat{\sigma}_T^2$ is not linear in $\boldsymbol{y}$.

Theorem 5.3 *In the homoskedastic regression model, $\hat{\sigma}_T^2$ is an unbiased estimator for σ_0^2.*

Proof: Recall that $\boldsymbol{I} - \boldsymbol{P}_X$ is orthogonal to span($\boldsymbol{X}$). Then,

$$\hat{\mathbf{e}} = (\boldsymbol{I}_T - \boldsymbol{P}_X)\boldsymbol{y} = (\boldsymbol{I}_T - \boldsymbol{P}_X)(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}) = (\boldsymbol{I}_T - \boldsymbol{P}_X)\boldsymbol{e},$$

and

$$\mathrm{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}|\boldsymbol{X}) = \mathrm{E}[\boldsymbol{e}'(\boldsymbol{I}_T - \boldsymbol{P}_X)\boldsymbol{e}|\boldsymbol{X}] = \mathrm{E}[\mathrm{trace}(\boldsymbol{e}\boldsymbol{e}'(\boldsymbol{I}_T - \boldsymbol{P}_X))|\boldsymbol{X}].$$

As the trace and expectation operators can be interchanged, we have that

$$\mathrm{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}|\boldsymbol{X}) = \mathrm{trace}(\mathrm{E}[\boldsymbol{e}\boldsymbol{e}'(\boldsymbol{I}_T - \boldsymbol{P}_X)|\boldsymbol{X}]) = \mathrm{trace}[\boldsymbol{D}(\boldsymbol{I}_T - \boldsymbol{P}_X)].$$

By the fact that, $\mathrm{trace}(\boldsymbol{I}_T - \boldsymbol{P}_X) = \mathrm{rank}(\boldsymbol{I}_T - \boldsymbol{P}_X) = T - k$ and $\boldsymbol{D} = \sigma^2 I_T$, it follows that $\mathrm{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}) = (T-k)\sigma_0^2$ and that

$$\mathrm{E}(\hat{\sigma}_T^2) = \mathrm{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}})/(T-k) = \sigma_0^2,$$

proving the unbiasedness of $\hat{\sigma}_T^2$. □

The OLS estimation for variance-covariance matrix of $\hat{\boldsymbol{\beta}}_T$ in the homoskedastic regression model becomes

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

which is unbiased for $\mathrm{var}(\hat{\boldsymbol{\beta}}_T) = \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ provided $\hat{\sigma}_T^2$ is unbiased for σ_0^2.

5.6 Gaussian Quasi-MLE and MVUE

In normal regression, $e_t|\boldsymbol{x}_t \sim N(0, \sigma^2)$ and then the likelihood for a single observation is

$$L_t(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2\right).$$

Then the log-likelihood function for the full sample $y^T = (y_1, \ldots, y_T)$ is

$$\begin{aligned} L_T(y^T; \boldsymbol{\beta}, \sigma^2) &= \sum_{t=1}^{T} \log L_t(y_t; \boldsymbol{\beta}, \sigma^2) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned}$$

The MLE $(\tilde{\boldsymbol{\beta}}_T, \tilde{\sigma}_T^2)$ maximize L_T. The FOCs of the maximization problem are

$$\begin{aligned} \frac{\nabla L_T}{\nabla \boldsymbol{\beta}} &= \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/\sigma^2 \overset{\text{set}}{=} 0, \\ \frac{\partial L_T}{\partial(\sigma^2)} &= -\frac{T}{2\sigma^2} + \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^4} \overset{\text{set}}{=} 0, \end{aligned}$$

which yield the MLEs of $\boldsymbol{\beta}_0$ and σ^2:

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_T &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \\ \tilde{\sigma}^2_T &= (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_T)'(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_T)/T = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/T. \end{aligned}$$

Clearly, the MLE $\tilde{\boldsymbol{\beta}}_T$ is the same as the OLS estimator $\hat{\boldsymbol{\beta}}_T$, but the MLE $\tilde{\boldsymbol{\beta}}^2_T$ is different from $\hat{\sigma}^2_T$. In fact, $\tilde{\sigma}^2_T$ is biased estimator because $\mathrm{E}(\tilde{\sigma}^2_T) = \sigma^2_0(T-k)/T \neq \sigma^2_0$.

Theorem 5.4 *(Minimum Variance Unbiased Estimator, MVUE) In normal regression models, the OLS estimators $\hat{\beta}_T$ and $\hat{\sigma}^2_T$ are the* minimum variance unbiased estimator *(MVUE).*

Consider a collection of independent random variables $z^T = (z_1, \ldots, z_T)$, where z_t has the density function $f_t(z_t, \theta)$ with θ a $r \times 1$ vector of parameters. Let the joint log-likelihood function of z^T be $L_T(z^T; \theta) = \log f^T(z^T; \theta)$. Then the *score function*

$$s^T(z^T; \theta) := \nabla \log f^T(z^T; \theta) = \frac{1}{f^T(z^T; \theta)} \nabla f^T(z^T; \theta)$$

is the $r \times 1$ vector of the first order derivatives of $\log f^T$ with respect to θ. Under regularity conditions, differentiation and integration can be interchanged. When the postulated density function f^T is the true density function of z^T, we have

$$\begin{aligned} \mathrm{E}[s^T(z^T; \theta)] &= \int \frac{1}{f^T(z^T; \theta)} \nabla f^T(z^T; \theta)\ f^T(z^T; \theta)\ dz^T \\ &= \nabla \left(\int f^T(z^T; \theta)\ dz^T \right) \\ &= 0. \end{aligned}$$

That is, $s^T(z^T; \theta)$ has mean zero. The variance of s^T is the *Fisher's information matrix*:

$$B_T(\theta) := \mathrm{var}[s^T(z^T; \theta)] = \mathrm{E}[s^T(z^T; \theta)\ s^T(z^T; \theta)'].$$

Consider the $r \times r$ *Hessian matrix* of the second order derivatives of $\log f^T$:

$$\begin{aligned} & H_T(z^T; \theta) \\ &:= \nabla^2 \log f^T(z^T; \theta) \\ &= \nabla \left(\frac{1}{f^T(z^T; \theta)} [\nabla f^T(z^T; \theta)]' \right) \\ &= \frac{1}{f^T(z^T; \theta)} \nabla^2 f^T(z^T; \theta) - \frac{1}{f^T(z^T; \theta)^2} [\nabla f^T(z^T; \theta)][\nabla f^T(z^T; \theta)]', \end{aligned}$$

where $\nabla^2 f = \nabla(\nabla f)'$. As

$$\int \frac{1}{f^T(z^T;\theta)} \nabla^2 f^T(z^T;\theta)\ f^T(z^T;\theta) dz^T = \nabla^2 \left(\int f^T(z^T;\theta) dz^T \right) = 0,$$

the expected value of the Hessian matrix becomes

$$\begin{aligned}
&\mathrm{E}[H_T(z^T;\theta)] \\
&= -\int \left(\frac{1}{f^T(z^T;\theta)^2} [\nabla f^T(z^T;\theta)][\nabla f^T(z^T;\theta)]' \right) f^T(z^T;\theta)\ dz^T \\
&= -E[s^T(z^T;\theta)\ s^T(z^T;\theta)] \\
&= -B_t(\theta).
\end{aligned}$$

This established the *information matrix equality*: $B_T(\theta) + \mathrm{E}[H_T(z^T;\theta)] = 0$. Suppose now $r = 1$ for simplicity so that both s^T and B_T are scalar. Let $\hat{\theta}_T$ denote an unbiased estimator. Then

$$\mathrm{cov}[s^T(z^T;\theta), \hat{\theta}] = \frac{\partial}{\partial\theta} \int \hat{\theta}_T\ f^T(z^T;\theta)\ dz^T = \frac{\partial}{\partial\theta} \mathrm{E}(\hat{\theta}_T) = 1,$$

by unbiasedness. By the celebrated Cauchy-Schwartz inequality:

$$\frac{\mathrm{cov}[s^T(z^T;\theta), \hat{\theta}_T]}{\mathrm{var}[s^T(z^T;\theta)]\mathrm{var}(\hat{\theta}_T)} = \frac{1}{\mathrm{var}[s^T(z^T;\theta)]\mathrm{var}(\hat{\theta}_T)} \leq 1.$$

It follows that $\mathrm{var}(\hat{\theta}_T) \geq 1/B_T(\theta)$. The RHS, $1/B_T(\theta)$, is also know as the *Cramér-Rao lower bound*. Thus, all unbiased estimators must have variance greater than or equal to the inverse of information. When θ is multi-dimensional, we have that $\mathrm{var}(\hat{\theta}_T) - B_T(\theta)^{-1}$ is a positive semi-definite matrix.

In our application, the inverse of the information matrix evaluated at the true parameters $\boldsymbol{\beta}_0$ and σ_0^2 can be easily calculated as

$$\left[\begin{array}{cc} \sigma_0^2(X'X)^{-1} & 0 \\ 0 & 2\sigma_0^4/T \end{array} \right].$$

As $\hat{\boldsymbol{\beta}}_T$ achieves the Cramér-Rao lower bound, it is efficient within the class of all unbiased estimators for $\boldsymbol{\beta}_0$, i.e., it is the MVUE. It can be shown that any other unbiased estimator of σ_0^2 has variance greater than or equal to that of $\hat{\sigma}_T^2$; hence $\hat{\sigma}_T^2$ is also the MVUE.

Consider the log-likelihood function of $\boldsymbol{y} = (y_1, \ldots, y_T)'$ to be

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

The corresponding Hessian matrix is

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

The information matrix is defined as $B_T(\boldsymbol{\beta}, \sigma^2) = -\mathrm{E}[H_T(\boldsymbol{\beta}, \sigma^2)]$. The $(1,2)$-th element of information matrix evaluated at true values $\boldsymbol{\beta}_0$ and σ_0 becomes

$$-\mathrm{E}[\frac{1}{\sigma_0^4}(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}_0)] = \frac{1}{\sigma^4}\mathrm{E}(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}) = \frac{1}{\sigma^4}(\boldsymbol{X}'\boldsymbol{X})\mathrm{E}(\boldsymbol{e}) = 0.$$

And, the $(2,2)$-th element of $B_T(\boldsymbol{\beta}, \sigma^2)$ evaluated at $\boldsymbol{\beta}_0$ and σ_0^2 is

$$\begin{aligned} & -\mathrm{E}\left(\frac{T}{2\sigma_0^4} - \frac{1}{\sigma_0^6}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)\right) \\ = & -\frac{T}{2\sigma_0^4} + \frac{1}{\sigma_0^6}\mathrm{E}(\boldsymbol{e}'\boldsymbol{e}) \\ = & -\frac{T}{2\sigma_0^4} + \frac{T\sigma_0^2}{\sigma_0^6} \\ = & \frac{-T + 2T}{2\sigma_0^4} = \frac{T}{2\sigma_0^4}. \end{aligned}$$

5.7 Distribution of $\hat{\boldsymbol{\beta}}_T$ in Normal Regression Models

In normal regression models,

$$\begin{aligned} y_t &= \boldsymbol{x}_t'\boldsymbol{\beta}_0 + e_t \\ e_t &\mid x_t \sim N(0, \sigma_0^2). \end{aligned}$$

Theorem 5.5 *In normal regression models,*

(a) $\hat{\boldsymbol{\beta}}_T \sim N(\boldsymbol{\beta}_0, \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$.

(b) $(T-k)\hat{\sigma}_T^2/\sigma_0^2 \sim \chi^2(T-k)$.

(c) $\hat{\sigma}_T^2$ *has mean* σ_0^2 *and variance* $2\sigma_0^4/(T-k)$.

Proof: As $\boldsymbol{e}|\boldsymbol{X} \sim N(0, \sigma_0^2\boldsymbol{I}_T)$, $\boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}_0, \sigma_0^2\boldsymbol{I}_T)$ and

$$\hat{\boldsymbol{\beta}}_T|\boldsymbol{X} \sim N(\boldsymbol{\beta}_0, \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

This establishes (a). To prove (b), we write $\hat{\boldsymbol{e}} = (\boldsymbol{I}_T - \boldsymbol{P}_X)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)$ and deduce

$$(T-k)\hat{\sigma}_T^2/\sigma_0^2 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/\sigma_0^2 = \boldsymbol{y}^{*\prime}(\boldsymbol{I}_T - \boldsymbol{P}_X)\boldsymbol{y}^*,$$

where $\boldsymbol{y}^* = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)/\sigma_0$. As $\boldsymbol{I}_T - \boldsymbol{P}_X$ is a symmetric and idempotent matrix with rank $T-k$, it is diagonalizable. Let $\boldsymbol{C}$ be the orthogonal matrix that diagonalizes $\boldsymbol{I}_T - \boldsymbol{P}_X$. Then, $\boldsymbol{C}'(\boldsymbol{I}_T - \boldsymbol{P}_X)\boldsymbol{C} = \boldsymbol{\Lambda}$. Then, $\boldsymbol{\Lambda}$ has $T-k$ eigenvalues equal to one and k eigenvalues equal to zero. Without loss of generality we can write

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

where $\boldsymbol{z} = \boldsymbol{C}'\boldsymbol{y}^*$. As $\boldsymbol{y}^* \sim N(\boldsymbol{0}, \boldsymbol{I}_T)$, $\boldsymbol{z}$ is also distributed as $N(\boldsymbol{0}, \boldsymbol{I}_T)$, so that its elements z_i are independent, standard normal random variables. Consequently,

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

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

6 Method of Moments Estimation

As mentioned previously, $E(\mathbf{x}_t e_t) = \mathbf{0}$ is hold by construction. Let β_0 denote the true value of β, then

$$E[\mathbf{x}_t(y_t - \mathbf{x}_t'\beta_0)] = \mathbf{0}.$$

Another way to write this is to define the "moment function"

$$g_t(\beta) = \mathbf{x}_t(y_t - \mathbf{x}_t'\beta)$$

and observe that

$$E[g_t(\beta_{\mathbf{0}})] = E[\mathbf{x}_t(y_t - \mathbf{x}_t'\beta_0)] = \mathbf{0}.$$

Note that $E[g_t(\beta)] \neq \mathbf{0}$ when $\beta \neq \beta_0$.

The empirical, or sample analog of a moment $E(X)$ is the sample moment $\sum x_t/T$. Similarly, the empirical analog of $E[g_t(\beta)]$ is

$$\begin{aligned} \bar{g}_T(\beta) &= \frac{1}{T}\sum_{t=1}^{T} g_t(\beta) \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{x}_t'\beta) \\ &= \frac{1}{T}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta), \end{aligned}$$

where $\mathbf{X}$ is the $T \times k$ matrix with the t-th row vector $\mathbf{x}_t'$.

The *method of moment estimator* (MME) of β is the vector $\hat{\beta}_T$ such that $\bar{g}_T(\hat{\beta}_T) = 0$. Thus $\hat{\beta}_T$ is defined to mimic, as closely as possible, the orthogonality property $E(\mathbf{x}_t e_t) = \mathbf{0}$. Thus

$$\mathbf{0} = \frac{1}{T}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\beta}_T)$$

which implies

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\hat{\beta}_T.$$

This equation is called the "normal equation". The solution is the MME of $\boldsymbol{\beta}$.

Proposition 6.1 *The MME solution to* $\bar{g}_T(\hat{\boldsymbol{\beta}}_T) = \mathbf{0}$ *is*

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

given $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ *exists.*

Define the *predicted (fitted) value* $\hat{y}_t = \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T$ and the *residual*

$$\begin{aligned} \hat{e}_t &= y_t - \hat{y}_t \\ &= y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T. \end{aligned}$$

In vector notation, $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_T$, $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_T$, and $\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}$.

Note that by definition,

$$\mathbf{0} = \bar{g}_T(\hat{\boldsymbol{\beta}}_T) = \frac{1}{T}(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_T) = \frac{1}{T}\boldsymbol{X}'\hat{\boldsymbol{e}}.$$

Thus

$$\boldsymbol{X}'\hat{\boldsymbol{e}} = 0$$

and the residual vector is orthogonal to the columns of $\boldsymbol{X}$. Since the first column of $\boldsymbol{X}$ is a vector of ones, $\boldsymbol{X}'\hat{\boldsymbol{e}} = \mathbf{0}$ implies that $\sum_{t=1}\hat{e}_t = \boldsymbol{\ell}'\hat{\boldsymbol{e}} = 0$.

As

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

and

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{P}_X\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}.$$

That is, $\hat{\boldsymbol{y}}$ is an orthogonal projection of $\boldsymbol{y}$ onto the space spanned by matrix $\boldsymbol{X}$ when $\boldsymbol{X}$ has full column rank and $\hat{\boldsymbol{e}}$ is also an orthogonal projection of $\boldsymbol{y}$ onto the orthogonal complement of the space spanned by $\boldsymbol{X}$. The orthogonal projection of $\boldsymbol{y}$ onto the space spanned by $\boldsymbol{X}$ is equivalent to the OLS (ordinary least squares) estimator of $\boldsymbol{y}$ regressing on $\boldsymbol{X}$.

7 Asymptotic Distribution Theory

In this section, properties of the OLS estimator are explored using "large sample" approximation. For these results, the condition $E(e_t/\boldsymbol{x}_t) =$ is not necessary. Rather, the orthogonal condition $E(\boldsymbol{x}_t e_t) = \mathbf{0}$ is sufficient.

7.1 Some Basic Mathematical Concepts

7.1.1 Some Inequalities

Recall that L_q-norms are defined as

$$\|X\|_q = (\mathrm{E}|X|^q)^{1/q}, \qquad X \in L^q.$$

We have the following inequalities.

(1) *Hölder's Inequality.*

Let $1 < p, q < \infty$ satisfy $1/p + 1/q = 1$ and suppose that $X \in L_p$ and $Y \in L_q$. Then

$$\mathrm{E}|XY| \leq \|X\|_p \|Y\|_q.$$

In particular, if $p = q = 2$, $\mathrm{E}|XY| \leq \|X\|_2 \|Y\|_2$. This is the Cauchy-Schwartz inequality.

(2) *Minkowski's Inequality.*

Let $p \geq 1$ and $X_i \in L_p$, $1 \leq i \leq n$. Then

$$\|X_1 + \cdots + X_n\|_p \leq \|X_1\|_p + \cdots + \|X_n\|_p.$$

The triangle inequality is a special case.

(3) *Chebyshev's Inequality.*

For $\epsilon > 0$ and $p > 0$,

$$\mathrm{P}\{|X| \geq \epsilon\} \leq \epsilon^{-p} \mathrm{E}|X|^p.$$

Taking $X = Y - E(Y)$, $p = 2$, the Chebyshev's inequality becomes

$$\mathrm{P}\{|Y - E(Y)| \geq \epsilon\} \leq \epsilon^{-2} \mathrm{E}|Y - E(Y)|^2.$$

Furthermore, taking $\epsilon = k\sigma_Y$, we have

$$\mathrm{P}\{|Y - E(Y)| \geq k\sigma_Y\} \leq \frac{\sigma_Y^2}{k^2\sigma_Y^2} = \frac{1}{k^2}.$$

(4) *Jensen's Inequality.*

$$\begin{array}{rcll} g(\mathrm{E}X) & \leq & \mathrm{E}g(X) & \text{if } g \text{ is convex,} \\ g(\mathrm{E}X) & \geq & \mathrm{E}g(X) & \text{if } g \text{ is concave.} \end{array}$$

Using (1), we find that $L_q \subseteq L_r$ if $q \leq r$; by (2), the L_q are linear spaces.

If $X \in L_k$, $\mathrm{E}X^k$ is the k-th moment and $\mathrm{E}(X - \mathrm{E}X)^k$ is the k-th central moment of X. If $X \in L_2$, the *variance* of X is the second central moment of X,

$$\mathrm{var}(X) = \mathrm{E}(X - \mathrm{E}X)^2.$$

The number

$$\mathrm{cov}(X, Y) = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y)$$

is the *covariance* of X and Y. If X and Y are vector-valued, square integrable random variables, we write $\mathrm{cov}(X, Y) = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y)' = [\mathrm{cov}(X_i, Y_j)]_{i,j}$ for the *covariance* between X and Y and $\mathrm{cov}(X) = \mathrm{cov}(X, X)$.

7.1.2 Modes of Convergence

Let $\{X_n\}_{n=1,2,\ldots}$ and X be $\mathbb{R}^d$-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$.

1. **Almost Sure Convergence (Convergence with Probability One)**
   We say that X_n converges to X almost surely if

   $$\mathrm{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$$

   and write $X_n \to X$ a.s.(P) or $X_n \to X$ w.p.1. Sometimes, $X_n(\omega)$ is said to converge *almost everywhere* (a.e.) in that space or that $X_n(\omega)$ is *strongly consistent* for b. This is a convergence concept analogous to nonstochastic convergence in the sense that $X_n(\omega) \to X(\omega)$ for all ω outside a P-null set.

Theorem 7.1 *(Komolgorov strong law of large number, SLLN): Let* $\bar{Z}_n \equiv n^{-1} \sum_{t=1}^{n} Z_t$*, where* $\{Z_t\}$ *is a sequence of i. i. d. random variables with* $\mathrm{E}(Z_t) = \mu < \infty$*. Then* $\bar{Z}_n \overset{a.s.}{\rightarrow} \mu$*.*

Proposition 7.1 *Given* $f : \mathcal{R}^k \rightarrow \mathcal{R}^l (k, l < \infty)$ *and any sequence* $\{b_n\}$ *such that* $b_n \overset{a.s.}{\rightarrow} b$*, where* b_n *and* b *are* $k \times 1$ *vectors, if* f *is continuous at* b*, then* $f(b_n) \overset{a.s.}{\rightarrow} f(b)$*.*

2. **Convergence in Probability**

We say that X_n converges to X in probability if for every $\epsilon > 0$,

$$\mathrm{P}(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

and write $X_n \overset{p}{\rightarrow} X$ or $\operatorname{plim} X_n = X$. Note that almost sure convergence implies convergence in probability. The converse is not true in general. Convergence in probability is also referred to as *weak consistency.*

Theorem 7.2 *(Chebshev weak law of large numbers, WLLN): Let* $\bar{Z}_n \equiv n^{-1} \sum_{t=1}^{n} Z_t$*, where* $\{Z_t\}$ *is a sequence of random variables such that* $\mathrm{E}(Z_t) = \mu$*,* $varZ_t = \sigma^2 < \infty$ *for all tand* $cov(Z_t, Z_\tau) = 0$ *for* $t \neq \tau$*. Then* $\bar{Z}_n \overset{a.s.}{\rightarrow} \mu$*.*

Proposition 7.2 *Let* $f : \mathcal{R}^k \rightarrow \mathcal{R}^l$ *and any sequence* $\{b_n\}$ *such that* $b_n \overset{p}{\rightarrow} b$*, where* b_n *and* b *are* $k \times 1$ *vectors, if* f *is continuous at* b*, then* $f(b_n) \overset{p}{\rightarrow} f(b)$*.*

3. **Convergence in the** q**-th Mean (**L_q**-convergence)**

If all X_n and X are in L_q, X_n is said to converge to X in the q-th mean if

$$\mathrm{E}|X_n - X|^q \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

denoted as $X_n \rightarrow^{L_q} X$. When $q = 2$, we have convergence in the quadratic mean, denoted as l.i.q.m. $X_n = X$. Note that if $q > r$, L_q-convergence implies L_r-convergence and that L_q-convergence implies convergence in probability. The converse is not true in general.

4. **Convergence in Distribution (Convergence in Law)**
Let F_n and F be the distribution functions of X_n and X, respectively. X_n is said to converge to X in distribution if

$$\int_{\mathbb{R}^d} g(x)\, dF_n(x) \to \int_{\mathbb{R}^d} g(x)\, dF(x) \quad \text{as } n \to \infty$$

for every bounded continuous function g. This is the case iff $\lim_n F_n(x) = F(x)$ for every continuity point x of F, or equivalently iff $\lim \varphi_n(\lambda) = \varphi(\lambda)$, where the φ are the corresponding characteristic functions. We write $X_n \to^d X$ or $X_n \to^d F$. We also say that F is the *limiting distribution* of X_n or that X_n is asymptotically distributed as F and write $X_n \sim^A F$. Note that convergence in probability implies convergence in distribution. The converse is not true in general.

7.1.3 Order Notations

Let $\{a_t\}$, $\{b_t\}$, $\{c_t\}$ be deterministic sequences.

1. If there is some $\Delta < \infty$ such that $|b_t/c_t| \leq \Delta$ for all sufficiently large t, we say that $\{b_t\}$ is (at most) of order $\{c_t\}$, symbolically $\{b_t\}$ is $O(c_t)$.

2. If $\lim_t |b_t/c_t| = 0$, we say that $\{b_t\}$ is of smaller order than $\{c_t\}$, symbolically $\{b_t\} = o(c_t)$.

Then, if $\{a_t\}$ is $O(t^r)$ and $\{b_t\}$ is $O(t^s)$, $\{a_t b_t\}$ is $O(t^{r+s})$ and $\{a_t + b_t\}$ is $O(t^{\max(r,s)})$. The same is true if O is replaced by o. In more details,

- $\{b_n\}$ is $O(n^a)$ if $\exists N \quad \text{s.t.} \quad \forall n \geq N, |b_n/n^a| \leq \triangle$ for some $\triangle < \infty$, i.e. b_n is is of order n^a.

- $\{b_n\}$ is $o(n^a)$ if for every $\epsilon > 0$, $\exists N \quad \text{s.t.} \quad \forall n \geq N, |b_n/n^a| \leq \epsilon$, i.e. b_n is is of order smaller than n^a.

- If $\{a_n\}$ is $O(n^r)$ and $\{b_n\}$ is $O(n^s)$, then $\{a_n b_n\}$ is $O(n^{r+s})$, $\{a_n + b_n\}$ is $O(n^q)$, where $q = \max\{r, s\}$.

- If $\{a_n\}$ is $o(n^r)$ and $\{b_n\}$ is $o(n^s)$, then $\{a_n b_n\}$ is $o(n^{r+s})$, $\{a_n + b_n\}$ is $o(n^q)$, where $q = \max\{r, s\}$.

- If $\{a_n\}$ is $O(n^r)$ and $\{b_n\}$ is $o(n^s)$, then $\{a_n b_n\}$ is $o(n^{r+s})$, $\{a_n + b_n\}$ is $O(n^q)$, where $q = \max\{r, s\}$.

For a sequence of random variables$\{X_t\}$, we use stochastic order notations.

1. $\{X_t\}$ is $O_{a.s.}(c_t)$ if $\{X_t(\omega)/c_t\}$ is $O(1)$ a.s.; $\{X_t\}$ is $o_{a.s.}(c_t)$ if $\{X_t(\omega)/c_t\}$ is $o(1)$ a.s. (i.e., $X_t/c_t \to 0$ a.s.).
2. $\{X_t\}$ is $O_{\mathrm{P}}(c_t)$ if for every $\epsilon > 0$ there is some finite Δ such that $\mathrm{P}(|X_t/c_t| \geq \Delta) \leq \epsilon$ for all t. $\{X_t\}$ is $o_{\mathrm{P}}(c_t)$ if $X_t/c_t \to^{\mathrm{P}} 0$.

If $\{X_t\}$ is $O_{\mathrm{P}}(1)$ ($o_{\mathrm{P}}(1)$), we say that $\{X_t\}$ is bounded (vanishing) in probability.

7.2 Consistency and Asymptotic Normality of OLS Estimators

Under the linear projection model:

$$\begin{aligned} y_t &= \boldsymbol{x}_t \boldsymbol{\beta}_0 + e_t \\ \mathrm{E}(\boldsymbol{x}_t e_t) &= \mathbf{0}, \end{aligned}$$

that is, $\boldsymbol{x}\boldsymbol{\beta}_0$ is the projection of y on the linear space of $\boldsymbol{x}$. Given $E(\boldsymbol{x}_t e_t) = 0$, $\boldsymbol{x}_t$ may be stochastic, but it must be uncorrelated with e_t and e_t is not required being heteroskedastic. When $\boldsymbol{x}_t$ does *not* include a lagged dependent variable, $E(\boldsymbol{x}_t e_t) = 0$ also permits serially correlated disturbances. Thus, the linear projection model is general enough to include many econometric models as special cases; e.g., the classical linear model, the general linear model, and models with lagged dependent variables as regressors.

Example 7.3 Consider an AR(p) model for y_t:

$$y_t = c + \psi_1 y_{t-1} + \cdots + \psi_p y_{t-p} + e_t. \tag{22}$$

It can be seen that by recursive substitution, y_t is a linear function of the current and past e_t. If $\{e_t\}$ is a white noise, then for $\boldsymbol{x}_t = (1 \;\; y_{t-1} \; \cdots \; y_{t-p})'$, we have

$$\mathrm{E}(\boldsymbol{x}_t e_t) = \mathbf{0},$$

by the white noise property of $\{e_t\}$. Thus, an AR(p) model with $\{e_t\}$ being a white noise satisfies $E(\boldsymbol{x}_t e_t) = \mathbf{0}$. On the other hand, suppose that $\{e_t\}$ is an MA(q) process:

$$e_t = u_t - \phi_1 u_{t-1} - \cdots - \phi_q u_{t-q},$$

where $\{u_t\}$ is a white noise with mean zero and variance σ_u^2. In this case, y_t is known as an ARMA(p,q) process. Note that

$$\mathrm{E}(e_t e_{t-i}) = -(\phi_i - \phi_{i+1}\phi_{i-1} - \cdots - \phi_q \phi_{q-i})\sigma_u^2, \qquad i = 1, \ldots, q,$$

with $\phi_0 = 1$ and $\phi_j = 0$ if $j < 0$. That is, e_t and e_{t-i} are correlated for $i = 1, \ldots, q$. It follows that

$$\mathrm{E}(\boldsymbol{x}_t e_t) \neq \mathbf{0}.$$

This is because (22) does not model the MA disturbances explicitly and is in fact an AR model with serially correlated disturbances.

As to the linear regression model:

$$\begin{aligned} y_t &= \boldsymbol{x}_t \boldsymbol{\beta}_0 + e_t \\ E(e_t | \boldsymbol{x}_t) &= 0, \end{aligned} \tag{23}$$

we have, by the law of iterated expectations

$$E(\boldsymbol{x}_t e_t) = E(E(\boldsymbol{x}_t e_t | \boldsymbol{x}_t)) = E(\boldsymbol{x}_t E(e_t | \boldsymbol{x}_t)) = \mathbf{0}.$$

That is the linear projection models are implied by linear regression models, i.e., the linear regression models are more restrictive than the linear projection models. Besides, under $E(e_t | \boldsymbol{x}_t)$,

$$E(e_t) = E[E(e_t | \boldsymbol{x}_t)] = 0,$$

that is, the unconditional mean is implied by the conditional mean. Furthermore,

$$E(e_t | \boldsymbol{x}_t) = E(y_t - \boldsymbol{x}_t' \boldsymbol{\beta}_0 | \boldsymbol{x}_t) = E(y_t | \boldsymbol{x}_t) - \boldsymbol{x}_t' \boldsymbol{\beta}_0 = 0,$$

that is, $E(y_t | \boldsymbol{x}_t) = \boldsymbol{x}_t' \boldsymbol{\beta}_0$ under linear regression models.

Although $E(e_t | \boldsymbol{x}_t) = 0$ is stronger than $E(\boldsymbol{x}_t e_t) = \mathbf{0}$, it still allows for the classical linear model, the general linear model, and models with stochastic regressors. Models with lagged dependent variables as regressors need not satisfy $E(e_t | \boldsymbol{x}_t) = 0$, however.

7.3 Consistency

Under the linear projection model, finding the best linear L_2 predictor of y amounts to finding the unknown parameter vector $\boldsymbol{\beta}_0$. We say that $\hat{\boldsymbol{\beta}}_T$ is *strongly (weakly) consistent* for $\boldsymbol{\beta}_0$ if $\hat{\boldsymbol{\beta}}_T \to \boldsymbol{\beta}_0$ a.s. (in probability) as sample size T becomes infinitely large. Consistency is clearly a desirable property because it asserts that $\hat{\boldsymbol{\beta}}_T$ will become arbitrarily close to the true parameter vector $\boldsymbol{\beta}_0$ in some probabilistic sense, provided that "enough" information (a sufficiently large sample) is available. Note that consistency is in sharp contrast with the unbiasedness property. While an unbiased estimator is "correct" on the average, its value is not necessarily close to the true parameter, no matter how large the sample is.

The OLS estimator of $\boldsymbol{\beta}_0$ can be expressed as

$$\hat{\boldsymbol{\beta}}_T = \left(\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{x}_t y_t\right) = \boldsymbol{\beta}_0 + \left(\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{x}_t e_t\right), \tag{24}$$

which is consistent for $\boldsymbol{\beta}^*$ provided that the second term on the right-hand side of (24) converges to zero almost surely (in probability).

A convenient approach to establish consistency is to write

$$\hat{\beta}_T = \boldsymbol{\beta}_0 + \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t e_t\right), \tag{25}$$

so that each term in the parenthesis is eventually governed by a suitable law of large numbers.

A sequence of integrable random variables $\{z_t\}$ is said to obey a strong law of large numbers (SLLN) if $\mathrm{E}(z_t) = \mu_t = O(1)$ such that

$$\frac{1}{T}\sum_{t=1}^{T}(z_t - \mu_t) \xrightarrow{\text{a.s.}} 0; \tag{26}$$

$\{z_t\}$ is said to obey a weak law of large numbers (WLLN) if almost sure convergence in (26) is replaced by convergence in probability. For a sequence of random vectors (matrices), a SLLN (WLLN) is defined elementwise. A SLLN (WLLN) asserts that the sample average of random variables essentially follows its mean behavior; random irregularities are eventually "wiped out" by averaging. Below are two well known strong laws.

Theorem 7.4 *(Kolmogorov's SLLN)* Let $\{Z_i\}$ be a sequence of i.i.d. random variables with mean μ. Then $n^{-1}\sum_{i=1}^{n} Z_i \xrightarrow{\text{a.s.}} \mu$.

Theorem 7.5 *(Chebyshev weak law of large number, WLLN)* Let $\bar{Z}_n \equiv= n^{-1}\sum_{i=1}^{n} Z_i$, where $\{Z_i\}$ is a sequence of random variables such that $E(Z_i) = \mu$, $\text{var}(Z_i) = \sigma^2 < \infty$ for all i and $\text{cov}(Z_i, Z_j) = 0$ for $i \neq j$. Then $\bar{Z}_n \to^p \mu$.

Theorem 7.6 *(Markov's SLLN)* Let $\{Z_t\}$ be a sequence of independent random variables with means $E(Z_t) = \mu_t$. If for some $\delta > 0$, $E|Z_t|^{1+\delta}$ are bounded for all t, then $T^{-1}\sum_{t=1}^{T} Z_t \xrightarrow{\text{a.s.}} \mu =: T^{-1}\sum_{t=1}^{T} \mu_t$.

As a non-stochastic sequence can be viewed as a sequence of independent random variables, it obeys Markov's SLLN if it is $O(1)$.

Definition 7.1 *(Definition 3.25, White (1984))* A one-to-one transformation T from Ω to Ω defined on $(\Omega, \mathcal{F}, P)$ is *measurable* provided that $T^{-1}(\mathcal{F}) \subset \mathcal{F}$.

Definition 7.2 *(Definition 3.27, White (1984))* A transformation T from Ω to Ω is *measure preserving* if it is measurable and if $P[T^{-1}(F)] = P(F)$ for all F in $\mathcal{F}$.

Definition 7.3 *(Definition 3.28, White (1984))* Let G_1 be the joint distribution function of the sequence $\{Z_1, Z_2, \ldots\}$, where Z_t is a $q \times 1$ vector, and let $G_{\tau+1}$ be the joint distribution of the sequence $\{Z_{\tau+1}, Z_{\tau+2}, \ldots\}$. The sequence $\{Z_t\}$ is *stationary* if and only if $G_1 = G_{\tau+1}$ for each $\tau \geq 1$.

Proposition 7.3 *(Proposition 3.29, White (1984))* Let Z be a random variable (i.e., $Z(\omega)$ is a measurable function) and T be a measurable-preserving transformation. Let $Z_1(\omega) = Z(\omega)$, $Z_2(\omega) = Z(T\omega), \ldots, Z_n(\omega) = Z(T^{n-1}\omega)$, for each $\omega \in \Omega$. Then $\{Z_t\}$ is a stationary sequence.

Proposition 7.4 *(Proposition 3.30, White (1984))* Let $\{Z_t\}$ be a stationary sequence. Then there exists a measure-preserving transformation T defined on $(\Omega, \mathcal{F}, P)$ such that $Z_1(\omega) = Z_1(\omega)$, $Z_2(\omega) = Z_1(T\omega)$, $Z_3(\omega) = Z_1(T^2\omega), \ldots, Z_n(\omega) = Z_1(T^{n-1}\omega)$ for all $\omega \in \Omega$.

Definition 7.4 *(Definition 3.33, White (1984))* Let $\{Z_t\}$ be a stationary sequence and let T be a measure-preserving transformation of Proposition 7.4 defined on $(\Omega, \mathcal{F}, P)$. Then $\{Z_t\}$ is *ergodic* if and only if any two events F and $G \in \mathcal{F}$, $\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} P(F \cap T^t G) = P(F)\, P(G)$.

Note that we can think of $T^t G$ as being the event G shifted t periods into the future, and since $P(T^t G) = P(G)$ when T is measure preserving, Proposition 7.4 says that an ergodic process (sequence) is one such that for any events F and G, F and $T^t G$ are independent on average in the limit. Thus, ergodicity can be thought of as a form of "average asymptotic independence".

Theorem 7.7 *(Ergodic Theorem)* Let $\{Z_t\}$ be a sequence of stationary ergodic scalar sequence with $E|Z_t| < \infty$ and with means $E(Z_t) = \mu_t$. Then $T^{-1} \sum_{t=1}^{T} Z_t \xrightarrow{\text{a.s.}} \mu = T^{-1} \sum_{t=1}^{T} \mu_t$.

Note from Theorem 7.6 that $E(Z_t)$ need not be a constant and that the average of $E(Z_t)$ need not converge. It is important to note from these examples that a SLLN (WLLN) holds under suitable regularity conditions. Typically, a sequence of random variables $\{z_t\}$ obeys a SLLN if these variables have certain bounded moments and, for each t, $\operatorname{corr}(z_{t+j}, z_t)$ converges to zero sufficiently fast when $j \to \infty$. For examples, weakly stationary AR(p) processes whose autocorrelations $\operatorname{corr}(z_{t+j}, z_t) \to 0$ exponentially fast when $j \to \infty$, MA(q) processes whose autocorrelations $\operatorname{corr}(z_{t+j}, z_t) = 0$ for all $j > q$, and weakly stationary ARMA(p, q) processes all obey a SLLN. Under suitable conditions (first established by McLeish (1975)), more general sequences of weakly dependent and heterogeneously distributed random variables, such as mixing sequences and mixingales, also obey a SLLN. In what follows, we shall not specify the regularity conditions under which a SLLN (WLLN) holds, see e.g., White (1984) and Davidson (1994) for more detailed, primitive conditions.

As $T^{-1} \sum_{t=1}^{T} \mu_t$ is $O(1)$, our definition implies that the the simple average $T^{-1} \sum_{t=1}^{T} z_t$ must be $O_{\text{a.s.}}(1)$. The examples below show that simple averages need not be bounded.

Example 7.8 Consider the sequences $\{t\}$ and $\{t^2\}$, $t = 1, 2, \ldots$. As

$$\sum_{t=1}^{T} t = T(T+1)/2, \qquad \sum_{t=1}^{T} t^2 = T(T+1)(2T+1)/6,$$

the simple averages $T^{-1}\sum_{t=1}^{T} t$ and $T^{-1}\sum_{t=1}^{T} t^2$ both diverge.

Example 7.9 Suppose that $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with mean μ and variance σ_ϵ^2. The sequence $\{t\epsilon_t\}$ is a sequence of independent (but not identically distributed) random variables with mean zero and unbounded $(1+\delta)$th moment and therefore does not obey Markov's SLLN. In fact,

$$\mathrm{var}(\sum_{t=1}^{T} t\epsilon_t) = \sum_{t=1}^{T} t^2 \mathrm{var}(\epsilon_t) = \sigma_\epsilon^2 T(T+1)(2T+1)/6,$$

which is $O(T^3)$. It follows that $\sum_{t=1}^{T} t\epsilon_t$ is $O_{\mathrm{P}}(T^{3/2})$. That is, $T^{-1}\sum_{t=1}^{T} t\epsilon_t$ also diverges.

Example 7.10 Suppose that y_t is a *random walk*: $y_t = y_{t-1}+\epsilon_t$, $t = 1, 2, \ldots$, where ϵ_t are i.i.d. with mean zero and variance σ_ϵ^2. Here, $y_t = \sum_{i=1}^{t} \epsilon_i$ has mean zero and unbounded variance $t\sigma_\epsilon^2$. It can be shown that $\mathrm{var}(\sum_{t=1}^{T} y_t) = O(T^3)$, and hence $\sum_{t=1}^{T} y_t = O_{\mathrm{P}}(T^{3/2})$. That is, $\{y_t\}$ does not obey a WLLN.

In what follows we write

$$\boldsymbol{M}_T = \frac{1}{T}\sum_{t=1}^{T} E(\boldsymbol{x}_t\boldsymbol{x}_t').$$

The result below shows that $\hat{\boldsymbol{\beta}}_T$ is strongly (weakly) consistent for $\boldsymbol{\beta}_0$.

Theorem 7.11 *Suppose that the following conditions hold:*

[B1] $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_0 + e_t$ *such that* $E(\boldsymbol{x}_t e_t) = \boldsymbol{0}$ *for* $t = 1, \ldots, T$*,*

[B2] $\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ *obeys a SLLN (WLLN) such that* $\boldsymbol{M}_T$ *are p.d. and for some* $\delta > 0$, $\det(\boldsymbol{M}_T) > \delta$ *for all* T *sufficiently large,*

[B3] $\{\boldsymbol{x}_t e_t\}$ *obeys a SLLN (WLLN).*

Then $\hat{\boldsymbol{\beta}}_T$ *exists a.s. (in probability) for all* T *sufficiently large and* $\hat{\boldsymbol{\beta}}_T \to \boldsymbol{\beta}_0$ *a.s. (in probability).*

Proof: Given [B2], we have

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t' - \boldsymbol{M}_T \xrightarrow{\text{a.s.}} \mathbf{0}.$$

Note that $\boldsymbol{M}_T$ is $O(1)$. As M_T is bounded away from singularity for all T sufficiently large, $\boldsymbol{M}_T^{-1}$ is also $O(1)$. Proposition 7.2 ensures that

$$\det\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'\right) - \det(\boldsymbol{M}_T) \xrightarrow{\text{a.s.}} \mathbf{0}.$$

Hence for all T sufficiently large, $\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'/T$ is almost surely invertible, so that $\hat{\boldsymbol{\beta}}_T$ exists almost surely. As the inverse function is continuous for all matrices that are nonsingular,

$$\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1} - \boldsymbol{M}_T^{-1} \xrightarrow{\text{a.s.}} \mathbf{0},$$

by Proposition 7.2. Given [B1] and [B3], we have

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t e_t \xrightarrow{\text{a.s.}} \mathrm{E}(\boldsymbol{x}_t e_t) = \mathbf{0}.$$

It follows from (25) that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0 &= \left[\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1} - \boldsymbol{M}_T^{-1}\right]\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t e_t\right) + \boldsymbol{M}_T^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t e_t\right) \\ &= o_{\text{a.s.}}(1) + o_{\text{a.s.}}(1), \end{aligned}$$

i.e., $\hat{\boldsymbol{\beta}}_T \to \boldsymbol{\beta}_0$ a.s. The conclusion for convergence in probability holds similarly. □

Remarks:

1. The conditions [B1]–[B3] are sufficient but not necessary. The OLS estimator may still be consistent even without the LLN effect; see Example 7.12 below.

2. Once the LLN effect sets in, all that matter for OLS consistency is $E(\boldsymbol{x}_t e_t) = \mathbf{0}$. It is evident from the preceding proof that when $\mathrm{E}(\boldsymbol{x}_t e_t) \neq \mathbf{0}$ (so that [B1] is violated), $\hat{\boldsymbol{\beta}}_T$ asymptotically behaves like $\boldsymbol{\beta}_0 + \boldsymbol{M}_T^{-1}\boldsymbol{c}$ and is therefore inconsistent for $\boldsymbol{\beta}_0$. In view of Example 7.3, it is readily seen that the OLS estimator is inconsistent when models contain lagged dependent variables as regressors *and* serially correlated e_t. More examples of inconsistent OLS estimators can be found in Section.

Example 7.12 Consider the simple time trend model

$$y_t = \alpha_0 + \boldsymbol{\beta}_0 t + e_t.$$

It is straightforward to show that the OLS estimators are

$$\begin{aligned}
\hat{\alpha}_T - \alpha_0 &= \frac{\sum_{t=1}^T t^2 \sum_{t=1}^T e_t - \sum_{t=1}^T t \sum_{t=1}^T t e_t}{T \sum_{t=1}^T t^2 - (\sum_{t=1}^T t)^2}, \\
\hat{\beta}_T - \beta_0 &= \frac{T \sum_{t=1}^T t e_t - \sum_{t=1}^T t \sum_{t=1}^T e_t}{T \sum_{t=1}^T t^2 - (\sum_{t=1}^T t)^2}.
\end{aligned}$$

In view of Examples 7.8 and 7.9, we have

$$\hat{\alpha}_T - \alpha_0 = \frac{O(T^3) o_{\mathrm{P}}(T) - O(T^2) o_{\mathrm{P}}(T^2)}{O(T^4) - O(T^4)} = o_{\mathrm{P}}(1).$$

Similarly, $\hat{\beta}_T \xrightarrow{\mathrm{P}} \beta_0$.

Suppose that there is conditional homoskedasticity $\mathrm{E}(e_t^2 | \boldsymbol{x}_t) = \sigma_0^2$. Then, e_t are also unconditional homoskedastic, i.e., $E(e_t^2) = \sigma_0^2$. It is then easy to verify that under suitable conditions,

$$\hat{\sigma}_T^2 = \frac{1}{T-k} \sum_{t=1}^T (y_t - \boldsymbol{x}_t' \hat{\boldsymbol{\beta}}_T)^2$$

converges σ_0^2 a.s. (in probability).

7.4 Asymptotic Normality

We say that $\hat{\boldsymbol{\beta}}_T$ is asymptotically normally distributed if the sequence of properly normalized $\hat{\boldsymbol{\beta}}_T$ converges in distribution to a multivariate normal random vector, i.e.,

$$\Sigma_T^{-1/2} \sqrt{T} (\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{I}_k),$$

for some nonstochastic $O(1)$ sequence $\{\Sigma_T\}$ with Σ_T a symmetric, p.d. matrix. In this definition, $\{\Sigma_T\}$ is not necessarily a convergent sequence, but if it is (say, $\Sigma_T \to \Sigma$), we also write

$$\sqrt{T} (\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \Sigma).$$

Here, Σ is the covariance matrix of the limiting normal distribution and will be referred to as the asymptotic covariance matrix of $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)$. It must be emphasized that if

$\hat{\boldsymbol{\beta}}_T$ is consistent for $\boldsymbol{\beta}_0$, it is degenerate at $\boldsymbol{\beta}_0$ in the limit and does *not* have a limiting normal distribution.

From (25), we have

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) = \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{x}_t \epsilon_t\right). \tag{27}$$

When Assumption [B2] in Theorem 7.11 holds, the first term on the right-hand side of (27) is essentially $\boldsymbol{M}_T^{-1}$ for large T. If we can show that

$$\Xi_T^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{x}_t \epsilon_t \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k), \tag{28}$$

for some nonstochastic $O(1)$ sequence $\{\Xi_T\}$ with Ξ_T a symmetric, p.d. matrix, then (27) would eventually behave like a linear transformation of $N(\boldsymbol{0}, \boldsymbol{I}_k)$.

A sequence of square-integrable random variables $\{\boldsymbol{z}_t\}$ in $\mathrm{I\!R}^d$ is said to obey a central limit theorem (CLT) if

$$\boldsymbol{\Sigma}_T^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (\boldsymbol{z}_t - \boldsymbol{\mu}_t) = \boldsymbol{\Sigma}_T^{-1/2} \sqrt{T}(\bar{\boldsymbol{z}}_T - \bar{\boldsymbol{\mu}}_T) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_d), \tag{29}$$

where $\boldsymbol{\mu}_t = \mathrm{E}(\boldsymbol{z}_t)$, $\bar{\boldsymbol{z}}_T = T^{-1}\sum_{t=1}^{T} \boldsymbol{z}_t$, $\bar{\boldsymbol{\mu}}_T = T^{-1}\sum_{t=1}^{T} \boldsymbol{\mu}_t$, and $\boldsymbol{\Sigma}_T = \mathrm{var}(T^{-1/2}\sum_{t=1}^{T} \boldsymbol{z}_t)$ is $O(1)$ and p.d. It is easy to verify that $\boldsymbol{\alpha}'\boldsymbol{z} \sim N(0,1)$ for any α such that $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$ if and only if $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I})$. Thus, (29) is equivalent to

$$\boldsymbol{\alpha}' \left(\boldsymbol{\Sigma}_T^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (\boldsymbol{z}_t - \boldsymbol{\mu}_t)\right) \xrightarrow{D} N(0,1), \tag{30}$$

for any $\boldsymbol{\alpha} \in \mathrm{I\!R}^d$ such that $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$, by the Cramér-Wold device.

Proposition 7.5 *(Central Limit Theorem): Given restriction on the dependence, heterogeneity, and moments of a scalar sequence* $\{Z_t\}$, $(\bar{Z_T} - \bar{\mu_T})/(\bar{\sigma_T}/\sqrt{T}) = \sqrt{T}(\bar{Z_T} - \bar{\mu_T})/(\bar{\sigma_T} \overset{A}{\sim} N(0,1)$, *where* $\bar{\mu_T} \equiv \mathrm{E}(\bar{Z_T})$ *and* $\bar{\sigma_T}^2/T \equiv varZ_T$.

Proposition 7.6 *(*Cram'er-Wold device*): Let* $\{b_n\}$ *be a sequence of random* $k \times 1$ *vectors and suppose that for any real* $k \times 1$ *vector* λ *such that* $\lambda'\lambda = 1$, $\lambda' b_n \overset{A}{\sim} \lambda' Z$, *where* Z *is a* $k \times 1$ *vector with joint distribution function* $F(z)$. *Then the limiting distribution function of* b_n *exists and equals* $F(z)$.

Theorem 7.13 (Lindeberg-Levy Central Limit Theorem)*: Let $\{Z_t\}$ be a sequence of i. i. d. random scalars. If $varZ_t \equiv \sigma^2 < \infty$, $\sigma^2 \neq 0$, then*

$$\sqrt{T}(\bar{Z_T} - \bar{\mu_T})/\bar{\sigma_T} = \sqrt{T}(\bar{Z_T} - \mu)/\sigma = \sqrt{T}\sum_{t=1}^{T}(Z_T - \mu)/\sigma \overset{A}{\sim} N(0,1).$$

Theorem 7.14 (Lindeberg-Feller)*: Let $\{Z_t\}$ be a sequence of independent random scalars with $\mathrm{E}(Z_t) \equiv \mu_t$, $varZ_t \equiv \sigma_t^2 < \infty, \sigma_t^2 \neq 0$, and distribution function $F_t(z)$. Then*

$$\sqrt{T}(\bar{Z}_T - \bar{\mu}_T)/\bar{\sigma}_T \overset{A}{\sim} N(0,1)$$

and

$$\lim_{n\to\infty} \bar{\sigma}_T^{-2}T^{-1}\sum_{t=1}^{T}\int_{(z-\mu_t)^2 > \epsilon T\bar{\sigma}_T^2}(z-\mu_T)^2 dF_t(z) = 0.$$

The last condition of this result is called Lindeberg condition. It essentially requires the average contribution of the extreme tails to the variance of Z_t to be zero in the limit. This implies that Z_t has to have "finite" variance. Since in general, the Lindeberg condition can be somewhat difficult to verify, so it is convenient to have a simpler condition that implies the Lindeberg condition.

Theorem 7.15 (Liapounov) *Let $\{Z_t\}$ be a sequence of independent random scalars with $\mathrm{E}(Z_t) = \mu_t$, $varZ_t = \sigma_t^2$, $\sigma_t^2 \neq 0$, and $\mathrm{E}|Z_t - \mu_t|^{2+\delta} < \Delta < \infty$ for some $\delta > 0$ and all t. If $\bar{\sigma}_T^2 > \delta' > 0$ for all T sufficiently large, then $\sqrt{T}(\bar{Z}_T - \bar{\mu}_T)/\bar{\sigma}_T \overset{A}{\sim} N(0,1)$.*

For obtaining an asymptotic normality result analogous to Theorem 5.3 for independent heterogeneous random variables, we have to apply the Cramér-Wold device to $T^{-1/2}\sum_{t=1}^{T}\lambda' V_T^{-1/2}X_t'\epsilon_t$.

Theorem 7.16 *Let $\{Z_t\}$ be a sequence of independent random scalars with $\mathrm{E}(Z_{Tt}) = \mu_{Tt}$, $varZ_{Tt} = \sigma_{Tt}^2, \sigma_{Tt}^2 \neq 0$, and $\mathrm{E}|Z_{Tt}|^{2+\delta} < \Delta < \infty$ for some $\delta > 0$ and all t. Define $\bar{Z}_T \equiv T^{-1}\sum_{t=1}^{T}Z_{Tt}$, $\bar{\mu}_T \equiv T^{-1}\sum_{t=1}^{T}\mu_{Tt}$ and $\bar{\sigma}_T^2 \equiv var\sqrt{T}\bar{Z}_T = T^{-1}\sum_{t=1}^{T}\sigma_{Tt}^2$. If $\bar{\sigma}_T^2 > \delta' > 0$ for all T sufficiently large, then $\sqrt{T}(\bar{Z}_T - \bar{\mu}_T)/\bar{\sigma}_T \overset{A}{\sim} N(0,1)$.*

A CLT ensures that the distribution of a suitably normalized average will be essentially close to that of the standard normal random variable, regardless of the distributions of the original random variables (apart from some regularity conditions). We shall not specify the regularity conditions under which a CLT holds, but we again note that a sequence of correlated and heterogeneously distributed random variables, such as stationary ARMA(p,q) processes, certain mixing sequences and mixingales, may obey a CLT.

The result below, shows that a linear transformation of an asymptotically normally distributed random vector is still asymptotically normally distributed; for a proof see White (1984, p. 67).

Lemma 7.17 *Let $\{\boldsymbol{z}_T\}$ be a sequence of random vectors in $\mathrm{I\!R}^d$ with mean zero and variance $\boldsymbol{V}_T$ such that both $\boldsymbol{V}_T$ and $\boldsymbol{V}_T^{-1}$ are $O(1)$ and $\boldsymbol{V}_T^{-1/2}\boldsymbol{z}_T \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_d)$. Let $\{\boldsymbol{A}_T\}$ be a non-stochastic $O(1)$ sequence of $m \times d$ matrices with full row rank for all T sufficiently large. Then,*

$$\boldsymbol{\Gamma}_T^{-1/2}\boldsymbol{A}_T\boldsymbol{z}_T \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_m),$$

where $\boldsymbol{\Gamma}_T = \boldsymbol{A}_T\Xi_T\boldsymbol{A}_T'$, and $\boldsymbol{\Gamma}_T$ and $\boldsymbol{\Gamma}_T^{-1}$ are $O(1)$.

In what follows we shall write

$$\Xi_T = \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t e_t\right).$$

We now state the asymptotic normality result for the OLS estimator.

Theorem 7.18 *Given Assumptions* [B1], *suppose that the following conditions hold:*

[B2′] *$\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ obeys a WLLN such that $\boldsymbol{M}_T$ are p.d. and for some $\delta > 0$, $\det(\boldsymbol{M}_T) > \delta$ for all T sufficiently large, and*

[B3′] *$\{\boldsymbol{x}_t e_t\}$ obeys a CLT such that $\Xi_T = O(1)$ is p.d. and for some $\delta > 0$, $\det(\Xi_T) > \delta$ for all T sufficiently large.*

Then, $\Sigma_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k)$, where $\Sigma_T = \boldsymbol{M}_T^{-1}\Xi_T\boldsymbol{M}_T^{-1}$. If, in addition, there exists a $\hat{\Xi}_T$ p.s.d. and symmetric such that $\hat{\Xi}_T - \Xi_T \xrightarrow{\mathrm{P}} \boldsymbol{0}$, then $\hat{\Sigma}_T - \Sigma_T \xrightarrow{\mathrm{P}} \boldsymbol{0}$, and

$$\hat{\Sigma}_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k),$$

where $\hat{\Sigma}_T = (\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'/T)^{-1}\hat{\Xi}_T(\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'/T)^{-1}$.

Proof: We first note that $\sum_{t=1}^{T} \boldsymbol{x}_t e_t/\sqrt{T}$ is $O_{\mathrm{P}}(1)$. Hence,

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t e_t \xrightarrow{\mathrm{P}} \mathbf{0}.$$

In the light of Theorem 7.11, $\hat{\boldsymbol{\beta}}_T$ exists and converges to $\boldsymbol{\beta}_0$ in probability. Observe that Σ_T and Σ_T^{-1} are $O(1)$ by [B2′] and [B3′]. From (27),

$$\begin{aligned}
&\Sigma_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \\
&= \Sigma_T^{-1/2}\boldsymbol{M}_T^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{x}_t e_t\right) + \\
&\qquad \Sigma_T^{-1/2}\left(\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} - \boldsymbol{M}_T^{-1}\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{x}_t e_t\right).
\end{aligned}$$

Clearly, the second term is $o_{\mathrm{P}}(1)$. Then it suffices to find the limiting distribution of the first term. Assumption [B3′] and Lemma 7.17 now ensure that

$$\Sigma_T^{-1/2}\boldsymbol{M}_T^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{x}_t e_t\right) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{I}_k).$$

This proves the first assertion. Given $\hat{\Xi}_T - \Xi_T \xrightarrow{\mathrm{P}} \mathbf{0}$, we have $\hat{\Sigma}_T - \Sigma_T \xrightarrow{\mathrm{P}} \mathbf{0}$, and hence

$$(\hat{\Sigma}_T^{-1/2} - \Sigma_T^{-1/2})\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \xrightarrow{\mathrm{P}} \mathbf{0},$$

It follows that these two expressions have the same limiting distribution. □

Remarks:

1. As $\hat{\boldsymbol{\beta}}_T$ converges to $\boldsymbol{\beta}_0$ at the rate $T^{-1/2}$, it is also said to be a root-T consistent estimator.

2. The second assertion of Theorem 7.18 shows that asymptotic normality is not affected when Σ_T is replaced by a consistent estimator $\hat{\Sigma}_T$. Hence, consistent estimation of Σ_T is crucial in ensuring asymptotic normality in practice. For instance, if $\hat{\Sigma}_T$ in Theorem 7.19 is used when heteroskedasticity is present, normalized OLS estimates will not converge in distribution to $N(\mathbf{0}, \boldsymbol{I}_k)$.

The examples below illustrate two leading cases of consistent estimates of Σ_T.

Example 7.19 Conditional homoskedasticity. Suppose that $\{(\boldsymbol{x}_t' \;\; e_t)'\}$ is an i.i.d. sequence and $E(e_t^2|\boldsymbol{x}_t) = \sigma_0^2$. For any continuous function g, $\{g(\boldsymbol{x}_t, e_t)\}$ is also an i.i.d. sequence; in particular, $\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ is an i.i.d. sequence. In this case,

$$\Xi_T = \frac{1}{T}\sum_{t=1}^{T} E(e_t^2\boldsymbol{x}_t\boldsymbol{x}_t') = E(e_t^2\boldsymbol{x}_t\boldsymbol{x}_t') = \sigma_0^2 E(\boldsymbol{x}_t\boldsymbol{x}_t') =: \sigma_0^2\boldsymbol{M},$$

which does not depend on T, and we can write Ξ_T as Ξ. The consistent estimation of Ξ and Σ now reduces to the consistent estimation of σ_0^2 and $\boldsymbol{M}$. The standard OLS estimator $\hat{\sigma}_T^2 = \sum_{t=1}^T \hat{e}_t^2/(T-k)$ is consistent for σ_0^2, and $T^{-1}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'$ is consistent for $\boldsymbol{M}$ by [B2′]. It follows that

$$\hat{\Xi}_T = \hat{\sigma}_T^2\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'\right), \qquad \hat{\Sigma}_T = \hat{\sigma}_T^2\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}.$$

Apart from the factor T, $\hat{\Sigma}_T$ is precisely the covariance matrix we obtained for the classical linear model.

Example 7.20 Conditional heteroskedasticity. Suppose that $\{(\boldsymbol{x}_t' \;\; e_t)'\}$ is an independent sequence and $E(e_t^2|\boldsymbol{x}_t) = \sigma_t^2$. As in the previous example, for any continuous function g, $\{g(\boldsymbol{x}_t, e_t)\}$ is also an independent sequence. Then,

$$\Xi_T = \frac{1}{T}\sum_{t=1}^{T} E(e_t^2\boldsymbol{x}_t\boldsymbol{x}_t') = \frac{1}{T}\sum_{t=1}^{T} E(\sigma_t^2\boldsymbol{x}_t\boldsymbol{x}_t').$$

Under this framework, it can be shown that with some additional conditions, a consistent estimator for Ξ_T is

$$\hat{\Xi}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'. \tag{31}$$

A consistent estimator for Σ_T is thus

$$\hat{\Sigma}_T = \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^T \hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'\right)\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}.$$

This is the heteroskedasticity-consistent covariance matrix estimator of White (1980), also known as White's covariance matrix estimator. A novel feature of this estimator is that it is consistent even when there is conditional heteroskedasticity of unknown form.

8 Linear Hypothesis Testing: Finite Sample and Large Sample Tests

After a model is estimated, it is important to conduct statistical tests to evaluate various economic or econometric hypotheses in terms of the model parameters. In this section we maintain the discussion in normal regression models and consider the linear hypothesis: $\boldsymbol{R}\boldsymbol{\beta}_0 = \boldsymbol{r}$, 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.

If the null hypothesis is correct, 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 probabilistically and is determined by the underlying distribution of the test statistics.

8.1 Finite Sample Tests

As shown in section 5.7, under the linear normal regression models, the sampling distribution of $\hat{\boldsymbol{\beta}}_T$ is

$$\hat{\boldsymbol{\beta}}_T \sim N(\boldsymbol{\beta}_0, \boldsymbol{V}_T),$$

where $\boldsymbol{V}_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\boldsymbol{D} = \text{var}(\boldsymbol{e}|\boldsymbol{X})$ for cases with heteroskedastic errors; and

$$\hat{\boldsymbol{\beta}}_T \sim N(\boldsymbol{\beta}_0, \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

for cases with homoskedastic errors in which $\boldsymbol{V}_T = \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

8.1.1 *t*- and *F*-Tests

If there is only a single hypothesis, the null hypothesis $\boldsymbol{R}\boldsymbol{\beta}_0 = \boldsymbol{r}$ 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. Under the null hypothesis and models with homoskedastic errors,

$$\frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\sigma_0[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}} = \frac{\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)}{\sigma_0[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}} \sim N(0, 1).$$

Although the left-hand side has a known distribution, it cannot be used as a test statistic because σ_0 is unknown. Replacing σ_0^2 by its OLS estimator $\hat{\sigma}_T^2$ yields an operational statistic:

$$\tau \;=\; \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\hat{\sigma}_T[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}}, \tag{32}$$

Theorem 8.1 *In the normal regression models with homoskedastic errors and under the null hypothesis that $\boldsymbol{R}\boldsymbol{\beta}_0 = \boldsymbol{r}$ with $\boldsymbol{R}$ a $1 \times k$ vector,*

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

where τ is given by (32).

Proof: Writing

$$\tau \;=\; \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\sigma_0[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}} \bigg/ \sqrt{\frac{\hat{\sigma}_T^2(T-k)/\sigma_0^2}{T-k}},$$

we can see that the numerator is distributed as $N(0,1)$, and that the square of the denominator is a central χ^2 random variable divided by its degrees of freedom $T-k$. The assertion follows if 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 normal random vectors with covariance matrix

$$\operatorname{cov}(\hat{\boldsymbol{e}}, \hat{\boldsymbol{\beta}}_T) = \mathrm{E}((\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{e}\boldsymbol{e}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = \sigma_0^2(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{0}.$$

Consequently, $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{e}}$, and hence $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$, are also independent. □

Thus, τ is also known as the t statistic. When the alternative hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_0 \neq \boldsymbol{r}$, this is a two-sided test; when the alternative hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_0 > \boldsymbol{r}$ (or $\boldsymbol{R}\boldsymbol{\beta}_0 < \boldsymbol{r}$), this is a one-sided test. For each test, it is typical to choose a small significance level α, say, 5%. Given α, the critical values $t_{\alpha/2}(T-k)$ and $t_{1-\alpha/2}(T-k)$ for the two-sided t test are such that

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

Hence, the event that $\tau > t_{1-\alpha/2}(T-k)$ or $\tau < t_{\alpha/2}(T-k)$ is unlikely under the null hypothesis (its probability α is small), and the null hypothesis is rejected at the significance level α when τ is either too large or too small relative to $t_{1-\alpha/2}(T-k)$ and $t_{\alpha/2}(T-k)$, respectively. The rejection/"acceptance" dichotomy is associate with the *Neyman-Pearson* approach to the hypothesis testing. The alternative approach, associated with *Fisher*, is to report the p-value which is defined as

$$p = P(t(T-k) \geq t | H_0 \, \text{is true}).$$

The null hypothesis is rejected when p-value is less than $\alpha/2$ or greater than $1-\alpha/2$.

The decision of rejection could be wrong, but the probability of committing such an error (type I error) will not exceed α. That is, the type I error is defined as

$$\text{P}(\text{Reject}\, H_0 | H_0 \, \text{is true}) \; = \; P(\tau \geq t_{1-\alpha/2}(T-k) \, \text{or} \, \tau \leq t_{\alpha/2}(T-k) | H_0 \, \text{is true}).$$

Similarly, for the hypothesis $\boldsymbol{R\beta}_0 > \boldsymbol{r}$ ($\boldsymbol{R\beta}_0 < \boldsymbol{r}$), the null hypothesis is rejected at the significance level α when τ is larger (smaller) than $t_{1-\alpha}(T-k)$ ($t_\alpha(T-k)$).

Example 8.2 Test a single coefficient equal to zero: $\beta_i = 0$. Here, $\boldsymbol{R}$ is the transpose of the ith Cartesian unit vector:

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

so that $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' = S^{ii}$ is the ith diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. The t statistic for this hypothesis, also known as the t ratio, is

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

When a t ratio rejects, it is said that the corresponding estimated coefficient is significantly different from zero; econometric packages usually report t ratios with the coefficient estimates.

Example 8.3 A single hypothesis involves two parameters: $\beta_i + \beta_j = 0$. Here, $\boldsymbol{R}$ is of the form

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

Hence, $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' = S^{ii} + 2S^{ij} + S^{jj}$, where S^{ij} is the (i,j)th element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, and

$$\tau = \frac{\hat{\beta}_{iT} + \hat{\beta}_{jT}}{\hat{\sigma}_T (S^{ii} + 2S^{ij} + S^{jj})^{1/2}} \sim t(T-k).$$

Several hypotheses can also be tested jointly. In this case, the null hypothesis $\boldsymbol{R\beta}_0 = \boldsymbol{r}$ is such that $\boldsymbol{R}$ is a matrix ($q \geq 2$) and $\boldsymbol{r}$ is a vector. As

$$[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_0 \sim N(\boldsymbol{0}, \boldsymbol{I}_q),$$

so that

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_0^2 \sim \chi^2(q).$$

Again, we can replace σ_0^2 by its OLS estimator $\hat{\sigma}_T^2$ to obtain an operational statistic:

$$\varphi \;=\; \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})}{\hat{\sigma}_T^2 q}. \tag{33}$$

Theorem 8.4 *Suppose the linear normal regression model with homoskedastic errors is considered. Then under the null hypothesis that $\boldsymbol{R\beta}_0 = \boldsymbol{r}$ with $\boldsymbol{R}$ a $q \times k$ matrix with rank $q < k$,*

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

where φ is given by (33)*.*

Proof: Note that

$$\varphi \;=\; \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/(\sigma_0^2 q)}{\hat{\sigma}_T^2/\sigma_0^2},$$

which is a ratio of two independent central χ^2 random variables, each divided by its degrees of freedom. □

The statistic φ is therefore 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 $F_\alpha(q, T-k)$ of the F table. Note that if there is only a single hypothesis, the F statistic is just the square of the corresponding t statistic. When φ rejects a joint null hypothesis, it suggests that there is evidence against at least one of its single hypothesis. Doing a joint test of several hypotheses is, however, different from testing these hypotheses separately.

Example 8.5 Joint null hypothesis: H_o: $\beta_1 = b_1$ and $\beta_2 = b_2$. The F statistic is

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

Remark: Consider the null hypothesis that s coefficients being zero. It can be shown that when the corresponding F statistic $\varphi > 1$ ($\varphi < 1$), dropping these s regressors will reduce (increase) $\bar{R}^2$.

8.1.2 An Alternative Approach

Given the constraint $\boldsymbol{R}\boldsymbol{\beta}_0 = \boldsymbol{r}$, the *constrained* OLS estimator can be obtained by finding the saddle point of the Lagrangian:

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

where $\boldsymbol{\lambda}$ is the $q \times 1$ vector of Lagrangian multipliers. By the first-order condition of minimizing Lagrangian:

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} L &= \frac{-2}{T}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{R}'\boldsymbol{\lambda} \overset{\text{set}}{=} 0 \\ \nabla_{\boldsymbol{\lambda}} L &= \boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r} \overset{\text{set}}{=} 0 \end{aligned}$$

or

$$\begin{bmatrix} \frac{\boldsymbol{X}'\boldsymbol{X}}{T} & \boldsymbol{R}' \\ \boldsymbol{R} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \frac{\boldsymbol{X}'\boldsymbol{y}}{T} \\ \boldsymbol{r} \end{bmatrix},$$

the solutions are

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

Note that the vector of constrained OLS residuals is

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

It follows that

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

where the second term on the right-hand side is nothing but the numerator of the F statistic (33). Let $\mathrm{ESS_c} = \ddot{\boldsymbol{\epsilon}}'\ddot{\boldsymbol{\epsilon}}$ denote the ESS of the constrained model and $\mathrm{ESS_u}$ denote the ESS of the unconstrained model. We have from (33) that

$$\varphi = \frac{\mathrm{ESS_c} - \mathrm{ESS_u}}{q\hat{\sigma}_T^2} = \frac{(\mathrm{ESS_c} - \mathrm{ESS_u})/q}{\mathrm{ESS_u}/(T-k)} = \frac{(R_\mathrm{u}^2 - R_\mathrm{c}^2)/q}{(1 - R_\mathrm{u}^2)/(T-k)}; \tag{34}$$

note that (33) and (34) are algebraically equivalent. In other words, the F test can be interpreted as the test of "loss of fit" because it compares the performance of the constrained and unconstrained models.

Example 8.6 Consider the unconstrained model: $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \epsilon_t$ with the hypothesis (constraint) $\beta_2 = \beta_3$. Then, the constrained model is

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

By estimating these two models separately, we obtain ESS_u and ESS_c, from which the F statistic can be easily computed.

Example 8.7 Test the null hypothesis that all the coefficients (except the constant term) equal zero. The resulting constrained model is $y_t = \beta_1 + \epsilon_t$, so that $R_\mathrm{c}^2 = 0$. Hence,

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

This test statistic is also routinely reported by most of econometric packages.

8.2 Large Sample Tests

For practical purposes, it is important to find suitable testing procedures and their distributions under the current framework. This is the topic to which we now turn. Specifically, we will be studying two large-sample tests for the linear hypothesis $\boldsymbol{R}\boldsymbol{\beta}_0 = \boldsymbol{r}$, where $\boldsymbol{R}$ is a $q \times k$ ($q < k$) nonstochastic matrix with rank q, and $\boldsymbol{r}$ is a pre-specified real vector.

8.3 t- and F-tests

As mentioned in the previous section, the t- and F-test statistics follow Student-t and F distributions, respectively, in the linear normal regression models with homoskedastic errors. However, parameter σ_0^2 does not exist in the linear regression model with heteroskedastic errors even normality being assumed for errors. That is $\mathrm{var}(\boldsymbol{e}|\boldsymbol{X}) = D$ but not $\sigma_0^2 I_T$ so that the result $(T-k)\hat{\sigma}_T^2/\sigma_0^2 \sim \chi^2(T-k)$ is invalid.

Given the linear normal regression models, we know that

$$\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0 \sim \mathbf{N}(\mathbf{0}, V_T),$$

where $V_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. And then,

$$\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \sim \mathbf{N}(\mathbf{0}, \boldsymbol{R}V_T\boldsymbol{R}'),$$

furthermore, for the null $H_0 : \boldsymbol{R}\boldsymbol{\beta}_0 = \boldsymbol{r}$ where $\boldsymbol{R}$ is a $q \times k$ matrix,

$$\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})[\boldsymbol{R}V_T\boldsymbol{R}']^{-1/2} \overset{H_0}{\sim} \mathbf{N}(\mathbf{0}, I_q).$$

For $q = 1$, the τ-test statistic is

$$\tau = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{[\boldsymbol{R}\hat{V}_T\boldsymbol{R}']^{1/2}},$$

where $\hat{V}_T$ is an estimator for V_T. As the sampling distribution of $\hat{V}_T$ is difficult to analyzed, so is the τ statistic.

For the linear projection model, the asymptotic distribution for $\hat{\boldsymbol{\beta}}_T$ is

$$\Sigma_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \xrightarrow{D} \mathbf{N}(\mathbf{0}, I_k).$$

Then

$$[\boldsymbol{R}\Sigma_T\boldsymbol{R}]^{-1/2}\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \xrightarrow{D} \mathbf{N}(\mathbf{0}, I_q)$$

and for the null $H_0 : \boldsymbol{R}\boldsymbol{\beta}_0 = \boldsymbol{r}$,

$$[\boldsymbol{R}\Sigma_T\boldsymbol{R}]^{-1/2}\sqrt{T}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}) \xrightarrow{D} \mathbf{N}(\mathbf{0}, I_q).$$

Define the τ-statistic for $q = 1$ as

$$\tau = [\boldsymbol{R}\hat{\Sigma}_T\boldsymbol{R}]^{-1/2}\sqrt{T}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}) \xrightarrow{D} \mathbf{N}(\mathbf{0}, I_q),$$

where $\hat{\Sigma}_T$ is a consistent estimator for Σ_T. Then the null is rejected if t is greater than $Z_{1-\alpha/2}$or less than $Z_{\alpha/2}$.

8.3.1 Wald Test

The consistency property of $\hat{\boldsymbol{\beta}}_T$ suggests that, under the null hypothesis, $\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) = \boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}$ should be close to zero when T becomes large. Thus, whether this difference is close to zero constitutes an evidence for or against the null hypothesis. This is the underlying idea of the Wald test.

Given the conditions of Theorem 7.18,

$$\boldsymbol{\Gamma}_T^{-1/2}\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \xrightarrow{D} \mathbf{N}(\mathbf{0}, \boldsymbol{I}_q), \tag{35}$$

where

$$\boldsymbol{\Gamma}_T = \boldsymbol{R}\Sigma_T\boldsymbol{R}' = \boldsymbol{R}\boldsymbol{M}_T^{-1}\Xi_T\boldsymbol{M}_T^{-1}\boldsymbol{R}'. \tag{36}$$

Clearly, (35) remains valid when $\boldsymbol{\Gamma}_T$ is replaced by its consistent estimator

$$\hat{\boldsymbol{\Gamma}}_T = \boldsymbol{R}\hat{\Sigma}_T\boldsymbol{R}' = \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\hat{\Xi}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}', \tag{37}$$

with $\hat{\Xi}_T$ a consistent estimator of $\Xi_T =$, where

$$\Xi_T = \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t e_t\right).$$

The Wald statistic is

$$\mathcal{W}_T = T(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'\hat{\boldsymbol{\Gamma}}_T^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}), \tag{38}$$

which is the inner product of (35) with $\boldsymbol{\Gamma}_T$ replaced by $\hat{\boldsymbol{\Gamma}}_T$ of (37).

Theorem 8.8 *Suppose that the conditions of Theorem* 7.18 *hold. Then under the null hypothesis,*

$$\mathcal{W}_T \xrightarrow{D} \chi^2(q).$$

where $\mathcal{W}_T$ *is given by* (38).

Proof: As $\hat{\boldsymbol{\Gamma}}_T^{-1/2}\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \xrightarrow{D} \mathbf{N}(\mathbf{0}, \boldsymbol{I}_q)$, it is immediate to have the result. □

Example 8.9 Consider testing s coefficients being zero: $\boldsymbol{R}\boldsymbol{\beta}_0 = \mathbf{0}$ with $\boldsymbol{R} = [\mathbf{0}\ \boldsymbol{I}_s]$. The Wald test statistic is

$$\mathcal{W}_T = T\hat{\boldsymbol{\beta}}_T'\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\hat{\Xi}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}}_T \xrightarrow{D} \chi^2(s).$$

When $\hat{\Xi}_T = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is consistent for Ξ_T, the Wald statistic becomes

$$\mathcal{W}_T = T\hat{\boldsymbol{\beta}}_T'\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}}_T/\hat{\sigma}_T^2,$$

which is just s times the standard F-statistic.

Remarks:

1. Provided that a consistent estimator for Ξ_T can be found, the Wald test is valid for a wide variety of models in which $(\boldsymbol{x}_t' \ e_t)'$ may be non-Gaussian, heteroskedastic, and serially correlated. It is often said that the Wald test can be made robust against heteroskedasticity and serial correlation by estimating Ξ_T properly.

2. If an inconsistent estimator $\hat{\Xi}_T$ is used, then $\hat{\Sigma}_T$ and $\hat{\boldsymbol{\Gamma}}_T$ are also inconsistent, and consequently, $\mathcal{W}_T$ will not have χ^2 distribution in the limit. In this case, the Wald test would reject too often when the null hypothesis is correct.

8.3.2 Lagrange Multiplier Test

We have learned from Section 8.1.2 that, given the constraint $\boldsymbol{R\beta} = \boldsymbol{r}$, the constrained OLS estimator can be obtained by finding the saddle point of the Lagrangian:

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

where $\boldsymbol{\lambda}$ is the $q \times 1$ vector of Lagrange multipliers. The underlying idea of the Lagrange Multiplier (LM) test is to test whether $\boldsymbol{\lambda}$ is sufficiently close to zero. Intuitively, $\boldsymbol{\lambda}$ can be interpreted as the "shadow price" of this constraint, and hence should be small when the constraint is valid (i.e., the null hypothesis is true).

It is easy to find the solutions to the Lagrangian as

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

Here, $\ddot{\boldsymbol{\beta}}_T$ is the constrained OLS estimator, and $\ddot{\boldsymbol{\lambda}}_T$ is the basic ingredient of the LM test. It follows that under the null hypothesis,

$$\boldsymbol{\Lambda}_T^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q), \tag{39}$$

where

$$\boldsymbol{\Lambda}_T = 4(\boldsymbol{R}\boldsymbol{M}_T^{-1}\boldsymbol{R}')^{-1}\boldsymbol{\Gamma}_T(\boldsymbol{R}\boldsymbol{M}_T^{-1}\boldsymbol{R}')^{-1}, \tag{40}$$

with $\boldsymbol{\Gamma}_T$ given by (36). Similar as before, this result remain valid if $\boldsymbol{\Lambda}_T$ is replaced by its consistent estimator:

$$\hat{\boldsymbol{\Lambda}}_T = 4[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}\ddot{\boldsymbol{\Gamma}}_T[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}, \tag{41}$$

with

$$\ddot{\boldsymbol{\Gamma}}_T = \boldsymbol{R}\ddot{\boldsymbol{D}}_T\boldsymbol{R}' = \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\ddot{\Xi}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}',$$

where $\ddot{\Xi}_T$ a consistent estimator of Ξ_T computed from the constrained regression model.

Let $\ddot{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\ddot{\boldsymbol{\beta}}_T$ denote the vector of constrained OLS residuals. It is easy to see that

$$\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r} = \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\ddot{\boldsymbol{\beta}}_T)/T = \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{X}'\ddot{\boldsymbol{\epsilon}}/T,$$

and hence that

$$\ddot{\boldsymbol{\lambda}}_T = 2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{X}'\ddot{\boldsymbol{\epsilon}}/T.$$

The LM test statistic is thus

$$\begin{aligned} \mathcal{LM}_T &= T\ddot{\boldsymbol{\lambda}}_T'\hat{\boldsymbol{\Lambda}}_T^{-1}\ddot{\boldsymbol{\lambda}}_T \\ &= T\ddot{\boldsymbol{\epsilon}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\ddot{\boldsymbol{\Gamma}}_T^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ddot{\boldsymbol{\epsilon}}, \end{aligned} \tag{42}$$

where the second equality follows from (41). In view of (42), we can see that only constrained estimation is needed to compute the LM statistic.

Theorem 8.10 *Suppose that the conditions of Theorem* 7.18 *hold. Then under the null hypothesis,*

$$\mathcal{LM}_T \xrightarrow{D} \chi^2(q),$$

where $\mathcal{LM}_T$ *is given by* (42)*.*

Proof: As $\hat{\boldsymbol{\Lambda}}_T^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q)$, the result is followed. □

Example 8.11 Consider again testing s coefficients being zero: $\boldsymbol{R\beta}_0 = \boldsymbol{0}$ with $\boldsymbol{R} = [\boldsymbol{0}\,\boldsymbol{I}_s]$. Accordingly, the original model can be written as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{X}_2\boldsymbol{b}_2 + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $T \times (k-s)$ and $T \times s$ matrices, respectively, and the constrained model is $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{\epsilon}$, so that the constrained OLS estimator is $\ddot{\boldsymbol{\beta}}_T = (\ddot{\boldsymbol{b}}_{1T}'\ \boldsymbol{0})'$, where

$$\ddot{\boldsymbol{b}}_{1T} = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y},$$

and the constrained OLS residual is $\ddot{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}_1 \ddot{\boldsymbol{b}}_{1T}$. The LM statistic can then be computed as (42). When $\ddot{\Xi}_T = \ddot{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is consistent for Ξ_T, where $\ddot{\sigma}_T^2 = \sum_{t=1}^T \ddot{e}_t^2/(T-k+s)$, we have

$$\ddot{\boldsymbol{\Gamma}}_T^{-1} = (\boldsymbol{R}\ddot{\Sigma}_T\boldsymbol{R}')^{-1} = (\ddot{\sigma}_T^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}')^{-1}.$$

It can be verified that by the Frisch-Waugh-Lovell Theorem,

$$\begin{aligned}
\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' &= [\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}, \\
\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' &= [\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1),
\end{aligned}$$

where $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$. The LM statistic now can be simplified as

$$\begin{aligned}
\mathcal{LM}_T &= \ddot{\boldsymbol{\epsilon}}(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2[\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\ddot{\epsilon}/\ddot{\sigma}_T^2 \\
&= \ddot{\boldsymbol{\epsilon}}\boldsymbol{X}_2[\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'\ddot{\epsilon}/\ddot{\sigma}_T^2,
\end{aligned}$$

because $\boldsymbol{X}_1'\ddot{\boldsymbol{\epsilon}} = \boldsymbol{0}$ so that $\boldsymbol{P}_1\ddot{\boldsymbol{\epsilon}} = \boldsymbol{0}$. It is then straightforward to see that

$$\mathcal{LM}_T = \frac{\ddot{\boldsymbol{\epsilon}}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ddot{\boldsymbol{\epsilon}}}{\ddot{\boldsymbol{\epsilon}}'\ddot{\boldsymbol{\epsilon}}/(T-k+s)} = (T-k+s)R^2,$$

where R^2 is the (non-centered) coefficient of determination of regressing $\ddot{\epsilon}$ on the complete data matrix X. If an ML estimator $\ddot{\sigma}_T^2 = \sum_{t=1}^T \ddot{\epsilon}_t^2/T$ is used, we simply have TR^2 as the test statistic.

Remarks:

1. The LM test is also applicable to models with $(\boldsymbol{x}_t' \;\; e_t)'$ being non-Gaussian, heteroskedastic, and serially correlated, as long as a consistent estimator for Ξ_T under the null hypothesis can be found. If $\ddot{\Xi}_T$ is inconsistent, then so are $\ddot{\Sigma}_T$, $\ddot{\boldsymbol{\Gamma}}_T$, and $\hat{\boldsymbol{\Lambda}}_T$. Consequently, the LM test will not have a χ^2 distribution in the limit.

2. While the Wald test requires estimating the unconstrained model, the LM test relies on constrained estimation. Thus, the Wald test is convenient when the constrained model is difficult to compute, such as a model with nonlinear constraints, and the LM test is easier to implement if the constrained model can be easily computed.

3. It can be shown that the Wald and LM statistics not only have the same limiting χ^2 distribution but also are asymptotically equivalent under the null hypothesis, i.e., $\mathcal{W}_T - \mathcal{LM}_T \xrightarrow{\mathrm{P}} 0$. If Ξ_T is known, these two statistics turn out to be algebraically equivalent. Note, however, that these two tests may result in conflicting statistical inferences. For instance, it can be shown that in the "ideal" case that there is no heteroskedasticity and serial correlation, $\mathcal{W}_T \geq \mathcal{LM}_T$; see e.g., Godfrey (1988) for more details.

8.4 Confidence Regions

A *confidence interval* for β_i, $(\underline{g}_\alpha, \overline{g}_\alpha)$ with the *confidence coefficient* $(1-\alpha)$ satisfies

$$\mathrm{P}\{\underline{g}_\alpha \leq \beta_i \leq \overline{g}_\alpha\} = 1-\alpha.$$

By (32), we have

$$\mathrm{P}\left\{-t_{\alpha/2}(T-k) \leq \frac{\hat{\beta}_{iT}-\beta_i}{\hat{\sigma}_T\sqrt{S^{ii}}} \leq t_{\alpha/2}(T-k)\right\} = 1-\alpha,$$

so that

$$\underline{g}_\alpha = \hat{\beta}_{iT} - t_{\alpha/2}(T-k)\hat{\sigma}_T\sqrt{S^{ii}}; \qquad \overline{g}_\alpha = \hat{\beta}_{iT} + t_{\alpha/2}(T-k)\hat{\sigma}_T\sqrt{S^{ii}},$$

where $t_{\alpha/2}(T-k)$ is the critical value of the (two-sided) t test at the significance level α.

Let A_1 denote the event that the confidence interval covers β_1 and A_2 denote the event that the confidence interval covers β_2. The intersection $A = A_1 \cap A_2$ is thus the event that a confidence "box" covers both coefficients. Suppose that $\mathrm{P}(A_1) = \mathrm{P}(A_2) = 90\%$. Then, $\mathrm{P}(A) \neq 90\%$ in general. When A_1 and A_2 are independent, $\mathrm{P}(A) = 81\%$, but when these two events are not independent, it becomes difficult to determine $\mathrm{P}(A)$. Hence, it would be difficult to find a proper confidence "box" based on individual confidence intervals. Alternatively, we can construct a confidence region using the result of the joint test (33). Specifically, the *confidence region* for $\boldsymbol{R}\boldsymbol{\beta}_o$ with the confidence coefficient $(1-\alpha)$ satisfies

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

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

Example 8.12 The confidence region for $(\beta_1 = b_1, \beta_2 = b_2)$. Suppose $T - k = 30$ and $\alpha = 0.05$, then $F_{0.05}(2, 30) = 3.32$. In view Example 8.5,

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

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

Remark: A point (β_1, β_2) 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 an illustrative example in Goldberger (1991, Chap. 19).

8.5 Power of the Tests

The *power* of a test is the probability of rejecting the null hypothesis when the null hypothesis is false. Let A denote the event that the test statistic is greater or less than the critical values and P_o and P_a denote the probability measure under the null and alternative hypotheses, respectively. Clearly, $\mathrm{P}_o(A)$ is the size (significance level) of the test, and $\mathrm{P}_a(A)$ is the power. Recall that a null hypothesis is rejected because the event A is unlikely under the null. Hence, a sensible test must be such that A is much more likely under the alternative hypothesis. That is, we would expect that the power of a sensible test, $\mathrm{P}_a(A)$, is greater than its size, $\mathrm{P}_o(A)$. In this section, we study the power performance of the tests discussed in the preceding section.

Theorem 8.13 *Suppose a normal linear regression model with homoskedastic errors is considered. Then under the hypothesis that $\boldsymbol{R}\boldsymbol{\beta}_0 - \boldsymbol{r} = \boldsymbol{\delta}$ with $\boldsymbol{R}$ a $q \times k$ matrix with rank $q < k$,*

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

where φ is given by (33) *and $\boldsymbol{D} = \sigma_0^2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']$.*

Proof: When $\boldsymbol{R}\boldsymbol{\beta}_0 - \boldsymbol{r} = \boldsymbol{\delta}$,

$$\begin{aligned} & [\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_0 \\ & \quad = \ [\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}[\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) + \boldsymbol{\delta}]/\sigma_0 \\ & \quad \sim \ N(\boldsymbol{D}^{-1/2}\boldsymbol{\delta}, \boldsymbol{I}_q). \end{aligned}$$

Hence,

$$\begin{aligned} & (\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_0^2 \\ & \qquad\qquad\qquad \sim \chi^{2*}(q;\ \boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}). \end{aligned}$$

As $(T-k)\hat{\sigma}_T^2/\sigma_0^2$ is still distributed as $\chi^2(T-k)$ by Theorem 5.5(b), the assertion follows because the numerator and denominator of φ are independent. □

Clearly, when the null hypothesis is correct, we have $\boldsymbol{\delta} = 0$ and $\varphi \sim F(q, T-k)$. Hence, Theorem 8.13 includes Theorem 8.4 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 8.1. The implication of Theorem 8.13 is that when $\boldsymbol{R}\boldsymbol{\beta}_o$ deviates farther from the hypothetical value $\boldsymbol{r}$, the non-centrality parameter $\boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}$ will increase, and so will the power. We illustrate this point using the following two examples, where the power are computed using the program GAUSS. Suppose first that the null distribution is $F(2, 20)$. Then the critical value at 5% is 3.49, and for the non-centrality parameter equal to $1, 3, 5$, the probabilities that φ exceeds 3.49 are approximately 12.1%, 28.2%, and 44.3%, respectively. Suppose now that the null distribution is $F(5, 60)$. Then the critical value at 5% is 2.37, and for the non-centrality parameter equal to $1, 3, 5$, 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.

9 Multicollinearity

In Section 4.2 we have seen that a linear specification suffers exact multicollinearity if the basic identifiability requirement (i.e., $\boldsymbol{X}$ is of full column rank) is not satisfied. In this case, the model parameters are not identified and the OLS estimator cannot be computed. This problem can be avoided if models are specified properly.

9.1 Near Multicollinearity

In practice, it is more common that the explanatory variables are related to some extent but do not satisfy an exact linear relationship. This is usually referred to as *near multicollinearity*. As long as there is no exact multicollinearity, the model parameters can still be identified, and the OLS estimator can be uniquely solved as (15) and remains the BLUE. Thus, near multicollinearity should cause no problems, at least theoretically. Nevertheless, we still see complaints about near multicollinearity in empirical studies.

In applications, one may find that the parameter estimates are very sensitive to small changes in data and that, while individual t ratios are all insignificant, the F statistic suggests that the model as a whole is highly significant. These problems 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 ith column $\boldsymbol{x}_i$. By the result of Theorem 4.1, the variance of $\hat{\beta}_{iT}$ is

$$\operatorname{var}(\hat{\beta}_{iT}) = \operatorname{var}([\boldsymbol{x}_i'(\boldsymbol{I}-\boldsymbol{P}_i)\boldsymbol{x}_i]^{-1}x_i'(\boldsymbol{I}-\boldsymbol{P}_i)\boldsymbol{\epsilon}) = \sigma_o^2[\boldsymbol{x}_i'(\boldsymbol{I}-\boldsymbol{P}_i)\boldsymbol{x}_i]^{-1},$$

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

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

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 highly related to other explanatory variables so that $R^2(i)$ is high, $\operatorname{var}(\hat{\beta}_{iT})$ would be large. Thus, $\hat{\beta}_{iT}$ are sensitive to data changes, and corresponding t ratios are likely to be insignificant. Near multicollinearity is not a necessary condition for these problems, however. Large $\operatorname{var}(\hat{\beta}_{iT})$ may be due to small variations of $\boldsymbol{x}_i$ and/or large σ_o^2.

Even when large var$(\hat{\beta}_i)$ 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 the coefficients 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 model. Other sophisticated statistical methods, such as the ridge estimator and principal component regressions, may also be used; details of these methods can be found in other econometrics textbooks.

9.2 Digress: Dummy Variables

A regression model 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_i denote the annual salary of college teacher i and x_i denote the years of teaching experience. Consider the dummy variable: $D_i = 1$ if i is a male and $D_i = 0$ if i is a female. Then, the model

$$y_i = \alpha_o + \alpha_1 D_i + \beta_o x_i + \epsilon_i$$

yields two regression lines with different intercepts. The "male" regression has the intercept $\alpha_o + \alpha_1$, and the "female" regression has the intercept α_o. We may test the hypothesis $\alpha_1 = 0$ to see whether there is a difference between (incoming) salaries of male and female teachers. This model can be expanded to incorporate the interaction term between D and x:

$$y_i = \alpha_0 + \alpha_1 D_i + \beta_0 x_i + \beta_1 (D_i x_i) + \epsilon_i.$$

This produces two regression lines with different intercepts and slopes. The slope of the "male" regression now is $\beta_0 + \beta_1$, and the slope of the "female" regression 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.

Suppose that we want to know whether the education level of the head of household affects the consumption pattern. We may classify the data into three groups: below high school, high school only, college or higher. Let $D_{1i} = 1$ if i has high school degree only and $D_{1i} = 0$ otherwise, and $D_{2i} = 1$ if i has college or higher degree and $D_{2i} = 0$ otherwise. Then, similar to the previous example, the following model,

$$y_i = \alpha_o + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta_o x_i + \epsilon_i,$$

yields three regression lines. The below-high-school regression has the intercept α_o, the high-school regression has the intercept $\alpha_o + \alpha_1$, and the college regression has the intercept $\alpha_o + \alpha_2$. Various interesting hypotheses can be tested based on this specification.

Remark: The preceding examples show that, when a model contains a constant term, the number of dummy variables is always one *less* than the number of categories that dummy variables try to classify. Otherwise, the model will have exact multicollinearity; this is the so-called "dummy variable trap."

10 Generalized Least Squares Theory

10.1 GLS Estimators

Suppose a linear regression model with heteroskedastic error is considered. That is

$$\begin{aligned} y_t &= \boldsymbol{x}_t'\boldsymbol{\beta}_0 + e_t \\ E(e_t|\boldsymbol{x}_t) &= 0 \\ E(\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}) &= \boldsymbol{D}. \end{aligned}$$

As $E(\boldsymbol{D}^{-1/2}\boldsymbol{e}|\boldsymbol{X}) = \boldsymbol{0}$ and

$$\begin{aligned} E[(\boldsymbol{D}^{-1/2}\boldsymbol{e})(\boldsymbol{D}^{-1/2}\boldsymbol{e})'] &= \boldsymbol{D}^{-1/2}E(\boldsymbol{e}\boldsymbol{e}')\boldsymbol{D}^{-1/2} \\ &= \boldsymbol{D}^{-1/2}\boldsymbol{D}\boldsymbol{D}^{-1/2} = I_T, \end{aligned}$$

we can transform the regression model as

$$\begin{aligned} \boldsymbol{D}^{-1/2}\boldsymbol{y} &= \boldsymbol{D}^{-1/2}\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{D}^{-1/2}\boldsymbol{e} \\ \boldsymbol{y}^* &= \boldsymbol{X}^*\boldsymbol{\beta}_0 + \boldsymbol{e}^*, \end{aligned}$$

which becomes a regression model with homoskedastic errors. Then the OLS estimator for $\boldsymbol{\beta}_0$ is written as

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_T &= (\boldsymbol{X}^{*'}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*'}\boldsymbol{y}^* \\ &= [(\boldsymbol{D}^{-1/2}\boldsymbol{X})'(\boldsymbol{D}^{-1/2}\boldsymbol{X})]^{-1}(\boldsymbol{D}^{-1/2}\boldsymbol{X})'\boldsymbol{D}^{-1/2}\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{y}). \end{aligned}$$

The estimator $\tilde{\boldsymbol{\beta}}_T$ is called the *generalized least squares* (GLS) estimator for $\boldsymbol{\beta}_0$ and is sometimes called the *Aitken estimator*.

Since $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}$,

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_T &= (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}[\boldsymbol{X}'\boldsymbol{D}^{-1}(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e})] \\ &= \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{e}). \end{aligned}$$

Since $\boldsymbol{D}$ is a unction of $\boldsymbol{X}$, $E(\tilde{\boldsymbol{\beta}}_T) = \boldsymbol{\beta}_)$ and

$$\begin{aligned} \mathrm{var}(\tilde{\boldsymbol{\beta}}_T|\boldsymbol{X}) &= E\{[(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{e})][(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{e})]'|\boldsymbol{X}\} \\ &= (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}. \end{aligned}$$

The class of unbiased estimators take the form

$$\check{\boldsymbol{\beta}}_T = A(\boldsymbol{X})'\boldsymbol{y} = A(\boldsymbol{X})'(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e})$$

with restriction $A(\boldsymbol{X})'\boldsymbol{X} = \boldsymbol{I}_k$, where $A(\boldsymbol{X}), T \times k$, is a function of $\boldsymbol{X}$ only. It is clear that OLS is the case $A(\boldsymbol{X}) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and GLS is the one $A(\boldsymbol{X}) = \boldsymbol{D}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}$. The variance of $\check{\boldsymbol{\beta}}_T$ is

$$\mathrm{var}(\check{\boldsymbol{\beta}}_T|\boldsymbol{X}) = A(\boldsymbol{X})'\boldsymbol{D}A(\boldsymbol{X}).$$

Theorem 10.1 (Gauss-Markov Theorem) The best (minimum-variance) liner unbiased estimator (BLUE) is GLS.

Proof:

Let $A^*(\boldsymbol{X}) = \boldsymbol{D}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}$ and then $\tilde{\boldsymbol{\beta}}_T = A^*(\boldsymbol{X})'\boldsymbol{y}$ is the GLS. Let $A(\boldsymbol{X})$ be any other $T \times k$ function of $\boldsymbol{X}$ and it can be represented as $A(\boldsymbol{X}) = A^*(\boldsymbol{X}) + \boldsymbol{C}$. Clearly, $\check{\boldsymbol{\beta}}_T = A(\boldsymbol{X})'\boldsymbol{y}$ is a linear estimator. Since

$$\begin{aligned} A(\boldsymbol{X})'\boldsymbol{y} &= A(\boldsymbol{X})'(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}) \\ &= (A^*(\boldsymbol{X}) + \boldsymbol{C})'(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}) \\ &= A^*(\boldsymbol{X})'\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{C}'\boldsymbol{X}\boldsymbol{\beta}_0 + A^*(\boldsymbol{X})'\boldsymbol{e} + \boldsymbol{C}'\boldsymbol{e}, \end{aligned}$$

it is clear that the estimator $A(\boldsymbol{X})'\boldsymbol{y}$ is unbiased for $\boldsymbol{\beta}_0$ when $\boldsymbol{C}'\boldsymbol{X} = \boldsymbol{0}$. As shown previously,

$$\begin{aligned} \mathrm{var}(\check{\boldsymbol{\beta}}_T|\boldsymbol{X}) &= A(\boldsymbol{X})'\boldsymbol{D}A(\boldsymbol{X}) \\ &= (A^*(\boldsymbol{X}) + \boldsymbol{C})'\boldsymbol{D}(A^*(\boldsymbol{X}) + \boldsymbol{C}) \\ &= A^*(\boldsymbol{X})'\boldsymbol{D}A^*(\boldsymbol{X}) + A^*(\boldsymbol{X})'\boldsymbol{D}\boldsymbol{C} + \boldsymbol{C}'\boldsymbol{D}A^*(\boldsymbol{X}) + \boldsymbol{C}'\boldsymbol{D}\boldsymbol{C} \\ &= A^*(\boldsymbol{X})'\boldsymbol{D}A^*(\boldsymbol{X}) + \boldsymbol{C}'\boldsymbol{D}\boldsymbol{C}, \end{aligned}$$

since

$$\begin{aligned} \boldsymbol{C}'\boldsymbol{D}A^*(\boldsymbol{X}) &= \boldsymbol{C}'\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1} \\ &= \boldsymbol{C}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1} = \boldsymbol{0}, \end{aligned}$$

given the estimator $A(\boldsymbol{X})'\boldsymbol{y}$ is unbiased for $\boldsymbol{\beta}_0$, $\boldsymbol{C}'\boldsymbol{X} = \boldsymbol{0}$. Therefore,

$$\begin{aligned} \mathrm{var}(\check{\boldsymbol{\beta}}_T - \mathrm{var}(\tilde{\boldsymbol{\beta}}_T &= A(\boldsymbol{X})'\boldsymbol{D}A(\boldsymbol{X}) - A^*(\boldsymbol{X})'\boldsymbol{D}A^*(\boldsymbol{X}) \\ &= \boldsymbol{C}'\boldsymbol{D}\boldsymbol{C} \end{aligned}$$

is positive definite. That is, the GLS is the most efficient estimator among linear unbiased estimators. □.

Theoretically speaking, the Gauss-Markov theorem is not a very powerful theorem, because the restriction to linear estimators is quite unnatural. That is, perhaps a "nonlinear" estimator can do even better. However, at least the theorem points out the inefficiency of OLS in regression models with heterokedastic errors. Chamberlain (*Journal of Econometrics*, 1987) established the general result that there is no regular consistent estimator can have a lower asymptotic variance than the GLS estimator in the regression model.

10.2 Feasible GLS

In the previous section, we showed that in regression model, OLS is inefficient relative to GLS, but the latter is infeasible. We discuss feasible approximate GLS estimation.

Suppose that the conditional variance takes the parametric form

$$\begin{aligned} \mathrm{var}(e_t|\boldsymbol{x}_t) &= \sigma_t^2 \\ &= \alpha_0 + \boldsymbol{z}_{t1}'\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}'\boldsymbol{z}_t, \end{aligned}$$

where $\boldsymbol{z}_{t1}$ is some $q \times 1$ function of $\boldsymbol{x}_t$. Typically, $\boldsymbol{z}_{t1}$ are squares (and perhaps levels) of some (or all) elements of $\boldsymbol{x}_t$. Let $\eta_t = e_t^2$. Then

$$E(\eta_t|\boldsymbol{x}_t) = \alpha_0 + \boldsymbol{z}_{t1}'\boldsymbol{\alpha}_1$$

and we have the regression equation

$$\begin{aligned} \eta_t &= \alpha_0 + \boldsymbol{z}_{t1}'\boldsymbol{\alpha}_1 + \xi_t \qquad (43) \\ E(\xi_t|\boldsymbol{x}_t) &= 0. \end{aligned}$$

Clearly, the conditional variance of ξ_t is

$$\begin{aligned} \mathrm{var}(\xi_t|\boldsymbol{x}_t) &= \mathrm{var}(e_t^2|\boldsymbol{x}_t) \\ &= E[(e_t^2 - E(e_t^2|\boldsymbol{x}_t))^2] \\ &= E(e_t^4|\boldsymbol{x}_t) - (E(e_t^2|\boldsymbol{x}_t))^2. \end{aligned}$$

When e_t is independent of $\boldsymbol{x}_t$ then

$$\mathrm{var}(\xi_t|\boldsymbol{x}_t) = E(e_t^4) - \sigma_0^4$$

and under normality it simplifies to

$$\text{var}(\xi_t|\boldsymbol{x}_t) = 2\sigma_0^4.$$

Suppose e_t (and thus η_t) were observed. Then we could estimate $\boldsymbol{\alpha}$ by OLS:

$$\hat{\boldsymbol{\alpha}}_T = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\eta \rightarrow^p \boldsymbol{\alpha}$$

and

$$\sqrt{T}(\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}) \xrightarrow{D} N(0, \boldsymbol{V}_{\boldsymbol{\alpha}}),$$

where

$$\boldsymbol{V}_{\boldsymbol{\alpha}} = [E(\boldsymbol{z}_t\boldsymbol{z}_t')]^{-1}E(\boldsymbol{z}_t\boldsymbol{z}_t'\xi_t^2)[E(\boldsymbol{z}_t\boldsymbol{z}_t')]^{-1}. \tag{44}$$

While e_t is not observed, we have the OLS residual $\hat{e}_t = y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T = e_t - \boldsymbol{x}_t'(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)$. Thus

$$\begin{aligned} \hat{\eta}_t - \eta_t &= \hat{e}_t^2 - e_t^2 \\ &= -2e_t\boldsymbol{x}_t'(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) + (\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)'\boldsymbol{x}_t\boldsymbol{x}_t'(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \\ &= \phi_t, \end{aligned}$$

say. Note that

$$\begin{aligned} &\frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{z}_t\phi_t \\ &= \frac{-2}{T}\sum_{t=1}^T \boldsymbol{z}_t e_t\boldsymbol{x}_t'\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) + \frac{1}{\sqrt{T}}\boldsymbol{z}_t(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)'\boldsymbol{x}_t\boldsymbol{x}_t'(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)\sqrt{T} \\ &\rightarrow^p \mathbf{0}. \end{aligned}$$

Let

$$\tilde{\boldsymbol{\alpha}}_T = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\hat{\eta} \tag{45}$$

be from OLS regression of $\hat{\eta}_t$ on $\boldsymbol{z}_t$. Then

$$\begin{aligned} \sqrt{T}(\tilde{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}) &= \sqrt{T}(\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}) + (T^{-1}\boldsymbol{Z}'\boldsymbol{Z})^{-1}T^{-1/2}\boldsymbol{Z}'\phi \\ &\xrightarrow{D} \mathbf{N}(0, \boldsymbol{V}_{\boldsymbol{\alpha}}). \end{aligned} \tag{46}$$

Thus the fact that η_t is replaced with $\hat{\eta}_t$ is asymptotic irrelevant. We may call (45) the *skedastic* regression, as it is estimating the conditional variance of the regression of y_t on $\boldsymbol{x}_t$. As shown that $\boldsymbol{\alpha}$ is consistently estimated by a simple procedure, and hence we can estimate $\sigma_t^2 = \boldsymbol{z}_t'\boldsymbol{\alpha}$ by $\tilde{\sigma}_t^2 = \boldsymbol{z}_t'\tilde{\boldsymbol{\alpha}}_T$.

Suppose that $\tilde{\sigma}_t^2 > 0$ for all t. Then set

$$\tilde{\boldsymbol{D}} = \mathrm{diag}(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_T^2)$$

and

$$\tilde{\boldsymbol{\beta}}_T = (\boldsymbol{X}'\tilde{\boldsymbol{D}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{D}}^{-1}\boldsymbol{y}.$$

This is the *feasible GLS*, or *FGLS*, estimator of $\boldsymbol{\beta}_0$.

Since there is not a unique specification for the conditional variance the FGLS estimator is not unique, and will depend on the model (and estimation method) for the skedastic regression. One typical problem with implementation of FGLS estimation is that in a linear regression specification, there is no guarantee that $\tilde{\sigma}_t^2 > 0$ for all t. If $\tilde{\sigma}_t^2 < 0$ for some t, then the FGLS estimator is not well defined. Furthermore, if $\tilde{\sigma}_t^2 \approx 0$ for some t, then the FGLS estimator will force the regression equation through the point $(y_t, \boldsymbol{x}_t)$, which is typically undesirable.

It is possible to show that if the skedastic regression correctly specified, then FGLS is asymptotically equivalent to GLS, that is

Theorem 10.2 If the skedastic regression is correctly specified,

$$\sqrt{T}\left(\tilde{\boldsymbol{\beta}}_{GLS} - \tilde{\boldsymbol{\beta}}_{FGLS}\right) \to^p \boldsymbol{0},$$

and thus

$$\sqrt{T}\left(\tilde{\boldsymbol{\beta}}_{FGLS} - \boldsymbol{\beta}_0\right) \xrightarrow{D} \mathbf{N}(\boldsymbol{0}, \boldsymbol{V}),$$

where

$$\boldsymbol{V} = [E(\sigma_t^2\boldsymbol{x}_t\boldsymbol{x}_t')]^{-1}.$$

10.3 Testing for Heteroskedasticity

The hypothesis of homoskedasticity is that $E(e_t^2|bx_t) = \sigma_0^2$, or equivalently that

$$H_0 : \boldsymbol{\alpha}_1 = \mathbf{0}$$

in the regression (43). We may therefore test this hypothesis by the estimation (45) and constructing a Wald statistic.

This hypothesis does not imply that ξ_t is independent of $\boldsymbol{x}_t$. Typically, however, we impose the stronger hypothesis and test the hypothesis that e_t is independent of $\boldsymbol{x}_t$, in which case ξ_t is independent of $\boldsymbol{x}_t$ and the asymptotic variance (44) for $\tilde{\boldsymbol{\alpha}}_T$ simplifies to

$$\boldsymbol{V}_{\boldsymbol{\alpha}} = [E(\boldsymbol{z}_t\boldsymbol{z}_t')]^{-1}E(\xi_t^2). \tag{47}$$

Hence the standard test of H_0 is a classic F (or Wald) test for exclusion of all regressors from the skedastic regression (45). The asymptotic distribution (46) and the asymptotic variance (47) under independence show that this test has an asymptotic chi-square distribution.

Theorem 10.3 Under H_0, and e_t independent of x_t, the Wald test of $H_0 : \boldsymbol{\alpha}_1 = \mathbf{0}$ is asymptotically $\chi^2(q)$.

11 Consistent Estimation of Covariance Matrices

We have seen in the preceding sections that the consistent estimation of Ξ_T is crucial for the asymptotic normality result and the limiting distribution of the Wald and LM tests. In the most general form, Ξ_T can be written as

$$\begin{aligned}
&\operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t e_t\right)\\
&= \frac{1}{T}\sum_{t=1}^{T}\operatorname{var}(\boldsymbol{x}_t e_t) + \frac{1}{T}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T}\mathrm{E}(x_{t-\tau}e_{t-\tau}e_t\boldsymbol{x}_t') + \mathrm{E}(\boldsymbol{x}_t e_t e_{t-\tau}x_{t-\tau}'). \qquad (48)
\end{aligned}$$

In this section we focus on the consistent estimation of this general covariance matrix.

Suppose that $\{(\boldsymbol{x}_t', e_t)'\}$ is an independent sequence so that possible serial correlations among $(\boldsymbol{x}_t'\ \ e_t)'$ are precluded. Then, Ξ_T in (48) reduces to

$$\Xi_T = \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t e_t\right) = \frac{1}{T}\sum_{t=1}^{T}\operatorname{var}(\boldsymbol{x}_t e_t). \qquad (49)$$

It has been shown that, given [B1]–[B3] and other suitable conditions, $\hat{\Xi}_T = \sum_{t=1}^{T}\hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'/T$ is consistent for Ξ_T. It should be clear that, as long as $\boldsymbol{x}_t e_t$ and $\boldsymbol{x}_\tau e_\tau$ are uncorrelated (but not necessarily independent) for all $t \neq \tau$, (48) again reduces to (49), for which $\hat{\Xi}_T = \sum_{t=1}^{T}\hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'/T$ is still consistent under suitable conditions.

When $\boldsymbol{x}_t e_t$ exhibits certain serial correlations, it is still possible to estimate (48) consistently. Let $m(T)$ denote a function of T which diverges to infinity with T but at a slower rate. Suppose that the correlations between $\boldsymbol{x}_t e_t$ and $\boldsymbol{x}_{t-\tau}e_{t-\tau}$ vanish sufficiently fast such that

$$\frac{1}{T}\sum_{\tau=m(T)+1}^{T-1}\sum_{t=\tau+1}^{T}\mathrm{E}(\boldsymbol{x}_t e_t e_{t-\tau}\boldsymbol{x}_{t-\tau}') \to \mathbf{0}.$$

That is, $\boldsymbol{x}_t e_t$ and $\boldsymbol{x}_{t-\tau}e_{t-\tau}$ are asymptotically uncorrelated in a proper way. Then for large T, Ξ_T can be well approximated by

$$\Xi_T^* = \frac{1}{T}\sum_{t=1}^{T}\operatorname{var}(\boldsymbol{x}_t e_t) + \frac{1}{T}\sum_{\tau=1}^{m(T)}\sum_{t=\tau+1}^{T}\mathrm{E}(\boldsymbol{x}_{t-\tau}e_{t-\tau}e_t\boldsymbol{x}_t') + \mathrm{E}(\boldsymbol{x}_t e_t e_{t-\tau}\boldsymbol{x}_{t-\tau}').$$

Estimating Ξ_T then amounts to estimating Ξ_T^*.

White (1984) notes that an estimator based on the sample counterpart of Ξ_T^*,

$$\check{\Xi}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{e}_t^2 \boldsymbol{x}_t \boldsymbol{x}_t' + \frac{1}{T}\sum_{\tau=1}^{m(T)} \sum_{t=\tau+1}^{T} \left(\boldsymbol{x}_{t-\tau}\hat{e}_{t-\tau}\hat{e}_t\boldsymbol{x}_t' + \boldsymbol{x}_t\hat{e}_t\hat{e}_{t-\tau}\boldsymbol{x}_{t-\tau}'\right),$$

is consistent for Ξ_T^* when heteroskedasticity and serial correlation are both present. A major problem with this naive estimator is that $\check{\Xi}_T$ need not be p.s.d. Newey & West (1987) show that with a suitable weighting function $w_{m(T)}(\tau)$,

$$\hat{\Xi}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{e}_t^2 \boldsymbol{x}_t \boldsymbol{x}_t' + \frac{1}{T}\sum_{\tau=1}^{T-1} w_{m(T)}(\tau) \sum_{t=\tau+1}^{T} \left(\boldsymbol{x}_{t-\tau}\hat{e}_{t-\tau}\hat{e}_t\boldsymbol{x}_t' + \boldsymbol{x}_t\hat{e}_t\hat{e}_{t-\tau}\boldsymbol{x}_{t-\tau}'\right) \qquad (50)$$

is guaranteed to be p.s.d. and remains consistent for Ξ_T^*. The estimator (50) is also known as the heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimator. It can be seen that when serial correlation is not considered, the HAC estimator reduces to White's estimator (31).

In particular, Newey & West (1987) adopt the so-called Bartlett kernel:

$$w_{m(T)}(\tau) = \begin{cases} 1-\tau/m(T), & \text{if } 0 \le \tau/m(T) \le 1, \\ 0, & \text{otherwise,} \end{cases}$$

and Gallant (1987) chooses the so-called Parzen kernel:

$$w_{m(T)}(\tau) = \begin{cases} 1-6[\tau/m(T)]^2 + 6[\tau/m(T)]^3, & \text{if } 0 \le \tau/m(T) \le 1/2, \\ 2[1-\tau/m(T)]^3, & \text{if } 1/2 \le \tau/m(T) \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Consider the Bartlett kernel where $w_{m(T)}(\tau) = 1 - \tau/m(T)$. For a fixed $m(T)$, it is decreasing in τ; hence a smaller weight is assigned when two random variables are separated for a long time period (i.e., τ is large). On the other hand, for a fixed τ, $w_{m(T)}(\tau) \to 1$ as $m(T) \to \infty$ and hence entails little loss asymptotically. In practice, a finite number of $m(T)$ must be chosen for computing $\hat{Xi}_T$. It is worth noting that a small $m(T)$ may result in substantial finite-sample bias. For other choices of weighting functions and a method of determining the approximation lags $m(T)$, we refer to Andrews (1991). Finally, we note that White's estimator (31) and the HAC estimator (50) are non-parametric estimators because they are constructed without postulating parametric models for heteroskedasticity and serial correlation.

12 Generalized Method of Moments

12.1 Endogeneity

We say that there is a problem of endogeneity in the linear model $y_t = \boldsymbol{z}_t'\boldsymbol{\beta}_0 + e_t$ if $\boldsymbol{\beta}_0$ is the parameter of interest and that $E(\boldsymbol{z}_t e_t) \neq 0$. This cannot happen if $\boldsymbol{\beta}_0$ is defined by linear projection, so requires a structural interpretation. The coefficient $\boldsymbol{\beta}_0$ must have meaning separately from the definition of a conditional mean or linear projection.

Example 1: Measurement error in the regressors. Suppose that $(y_t, \boldsymbol{x}_t^*)$ are joint random variables, $E(y_t|\boldsymbol{x}_t^*) = \boldsymbol{x}_t^{*'}\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_0$ is the parameter of interest, and $\boldsymbol{x}_t^*$ is **not** observed. Instead, variables $\boldsymbol{x}_t = \boldsymbol{x}_t^* + \boldsymbol{u}_t$ are observed where $\boldsymbol{u}_t$ is an $k \times 1$ measurement error, independent of y_t and $\boldsymbol{x}_t^*$. Then

$$\begin{aligned} y_t &= \boldsymbol{x}_t^{*'}\boldsymbol{\beta}_0 + e_t \\ &= (\boldsymbol{x}_t - \boldsymbol{u}_t)'\boldsymbol{\beta}_0 + e_t \\ &= \boldsymbol{x}_t'\boldsymbol{\beta}_0 + e_t - \boldsymbol{u}_t'\boldsymbol{\beta}_0 \\ &= \boldsymbol{x}_t'\boldsymbol{\beta}_0 + v_t. \end{aligned}$$

The problem is that

$$E(\boldsymbol{x}_t v_t) = E[(\boldsymbol{x}_t^* + \boldsymbol{u}_t)(e_t - \boldsymbol{u}_t'\boldsymbol{\beta}_0)] = -E[(\boldsymbol{u}_t\boldsymbol{u}_t')\boldsymbol{\beta}_0] \neq 0.$$

This is called *measurement error bias.*

Example 2: Supply and Demand. The variables q_i and p_i (quantity and price) are determined jointly by the demand equation

$$q_i = -\beta_1 p_i + e_{1i}$$

and the supply equation

$$q_i = \beta_2 p_i + e_{2i}.$$

Assume that $\boldsymbol{e}_i = (e_{1i} \quad e_{2i})'$ is i.i.d. and $E(\boldsymbol{e}_i) = \boldsymbol{0}$ and $E(\boldsymbol{e}_i\boldsymbol{e}_i') = \boldsymbol{I}_2$. The question is, if we regress q_i on p_i, what happens?

It is helpful to solve for q_i and p_i in terms of the errors. In matrix notation,

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix} \begin{bmatrix} q_i \\ p_i \end{bmatrix} = \begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix}$$

so

$$\begin{aligned} \begin{bmatrix} q_i \\ p_i \end{bmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix} \\ &= \frac{1}{\beta_1 + \beta_2} \begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\beta_2}{\beta_1+\beta_2} e_{1i} + \frac{\beta_1}{\beta_1+\beta_2} e_{2i} \\ \frac{1}{\beta_1+\beta_2}(e_{1i} - e_{2i}) \end{bmatrix}. \end{aligned}$$

The regression of q_i on p_i yields

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n p_i q_i}{\sum_{i=1}^n p_i^2} \\ &\to^p \frac{E(p_i q_i)}{E(p_i^2)} \\ &= \frac{E\left[\left(\frac{1}{\beta_1+\beta_2}(e_{1i} - e_{2i})\right)\left(\frac{\beta_2}{\beta_1+\beta_2} e_{1i} + \frac{\beta_1}{\beta_1+\beta_2} e_{2i}\right)\right]}{E\left[\left(\frac{1}{\beta_1+\beta_2}(e_{1i} - e_{2i})\right)^2\right]} \\ &= \frac{\beta_2 - \beta_1}{2}, \end{aligned}$$

which does not equal either β_1 and β_2. This is called *simultaneous equation bias.*

Example 3. Models with lagged dependent variables and serially correlated errors. Suppose the model is

$$y_t = \boldsymbol{w}_t' \delta_0 + \alpha_0 y_{t-1} + e_t, \quad E(\boldsymbol{w}_t e_t) = \mathbf{0}$$

$$e_t = \rho_0 e_{t-1} + v_t, \quad E(e_{t-1} v_t) = 0.$$

Let $\boldsymbol{x}_t = (\boldsymbol{w}_t, y_{t-1})'$ and $\boldsymbol{\beta}_0 = (\delta', \alpha_0)'$. Then the model is

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta}_0 + e_t,$$

but with $E(\boldsymbol{x}_t e_t) = E[(\boldsymbol{w}_t, y_{t-1})' e_t] = [\mathbf{0}', E(y_{t-1} e_t)]'$. If we assume $E(y_{t-1} v_t) = 0$, $E(y_{t-1} e_{t-1}) = E(y_t e_t)$, and $E(e_t^2) = \sigma_0^2$, it can be shown that

$$E(y_{t-1} e_t) = \frac{\sigma_0^2 \rho_0}{1 - \alpha_0 \rho_0}.$$

Thus $E(\boldsymbol{x}_t e_t) \neq \mathbf{0}$ so that $E(e|\boldsymbol{X}) \neq 0$.

12.2 Instrument Variables

Let the equation of interest be

$$y_t = \boldsymbol{z}_t \boldsymbol{\beta}_0 + e_t \tag{51}$$

where $\boldsymbol{z}_t$ is $k \times 1$, and assume $E(\boldsymbol{z}_t e_t) \neq \mathbf{0}$ so that there is a problem of *endogeneity.* We call (51) the *structural equation.* In matrix notation, this can be written as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_0 + \boldsymbol{e}. \tag{52}$$

Any solution to the problem of endogeneity requires additional information which we call *instrument.*

Definition 12.1 The $l \times 1$ random vector $\boldsymbol{x}_t$ is an *instrument variable* for (51) if $E(\boldsymbol{x}_t e_t) = \mathbf{0}$.

In a typical set-up, some regressors in $\boldsymbol{z}_t$ will be uncorrelated with e_t (for example, at least the intercept). Thus we make the partition

$$\boldsymbol{z}_t = \begin{pmatrix} \boldsymbol{z}_{t1} \\ \boldsymbol{z}_{t2} \end{pmatrix} \tag{53}$$

where $\boldsymbol{z}_{t1}$ is $k_1 \times 1$ with $E(\boldsymbol{z}_{t1} e_t) = \mathbf{0}$ and $\boldsymbol{z}_{t2}$ is $k_2 \times 1$ with $E(\boldsymbol{z}_{t2} e_t) \neq \mathbf{0}$. We call $\boldsymbol{z}_{t1}$ exogenous and $\boldsymbol{z}_{t2}$ endogenous. By the definition (12.1), $\boldsymbol{z}_{t1}$ is an instrument variable for (51), so should be included in $\boldsymbol{x}_t$. So we have partition

$$\boldsymbol{x}_t = \begin{pmatrix} \boldsymbol{z}_{t1} \\ \boldsymbol{x}_{t2} \end{pmatrix} \tag{54}$$

where $\boldsymbol{z}_{t1} = \boldsymbol{x}_{t1}$ are the *included exogenous variables*, and $\boldsymbol{z}_{t2}$ ($l_2 \times 1$) are the *excluded exogenous variables.* That is $\boldsymbol{x}_{t2}$ are variables which could be included in the equation for y_t (in the sense that they are uncorrelated with e_t) yet can be *excluded*, as they would have true zero coefficients in the equation. We say that the model is *just-identified* if $l = k$ (i.e., $l_2 = k_2$) and *over-identified* if $l > k$ (i.e., $l_2 > k_2$). If $l < k$ then the model is not identified. Note that any solution to the problem of endogeneity requires instruments. This does not mean that valid instruments actually exist.

12.3 Reduced Form

The reduced form relationship between the variables or "regressors" $\boldsymbol{z}_t$ and the instruments $\boldsymbol{x}_t$ is found by linear projection. Let

$$\Gamma = E(\boldsymbol{x}_t\boldsymbol{x}_t')^{-1}E(\boldsymbol{x}_t\boldsymbol{z}_t')$$

be the $l \times k$ matrix of coefficients from a projection of $\boldsymbol{z}_t$ on $\boldsymbol{x}_t$, and define

$$\boldsymbol{u}_t = \boldsymbol{z}_t - \boldsymbol{x}_t'\Gamma$$

as the projection errors. Then the reduced form linear relationship between $\boldsymbol{z}_t$ and $\boldsymbol{x}_t$ is

$$\boldsymbol{z}_t = \boldsymbol{x}_t\Gamma + \boldsymbol{u}_t. \tag{55}$$

In matrix notation, (55) can be written as

$$\boldsymbol{Z} = \boldsymbol{X}\Gamma + \boldsymbol{u} \tag{56}$$

where $\boldsymbol{u}$ in $T \times k$.

By construction,

$$E(\boldsymbol{x}_t\boldsymbol{u}_t') = \boldsymbol{0},$$

so (55) is a projection and can be estimated by OLS:

$$\begin{aligned} \boldsymbol{Z} &= \boldsymbol{X}\hat{\Gamma} + \hat{\boldsymbol{u}} \\ \hat{\Gamma} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Z}). \end{aligned}$$

Substituting (56) into (52), we have

$$\begin{aligned} \boldsymbol{y} &= (\boldsymbol{X}\Gamma + \boldsymbol{u})\boldsymbol{\beta}_0 + \boldsymbol{e} \\ &= \boldsymbol{X}\lambda + \boldsymbol{v}, \end{aligned} \tag{57}$$

where

$$\lambda = \Gamma\boldsymbol{\beta}_0 \tag{58}$$

and

$$\boldsymbol{v} = \boldsymbol{u}\boldsymbol{\beta}_0 + \boldsymbol{e}.$$

Observe that

$$E(\boldsymbol{x}_t v_i) = E(\boldsymbol{x}_t \boldsymbol{u}_i')\boldsymbol{\beta}_0 + E(\boldsymbol{x}_t e_i) = 0.$$

Thus (57) is a projection equation and may be estimated by OLS. This is

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\hat{\lambda} + \hat{\boldsymbol{v}}, \\ \hat{\lambda} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \end{aligned}$$

The equation (57) is the reduced form for $\boldsymbol{y}$. (56) and (57) together are the *reduced form equations* for the system.

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\lambda + \boldsymbol{v} \\ \boldsymbol{Z} &= (\boldsymbol{X}\Gamma + \boldsymbol{u}. \end{aligned}$$

12.4 Identification

The structural parameter $\boldsymbol{\beta}_0$ relates to (λ, Γ) through (58). The parameter $\boldsymbol{\beta}_0$ is *identified*, meaning that it can be recovered from the reduced form, if

$$\operatorname{rank}(\Gamma) = k. \tag{59}$$

Assume that (59) holds. If $l = k$, then $\boldsymbol{\beta}_0 = \Gamma^{-1}\lambda$. If $l > k$, then for any $\boldsymbol{W}$, $\boldsymbol{\beta}_0 = (\Gamma'\boldsymbol{W}\Gamma)^{-1}\Gamma'\boldsymbol{W}\lambda$. If (59) is not satisfied, $\boldsymbol{\beta}_0$ can not be recovered from $\lambda, \Gamma)$. Note that a necessary (although not sufficient) condition for (59) is $l \geq k$.

Since $\boldsymbol{X}$ and $\boldsymbol{Z}$ have the common variables $\boldsymbol{X}_1$, i.e., $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $\boldsymbol{Z} = [\boldsymbol{X}_1, \boldsymbol{Z}_2]$, we can partition Γ as

$$\begin{aligned} \Gamma &= \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{I} & \Gamma_{12} \\ \underset{\sim}{0} & \Gamma_{22} \end{bmatrix}, \end{aligned}$$

(56) can be rewritten as

$$\begin{aligned} \boldsymbol{Z}_1 &= \boldsymbol{X}_1 \\ \boldsymbol{Z}_2 &= \boldsymbol{X}_1\Gamma_{12} + \boldsymbol{X}_2\Gamma_{22} + \boldsymbol{u}_2. \end{aligned} \tag{60}$$

$\boldsymbol{\beta}_0$ is identified if $\operatorname{rank}(\Gamma) = k$, which is true if and only if $\operatorname{rank}(\Gamma_{22}) = k_2$. Thus the key to identification of the model rests on the $l_2 \times k_2$ matrix Γ_{22} in (60).

12.5 Instrument Variables Estimation

Suppose the model is just-identified ($k = l$). Then $\boldsymbol{\beta}_0 = \Gamma^{-1}\lambda$. This suggests the Indirect Least Squares (ILS) estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{IV}} &= \hat{\Gamma}^{-1}\hat{\lambda} \\ &= \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\right]^{-1}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\ &= (\boldsymbol{X}'\boldsymbol{Z})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{X}'\boldsymbol{y}.\end{aligned}$$

$\hat{\boldsymbol{\beta}}_{\text{IV}}$ is also called the *instrument variables estimator* of $\boldsymbol{\beta}_0$, where $\boldsymbol{X}$ is used as an instrument for $\boldsymbol{Z}$

Since $(\hat{\lambda}, \hat{\Gamma}) \to^p (\lambda, \Gamma)$ and Γ is invertible, $\hat{\boldsymbol{\beta}}_{\text{IV}}$ is consistent by continuous mapping theory. A more direct way to see consistency is as follows.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{IV}} &= (\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{X}'(\boldsymbol{Z}\boldsymbol{\beta}_0 + \boldsymbol{e}) \\ &= \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{X}'\boldsymbol{e} \to^p \boldsymbol{\beta}_0\end{aligned}$$

as $(\boldsymbol{X}'\boldsymbol{Z})/T \to^p \boldsymbol{M}_T$ a positive definite matrix and $\boldsymbol{X}'\boldsymbol{e}/T \to^p \sum_{t=1}^T E(\boldsymbol{x}_t e_t)/T = 0$ given $E(\boldsymbol{x}_t e_t) = 0$.

We can also derive the IV estimator $\hat{\boldsymbol{\beta}}_{\text{IV}}$ as a MME. As $E(\boldsymbol{x}_t e_t) = 0$ which suggests the moment equation

$$g_t(\boldsymbol{\beta}) = \boldsymbol{x}_t(y_t - \boldsymbol{z}_t'\boldsymbol{\beta}) \tag{61}$$

so that at the true value $\boldsymbol{\beta}_0$ we have the equality

$$E[g_t(\boldsymbol{\beta}_0)] = E[\boldsymbol{x}_t(y_t - \boldsymbol{z}_t'\boldsymbol{\beta}_0)] = E[\boldsymbol{x}_t e_t] = \mathbf{0}. \tag{62}$$

The sample analog is

$$\bar{g}_T(\boldsymbol{\beta}) = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t(y_t - \boldsymbol{z}_t\boldsymbol{\beta}) = \frac{1}{T}(\boldsymbol{X}\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\beta}). \tag{63}$$

The MME sets $\bar{g}_T = 0$, i.e.,

$$\bar{g}_T(\boldsymbol{\beta}) = \frac{1}{T}(\boldsymbol{X}\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\beta}) \overset{=}{\text{set}}\ 0$$

which yields $\hat{\boldsymbol{\beta}}_{\text{IV}} = (\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Note that the consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}_{\text{IV}}$ are discussed detail in White (1984).

12.6 GMM Estimator

In the overidentified case ($l > k$), the simple IV estimator described above does not exist. Since $l > k$, $\bar{g}_T(\boldsymbol{\beta})$ is $l \times 1$ while $\boldsymbol{\beta}_0$ is $k \times 1$, so there is (in general) no $\hat{\boldsymbol{\beta}}_{\text{IV}}$ such that $\bar{g}_T(\hat{\boldsymbol{\beta}}_{\text{IV}}) = 0$. Therefore, the MME has to be generalized. The idea of the generalized method of moments (GMM) is to set this vector "close" to zero.

For some $l \times l$ weight matrix $\boldsymbol{W}_T$ which is positive definite, let

$$J(\boldsymbol{\beta}) = T\bar{g}_T(\boldsymbol{\beta})'\boldsymbol{W}_T\bar{g}_T(\boldsymbol{\beta}).$$

This is a non-negative measure of the "length" of the weighted vector $\bar{g}_T(\boldsymbol{\beta})$. The GMM estimator is minimizes $J(\boldsymbol{\beta})$. The first order conditions for the minimization is

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} T\bar{g}_T(\boldsymbol{\beta})'\boldsymbol{W}_T\bar{g}_T(\boldsymbol{\beta}) \\
&= \nabla_{\boldsymbol{\beta}} T\left[\frac{1}{T}(\boldsymbol{X}\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\beta})\right]' \boldsymbol{W}_T \left[\frac{1}{T}(\boldsymbol{X}\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\beta})\right] \\
&= \frac{1}{T}\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{W}_T(\boldsymbol{X}\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\beta}) \overset{=}{\text{set}} 0,
\end{aligned}$$

so the GMM estimator is

$$\hat{\beta}_{\text{GMM}} = (\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{W}_T\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{W}_T\boldsymbol{X}'\boldsymbol{y}.$$

12.7 2SLS Estimator

Suppose $\boldsymbol{W}_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, then the GMM estimator becomes

$$\hat{\beta}_{\text{GMM}} = (\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Writing

$$\begin{aligned}
\boldsymbol{P}_X &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\
\hat{\boldsymbol{Z}} &= \boldsymbol{P}_X\boldsymbol{Z} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z},
\end{aligned}$$

then

$$\begin{aligned}
\hat{\beta}_{\text{GMM}} &= (\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{Z}'\boldsymbol{P}_X\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{P}_X\boldsymbol{y} \\
&= (\boldsymbol{Z}'\boldsymbol{P}_X'\boldsymbol{P}_X\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{P}_X\boldsymbol{y} \\
&= (\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}})^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{y}.
\end{aligned}$$

The above formula can be considered as

1. First regress $\boldsymbol{Z}$ on $\boldsymbol{X}$, i.e., $\hat{\Gamma}_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}$ and $\hat{\boldsymbol{Z}} = \boldsymbol{X}'\hat{\Gamma}_T = \boldsymbol{P}_X\boldsymbol{Z}$.
2. Second, regress $\boldsymbol{y}$ on $\hat{\boldsymbol{Z}}$, i.e., $\hat{\boldsymbol{\beta}}_{\text{GMM}} = (\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}})^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{y}$.

That is $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ can be obtained by two-stage regressions. Therefore, the 2SLS (2 Stage Least Squares) estimator is the GMM estimator given $\boldsymbol{W}_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, i.e,

$$\hat{\boldsymbol{\beta}}_{\text{2SLS}} = (\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Recall $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$ and $\boldsymbol{X} = [\boldsymbol{Z}_1, \boldsymbol{X}_2]$, then

$$\begin{aligned}\hat{\boldsymbol{Z}} &= [\hat{\boldsymbol{Z}}_1, \hat{\boldsymbol{Z}}_2] \\ &= [\boldsymbol{P}_X\boldsymbol{Z}_1, \boldsymbol{P}_X\boldsymbol{Z}_2] \\ &= [\boldsymbol{Z}_1, \boldsymbol{P}_X\boldsymbol{Z}_2] \\ &= [\boldsymbol{Z}_1, \hat{\boldsymbol{Z}}_2],\end{aligned}$$

since $\boldsymbol{Z}_1$ lies in the span of $\boldsymbol{X}$. Thus in the second stage, we regress $\boldsymbol{y}$ on $\boldsymbol{Z}_1$ and $\hat{\boldsymbol{Z}}_2$. So only the endogenous variables $\boldsymbol{Z}_2$ are replaced by their fitted values:

$$\hat{\boldsymbol{Z}}_2 = \boldsymbol{X}_1\hat{\Gamma}_{12} + \boldsymbol{X}_2\hat{\Gamma}_{22}.$$

12.8 Distribution of GMM Estimator

Assume that $\boldsymbol{W}_T \to^p \boldsymbol{W}$ is positive definite. Let

$$\frac{1}{T}\sum_{t=1}^{T} E(\boldsymbol{x}_t\boldsymbol{z}_t') = \frac{\boldsymbol{Z}'\boldsymbol{X}}{T} = \boldsymbol{M}_T$$

and

$$\Xi_T = \text{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t e_t\right).$$

Then

$$\left(\frac{\boldsymbol{Z}'\boldsymbol{X}}{T}\right)\boldsymbol{W}_T\left(\frac{\boldsymbol{X}'\boldsymbol{Z}}{T}\right) \to^p \boldsymbol{M}_T'\boldsymbol{W}\boldsymbol{M}_T$$

and given suitable conditions

$$\frac{\boldsymbol{X}'\boldsymbol{e}}{\sqrt{T}} \to^d N(0, \Xi_T).$$

The asymptotic normality of $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ can be obtained as follows. Since

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{GMM}} &= (\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{W}_T\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{W}_T\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{W}_T\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{W}_T\boldsymbol{X}'(\boldsymbol{Z}\boldsymbol{\beta}_0+\boldsymbol{e}) \\
&= \boldsymbol{\beta}_0 + (\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{W}_T\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{W}_T\boldsymbol{X}'\boldsymbol{e},
\end{aligned}$$

we have

$$\begin{aligned}
\sqrt{T}(\hat{\boldsymbol{\beta}}_{\text{GMM}}-\boldsymbol{\beta}_0) &= \left(\frac{\boldsymbol{Z}'\boldsymbol{X}}{T}\boldsymbol{W}_T\frac{\boldsymbol{X}'\boldsymbol{Z}}{T}\right)^{-1}\frac{\boldsymbol{Z}'\boldsymbol{X}}{T}\boldsymbol{W}_T\frac{\boldsymbol{X}'\boldsymbol{e}}{\sqrt{T}} \\
&\to^d (\boldsymbol{M}_T'\boldsymbol{W}\boldsymbol{M}_T)^{-1}\boldsymbol{M}_T'\boldsymbol{W}N(0,\Xi_T) \\
&\quad = N(0,(\boldsymbol{M}_T'\boldsymbol{W}\boldsymbol{M}_T)^{-1}\boldsymbol{M}_T'\boldsymbol{W}\Xi_T\boldsymbol{W}\boldsymbol{M}_T(\boldsymbol{M}_T'\boldsymbol{W}\boldsymbol{M}_T)^{-1}).
\end{aligned}$$

12.9 Optimal Weight Matrix

The optimal weight matrix $\boldsymbol{W}_0$ is one which minimizes $\hat{\Sigma}_{\text{GMM}} = \text{var}(\sqrt{T}(\hat{\boldsymbol{\beta}}_{\text{GMM}}-\boldsymbol{\beta}_0))$. This turns out to be $\boldsymbol{W}_0 = \Xi_T^{-1}$. This yields the *efficient GMM* estimator:

$$\hat{\boldsymbol{\beta}}_{\text{GMM}} = (\boldsymbol{Z}'\boldsymbol{X}\Xi_T^{-1}\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\Xi_T^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Then, the variance-covariance matrix of the efficient GMM estimator becomes

$$\begin{aligned}
\text{varvar}(\sqrt{T}(\hat{\boldsymbol{\beta}}_{\text{GMM}}-\boldsymbol{\beta}_0)) &= (\boldsymbol{M}_T'\boldsymbol{W}_0\boldsymbol{M}_T)^{-1}\boldsymbol{M}_T'\boldsymbol{W}_0\Xi_T\boldsymbol{W}_0\boldsymbol{M}_T(\boldsymbol{M}_T'\boldsymbol{W}_0\boldsymbol{M}_T)^{-1} \\
&= (\boldsymbol{M}_T'\Xi_T^{-1}\boldsymbol{M}_T)^{-1}\boldsymbol{M}_T'\Xi_T^{-1}\Xi_T\Xi_T^{-1}\boldsymbol{M}_T(\boldsymbol{M}_T'\Xi_T^{-1}\boldsymbol{M}_T)^{-1} \\
&= (\boldsymbol{M}_T'\Xi_T^{-1}\boldsymbol{M}_T)^{-1}.
\end{aligned}$$

This estimator is efficient only in the sense that it is the best (asymptotically) in the class of GMM estimators with this set of moment conditions. $\boldsymbol{W}_0 = \Xi_T^{-1}$ is not known in practice, but it can be estimated consistently. For any $\hat{\boldsymbol{W}}_T \to^p \boldsymbol{W}_0$, we still call $\hat{\boldsymbol{\beta}}_{\text{GMM}} = (\boldsymbol{Z}'\boldsymbol{X}\hat{\boldsymbol{W}}_T\boldsymbol{X}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\hat{\boldsymbol{W}}_T\boldsymbol{X}'\boldsymbol{y}$ the efficient GMM estimator, as it has the same asymptotic distribution. In the special case that $E(e_t^2|\boldsymbol{x}_t) = \sigma_0^2$ (homokedasticity), then

$$\begin{aligned}
\boldsymbol{W}_0 &= \Xi_T^{-1} = \sigma_0^2\left(\frac{1}{T}\sum_{t=1}^{T}E(\boldsymbol{x}_t\boldsymbol{x}_t')\right)^{-1} \\
&\propto \left(\frac{1}{T}\sum_{t=1}^{T}E(\boldsymbol{x}_t\boldsymbol{x}_t')\right)^{-1}.
\end{aligned}$$

Recall that 2SLS sets $\boldsymbol{W}_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ which is consistent to $\sum_{t=1}^{T} E(\boldsymbol{x}_t\boldsymbol{x}_t')/T$. Thus, under the homoskedasticity assumption 2SLS is *asymptotically efficient.* In general, however, if $l > k$, then 2SLS is asymptotic inefficient.

12.10 Estimation of the Efficient Weight Matrix

Let $\hat{e}_t = y_t - \boldsymbol{z}_t'\hat{\boldsymbol{\beta}}_{\text{2SLS}}$ and then set $\hat{g}_t = \boldsymbol{x}_t\hat{e}_t$ and $\hat{\mathbf{g}}$ be the associated $T \times l$ matrix, $\bar{g}_T = \sum_{t=1}^{T} \hat{g}_t/T$. The efficient GMM estimator is either

$$\hat{\boldsymbol{\beta}}_{\text{GMM}} = \left(\boldsymbol{Z}'\boldsymbol{X}(\hat{\mathbf{g}}'\hat{\mathbf{g}})^{-1}\boldsymbol{X}'\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\boldsymbol{X}(\hat{\mathbf{g}}'\hat{\mathbf{g}})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

or (suggested by Alstair Hall, *Econometrica*, 2000)

$$\hat{\boldsymbol{\beta}}_{\text{GMM}} = \left(\boldsymbol{Z}'\boldsymbol{X}(\hat{\mathbf{g}}'\hat{\mathbf{g}} - T\bar{g}_T\bar{g}_T')^{-1}\boldsymbol{X}'\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\boldsymbol{X}(\hat{\mathbf{g}}'\hat{\mathbf{g}} - T\bar{g}_T\bar{g}_T')^{-1}\boldsymbol{X}'\boldsymbol{y}$$

and their variance-covariance matrices are $(\boldsymbol{Z}'\boldsymbol{X}(\hat{\mathbf{g}}'\hat{\mathbf{g}})^{-1}\boldsymbol{X}'\boldsymbol{Z})^{-1}$ and $(\boldsymbol{Z}'\boldsymbol{X}(\hat{\mathbf{g}}'\hat{\mathbf{g}}-T\bar{g}_T\bar{g}_T')^{-1}\boldsymbol{X}'\boldsymbol{Z})^{-1}$, respectively. Note that in most cases, when we say "GMM", we actually mean "efficient GMM", as there is little point in using an inefficient GMM estimator, and it is so easy to compute.

13 Nonlinear Regression Models

We say that the regression function $g(\boldsymbol{x}, \boldsymbol{\theta}) = E(y_t|\boldsymbol{x}_t = \boldsymbol{x})$ is nonlinear in the parameter if it cannot be written as $g(\boldsymbol{x}, \boldsymbol{\theta}) = z(\boldsymbol{x})'\boldsymbol{\theta}$ for some function $z(\boldsymbol{x})$. Examples of nonlinear regression function include

$$\begin{aligned}
g(x,\theta) &= \theta_1 + \theta_2 \frac{x}{1+\theta_3 x} \\
g(x,\theta) &= \theta_1 + \theta_2 x^{\theta_3} \\
g(x,\theta) &= \theta_1 + \theta_2 \exp(\theta_3 x) \\
g(x,\theta) &= \theta_1 + \theta_2 x + (\theta_3 + \theta_4 x)\Phi\left(\frac{x-\theta_5}{\theta_6}\right) \\
g(x,\theta) &= \theta_1 + \theta_2 x + \theta_4 (x-\theta_3)\, 1(x > \theta_3) \\
g(x,\theta) &= (\theta_1 + \theta_2 x)\, 1(x < \theta_3) + (\theta_4 + \theta_5 x)\, 1(x < \theta_3) \\
g(x,\theta) &= G(x'\theta), \quad G\,\text{known.}
\end{aligned}$$

13.1 NLLS Estimation

The least squares estimator $\hat{\boldsymbol{\theta}}_T$ minimizes the sum-of-squared-errors

$$S_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} (y_t - g(\boldsymbol{x}_t, \boldsymbol{\theta}))^2.$$

When the regression function is nonlinear, we call this the *nonlinear least squares (NLLS) estimator*. The NLLS residuals are $\hat{e}_t = y_t - g(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_T)$. A common method to minimize the function $S_T(\boldsymbol{\theta})$ is the Gauss-Newton method or one of its variants. When $g(\boldsymbol{x}_t, \boldsymbol{\theta})$ is differentiable, then the FOC for minimization are

$$\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_T)\hat{e}_t = \mathbf{0}.$$

13.2 Concentration

A major simplification can be achieved through "concentration." This can be done when we partition $\boldsymbol{\theta} = (\boldsymbol{\beta}', \gamma')'$ so that

$$g(\boldsymbol{x}_t, \boldsymbol{\theta}) = \boldsymbol{\beta}'\boldsymbol{x}_t(\gamma)$$

where $\boldsymbol{x}_t(\gamma)$ is a $k \times 1$ function of $\boldsymbol{x}_t$ and γ. In all the examples, this can be done with γ of much smaller dimension than $\boldsymbol{\beta}$. In many cases, γ is scalar.

The SSE function is $S_T(\boldsymbol{\theta}) = S_T(\boldsymbol{\beta}, \gamma)$ and thus

$$\min_{\boldsymbol{\theta}} S_T(\boldsymbol{\theta}) = \min_{\gamma} \min_{\boldsymbol{\beta}} S_T(\boldsymbol{\beta}, \gamma). \tag{64}$$

Since $\boldsymbol{\beta}$ enters the model linearly, we see that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_T(\gamma) &= \arg\min_{\boldsymbol{\beta}} S_T(\boldsymbol{\beta}, \gamma) \\ &= [\boldsymbol{X}(\gamma)'\boldsymbol{X}(\gamma)]^{-1}\boldsymbol{X}(\gamma)'\boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{X}(\gamma)$ is the $T \times k$ matrix of the stacked $\boldsymbol{x}_t(\gamma)'$.

Now set

$$S_T(\gamma) = S_T(\hat{\boldsymbol{\beta}}_T(\gamma), \gamma)$$

which is the "concentrated" sum of squared errors. We have

$$\begin{aligned} \hat{\gamma}_T &= \arg\min_{\gamma} S_T(\gamma) = \arg\min_{\gamma} S_T(\hat{\boldsymbol{\beta}}_T, \gamma) \\ \hat{\boldsymbol{\beta}}_T &= \hat{\boldsymbol{\beta}}_T(\hat{\gamma}_T). \end{aligned}$$

The pair $(\hat{\boldsymbol{\beta}}_T, \hat{\gamma}_T)$ are the joint NLLS estimates of $(\boldsymbol{\beta}, \gamma)$.

The main benefit of concentration is that the dimension of the numerical optimization is typically reduced dramatically. When γ is scalar, the final minimization over γ can be done by a grid search, for example.

13.3 Computation Using Linearization

A linearization regression can also be used to find the NLLS estimator $\hat{\boldsymbol{\theta}}_T$. It is an iterative technique, meaning that we start with an initial guess $\hat{\boldsymbol{\theta}}_{1T}$, and then define an iteration rule $\hat{\boldsymbol{\theta}}_{jT} \rightarrow \hat{\theta}_{j+1T}$, stopping when the iteration "converge", meaning in practice that the difference $\|\hat{\boldsymbol{\theta}}_{j+1,T} - \hat{\boldsymbol{\theta}}_{jT}\|$ is smaller than some pre-specified level.

We now define the iteration rule

$$\hat{\boldsymbol{\theta}}_{j+1,T} = \hat{\boldsymbol{\theta}}_{jT} + \boldsymbol{d}_j. \tag{65}$$

where the "direction" $\boldsymbol{d}_j$ is a function of $\hat{\boldsymbol{\theta}}_{jT}$. Let

$$\begin{aligned} g_{\boldsymbol{\theta}_{tT}}(j) &= g_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_{jT}) \\ \hat{e}_t(j) &= y_t - g(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_{jT}), \end{aligned}$$

and

$$\boldsymbol{d}_j = \left(\sum_{t=1}^{T} g_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_{jT})\, g_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_{jT})' \right)^{-1} \left(\sum_{t=1}^{T} g_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_{jT}) \hat{e}_t(j) \right).$$

Convergence requires $\boldsymbol{d}_j = \mathbf{0}$, which requires $\sum_{t=1}^{T} g_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_{jT}) \hat{e}_t(j) = \mathbf{0}$, which is the same as the first-order condition for NLLS minimization (64). Thus if (65) converges, it yields the NLLS estimator.

One problem is that the updating rule (65) may tend to overshoot and thus fail to converge. This algorithm can be easily modified to correct for this, by substituting for (65) the rule

$$\hat{\boldsymbol{\theta}}_{j+1,T} = \hat{\boldsymbol{\theta}}_{jT} + \lambda \boldsymbol{d}_j,$$

where $\lambda > 0$ is a scalar "step length". Rules for determining the step length are discussed in the numerical optimization literature. The goal is to find λ so that $S_T^*(\lambda) = S_T(\hat{\boldsymbol{\theta}}_{jT} + \lambda \boldsymbol{d}_j)$ is minimized. One simple rule is the "half" rule. Essentially, try the sequence $\lambda =$

$1, 1/2, 1/4, \ldots$, untill a value of λ is found which reduces the criterion $S_T^*(\lambda)$. Specifically, first compute $S_T^*(1)$. If $S_T^*(1) < S_T^*(0) = S_T(\hat{\boldsymbol{\theta}}_{jT})$, then set $\hat{\boldsymbol{\theta}}_{j+1,T} = \hat{\boldsymbol{\theta}}_{jT} + \boldsymbol{d}_j$. If not, compute $S_T^*(1/2)$. If $S_T^*(1/2) < S_T^*(0)$, then set $\hat{\theta}_{j+1} = \hat{\boldsymbol{\theta}}_{jT} + (1/2)\boldsymbol{d}_j$. This is continued until a value of λ yields an "improvement" in the criterion.

13.4 Asymptotic Distribution

Let $g_{\boldsymbol{\theta}_t} = g_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{\theta}_0)$.

Theorem 13.1 *If the model is identified and $g(\boldsymbol{x}, \boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$,*

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \to^d N(0, \boldsymbol{V}_T)$$

where

$$\boldsymbol{V} = [E(g_{\boldsymbol{\theta}_t} g'_{\boldsymbol{\theta}_t})]^{-1}[E(g_{\boldsymbol{\theta}_t} g'_{\boldsymbol{\theta}_t} e_t^2)][E(g_{\boldsymbol{\theta}_t} g'_{\boldsymbol{\theta}_t})]^{-1}.$$

Proof: Let

$$y_t^0 = e_t + g'_{\boldsymbol{\theta}_t}\boldsymbol{\theta}_0.$$

For $\boldsymbol{\theta}$ close to the true value θ_0, by a first-order Taylor series approximation,

$$g(\boldsymbol{x}_t, \boldsymbol{\theta}) \approx g(\boldsymbol{x}_t, \boldsymbol{\theta}_0) + g'_{\boldsymbol{\theta}_t}(\boldsymbol{\theta} - \theta_0).$$

Thus

$$\begin{aligned} y_t - g(\boldsymbol{x}_t, \boldsymbol{\theta}_0) &\approx (e_t + g(\boldsymbol{x}_t, \boldsymbol{\theta}_0)) - (g(\boldsymbol{x}_t, \boldsymbol{\theta}_0) + g'_{\boldsymbol{\theta}_t}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) \\ &= e_t - g'_{\boldsymbol{\theta}_t}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= y_t^0 - g'_{\boldsymbol{\theta}_t}\boldsymbol{\theta}. \end{aligned}$$

Hence

$$\begin{aligned} \hat{\boldsymbol{\theta}}_T &= \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^T (y_t - g(\boldsymbol{x}_t, \boldsymbol{\theta}))^2 \\ &\approx \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^T (y_t^0 - g'_{\boldsymbol{\theta}_t}\boldsymbol{\theta})^2 \\ &= \left(\sum_{t=1}^T g_{\boldsymbol{\theta}_t} g'_{\boldsymbol{\theta}_t}\right)^{-1} \left(\sum_{t=1}^T g_{\boldsymbol{\theta}_t} y_t^0\right) \\ &= \boldsymbol{\theta}_0 + \left(\sum_{t=1}^T g_{\boldsymbol{\theta}_t} g'_{\boldsymbol{\theta}_t}\right)^{-1} \left(\sum_{t=1}^T g_{\boldsymbol{\theta}_t} e_t\right). \end{aligned}$$

This is a linear regression formula so the asymptotic distribution follows from the theory for OLS. An estimate of the variance-covariance matrix for $\hat{\boldsymbol{\theta}}_T$ is

$$\hat{\boldsymbol{V}} = \left(\sum_{t=1}^{T} \hat{g}_{\boldsymbol{\theta}_t} \hat{g}'_{\boldsymbol{\theta}_t}\right)^{-1} \left(\sum_{t=1}^{T} \hat{g}_{\boldsymbol{\theta}_t} \hat{g}'_{\boldsymbol{\theta}_t} \hat{e}_t^2\right)^{-1} \left(\sum_{t=1}^{T} \hat{g}_{\boldsymbol{\theta}_t} \hat{g}'_{\boldsymbol{\theta}_t}\right)^{-1}$$

where $\hat{g}_{\boldsymbol{\theta}_t} = g_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_T)$ and $\hat{e}_t = y_t - g(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_T)$.

14 Regression Models with Limited Dependent Variables

In the previous sections, the dependent variable in regression models is a quantitative random variable. What happens if we want to use multiple regression to *explain* a qualitative dependent variable.

14.1 A Binary Dependent Variable: the Linear Probability Models

Consider the regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e},$$

where the components in $\boldsymbol{y}$ is a binary variable, i.e, $y_t = 1$ or 0 only. Given $E(\boldsymbol{e}|\boldsymbol{X}) = \boldsymbol{0}$, then we have the conditional mean of y as

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_0.$$

When y_t is a binary variable taking on values zero and one (i.e., a Bernoulli random variable with parameter p which is the probability value of occurring "success"), it is always true that $P(y_t = 1|\boldsymbol{x}_t)) = E(y_t|\boldsymbol{x}_t)$. That is,

$$P(y_t = 1|\boldsymbol{x}_t)) = E(y_t|\boldsymbol{x}_t) = \beta_{10} + \beta_{20}x_{t2} + \cdots + \beta_{k0}x_{tk}, \tag{66}$$

which say that the probability of "success" is a linear function of the $\boldsymbol{x}_t$. Equation (66) is also called the *response probability.* Note that $P(y_t = 0|\boldsymbol{x}_t)) = 1 - P(y_t = 1|\boldsymbol{x}_t))$ is also a linear function of $\boldsymbol{x}_t$. The multiple linear regression model with a binary dependent variable is called the *linear probability model* because the response probability is linear in the parameters β_{j0} which measures the change in the probability of success when x_{tj} changes, holding other factors fixed:

$$\triangle P(y_t = 1|\boldsymbol{x}_t) = \beta_{j0}\triangle x_{tj}.$$

Given the condition of full column rank of $\boldsymbol{X}$, the parameters $\boldsymbol{\beta}_)$ in a linear probability model can be estimated by OLS and then the estimated equation is obtained:

$$\hat{y}_t = \hat{\beta}_{1T} + \hat{\beta}_{2T} x_{t2} + \cdots + \hat{\beta}_{kT} x_{tk},$$

where $\hat{y}_t$ is the predicted probability of success. However, some drawbacks of the OLS estimation for the linear probability model:

1. $\hat{y}_t$ could be less than 0 or greater than 1.

2. A probability cannot be linearly related to the independent variables for all their possible values. For example, $\hat{\beta}_{jT} = 0.262$ and $\triangle x_{tj} = 4$, then $\triangle P(y_t = 1) = 0.262 \times 4 = 1.048$ which is greater than 1.

14.2 Logit and Probit Models for Binary Response

The linear probability model is simple to estimate and use, but it has two most important disadvantages: the fitted probabilities can be less than zero or greater than one and the partial effect of any explanatory variable is constant. These limitations can be overcome by using more sophisticated binary response models.

In a binary response model, interest lies primary in the response probability and it can be transformed as

$$\begin{aligned} P(y=1|\boldsymbol{X}) &= P(y=1|\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_k) \\ &= G(\beta_{10} + \beta_{20}\boldsymbol{x}_2 + \cdots + \beta_{k0}\boldsymbol{x}_k) = G(\boldsymbol{X}\boldsymbol{\beta}_0), \end{aligned} \tag{67}$$

where $0 < G(z) < 1$ for any real numbers z. Various nonlinear functions have been suggested for the function G to make sure that the probabilities are between zero and one. The common used functions are logist and Gaussian density functions. In the *logit model*, G is the logist function:

$$G(z) = \exp(z)/[1 + \exp(z)].$$

This is the cumulative distribution function for a *standard logistic random variable.* In the *probit model*, G is the standard Gaussian cumulative distribution function:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-v^2/2) dv.$$

Logit and probit models can be derived from an *latent variable model.* Let y^* be an unobserved, or latent, variable, defined by

$$y^* = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{e}, y = 1[y^* > 0], \tag{68}$$

where $1[\cdot]$ i an indication function. Therefore, $y = 1$ if $y^* > 0$ and $y = 1$ if $y^* \leq 0$. Assume $\boldsymbol{e}$ is independent of $\boldsymbol{X}$ and that $\boldsymbol{e}$ has the standard logistic distribution or the standard normal distribution. In either case, $\boldsymbol{e}$ is symmetrically distributed about zero, then $1 - G(-z) = G(z)$ for all real numbers z. From (68), we can derive the response probability for y:

$$\begin{aligned} P(y = 1|\boldsymbol{X}) &= P(y^* > 0|\boldsymbol{X}) = P(e > -\boldsymbol{X}\boldsymbol{\beta}_0|\boldsymbol{X}) \\ &= 1 - G(-\boldsymbol{X}\boldsymbol{\beta}_0) = G(\boldsymbol{X}\boldsymbol{\beta}_0), \end{aligned}$$

which is exactly the same as (67).

The partial effect of roughly continuous variable on the response probability is derived as

$$\begin{aligned} \frac{\partial P(y = 1|\boldsymbol{X})}{\partial \boldsymbol{x}_j} &= \frac{\partial G(\boldsymbol{X}\boldsymbol{\beta}_0)}{\partial \boldsymbol{x}_j} \\ &= \frac{dG(z)}{dz}\frac{\partial \boldsymbol{X}\boldsymbol{\beta}_0}{\partial \boldsymbol{x}_j} \\ &= g(\boldsymbol{X}\boldsymbol{\beta}_0)\beta_{j0}. \end{aligned}$$

Because G is the cdf of a continuous random variable, g is the probability density function. In the logit and probit cases, $G(\cdot)$ is a strictly increasing cdf, and so $g(z) > 0$ for all z. Therefore, the partial effect of $\boldsymbol{x}_j$ on $P(y = 1|\boldsymbol{X})$ depends on $\boldsymbol{X}$ through the positive quantity $g(\boldsymbol{X}\boldsymbol{\beta}_0)$, which means that the partial effect always has the same sign as β_{j0}. Besides, the *relative* effects of any two continuous explanatory variables do not depend on $\boldsymbol{X}$ but on β_{j0}/β_{h0}.

As $P(y_t = 1|\boldsymbol{x}_t) = G(\boldsymbol{x}_t'\boldsymbol{\beta}_0)$ and $P(y_t = 0|\boldsymbol{x}_t) = 1 - G(\boldsymbol{x}_t'\boldsymbol{\beta}_0)$, the density of y_t given $\boldsymbol{x}_t$ can be written as

$$f(y_t|\boldsymbol{x}_t; \boldsymbol{\beta}_0) = [G(\boldsymbol{x}_t'\boldsymbol{\beta}_0)]^{y_t}\,[1 - G(\boldsymbol{x}_t'\boldsymbol{\beta}_0)]^{1-y_t}, \quad y_t = 0, 1.$$

The log-likelihood for observation t is

$$\ell_t(\boldsymbol{\beta}_0) = y_t \log[G(\boldsymbol{x}_t'\boldsymbol{\beta}_0)] + (1 - y_t) \log[1 - G(\boldsymbol{x}_t'\boldsymbol{\beta}_0)].$$

Because $G(\cdot)$ is strictly between zero and one for logit and probit, $\ell_t(\boldsymbol{\beta}_0)$ is well-defined for all values of $\boldsymbol{\beta}_0$. Then, the log-likelihood for a sample of T observations is $\mathcal{L}(\boldsymbol{\beta}_0) = \sum_{t=1}^T \ell_t(\boldsymbol{\beta}_0)$ and the maximum likelihood estimator $\tilde{\boldsymbol{\beta}}_T$ for $\boldsymbol{\beta}_0$ is obtained by maximizing $\mathcal{L}(\boldsymbol{\beta}_0)$. The asymptotic variance-covariance matrix of $\tilde{\boldsymbol{\beta}}_T$ is

$$\hat{\text{Avar}}(\tilde{\boldsymbol{\beta}}_T) = \left(\sum_{t=1}^T \frac{[g(\boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}_T)]^2 \boldsymbol{x}_t \boldsymbol{x}_t'}{G(\boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}_T)[1 - G(\boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}_T)]} \right)^{-1}.$$

Several measures of goodness-of-fit have been suggested for the logit and probit models. The *percent correctly predicted* is computed as follows. For each t, denote $\tilde{y}_t = 1$ if $G(\boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}_T > 0.5$ and $\tilde{y}_t = 0$ if $G(\boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}_T \leq 0.5$. Then the percent correctly predicted is computed as $\sum_{t=1}^T 1(\tilde{y}_t = y_t)/T$. That is the percentage of times of predicted y_t matches the actual y_t, i.e, $\tilde{y}_t = y_t$. *Pseudo R-squared* measure for binary response models has been suggested by McFadden (1974) as $1 - \mathcal{L}_u/\mathcal{L}_0$, where

$$\begin{aligned}
\mathcal{L}_u &= \sum_{t=1}^T y_t \log[G(\boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}_T)] + (1 - y_t) \log[1 - G(\boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}_T)] \\
\mathcal{L}_0 &= \sum_{t=1}^T y_t \log[G(\tilde{\beta}_{1T})] + (1 - y_t) \log[1 - G(\tilde{\beta}_{1T})],
\end{aligned}$$

in which $\tilde{\beta}_{1T}$ is the MLE for β_{10} under $\beta_{j0} = 0, \forall j \geq 2$. Some other measures for goodness-of-fit have been suggested by Maddala (1983).

15 The Bootstrap

15.1 An Example

Consider the DGP of an i.i.d. sample $\{y_i, x_{i1}, x_{i2}, i = 1, \ldots, n\}$ by the linear Gaussian regression

$$\begin{aligned} y_i &= \beta_0 + \beta_{10} x_{i1} + \beta_{20} x_{i2} + e_i \\ \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} &\sim N(0, \boldsymbol{I}_2) \\ e_i &\sim N(0, \sigma_0^2). \end{aligned}$$

We set $\sigma_0 = 3, \beta_0 = 0, \beta_{10} = 1, \beta_{20} = 0.5$, and $n = 300$.

15.2 Definition of the Bootstrap

Let $\boldsymbol{w}_i = (y_i, x_i')'$ and let $F(\boldsymbol{w}) = P(\boldsymbol{w}_i \leq \boldsymbol{w})$ be the cumulative distribution function (CDF) of $\boldsymbol{w}_i$. Let $\boldsymbol{\beta} = (Ex_i x_i')^{-1} E(x_i y_i)$ be the regression slope and $\theta = h(\boldsymbol{\beta})$ be some parameter of interest. Let $\boldsymbol{F}_0, \boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0 = h(\boldsymbol{\beta}_0)$ denote the true values of $\boldsymbol{F}, \boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

It will be helpful to think of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ as a function of $\boldsymbol{F}$ as

$$\begin{aligned} \boldsymbol{\beta} &= (Ex_i x_i')^{-1} E(x_i y_i) \\ &= \left(\int \boldsymbol{x}\boldsymbol{x}' d\boldsymbol{F}(\boldsymbol{w}) \right)^{-1} \left(\int \boldsymbol{x} y d\boldsymbol{F}(\boldsymbol{w}) \right) :\equiv \boldsymbol{\beta}(\boldsymbol{F}). \end{aligned}$$

The true values of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ satisfy $\boldsymbol{\beta}_0 = \boldsymbol{\beta}(\boldsymbol{F}_0)$ and $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{F}_0)$.

Let $\hat{\boldsymbol{\beta}}_n$ be the OLS estimator and $\hat{\boldsymbol{\theta}}_n = h(\hat{\boldsymbol{\beta}}_n)$. Let $T_n = T_n(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n, \boldsymbol{\theta})$ be a statistic of interest. Let $G_n(t, \boldsymbol{F}) = P(T_n \leq t | \boldsymbol{F})$ be the exact CDF of T_n when the data are sampled from the distribution $\boldsymbol{F}$. The exact distribution G_n is a function only of $\boldsymbol{F}$, because the distribution of T_n depends only on the distribution of $\boldsymbol{w}_i$, which is $\boldsymbol{F}$, and the parameter $\boldsymbol{\theta}$, which is also determined by $\boldsymbol{F}$. The true CDF of T_n is $G_n(t, \boldsymbol{F}_0)$, which is unknown since $\boldsymbol{F}_0$ is unknown.

The bootstrap, an idea attributed to Efron (1979), is to use the empirical distribution of the data $\{y_i, x_{i1}, x_{i2}, i = 1, \ldots, n\}$ to estimate $\boldsymbol{F}_0$ and hence $G_n(t,_{\mathbf{0}})$. In many cases, the bootstrap achieves a much better approximation than asymptotic methods.

For any estimate $\hat{\boldsymbol{F}}_n$ of $\boldsymbol{F}_0$, the bootstrap estimator of $G_n(t, \boldsymbol{F}_0)$ is $\hat{G}_n(t) = G_n(t, \hat{\boldsymbol{F}}_n)$. Bootstrap inference is based on $\hat{G}_n(t)$. The most common choice for $\hat{\boldsymbol{F}}_n$ is the empirical distribution function (EDF) of the sampled data which will be defined in the next section. In this case $\hat{G}_n(t)$ is called *nonparametric bootstrap*. Some other estimates of $\hat{\boldsymbol{F}}_n$ are possible and will be discussed later.

The bootstrap distribution substitutes $\hat{\boldsymbol{F}}_n$ for $\boldsymbol{F}_0$ in the formula $G_n(t, \boldsymbol{F})$. As such, it not only pretends that the distribution of $\boldsymbol{w}_i$ is $\hat{\boldsymbol{F}}_n$ rather than $\boldsymbol{F}_0$, but it also pretends that the true value of the parameter is $\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(\hat{\boldsymbol{F}}_n)$, rather than $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{F}_0)$. We call $\hat{\boldsymbol{\theta}}_n$ the *bootstrap parameter estimate*.

Let $\boldsymbol{w}_i^*$ be a random variable with distribution $\hat{\boldsymbol{F}}_n$ and $T_n^* = T_n(\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_n^*, \hat{\boldsymbol{\theta}}_n)$ be a random variable with distribution $\hat{G}_n^*$. That is

$$\begin{aligned} P(\boldsymbol{w}_i^* \leq \boldsymbol{w}) &= \hat{\boldsymbol{F}}_n(\boldsymbol{w}) \\ P(T_n^* \leq t) &= \hat{G}_n^*(t). \end{aligned}$$

T_n^* is the correct analog of T_n when the true CDF is $\hat{\boldsymbol{F}}_n$, as the data $\boldsymbol{w}_i^*$ are sampled from the CDF $\hat{\boldsymbol{F}}_n$, and the parameter $\hat{\boldsymbol{\theta}}_n$ is determined by $\hat{\boldsymbol{F}}_n$.

15.3 The Empirical Distribution Function

Note that $\boldsymbol{F}_0(\boldsymbol{w}) = P(\boldsymbol{w}_i \leq \boldsymbol{w}) = E[I(\boldsymbol{w}_i \leq \boldsymbol{w})]$, where $I(\cdot)$ is the indicator function, so $\boldsymbol{F}_0(\boldsymbol{w})$ can be expressed as a population moment. A natural estimate is therefore the corresponding sample moment:

$$\hat{\boldsymbol{F}}_n(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{w}_i \leq \boldsymbol{w}).$$

$\hat{\boldsymbol{F}}_n(\boldsymbol{w})$ is called the *empirical distribution function* (EDF). $\hat{\boldsymbol{F}}_n$ is a nonparametric estimate of $\boldsymbol{F}_0$. Note that $\boldsymbol{F}_0$ may be either discrete or continuous, $\hat{\boldsymbol{F}}_n$ is by construction a (discontinuous) step function.

For ant $\boldsymbol{w}$, $1(\boldsymbol{w}_i \leq \boldsymbol{w})$ is an i.i.d. random variable with expectation

$$E[1(\boldsymbol{w}_i \leq \boldsymbol{w})] = \int_{-\infty}^{\infty} 1(\boldsymbol{w}_i \leq \boldsymbol{w}) d\boldsymbol{F}_0 = \boldsymbol{F}_0(\boldsymbol{w}).$$

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, **59**, 817–858.

Davidson, J. (1994). *Stochastic Limit Theory*, New York: Oxford University Press.

Gallant, A. R. (1987). *Nonlinear Statistical Models*, New York: Wiley.

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and other Approaches*, New York: Cambridge University Press.

Newey, W. K., & K. West (1987). A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, **55**, 703–708.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, **48**, 817–838.

White, H. (1984). *Asymptotic Theory for Econometricians*, Orlando, FL: Academic Press.