Chapter 6

Asymptotic Least Squares Theory: Part I

We have shown that the OLS estimator and related tests have good finite-sample properties under the classical conditions. These conditions are, however, quite restrictive in practice, as discussed in Section 3.6. It is therefore natural to ask the following questions. First, to what extent may we relax the classical conditions so that the OLS method has broader applicability? Second, what are the properties of the OLS method under more general conditions? The purpose of this chapter is to provide some answers to these questions. In particular, we shall allow explanatory variable to be random variables, possibly weakly dependent and heterogeneously distributed. This relaxation permits applications of the OLS method to various data and models, but it also renders the analysis of finite-sample properties difficult. Nonetheless, it is relatively easy to analyze the asymptotic performance of the OLS estimator and construct large-sample tests. As the asymptotic results are valid under more general conditions, the OLS method remains a useful tool for a wide variety of applications.

6.1 When Regressors are Stochastic

Given the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, suppose now that $\boldsymbol{X}$ is stochastic. In this case, [A2](i) can never hold because $\boldsymbol{X}\boldsymbol{\beta}_o$ is random and can not be $\mathbb{E}(\boldsymbol{y})$. Even when a condition on $\mathbb{E}(\boldsymbol{y})$ is imposed, we are still unable to evaluate

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

because $\hat{\boldsymbol{\beta}}_T$ now is a complex function of the elements of $\boldsymbol{y}$ and $\boldsymbol{X}$. Similarly, a condition on $\mathrm{var}(\boldsymbol{y})$ is of little use for calculating $\mathrm{var}(\hat{\boldsymbol{\beta}}_T)$.

To ensure unbiasedness, it is typical to assume that $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ for some $\boldsymbol{\beta}_o$, instead of [A2](i). Under this condition,

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

by the law of iterated expectations (Lemma 5.9). Yet the condition $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ may not always be realistic. To see this, let $\boldsymbol{x}_t$ denote the t^{th} column of $\boldsymbol{X}'$ and write the t^{th} element of $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ as

$$\mathbb{E}(y_t \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) = \boldsymbol{x}_t'\boldsymbol{\beta}_o, \qquad t = 1, 2, \ldots, T.$$

Consider the simple AR(1) specification for time series data such that $\boldsymbol{x}_t$ contains only one regressor y_{t-1}:

$$y_t = \beta y_{t-1} + e_t, \qquad t = 2, \ldots, T.$$

While $\mathbb{E}(y_t \mid y_1, \ldots, y_{T-1}) = y_t$ for $t = 2, \ldots, T$ by Lemma 5.10, the aforementioned condition for this specification reads:

$$\mathbb{E}(y_t \mid y_1, \ldots, y_{T-1}) = \beta_o y_{t-1},$$

for some β_o. This amounts to requiring $y_t = \beta_o y_{t-1}$ with probability one so that y_t must be determined by its immediate past value without any random disturbance. If, however, $\{y_t\}$ is indeed an AR(1) process: $y_t = \beta_o y_{t-1} + \epsilon_t$ and ϵ_t has a continuous distribution, the event that $y_t = \beta_o y_{t-1}$ (i.e., $\epsilon_t = 0$) can occur only with probability zero, which violates the imposed condition.

Suppose that $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ and $\operatorname{var}(\boldsymbol{y} \mid \boldsymbol{X}) = \sigma_o^2\boldsymbol{I}_T$. It is easy to see that

$$\begin{aligned}
\operatorname{var}(\hat{\boldsymbol{\beta}}_T) &= \mathbb{E}\big[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\big] \\
&= \mathbb{E}\big[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\operatorname{var}(\boldsymbol{y} \mid \boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\big] \\
&= \sigma_o^2\,\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})^{-1},
\end{aligned}$$

which is not exactly the same as the variance-covariance matrix when $\boldsymbol{X}$ is non-stochastic, cf. Theorem 3.4(c). The condition on $\operatorname{var}(\boldsymbol{y} \mid \boldsymbol{X})$, again, is not always a reasonable one. For example, as in the previous example that $\boldsymbol{x}_t = y_{t-1}$, we have $y_t = \beta_o y_{t-1}$ with probability one, so that the conditional variance must be zero, rather than a positive constant σ_o^2.

The discussions above show that the conditions on $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X})$ and $\operatorname{var}(\boldsymbol{y} \mid \boldsymbol{X})$ may not hold when $\boldsymbol{x}_t$ are random vectors. Without such conditions, it is difficult, if not impossible, to evaluate the mean and variance of the OLS estimator. Moreover, when $\boldsymbol{X}$ is stochastic, $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ need not be normally distributed even when $\boldsymbol{y}$ is. Consequently, the results for hypothesis testing discussed in Section 3.3 become invalid.

6.2 Asymptotic Properties of the OLS Estimators

Suppose that we observe the data $(y_t \ \boldsymbol{w}_t')'$, where y_t is the variable of interest (dependent variable), and $\boldsymbol{w}_t$ is an $m \times 1$ vector of "exogenous" variables. By exogenous variables we mean those variables whose random behaviors are not explicitly modeled. Let $\mathcal{W}^t$ denote the collection of random vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t$ and $\mathcal{Y}^t$ the collection of $y_1, \ldots, y_t$. The set $\{\mathcal{Y}^{t-1}, \mathcal{W}^t\}$ generates a σ-algebra that is understood as the information set up to time t. What we would like to do is to account for the behavior of y_t based on this information set.

We first determine a $k \times 1$ vector of regressors $\boldsymbol{x}_t$ from the information set $\{\mathcal{Y}^{t-1}, \mathcal{W}^t\}$. The chosen $\boldsymbol{x}_t$ may include lagged dependent variables (taken from $\mathcal{Y}^{t-1}$) as well as current and lagged exogenous variables (taken from $\mathcal{W}^t$). The resulting linear specification is

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t, \qquad t = 1, 2, \ldots, T, \tag{6.1}$$

which is just the t^{th} observation of the more familiar expression $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with $\boldsymbol{x}_t$ the t^{th} column of $\boldsymbol{X}'$. The expression (6.1) is more intuitive because it explicitly relates the t^{th} observation of y to the t^{th} observation of all explanatory variables. The OLS estimator of the specification (6.1) now can be expressed as

$$\hat{\boldsymbol{\beta}}_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\sum_{t=1}^{T} \boldsymbol{x}_t y_t\right). \tag{6.2}$$

The right-hand side of the second equality is useful in subsequent asymptotic analysis.

6.2.1 Consistency

The OLS estimator $\hat{\boldsymbol{\beta}}_T$ is said to be *strongly (weakly) consistent* for the parameter vector $\boldsymbol{\beta}^*$ if $\hat{\boldsymbol{\beta}}_T \xrightarrow{\text{a.s.}} \boldsymbol{\beta}^*$ ($\hat{\boldsymbol{\beta}}_T \xrightarrow{\mathbb{P}} \boldsymbol{\beta}^*$) as T tends to infinity. Consistency in effect requires $\hat{\boldsymbol{\beta}}_T$ to be eventually close to $\boldsymbol{\beta}^*$ in a proper probabilistic sense when "enough" information (a sufficiently large sample) becomes available. Note that consistency is in sharp contrast with unbiasedness. While an unbiased estimator of $\boldsymbol{\beta}^*$ is "correct" on average, there is no guarantee that its values will be close to $\boldsymbol{\beta}^*$, no matter how large the sample is.

To analyze the limiting behavior of $\hat{\boldsymbol{\beta}}_T$, we impose the following conditions.

[B1] $\{(y_t \ \boldsymbol{w}_t')'\}$ is a sequence of random vectors, and $\boldsymbol{x}_t$ is a random vector containing some elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$.

(i) $\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ obeys a SLLN (WLLN) with the almost sure (probability) limit

$$\boldsymbol{M}_{xx} := \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'),$$

which is a nonsingular matrix.

(ii) $\{\boldsymbol{x}_t y_t\}$ obeys a SLLN (WLLN) with the almost sure (probability) limit

$$\boldsymbol{m}_{xy} := \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t y_t).$$

[B2] There exists a $\boldsymbol{\beta}_o$ such that $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_o + \epsilon_t$ with $\mathbb{E}(\boldsymbol{x}_t\epsilon_t) = \mathbf{0}$ for all t.

[B1] and [B2] are are quite different from the classical conditions. Compared with [A1], the condition [B1] now explicitly allows $\boldsymbol{x}_t$ to be a random vector which may contain some lagged dependent variables ($y_{t-j}, j \geq 1$) as well as current and past exogenous variables ($\boldsymbol{w}_{t-j}, j \geq 0$). [B1] also admits non-stochastic regressors which can be viewed as degenerate random variables. Compared with [A2](ii), [B1] allows the random data to exhibit certain forms of dependence and heterogeneity. It does not rule out serially correlated y_t and $\boldsymbol{x}_t$, nor does it restrict y_t to be unconditionally homoskedastic ($\mathrm{var}(y_t)$ being a constant) or conditionally homoskedastic ($\mathrm{var}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t)$ being a constant). What really matters is that the data must be well behaved in the sense that they are governed by some SLLN (WLLN). Thus, the deterministic time trend t and random walks are excluded under [B1]; see Examples 5.29 and 5.31.

Similar to [A2](i), [B2] may be interpreted as a condition of correct specification. Here, $\epsilon_t = e_t(\boldsymbol{\beta}_o)$ is known as the *disturbance* term, and $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ is the orthogonal projection of y_t onto the space of all linear functions of $\boldsymbol{x}_t$ and also known as a *linear projection* of y_t. A sufficient condition for [B2] is that $\boldsymbol{x}_t'\boldsymbol{\beta}$ is the correct specification of the conditional mean function, i.e., there exists a $\boldsymbol{\beta}_o$ such that

$$\mathbb{E}\big(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t\big) = \boldsymbol{x}_t'\boldsymbol{\beta}_o,$$

or $\mathbb{E}\big(\epsilon_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t\big) = 0$. This implies [B2] because, by the law of iterated expectations,

$$\mathbb{E}(\boldsymbol{x}_t\epsilon_t) = \mathbb{E}\big[\boldsymbol{x}_t\, \mathbb{E}\big(\epsilon_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t\big)\big] = \mathbf{0}.$$

Recall that the conditional mean function of y_t is the orthogonal projection of y_t onto the space of all measurable (not necessarily linear) functions of $\boldsymbol{x}_t$ and hence is not a linear function in general. Yet, when the conditional mean is indeed linear in $\boldsymbol{x}_t$ (for example, when y_t and $\boldsymbol{x}_t$ are jointly normally distributed), it must also be the linear projection. The converse is not true in general, however.

To analyze the behavior of the OLS estimator, we proceed as follows. By [B1], $\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ obeys a SLLN (WLLN):

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t' \to \boldsymbol{M}_{xx} \quad \text{a.s. (in probability)},$$

where $\boldsymbol{M}_{xx}$ is nonsingular. Note that matrix inversion is a continuous function of invertible matrices. By Lemma 5.13 (Lemma 5.17), almost sure convergence (convergence in probability) carries over under continuous transformations, so that

$$\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1} \to \boldsymbol{M}_{xx}^{-1} \quad \text{a.s. (in probability)}.$$

This, together with [B1](ii), immediately implies that the OLS estimator (6.2) is

$$\hat{\boldsymbol{\beta}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t y_t\right) \to \boldsymbol{M}_{xx}^{-1}\boldsymbol{m}_{xy} \quad \text{a.s. (in probability)}.$$

Consider the special case that $\mathbb{E}(\boldsymbol{x}_t y_t)$ and $\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$ are constants. Then, $\boldsymbol{m}_{xy} = \mathbb{E}(\boldsymbol{x}_t y_t)$ and $\boldsymbol{M}_{xx} = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$. When [B2] holds,

$$\mathbb{E}(\boldsymbol{x}_t y_t) = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')\boldsymbol{\beta}_o,$$

so that $\boldsymbol{\beta}_o = \boldsymbol{\beta}^*$. This shows that the parameter $\boldsymbol{\beta}_o$ of the linear projection function is indeed the almost sure (probability) limit of the OLS estimator. We have established the following consistency result.

Theorem 6.1 *Consider the linear specification* (6.1).

(i) *When* [B1] *holds,* $\hat{\boldsymbol{\beta}}_T$ *is strongly (weakly) consistent for* $\boldsymbol{\beta}^* = \boldsymbol{M}_{xx}^{-1}\boldsymbol{m}_{xy}$.

(ii) *When* [B1] *and* [B2] *hold,* $\boldsymbol{\beta}_o = \boldsymbol{M}_{xx}^{-1}\boldsymbol{m}_{xy}$ *so that* $\hat{\boldsymbol{\beta}}_T$ *is strongly (weakly) consistent for* $\boldsymbol{\beta}_o$.

The first assertion states that the OLS estimator is strongly (weakly) consistent for some parameter vector $\boldsymbol{\beta}^*$, provided that the behaviors of $\boldsymbol{x}_t\boldsymbol{x}_t'$ and $\boldsymbol{x}_t y_t$ are governed by proper laws of large numbers. Note that this conclusion holds without [B2], the condition of correct specification. When [B2] is also satisfied, the second assertion indicates that the limit of the OLS estimator is the parameter vector of the linear projection. Thus, [B1] assures convergence of the OLS estimator, whereas [B2] determines to which parameter the OLS estimator converges.

As an example, we show below that Theorem 6.1 holds under some specific conditions on data. This result may be applied to models with cross section data that are independent over t.

Corollary 6.2 *Given the linear specification* (6.1), *suppose that* $(y_t\ \boldsymbol{x}_t')'$ *are independent random vectors with bounded* $(2+\delta)^{\,th}$ *moment for any* $\delta > 0$. *If* $\boldsymbol{M}_{xx}$ *and* $\boldsymbol{m}_{xy}$ *defined in* [B1] *exist, the OLS estimator* $\hat{\boldsymbol{\beta}}_T$ *is strongly consistent for* $\boldsymbol{\beta}^* = \boldsymbol{M}_{xx}^{-1}\boldsymbol{m}_{xy}$. *If* [B2] *also holds,* $\hat{\boldsymbol{\beta}}_T$ *is strongly consistent for* $\boldsymbol{\beta}_o$ *defined in* [B2].

Proof: By the Cauchy-Schwartz inequality (Lemma 5.5), the i^{th} element of $\boldsymbol{x}_t y_t$ is such that

$$\mathbb{E}\,|x_{ti}y_t|^{1+\delta} \le \left[\mathbb{E}\,|x_{ti}|^{2(1+\delta)}\right]^{1/2}\left[\mathbb{E}\,|y_t|^{2(1+\delta)}\right]^{1/2} \le \Delta,$$

for some $\Delta > 0$. Similarly, each element of $\boldsymbol{x}_t\boldsymbol{x}_t'$ also has bounded $(1+\delta)^{\text{th}}$ moment. Then, $\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ and $\{\boldsymbol{x}_t\boldsymbol{y}_t\}$ obey Markov's SLLN by Lemma 5.26 with the respective almost sure limits $\boldsymbol{M}_{xx}$ and $\boldsymbol{m}_{xy}$. The assertions now follow from Theorem 6.1. □

For other types of data, we do not explicitly specify the sufficient conditions that ensure OLS consistency; see White (2001) for such conditions and Section 5.5 for related discussions. The example below is an illustration of OLS consistency when the data are weakly stationary.

Example 6.3 Given the simple AR(1) specification

$$y_t = \alpha y_{t-1} + e_t,$$

suppose that $\{y_t^2\}$ and $\{y_t y_{t-1}\}$ obey a SLLN (WLLN). Let $y_0 = 0$. Then by Theorem 6.1(i), the OLS estimator of α is such that

$$\hat{\alpha}_T \to \frac{\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(y_t y_{t-1})}{\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(y_{t-1}^2)} \quad \text{a.s. (in probability)},$$

provided that the above limits exist.

When $\{y_t\}$ follows a stationary AR(1) process:

$$y_t = \alpha_o y_{t-1} + u_t, \quad |\alpha_o| < 1,$$

where u_t are i.i.d. with mean zero and variance σ_u^2, we have $\mathbb{E}(y_t) = 0$, $\operatorname{var}(y_t) = \sigma_u^2/(1-\alpha_o^2)$ and $\operatorname{cov}(y_t, y_{t-1}) = \alpha_o \operatorname{var}(y_t)$. In this case, it is typically true that $\{y_t^2\}$ and $\{y_t y_{t-1}\}$ obey a SLLN (WLLN). It follows that

$$\hat{\alpha}_T \to \frac{\operatorname{cov}(y_t, y_{t-1})}{\operatorname{var}(y_t)} = \alpha_o, \quad \text{a.s. (in probability)}.$$

Alternatively, this result may be verified by noting that $\mathbb{E}(y_{t-1}u_t) = 0$ so that αy_{t-1} is a correct specification for the linear projection of y_t. Theorem 6.1(ii) now ensures $\hat{\alpha}_T \to \alpha_o$ a.s. (in probability). □

Remark: If for some $\boldsymbol{\beta}_o$ such that $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ is not the linear projection of y_t, $\mathbb{E}(\boldsymbol{x}_t\epsilon_t) \neq \boldsymbol{0}$, and

$$\mathbb{E}(\boldsymbol{x}_t y_t) = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')\boldsymbol{\beta}_o + \mathbb{E}(\boldsymbol{x}_t\epsilon_t).$$

Then, $\boldsymbol{m}_{xy} = \boldsymbol{M}_{xx}\boldsymbol{\beta}_o + \boldsymbol{m}_{x\epsilon}$, where

$$\boldsymbol{m}_{x\epsilon} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t\epsilon_t).$$

The almost sure (probability) limit of the OLS estimator becomes

$$\boldsymbol{\beta}^* = \boldsymbol{M}_{xx}^{-1}\boldsymbol{m}_{xy} = \boldsymbol{\beta}_o + \boldsymbol{M}_{xx}^{-1}\boldsymbol{m}_{x\epsilon},$$

rather than $\boldsymbol{\beta}_o$. The following examples illustrate.

Example 6.4 Consider the specification

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t,$$

where $\boldsymbol{x}_t'$ is $k_1 \times 1$. Suppose that

$$\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o + \boldsymbol{z}_t'\boldsymbol{\gamma}_o,$$

where $\boldsymbol{z}_t$ ($k_2 \times 1$) also contains the elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$ that are distinct from the elements of $\boldsymbol{x}_t$. This is an example that a specification omits relevant variables. When [B1] holds, $\hat{\boldsymbol{\beta}}_T \to \boldsymbol{M}_{xx}^{-1}\boldsymbol{m}_{xy}$ a.s. (in probability) by Theorem 6.1(i). Writing

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_o + \boldsymbol{z}_t'\boldsymbol{\gamma}_o + \epsilon_t = \boldsymbol{x}_t'\boldsymbol{\beta}_o + u_t,$$

where $\epsilon_t = y_t - \mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t)$ and $u_t = \boldsymbol{z}_t'\boldsymbol{\gamma}_o + \epsilon_t$, we have $\mathbb{E}(\boldsymbol{x}_t u_t) = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{z}_t')\boldsymbol{\gamma}_o$. When $\mathbb{E}(\boldsymbol{x}_t\boldsymbol{z}_t')\boldsymbol{\gamma}_o$ is non-zero, $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ is not the linear projection of y_t. Thus, the OLS estimator of $\boldsymbol{\beta}$ need not converge to $\boldsymbol{\beta}_o$. In fact, setting $\boldsymbol{M}_{xz} := \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\mathbb{E}(\boldsymbol{x}_t\boldsymbol{z}_t')$, we have the almost sure (probability) limit of $\hat{\boldsymbol{\beta}}_T$:

$$\boldsymbol{M}_{xx}^{-1}\boldsymbol{m}_{xy} = \boldsymbol{\beta}_o + \boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{xz}\boldsymbol{\gamma}_o,$$

which is not $\boldsymbol{\beta}_o$ in general. Consistency for $\boldsymbol{\beta}_o$ would hold when the elements of $\boldsymbol{x}_t$ are orthogonal to those of $\boldsymbol{z}_t$, i.e., $\mathbb{E}(\boldsymbol{x}_t\boldsymbol{z}_t') = \boldsymbol{0}$. In this case, $\boldsymbol{M}_{xz} = \boldsymbol{0}$ so that $\hat{\boldsymbol{\beta}}_T \to \boldsymbol{\beta}_o$ almost surely (in probability). That is, $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ is the linear projection of y_t onto the space of all linear functions of $\boldsymbol{x}_t$ when z_t are orthogonal to $\boldsymbol{x}_t$. □

Example 6.5 Given the simple AR(1) specification:

$$y_t = \alpha y_{t-1} + e_t,$$

suppose that y_t are generated according to

$$y_t = \alpha_o y_{t-1} + \epsilon_t, \quad |\alpha_o| < 1,$$

where $\epsilon_t = u_t - \pi_o u_{t-1}$ with $|\pi_o| < 1$, and $\{u_t\}$ is a white noise with mean zero and variance σ_u^2. A process so generated is a weakly stationary ARMA(1,1) process (*autoregressive and moving average process* of order (1,1)). As in Example 6.3, when $\{y_t^2\}$ and $\{y_t y_{t-1}\}$ obey a SLLN (WLLN), $\hat{\alpha}_T$ converges to $\operatorname{cov}(y_t, y_{t-1})/\operatorname{var}(y_{t-1})$ almost surely (in probability). Note, however, that $\alpha_o y_{t-1}$ in this case is not the linear projection of y_t because y_{t-1} depends on $\epsilon_{t-1} = u_{t-1} - \pi_o u_{t-2}$ and

$$\mathbb{E}(y_{t-1}\epsilon_t) = \mathbb{E}[y_{t-1}(u_t - \pi_o u_{t-1})] = -\pi_o \sigma_u^2.$$

The limit of $\hat{\alpha}_T$ now reads

$$\frac{\operatorname{cov}(y_t, y_{t-1})}{\operatorname{var}(y_{t-1})} = \frac{\alpha_o \operatorname{var}(y_{t-1}) + \operatorname{cov}(\epsilon_t, y_{t-1})}{\operatorname{var}(y_{t-1})} = \alpha_o - \frac{\pi_o \sigma_u^2}{\operatorname{var}(y_{t-1})}.$$

The OLS estimator is therefore inconsistent for α_o unless $\pi_o = 0$ (i.e., $\epsilon_t = u_t$ are serially uncorrelated), in contrast with Example 6.3. Inconsistency here is, again, due to the fact that $\alpha_o y_{t-1}$ is not the linear projection of y_t. This failure arises because ϵ_t are serially correlated with ϵ_{t-1} and hence are correlated with the lagged dependent variable y_{t-1}.

The conclusion holds more generally. Consider the specification that includes a lagged dependent variable as a regressor:

$$y_t = \alpha y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta} + e_t.$$

Suppose that y_t are generated as $y_t = \alpha_o y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta}_o + \epsilon_t$ such that ϵ_t are serially correlated. The OLS consistency again breaks down because $\alpha_o y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta}_o$ is not the linear projection, a consequence of the joint presence of a lagged dependent variable and serially correlated disturbances. □

6.2.2 Asymptotic Normality

We say that $\hat{\boldsymbol{\beta}}_T$ is *asymptotically normally distributed* (about $\boldsymbol{\beta}_o$) if

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0},\ \boldsymbol{D}_o),$$

where $\boldsymbol{D}_o$ is a positive-definite matrix. That is, the sequence of properly normalized $\hat{\boldsymbol{\beta}}_T$ converges in distribution to a multivariate normal random vector. The matrix $\boldsymbol{D}_o$ is the variance-covariance matrix of the limiting normal distribution and hence known as the *asymptotic variance-covariance matrix* of $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)$. Equivalently, we may also express asymptotic normality by

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

It should be emphasized that asymptotic normality here is referred to $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)$ rather than $\hat{\boldsymbol{\beta}}_T$; the latter has only a degenerate distribution at $\boldsymbol{\beta}_o$ in the limit by strong (weak) consistency.

When $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)$ has a limiting distribution, it is $O_{\mathbb{P}}(1)$ by Lemma 5.24. Therefore, $\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o$ is necessarily $O_{\mathbb{P}}(T^{-1/2})$; that is, $\hat{\boldsymbol{\beta}}_T$ tend to $\boldsymbol{\beta}_o$ at the rate $T^{-1/2}$. Thus, the asymptotic normality result tells us not only (weak) consistency but also the *rate of convergence* to $\boldsymbol{\beta}_o$. An estimator that is consistent at the rate $T^{-1/2}$ is referred to as a "$\sqrt{T}$-consistent" estimator. For standard cases in econometrics, estimators are typically $\sqrt{T}$-consistent. There are consistent estimators that converge more quickly; we will discuss such estimators in Chapter 7.

Given the specification $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$ and [B2], define

$$\boldsymbol{V}_T := \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right),$$

where ϵ_t is specified in [B2]. We now impose an additional condition.

[B3] For ϵ_t in [B2], $\{\boldsymbol{V}_o^{-1/2}\boldsymbol{x}_t\epsilon_t\}$ obeys a CLT, where $\boldsymbol{V}_o = \lim_{T\to\infty}\boldsymbol{V}_T$ is positive-definite.

To establish asymptotic normality, we express the normalized OLS estimator as

$$\begin{aligned}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) &= \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right)\\ &= \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\boldsymbol{V}_o^{1/2}\left[\boldsymbol{V}_o^{-1/2}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right)\right].\end{aligned} \tag{6.3}$$

By [B1](i), the first term on the right-hand side of (6.3) converges to $\boldsymbol{M}_{xx}^{-1}$ almost surely (in probability). Then by [B3],

$$\boldsymbol{V}_o^{-1/2}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right) \xrightarrow{D} \mathcal{N}(\boldsymbol{0},\ \boldsymbol{I}_k).$$

In view of (6.3), we have from Lemma 5.22 that

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) \xrightarrow{D} \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o^{1/2}\mathcal{N}(\boldsymbol{0},\ \boldsymbol{I}_k) \overset{d}{=} \mathcal{N}(\boldsymbol{0},\ \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}),$$

where $\overset{d}{=}$ stands for equality in distribution. This proves the following asymptotic normality result.

Theorem 6.6 *Given the linear specification* (6.1)*, suppose that* [B1](i)*,* [B2] *and* [B3] *hold. Then,*

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0},\ \boldsymbol{D}_o),$$

where $\boldsymbol{D}_o = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}$.

Theorem 6.6 is also stated without specifying the conditions that ensure the effect of a CLT. We note that it may hold for weakly dependent and heterogeneously distributed data, as long as these data obey a proper CLT. This result differs from the normality property described in Theorem 3.7(a), in that the latter gives an exact distribution but is valid only when y_t are independent, normal random variables and $\boldsymbol{x}_t$ are non-stochastic.

The corollary below specializes on independent data and may be applied to models with cross section data.

Corollary 6.7 *Given the linear specification* (6.1)*, suppose that* $(y_t\ \boldsymbol{x}_t')'$ *are independent random vectors with bounded* $(4+\delta)^{\,th}$ *moment for any* $\delta > 0$ *and that* [B2] *holds. If* $\boldsymbol{M}_{xx}$ *defined in* [B1] *and* $\boldsymbol{V}_o$ *defined in* [B3] *exist,*

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0},\ \boldsymbol{D}_o),$$

where $\boldsymbol{D}_o = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}$.

Proof: Let $z_t = \boldsymbol{\lambda}'\boldsymbol{x}_t\epsilon_t$, where $\boldsymbol{\lambda}$ is a column vector such that $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$. If $\{z_t\}$ obeys a CLT, then $\{\boldsymbol{x}_t\epsilon_t\}$ obeys a multivariate CLT by the Cramér-Wold device (Lemma 5.18). Clearly, z_t are independent random variables because $\boldsymbol{x}_t\epsilon_t = \boldsymbol{x}_t(y_t - \boldsymbol{x}_t\boldsymbol{\beta}_o)$ are. We will show that z_t satisfy the conditions imposed in Lemma 5.36 and hence obey Liapunov's CLT. First, z_t have mean zero under [B2] and $\operatorname{var}(z_t) = \boldsymbol{\lambda}'[\operatorname{var}(\boldsymbol{x}_t\epsilon_t)]\boldsymbol{\lambda}$. By data independence,

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

The average of $\operatorname{var}(z_t)$ is then

$$\frac{1}{T}\sum_{t=1}^{T}\operatorname{var}(z_t) = \boldsymbol{\lambda}'\boldsymbol{V}_T\boldsymbol{\lambda} \to \boldsymbol{\lambda}\boldsymbol{V}_o\boldsymbol{\lambda}.$$

By the Cauchy-Schwartz inequality (Lemma 5.5),

$$\mathbb{E}\,|x_{ti}y_t|^{2+\delta} \le \left[\mathbb{E}\,|x_{ti}|^{2(2+\delta)}\right]^{1/2}\left[\mathbb{E}\,|y_t|^{2(2+\delta)}\right]^{1/2} \le \Delta,$$

for some $\Delta > 0$. Similarly, $x_{ti}x_{tj}$ have bounded $(2+\delta)^{\,\text{th}}$ moment. It follows that $x_{ti}\epsilon_t$ (which is an element of $\boldsymbol{x}_t y_t - \boldsymbol{x}_t\boldsymbol{x}_t'\boldsymbol{\beta}_o$) and z_t (which is a weighted sum of $x_{ti}\epsilon_t$) also have

bounded $(2+\delta)^{\text{th}}$ moment by Minkowski's inequality (Lemma 5.7). We may now invoke Lemma 5.36 and conclude that

$$\frac{1}{\sqrt{T(\boldsymbol{\lambda}'\boldsymbol{V}_o\boldsymbol{\lambda})}}\sum_{t=1}^{T} z_t \xrightarrow{D} \mathcal{N}(0,\, 1).$$

Then by the Cramér-Wold device,

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

as required by [B3]. The assertion follows from Theorem 6.6. □

The example below illustrates that the OLS estimator may or may not have a asymptotic normal distribution, depending on data characteristics.

Example 6.8 Consider the the AR(1) specification:

$$y_t = \alpha y_{t-1} + e_t.$$

Case 1: $\{y_t\}$ is a stationary AR(1) process: $y_t = \alpha_o y_{t-1} + u_t$ with $|\alpha_o| < 1$, where u_t are i.i.d. random variables with mean zero and variance σ_u^2. From Example 6.3 we know that $\mathbb{E}(y_{t-1}u_t) = 0$ and that αy_{t-1} is a correct specification. It can also be seen that

$$\operatorname{var}(y_{t-1}u_t) = \mathbb{E}(y_{t-1}^2)\,\mathbb{E}(u_t^2) = \sigma_u^4/(1-\alpha_o^2),$$

and $\operatorname{cov}(y_{t-1}u_t,\, y_{t-1-j}u_{t-j}) = 0$ for all $j > 0$. It is typically true that $\{y_{t-1}u_t\}$ obeys a CLT so that

$$\frac{\sqrt{1-\alpha_o^2}}{\sigma_u^2\sqrt{T}}\sum_{t=1}^{T} y_{t-1}u_t \xrightarrow{D} \mathcal{N}(0,\, 1).$$

As $\sum_{t=1}^{T} y_{t-1}^2/T$ converges to $\sigma_u^2/(1-\alpha_o^2)$, we have from Theorem 6.6 that

$$\frac{\sqrt{1-\alpha_o^2}}{\sigma_u^2}\,\frac{\sigma_u^2}{1-\alpha_o^2}\,\sqrt{T}(\hat{\alpha}_T - \alpha_o) = \frac{1}{\sqrt{1-\alpha_o^2}}\sqrt{T}(\hat{\alpha}_T - \alpha_o) \xrightarrow{D} \mathcal{N}(0,\, 1),$$

or equivalently, $\sqrt{T}(\hat{\alpha}_T - \alpha_o) \xrightarrow{D} \mathcal{N}(0,\, 1-\alpha_o^2)$.

Case 2: $\{y_t\}$ is a random walk:

$$y_t = y_{t-1} + u_t.$$

We observe from Example 5.32 that $\operatorname{var}(T^{-1/2}\sum_{t=1}^{T} y_{t-1}u_t)$ is $O(T)$ and hence diverges with T. Moreover, Example 5.38 shows that $\{y_{t-1}u_t\}$ does not obey a CLT. Theorem 6.6 is therefore not applicable, and there is no guarantee that normalized $\hat{\alpha}_T$ is asymptotically normally distributed. □

When $\boldsymbol{V}_o$ is unknown, let $\widehat{\boldsymbol{V}}_T$ denote a symmetric and positive definite matrix that is weakly consistent for $\boldsymbol{V}_o$. Then, a weakly consistent estimator of $\boldsymbol{D}_o$ is

$$\widehat{\boldsymbol{D}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\widehat{\boldsymbol{V}}_T\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1},$$

and $\widehat{\boldsymbol{D}}_T^{-1/2} \xrightarrow{\mathbb{P}} \boldsymbol{D}_o^{-1/2}$. It follows from Theorem 6.6 and Lemma 5.19 that

$$\widehat{\boldsymbol{D}}_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) \xrightarrow{D} \boldsymbol{D}_o^{-1/2}\mathcal{N}(\boldsymbol{0}, \boldsymbol{D}_o) \overset{d}{=} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k).$$

The shows that Theorem 6.6 remains valid when the asymptotic variance-covariance matrix $\boldsymbol{D}_o$ is replaced by a weakly consistent estimator $\widehat{\boldsymbol{D}}_T$. This conclusion is stated below; note that $\widehat{\boldsymbol{D}}_T$ does not have to be a strongly consistent estimator here.

Theorem 6.9 *Given the linear specification* (6.1)*, suppose that* [B1](i)*,* [B2] *and* [B3] *hold. Then,*

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

where $\widehat{\boldsymbol{D}}_T = (\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'/T)^{-1}\widehat{\boldsymbol{V}}_T(\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'/T)^{-1}$ *and* $\widehat{\boldsymbol{V}}_T \xrightarrow{\mathbb{P}} \boldsymbol{V}_o$.

Remark: It is practically important to find a consistent estimator for $\boldsymbol{V}_o$ and hence a consistent estimator for $\boldsymbol{D}_o$. Normalizing the OLS estimator with an inconsistent estimator of $\boldsymbol{D}_o$ will, in general, destroy asymptotic normality.

6.3 Consistent Estimation of Covariance Matrix

We have seen in the preceding section that a consistent estimator of $\boldsymbol{D}_o = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}$ is crucial for the asymptotic normality result. The matrix $\boldsymbol{M}_{xx}$ can be consistently estimated by its sample counterpart $T^{-1}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'$; it remains to find a consistent estimator of $\boldsymbol{V}_o$. Recall that $\boldsymbol{V}_o = \lim_{T\to\infty}\boldsymbol{V}_T$, where

$$\boldsymbol{V}_o = \lim_{T\to\infty}\boldsymbol{V}_T = \lim_{T\to\infty}\operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right).$$

More specifically, we can write

$$\boldsymbol{V}_o = \lim_{T\to\infty}\sum_{j=-T+1}^{T-1}\boldsymbol{\Gamma}_T(j), \tag{6.4}$$

with

$$\boldsymbol{\Gamma}_T(j) = \begin{cases} \frac{1}{T}\sum_{t=j+1}^{T}\mathbb{E}(\boldsymbol{x}_t\epsilon_t\epsilon_{t-j}\boldsymbol{x}_{t-j}'), & j = 0, 1, 2, \ldots, \\ \frac{1}{T}\sum_{t=-j+1}^{T}\mathbb{E}(\boldsymbol{x}_{t+j}\epsilon_{t+j}\epsilon_t\boldsymbol{x}_t'), & j = -1, -2, \ldots. \end{cases}$$

Note that $\mathbb{E}(\boldsymbol{x}_t\epsilon_t\epsilon_{t-j}\boldsymbol{x}_{t-j}')$ are not the same as $\mathbb{E}(\boldsymbol{x}_{t-j}\epsilon_{t-j}\epsilon_t\boldsymbol{x}_t')$ in general.

When $\{\boldsymbol{x}_t\epsilon_t\}$ is a weakly stationary process such that $\mathbb{E}(\boldsymbol{x}_t\epsilon_t\epsilon_{t-j}\boldsymbol{x}_{t-j}')$ depend only on the time difference $|j|$ but not on t,

$$\boldsymbol{\Gamma}_T(j) = \boldsymbol{\Gamma}_T(-j) = \mathbb{E}(\boldsymbol{x}_t\epsilon_t\epsilon_{t-j}\boldsymbol{x}_{t-j}'), \quad j = 0, 1, 2, \ldots,$$

which are independent of T and may be written as $\boldsymbol{\Gamma}(j)$. It follows that $\boldsymbol{V}_o$ simplifies to

$$\boldsymbol{V}_o = \boldsymbol{\Gamma}(0) + \lim_{T\to\infty} 2\sum_{j=1}^{T-1}\boldsymbol{\Gamma}(j). \tag{6.5}$$

The presence of $\boldsymbol{\Gamma}_T(j)$, $j \neq 0$, in (6.4) (or $\boldsymbol{\Gamma}(j)$, $j \neq 0$, in (6.5)) renders the estimation of $\boldsymbol{V}_o$ practically difficult.

6.3.1 When Serial Correlations Are Absent

When $\boldsymbol{x}_t\epsilon_t$ are serially uncorrelated (but not necessarily independent over t), $\boldsymbol{\Gamma}_T(j)$ in (6.4) are all zero for $j \neq 0$, so that

$$\boldsymbol{V}_o = \lim_{T\to\infty}\boldsymbol{\Gamma}_T(0) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t'). \tag{6.6}$$

Estimation of $\boldsymbol{V}_o$ is then relatively simple.

Note that $\boldsymbol{x}_t\epsilon_t$ would be serially uncorrelated if $\{\epsilon_t\}$ is a *martingale difference sequence* with respect to the σ-algebras generated by $(\mathcal{Y}^{t-1}, \mathcal{W}^t)$, i.e., $\mathbb{E}\big(\epsilon_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t\big) = 0$. In this case, $\mathbb{E}(\boldsymbol{x}_t\epsilon_t) = \boldsymbol{0}$ and, for any $t \neq \tau$,

$$\mathbb{E}(\boldsymbol{x}_t\epsilon_t\epsilon_\tau\boldsymbol{x}_\tau') = \mathbb{E}\big[\boldsymbol{x}_t\,\mathbb{E}\big(\epsilon_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t\big)\epsilon_\tau\boldsymbol{x}_\tau'\big] = \boldsymbol{0}.$$

That is, $\{\boldsymbol{x}_t\epsilon_t\}$ is a sequence of uncorrelated, zero-mean random vectors.

A consistent estimator of $\boldsymbol{V}_o$ in (6.6) is its sample counterpart:

$$\widehat{\boldsymbol{V}}_T = \frac{1}{T}\sum_{t=1}^{T}\hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'. \tag{6.7}$$

To see this, we write $\hat{e}_t = \epsilon_t - \boldsymbol{x}_t'(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)$ and obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}[\hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t' - \mathbb{E}(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t')] \\
&\quad= \frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t' - \mathbb{E}(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t')\right) - \frac{2}{T}\sum_{t=1}^{T}\left(\epsilon_t\boldsymbol{x}_t'(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)\boldsymbol{x}_t\boldsymbol{x}_t'\right) + \\
&\qquad\frac{1}{T}\sum_{t=1}^{T}\left((\hat{\beta}_T - \boldsymbol{\beta}_o)'\boldsymbol{x}_t\boldsymbol{x}_t'(\hat{\beta}_T - \boldsymbol{\beta}_o)\boldsymbol{x}_t\boldsymbol{x}_t'\right).
\end{aligned}$$

The first term on the right-hand side would converge to zero in probability if $\{\epsilon_t^2 \boldsymbol{x}_t \boldsymbol{x}_t'\}$ obeys a WLLN. The second term on the right-hand side also vanishes because $\hat{\boldsymbol{\beta}}_T \xrightarrow{\mathbb{P}} \boldsymbol{\beta}_o$ and $T^{-1}\sum_{t=1}^T \epsilon_t \boldsymbol{x}_t' \boldsymbol{x}_t \boldsymbol{x}_t'$ converges in probability by a suitable WLLN. Similarly, the third term also vanishes in the limit provided that the average of $\boldsymbol{x}_t \boldsymbol{x}_t' \boldsymbol{x}_t \boldsymbol{x}_t'$ converges in probability. It follows that

$$\frac{1}{T}\sum_{t=1}^T [\hat{e}_t^2 \boldsymbol{x}_t \boldsymbol{x}_t' - \mathbb{E}(\epsilon_t^2 \boldsymbol{x}_t \boldsymbol{x}_t')] \xrightarrow{\mathbb{P}} \mathbf{0},$$

proving weak consistency of the estimator (6.7).

The estimator (6.7) is practically useful because it permits *conditional heteroskedasticity* of an unknown form, i.e., $\mathbb{E}(\epsilon_t^2 \mid \mathcal{Y}^{t-1}, \mathcal{W}^t)$ changes with t but does not have an explicit functional form. This estimator is therefore known as a *heteroskedasticity-consistent* covariance matrix estimator. A consistent estimator of $\boldsymbol{D}_o$ is then

$$\widehat{\boldsymbol{D}}_T = \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^T \hat{e}_t^2 \boldsymbol{x}_t \boldsymbol{x}_t'\right) \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1}. \tag{6.8}$$

The estimator (6.8) was proposed by Eicker (1967) and White (1980) and also known as the Eicker-White covariance matrix estimator.

If, in addition, ϵ_t are also *conditionally homoskedastic*:

$$\mathbb{E}\big(\epsilon_t^2 \mid \mathcal{Y}^{t-1}, \mathcal{W}^t\big) = \sigma_o^2,$$

(6.6) can be further simplified as

$$\begin{aligned} \boldsymbol{V}_o &= \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[\mathbb{E}\big(\epsilon_t^2 \mid \mathcal{Y}^{t-1}, \mathcal{W}^t\big)\boldsymbol{x}_t \boldsymbol{x}_t'\big] \\ &= \sigma_o^2 \left(\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t')\right) \\ &= \sigma_o^2\, \boldsymbol{M}_{xx}. \end{aligned} \tag{6.9}$$

The asymptotic variance-covariance matrix of $\sqrt{T}(\hat{\beta}_T - \boldsymbol{\beta}_o)$ is then

$$\boldsymbol{D}_o = \boldsymbol{M}_{xx}^{-1} \boldsymbol{V}_o \boldsymbol{M}_{xx}^{-1} = \sigma_o^2\, \boldsymbol{M}_{xx}^{-1}.$$

As $\boldsymbol{M}_{xx}$ can be consistently estimated by its sample counterpart, it remains to estimate σ_o^2. Exercise 6.6 shows that $\hat{\sigma}_T^2 = \sum_{t=1}^T \hat{e}_t^2/(T-k)$ is consistent for σ_o^2, where $\hat{e}_t$ are the OLS residuals. A consistent estimator of this $\boldsymbol{D}_o$ is

$$\widehat{\boldsymbol{D}}_T = \hat{\sigma}_T^2 \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1}. \tag{6.10}$$

Note that, apart from the factor T, $\widehat{\boldsymbol{D}}_T$ in this case is also the covariance matrix estimator for $\hat{\boldsymbol{\beta}}_T$ in the classical least squares theory.

While the estimator (6.10) is inconsistent under conditional heteroskedasticity, the Eicker-White estimator is "robust" and preserves consistency when heteroskedasticity is present and of an unknown form. It should be noted that, under conditional homoskedasticity, the Eicker-White estimator remains consistent but may suffer from some efficiency loss.

6.3.2 When Serial Correlations Are Present

When $\boldsymbol{x}_t\epsilon_t$ exhibit serial correlations, it is still possible to estimate (6.4) and (6.5) consistently. Let $\ell(T)$ denote a function of T that diverges with T such that

$$\boldsymbol{V}_T^\dagger = \sum_{j=-\ell(T)}^{\ell(T)} \boldsymbol{\Gamma}_T(j) \to \boldsymbol{V}_o,$$

as T tends to infinity. It is then natural to estimate $\boldsymbol{V}_T^\dagger$ by its sample counterpart:

$$\widehat{\boldsymbol{V}}_T^\dagger = \sum_{j=-\ell(T)}^{\ell(T)} \widehat{\boldsymbol{\Gamma}}_T(j),$$

with the sample autocovariances:

$$\widehat{\boldsymbol{\Gamma}}_T(j) = \begin{cases} \frac{1}{T}\sum_{t=j+1}^{T} \boldsymbol{x}_t\hat{e}_t\hat{e}_{t-j}\boldsymbol{x}_{t-j}', & j = 0, 1, 2, \ldots, \\ \frac{1}{T}\sum_{t=-j+1}^{T} \boldsymbol{x}_{t+j}\hat{e}_{t+j}\hat{e}_t\boldsymbol{x}_t', & j = -1, -2, \ldots. \end{cases}$$

The estimator $\widehat{\boldsymbol{V}}_T^\dagger$ approximates $\boldsymbol{V}_T^\dagger$ and would be consistent for $\boldsymbol{V}_o$ provided that $\ell(T)$ does not grow too fast with T.

A problem with $\widehat{\boldsymbol{V}}_T^\dagger$ is that it need not be a positive semi-definite matrix and hence may not be a proper variance-covariance matrix. A consistent estimator that is also positive semi-definite is the following non-parametric kernel estimator:

$$\widehat{\boldsymbol{V}}_T^\kappa = \sum_{j=-T+1}^{T-1} \kappa\Big(\frac{j}{\ell(T)}\Big)\widehat{\boldsymbol{\Gamma}}_T(j), \tag{6.11}$$

where κ is a kernel function and $\ell(T)$ is its bandwidth. The kernel function and its bandwidth jointly determine the weights assigned to $\widehat{\boldsymbol{\Gamma}}_T(j)$. This estimator is known as a *heteroskedasticity and autocorrelation-consistent* (HAC) covariance matrix estimator.

The HAC estimator was originated from spectral estimation in the time series literature and was brought to the econometrics literature by Newey and West (1987) and Gallant (1987). The resulting consistent estimator of $\boldsymbol{D}_o$ is

$$\widehat{\boldsymbol{D}}_T^\kappa = \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1} \widehat{\boldsymbol{V}}_T^\kappa \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}, \tag{6.12}$$

with $\widehat{\boldsymbol{V}}_T^\kappa$ given by (6.11), cf. the Eicker-White estimator (6.8). The estimator (6.12) is usually referred to as the Newey-West covariance matrix estimator in the literature.

The kernel function κ is typically required to satisfy: $|\kappa(x)| \leq 1$, $\kappa(0) = 1$, $\kappa(x) = \kappa(-x)$ for all $x \in \mathbb{R}$, $\int |\kappa(x)|\,\mathrm{d}x < \infty$, κ is continuous at 0 and at all but a finite number of other points in $\mathbb{R}$, and

$$\int_{-\infty}^{\infty} \kappa(x)e^{-ix\omega}\,\mathrm{d}x \geq 0, \quad \forall \omega \in \mathbb{R}.$$

Below are some commonly used kernel functions:
(i) Bartlett kernel (Newey and West, 1987):

$$\kappa(x) = \begin{cases} 1-|x|, & |x| \leq 1, \\ 0, & \text{otherwise;} \end{cases}$$

(ii) Parzen kernel (Gallant, 1987):

$$\kappa(x) = \begin{cases} 1-6x^2+6|x|^3, & |x| \leq 1/2, \\ 2(1-|x|)^3, & 1/2 \leq |x| \leq 1, \\ 0, & \text{otherwise;} \end{cases}$$

(iii) Quadratic spectral kernel (Andrews, 1991):

$$\kappa(x) = \frac{25}{12\pi^2x^2}\left(\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(6\pi x/5)\right);$$

(iv) Daniel kernel (Ng and Perron, 1996):

$$\kappa(x) = \frac{\sin(\pi x)}{\pi x}.$$

These kernels are all symmetric about the vertical axis, where the first two kernels have a bounded support $[-1, 1]$, but the other two have unbounded support. These kernel functions with non-negative x are depicted in Figure 6.1.

It can be seen from Figure 6.1 that the magnitudes of all kernel weights are all less than one. For the Bartlett and Parzen kernels, the weight assigned to $\widehat{\boldsymbol{\Gamma}}_T(j)$ decreases with $|j|$ and becomes zero for $|j| \geq \ell(T)$. Hence, $\ell(T)$ in these functions is also known

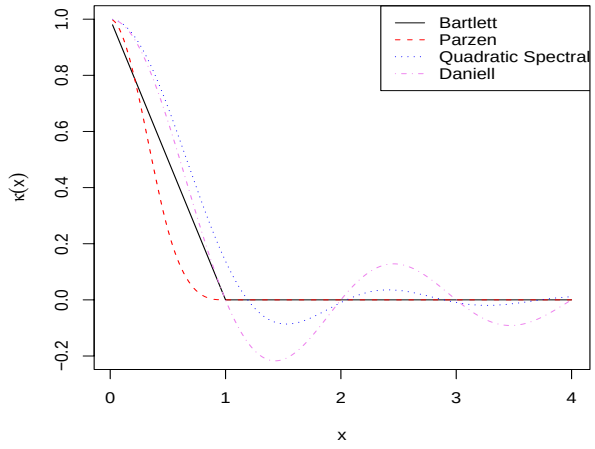

Figure 6.1: The Bartlett, Parzen, quandratic spectral, and Daniel kernels.

as a *truncation lag* parameter. For the quadratic spectral and Daniel kernels, there is no truncation, but the weights first decline and then exhibit damped sine waves for large $|j|$. The kernel weighting scheme brings bias to the estimated autocovariances. Yet, the kernel function entails little asymptotic bias because, for a given j, the kernel weights tends to unity asymptotically when $\ell(T)$ diverges with T. This is why the consistency of $\widehat{\boldsymbol{V}}_T^{\kappa}$ is not affected. Such bias, however, may not be negligible in finite samples, especially when $\ell(T)$ is small.

Both the Eicker-White estimator (6.8) and the Newey-West estimator (6.12) are non-parametric in the sense that they do not rely on any parametric model of conditional heteroskedasticity and serial correlations. Comparing to the Eicker-White estimator, the Newey-West estimator is robust to both conditional heteroskedasticity of ϵ_t and serial correlations of $\boldsymbol{x}_t\epsilon_t$. Yet, the latter would be less efficient than the former if $\boldsymbol{x}_t\epsilon_t$ are not serially correlated.

Remark: Andrews (1991) analyzed the estimator (6.12) with the Bartlett, Parzen and quadratic spectral kernels. It was shown that the estimator with the Bartlett kernel has the rate of convergence $O(T^{-1/3})$, whereas the other two kernels yield a faster rate of convergence, $O(T^{-2/5})$. Moreover, it is found that the quadratic spectral kernel is 8.6% more efficient asymptotically than the Parzen kernel, while the Bartlett kernel is the least efficient. These two results together suggest that the quadratic spectral kernel is to be preferred in HAC estimation, at least asymptotically. Andrews (1991) also proposed an "automatic" method to determine the desired bandwidth $\ell(T)$; we omit the details.

6.4 Large-Sample Tests

After learning the asymptotic properties of the OLS estimator under more general conditions [B1]–[B3], we are now able to construct tests for the parameters of interest and derive their limiting distributions. In this section, we will concentrate on three large-sample tests for the linear hypothesis

$$H_0\colon\ \boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r},$$

where $\boldsymbol{R}$ is a $q \times k$ ($q < k$) nonstochastic matrix and $\boldsymbol{r}$ is a pre-specified real vector, as in Section 3.3. We, again, require $\boldsymbol{R}$ to have rank q so as to exclude "redundant" hypotheses, the hypotheses that are linearly dependent on the other hypotheses.

6.4.1 Wald Test

Given that the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is consistent for some parameter vector $\boldsymbol{\beta}_o$, one would expect that $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ is "close" to $\boldsymbol{R}\boldsymbol{\beta}_o$ when T becomes large. As $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$ under the null hypothesis, whether $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ is sufficiently "close" to $\boldsymbol{r}$ constitutes an evidence for or against the null hypothesis. The *Wald test* is based on this intuition, and its key ingredient is the difference between $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ and the hypothetical value $\boldsymbol{r}$.

When [B1](i), [B2] and [B3] hold, we have learned from Theorem 6.6 that

$$\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}'),$$

where $\boldsymbol{D}_o = \boldsymbol{R}\boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}\boldsymbol{R}'$, or equivalently,

$$(\boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}')^{-1/2}\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q).$$

Letting $\widehat{\boldsymbol{V}}_T$ be a consistent estimator of $\boldsymbol{V}_o$,

$$\widehat{\boldsymbol{D}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\widehat{\boldsymbol{V}}_T\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}$$

is a consistent estimator for $\boldsymbol{D}_o$. We have the following asymptotic normality result based on $\widehat{\boldsymbol{D}}_T$:

$$(\boldsymbol{R}\widehat{\boldsymbol{D}}_T\boldsymbol{R}')^{-1/2}\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q). \tag{6.13}$$

As $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$ under the null hypothesis, the Wald test statistic is the inner product of (6.13):

$$\mathcal{W}_T = T(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'(\boldsymbol{R}\widehat{\boldsymbol{D}}_T\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}). \tag{6.14}$$

The result below follows directly from the continuous mapping theorem (Lemma 5.20).

Theorem 6.10 *Given the linear specification* (6.1)*, suppose that* [B1](i), [B2] *and* [B3] *hold. Then under the null hypothesis,*

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

where $\mathcal{W}_T$ *is given by* (6.14) *and* q *is the number of hypotheses.*

The Wald test has much wider applicability because it is valid for a wide variety of data which may be non-Gaussian, heteroskedastic, and/or serially correlated. What really matter here are two things: (1) asymptotic normality of the OLS estimator, and (2) a consistent estimator of $\boldsymbol{V}_o$. When an inconsistent estimator of $\boldsymbol{V}_o$ is used in the test statistic, $\widehat{\boldsymbol{D}}_T$ is inconsistent so that the resulting Wald statistic does not have a limiting χ^2 distribution.

Example 6.11 Given the linear specification

$$y_t = \boldsymbol{x}_{1,t}'\boldsymbol{b}_1 + \boldsymbol{x}_{2,t}'\boldsymbol{b}_2 + e_t,$$

where $\boldsymbol{x}_{1,t}$ is $(k-s)\times 1$ and $\boldsymbol{x}_{2,t}$ is $s\times 1$, suppose that $\boldsymbol{x}_{1,t}'\boldsymbol{b}_1 + \boldsymbol{x}_{2,t}'\boldsymbol{b}_2$ is the correct specification for the linear projection with $\boldsymbol{\beta}_o = [\boldsymbol{b}_{1,o}'\ \boldsymbol{b}_{2,o}']'$. An interesting hypothesis is whether the correct specification is of a simpler form: $\boldsymbol{x}_{1,t}'\boldsymbol{b}_1$. This amounts to testing the hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{0}$, where $\boldsymbol{R} = [\boldsymbol{0}_{s\times(k-s)}\ \boldsymbol{I}_s]$. The Wald test statistic for this hypothesis reads

$$\mathcal{W}_T = T\hat{\boldsymbol{\beta}}_T'\boldsymbol{R}'\big(\boldsymbol{R}\widehat{\boldsymbol{D}}_T\boldsymbol{R}'\big)^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}}_T \xrightarrow{D} \chi^2(s),$$

where $\widehat{\boldsymbol{D}}_T = (\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\hat{\boldsymbol{V}}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}$. The exact form of $\mathcal{W}_T$ depends on $\widehat{\boldsymbol{D}}_T$.

In particular, when $\widehat{\boldsymbol{V}}_T = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is a consistent estimator for $\boldsymbol{V}_o$, $\widehat{\boldsymbol{D}}_T = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}$ is consistent for $\boldsymbol{D}_o$, and the Wald statistic becomes

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

which is s times the standard F statistic discussed in Section 3.3.1. Further, if the null hypothesis is the i^{th} coefficient being zero, $\boldsymbol{R}$ is the i^{th} Cartesian unit vector $\boldsymbol{c}_i$, and the Wald statistic is

$$\mathcal{W}_T = T\hat{\boldsymbol{\beta}}_{i,T}^2/\hat{d}_{ii} \xrightarrow{D} \chi^2(1),$$

where $\hat{d}_{ii}$ is the i^{th} diagonal element of $\hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}$. Thus,

$$\sqrt{T}\hat{\beta}_{i,T}/\sqrt{\hat{d}_{ii}} \xrightarrow{D} \mathcal{N}(0,1), \tag{6.15}$$

where $(\hat{d}_{ii}/T)^{1/2}$ is the OLS standard error for $\hat{\beta}_{i,T}$. One can easily identify that the left-hand side of (6.15) is the standard t ratio discussed in Example 3.10 in Section 3.3. The difference is that the critical values of the t ratio should be taken from $\mathcal{N}(0,1)$, rather than a t distribution. When $\widehat{\boldsymbol{D}}_T = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}$ is inconsistent for $\boldsymbol{D}_o$, the t ratio can be robustified by choosing the i^{th} diagonal element of the Eicker-White or the Newey-West estimator $\widehat{\boldsymbol{D}}_T$ as $\hat{d}_{ii}$ in (6.15). The resulting $(\hat{d}_{ii}/T)^{1/2}$ is also known as the Eicker-White or the Newey-West standard error for $\hat{\beta}_{i,T}$. In other words, the significance of the i^{th} coefficient should be tested using the t ratio with a consistent standard error. □

Remark: The F-version of the Wald test is valid only when $\widehat{\boldsymbol{V}}_T = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is consistent for $\boldsymbol{V}_o$. As we have seen, this is the case when, e.g., $\{\epsilon_t\}$ is a martingale difference sequence and conditionally homoskedastic. Otherwise, this estimator need not be consistent for $\boldsymbol{V}_o$ and hence renders the F-version of the Wald test invalid. Nevertheless, the Wald test that involves a consistent $\widehat{\boldsymbol{D}}_T$ is still valid with a limiting χ^2 distribution.

6.4.2 Lagrange Multiplier Test

From Section 3.3.3 we have seen that, given the constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, the constrained OLS estimator can be obtained by finding the saddle point of the Lagrangian:

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

where $\boldsymbol{\lambda}$ is the $q \times 1$ vector of Lagrange multipliers. The underlying idea of the *Lagrange Multiplier (LM) test* of this constraint is to check 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); otherwise, $\boldsymbol{\lambda}$ ought to be "large." Again, the closeness between $\boldsymbol{\lambda}$ and zero must be determined by the distribution of the estimator of $\boldsymbol{\lambda}$.

The solutions to the Lagrangian above can be expressed as

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

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

Here, $\ddot{\boldsymbol{\beta}}_T$ denotes the constrained OLS estimator of $\boldsymbol{\beta}$, and $\ddot{\boldsymbol{\lambda}}_T$ is the basic ingredient of the LM test. Under the null hypothesis, the asymptotic normality of $\sqrt{T}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})$ now implies

$$\sqrt{T}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} 2\big(\boldsymbol{R}\boldsymbol{M}_{xx}^{-1}\boldsymbol{R}'\big)^{-1}\mathcal{N}(\boldsymbol{0}, \boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}'),$$

where $\boldsymbol{D}_o = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}$, or equivalently,

$$\sqrt{T}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}_o),$$

where $\boldsymbol{\Lambda}_o = 4(\boldsymbol{R}\boldsymbol{M}_{xx}^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}')(\boldsymbol{R}\boldsymbol{M}_{xx}^{-1}\boldsymbol{R}')^{-1}$. Equivalently, we have

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

which remains valid when $\boldsymbol{\Lambda}_o$ is replaced by a consistent estimator.

Let $\ddot{\boldsymbol{V}}_T$ be a consistent estimator of $\boldsymbol{V}_o$ based on the constrained estimation result. A consistent estimator of $\boldsymbol{\Lambda}_o$ is

$$\ddot{\boldsymbol{\Lambda}}_T = 4\big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\big]^{-1}\big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\ddot{\boldsymbol{V}}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\big] \big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\big]^{-1}.$$

It follows that

$$\ddot{\boldsymbol{\Lambda}}_T^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q). \tag{6.16}$$

The inner product of the left-hand side of (6.16) yields the LM statistic:

$$\mathcal{LM}_T = T\ddot{\boldsymbol{\lambda}}_T'\ddot{\boldsymbol{\Lambda}}_T^{-1}\ddot{\boldsymbol{\lambda}}_T. \tag{6.17}$$

The result below is a direct consequence of (6.16) and the continuous mapping theorem.

Theorem 6.12 *Given the linear specification* (6.1)*, suppose that* [B1](i), [B2] *and* [B3] *hold. Then under the null hypothesis,*

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

where $\mathcal{LM}_T$ *is given by* (6.17) *and* q *is the number of hypotheses.*

Similar to the Wald test, the LM test is also valid for a wide variety of data which may be non-Gaussian, heteroskedastic, and serially correlated. The asymptotic normality of the OLS estimator and consistent estimation of $\boldsymbol{V}_o$ remain crucial for the validity of the LM test. If an inconsistent estimator of $\boldsymbol{V}_o$ is used to construct $\ddot{\boldsymbol{\Lambda}}_T$, the resulting LM test will not have a limiting χ^2 distribution.

To implement the LM test, we write the vector of constrained OLS residuals as $\ddot{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\ddot{\boldsymbol{\beta}}_T$ and observe that

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

Thus, $\ddot{\boldsymbol{\lambda}}_T$ is

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

so that the LM test statistic can be computed as

$$\begin{aligned}\mathcal{LM}_T = T\ddot{\boldsymbol{e}}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\ddot{\boldsymbol{V}}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\big]^{-1} \\ \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ddot{\boldsymbol{e}}.\end{aligned} \tag{6.18}$$

This expression shows that, aside from matrix multiplication and matrix inversion, only constrained estimation is needed to compute the LM statistic. This is in sharp contrast with the Wald test which requires unconstrained estimation.

Remark: From (6.14) and (6.18) it is easy to see that the Wald and LM tests have distinct numerical values because they employ different consistent estimators of $\boldsymbol{V}_o$. Therefore, these two tests are asymptotically equivalent under the null hypothesis, i.e.,

$$\mathcal{W}_T - \mathcal{LM}_T \xrightarrow{\mathbb{P}} 0.$$

If $\boldsymbol{V}_o$ is known and does not have to be estimated, the Wald and LM tests would be algebraically equivalent. As these two tests have different statistics in general, they may result in conflicting inferences in finite samples.

Example 6.13 Analogous to Example 6.11, we are interested in testing whether additional s regressors should be added to the (constrained) specification:

$$y_t = \boldsymbol{x}_{1,t}'\boldsymbol{b}_1 + e_t.$$

The unconstrained specification is

$$y_t = \boldsymbol{x}_{1,t}'\boldsymbol{b}_1 + \boldsymbol{x}_{2,t}'\boldsymbol{b}_2 + e_t,$$

and the null hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{0}$ with $\boldsymbol{R} = [\boldsymbol{0}_{s\times(k-s)}\ \boldsymbol{I}_s]$. The constrained OLS estimator is $\ddot{\boldsymbol{\beta}}_T = (\ddot{\boldsymbol{b}}_{1,T}'\ \boldsymbol{0}')'$, with

$$\ddot{\boldsymbol{b}}_{1,T} = \left(\sum_{t=1}^T \boldsymbol{x}_{1,t}\boldsymbol{x}_{1,t}'\right)^{-1}\sum_{t=1}^T \boldsymbol{x}_{1,t}y_t = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}.$$

The LM statistic now can be computed as (6.18) with $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and $\ddot{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}_1\ddot{\boldsymbol{b}}_{1,T}$.

Consider now the special case that $\ddot{\boldsymbol{V}}_T = \ddot{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is consistent for $\boldsymbol{V}_o$ under the null hypothesis, where $\ddot{\sigma}_T^2 = \sum_{t=1}^T \ddot{e}_t^2/(T-k+s)$. Then, the LM test in (6.18) reads

$$\mathcal{LM}_T = T\ddot{\boldsymbol{e}}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\big]^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ddot{\boldsymbol{e}}/\ddot{\sigma}_T^2.$$

Applying the formula for the inverse of a partitioned matrix (Section 1.4), it is not too difficult to show that

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

where $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$. As $\boldsymbol{X}_1'\ddot{\boldsymbol{e}} = \boldsymbol{0}$ and $(\boldsymbol{I} - \boldsymbol{P}_1)\ddot{\boldsymbol{e}} = \ddot{\boldsymbol{e}}$, the LM statistic becomes

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

The fact $\ddot{\boldsymbol{e}}'\boldsymbol{X}_2\boldsymbol{R} = [\boldsymbol{0}_{1\times(k-s)}\ \ddot{\boldsymbol{e}}'\boldsymbol{X}_2] = \ddot{\boldsymbol{e}}'\boldsymbol{X}$ then leads to a simple form of the LM test:

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

where R^2 is the (non-centered) coefficient of determination of the auxiliary regression of $\ddot{\boldsymbol{e}}$ on $\boldsymbol{X}$. If $\ddot{\sigma}_T^2 = \sum_{t=1}^T \ddot{e}_t^2/T$ is used in the statistic, the LM test is simply TR^2. Thus, the LM test in this case can be easily obtained by running an auxiliary regression.

It must be emphasized that the simple TR^2 version of the LM statistic is valid only when $\ddot{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is a consistent estimator of $\boldsymbol{V}_o$; otherwise, TR^2 need not have a limiting χ^2 distribution. For example, if the LM statistic is based on the heteroskedasticity-consistent covariance matrix estimator:

$$\ddot{\boldsymbol{V}}_T = \frac{1}{T}\sum_{t=1}^T \ddot{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t',$$

it cannot be simplified to TR^2. □

Comparing Example 6.13 and Example 6.11, we can see that, while the LM test checks whether additional s regressors should be incorporated into a simpler (constrained) specification, the Wald test checks whether s regressors are redundant and should be excluded from a more complete (unconstrained) specification. The LM test thus permits testing a specification "from specific to general" (bottom up), and the Wald test evaluates a specification "from general to specific" (top down).

6.4.3 Likelihood Ratio Test

Another approach to hypothesis testing is to construct tests under the likelihood framework. In this section, we will not discuss the general, likelihood-based tests but focus only on a special case, the *likelihood ratio (LR) test* under the conditional normality assumption. We note that both the Wald and LM tests can also be derived under the same framework.

Recall from Section 3.2.3 that the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is also the MLE $\tilde{\boldsymbol{\beta}}_T$ that maximizes

$$L_T(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{T}\sum_{t=1}^T \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}.$$

When $\boldsymbol{x}_t$ are stochastic, this log-likelihood function is understood as the average of

$$\log f(y_t \mid \boldsymbol{x}_t; \boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\sigma^2},$$

where f is the conditional normal density function of with the conditional mean $\boldsymbol{x}_t'\boldsymbol{\beta}$ and the conditional variance σ^2.

When there is no constraint, $\tilde{\boldsymbol{\beta}}_T = \hat{\boldsymbol{\beta}}_T$ is the unconstrained MLE of $\boldsymbol{\beta}$. The unconstrained MLE of σ^2 is

$$\tilde{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T} \hat{e}_t^2,$$

where $\hat{e}_t = y_t - \boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}_T$ are the unconstrained residuals which are also the OLS residuals. Given the constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, let $\ddot{\boldsymbol{\beta}}_T$ denote the constrained MLE of $\boldsymbol{\beta}$. Then $\ddot{e}_t = y_t - \boldsymbol{x}_t'\ddot{\boldsymbol{\beta}}_T$ are the constrained residuals, and the constrained MLE of σ^2 is

$$\ddot{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T} \ddot{e}_t^2.$$

The LR test is based on the difference between the constrained and unconstrained L_T:

$$\mathcal{LR}_T = -2T\big(L_T(\ddot{\beta}_T, \ddot{\sigma}_T^2) - L_T(\tilde{\beta}_T, \tilde{\sigma}_T^2)\big) = T\log\left(\frac{\ddot{\sigma}_T^2}{\tilde{\sigma}_T^2}\right). \tag{6.19}$$

If the null hypothesis is true, two log-likelihood values should not be much different so that the likelihood ratio is close to one and $\mathcal{LR}_T$ is close to zero; otherwise, $\mathcal{LR}_T$ is positive. In contrast with the Wald and LM tests, the LR test has a disadvantage in practice because it requires estimating both constrained and unconstrained likelihood functions.

Writing the vector of $\ddot{e}_t$ as $\ddot{\boldsymbol{e}} = \boldsymbol{X}(\tilde{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T) + \hat{\boldsymbol{e}}$ and noting that $\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}$, we have

$$\ddot{\sigma}_T^2 = \tilde{\sigma}_T^2 + (\tilde{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T)'(\boldsymbol{X}'\boldsymbol{X}/T)(\tilde{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T).$$

In Section 6.4.2 we also find that

$$\ddot{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_T = -(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\big]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}).$$

It follows that

$$\ddot{\sigma}_T^2 = \tilde{\sigma}_T^2 + (\boldsymbol{R}\tilde{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}}_T - \boldsymbol{r}),$$

and that

$$\mathcal{LR}_T = T\,\log\big(1 + \underbrace{(\boldsymbol{R}\tilde{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}}_T - \boldsymbol{r})/\tilde{\sigma}_T^2}_{=:\,a_T}\big).$$

Owing to the consistency of the OLS estimator, $a_T \to 0$ almost surely (in probability). The mean value expansion of $\log(1 + a_T)$ about $a_T = 0$ is $(1 + a_T^\dagger)^{-1} a_T$, where $a_T^\dagger$ lies between a_T and 0 and hence also converges to zero almost surely (in probability). Note that Ta_T is exactly the Wald statistic with $\widehat{\boldsymbol{V}}_T = \tilde{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ and converges in distribution. The LR test statistic now can be written as

$$\mathcal{LR}_T = T(1 + a_T^\dagger)^{-1} a_T = Ta_T + o_{\mathrm{I\!P}}(1).$$

This shows that $\mathcal{LR}_T$ is asymptotically equivalent to Ta_T. Then, provided that $\widehat{\boldsymbol{V}}_T = \tilde{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is consistent for $\boldsymbol{V}_o$, $\mathcal{LR}_T$ also has a $\chi^2(q)$ distribution in the limit by Lemma 5.21.

Theorem 6.14 *Given the linear specification* (6.1)*, suppose that* [B1](i), [B2] *and* [B3] *hold and that* $\tilde{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ *is consistent for* $\boldsymbol{V}_o$*. Then under the null hypothesis,*

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

where $\mathcal{LR}_T$ *is given by* (6.19) *and* q *is the number of hypotheses.*

Remarks:

1. When $\tilde{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is consistent for $\boldsymbol{V}_o$, three large-sample tests (the LR, Wald and LM tests) are asymptotically equivalent under the null hypothesis. Otherwise, the LR test (6.19) may not even have a limiting χ^2 distribution. T

2. The LR test (6.19) can *not* be made robust to conditional heteroskedasticity and serial correlation. This should not be too surprising because the log-likelihood function postulated here is unable to accommodate heterogeneity and/or correlations over time.

3. When the Wald test involves $\widehat{\boldsymbol{V}}_T = \tilde{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ and the LM test uses $\ddot{\boldsymbol{V}}_T = \ddot{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$, it can be shown that

   $$\mathcal{W}_T \geq \mathcal{LR}_T \geq \mathcal{LM}_T;$$

   see Exercises 6.11 and 6.12. This is not an asymptotic result; conflicting inferences in finite samples therefore may arise when the critical values are between two statistics. See Godfrey (1988) for more details.

6.4.4 Power of the Tests

In this section we analyze the power property of the aforementioned tests under the alternative hypothesis that $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} \neq \boldsymbol{0}$.

We first consider the case that $\boldsymbol{D}_o$, the asymptotic variance-covariance matrix of $T^{1/2}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)$, is known. Recall that when $\boldsymbol{D}_o$ is known, the Wald statistic is

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

which is algebraically equivalent to the LM statistic. Under the alternative that $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r} + \boldsymbol{\delta}$,

$$\sqrt{T}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}) = \sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) + \sqrt{T}\boldsymbol{\delta},$$

where the first term on the right-hand side converges in distribution and hence is $O_{\mathbb{P}}(1)$. This implies that $\mathcal{W}_T$ must diverge at the rate T under the alternative hypothesis; in fact,

$$\frac{1}{T}\,\mathcal{W}_T \xrightarrow{\mathbb{P}} \boldsymbol{\delta}'(\boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}')^{-1}\boldsymbol{\delta}.$$

Consequently, for any critical value c, $\mathbb{P}(\mathcal{W}_T > c) \to 1$ when T tends to infinity; that is, the Wald test can reject the null hypothesis with probability approaching one. The Wald and LM tests in this case are therefore *consistent* tests.

When $\boldsymbol{D}_o$ is unknown, the estimator $\hat{\boldsymbol{D}}_T$ in the Wald test is computed from the unconstrained specification and is still consistent for $\boldsymbol{D}_o$ under the alternative. Analogous to the previous conclusion, we have

$$\frac{1}{T}\,\mathcal{W}_T \xrightarrow{\mathbb{P}} \boldsymbol{\delta}'(\boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}')^{-1}\boldsymbol{\delta},$$

showing that the Wald test is still consistent. On the other hand, the estimator $\ddot{\boldsymbol{D}}_T = (\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\ddot{\boldsymbol{V}}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}$ is computed from the constrained specification and need not be consistent for $\boldsymbol{D}_o$ under the alternative. It is not too difficult to see that, as long as $\ddot{\boldsymbol{D}}_T$ is bounded in probability, the LM test is also consistent because

$$\frac{1}{T}\,\mathcal{LM}_T = O_{\mathbb{P}}(1).$$

These consistency results ensure that the Wald and LM tests can detect any deviation, however small, from the null hypothesis when there is a sufficiently large sample.

6.5 Digression: Instrumental Variable Estimator

We have seen that the OLS estimator may lose consistency when the postulated specification is not a linear projection. This may happen when a model (i) omits relevant regressors

(Example 6.4, (ii) includes lagged dependent variables as regressors together with serially correlated errors (Example 6.5), or (iii) involve regressors that are measured with errors (Exercise 6.8). Indeed, inconsistency is not uncommon in practice. For example, when the dependent variable and regressors are jointly determined at the same time, the OLS estimator is inconsistent because the regressors are necessarily correlated with errors. This is known as a problem of *simultaneity*. There are other cases in which the OLS estimator loses consistency; we omit the details.

To obtain consistency for $\boldsymbol{\beta}_o$ in $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_o + \epsilon_t$, let $\boldsymbol{z}_t$ be a k-dimensional vector of variables taken from the information set $(\mathcal{Y}^{t-1}, \mathcal{W}^t)$ such that $\mathbb{E}(\boldsymbol{z}_t\epsilon_t) = \mathbf{0}$ and $\boldsymbol{z}_t$ are correlated with $\boldsymbol{x}_t$ in the sense that $\mathbb{E}(\boldsymbol{z}_t\boldsymbol{x}_t')$ is not singular. The sample counterpart of $\mathbb{E}(\boldsymbol{z}_t\epsilon_t) = \mathbb{E}[\boldsymbol{z}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_o)] = \mathbf{0}$ is

$$\frac{1}{T}\sum_{t=1}^{T}[\boldsymbol{z}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})] = \mathbf{0},$$

which is a system of k equations with k unknowns. Solving this system for $\boldsymbol{\beta}$ we obtain

$$\hat{\boldsymbol{\beta}}_{T,\mathrm{IV}} = \left(\sum_{t=1}^{T}\boldsymbol{z}_t\boldsymbol{x}_t'\right)^{-1}\left(\sum_{t=1}^{T}\boldsymbol{z}_t y_t\right).$$

When $\boldsymbol{z}_t\boldsymbol{x}_t'$ and $\boldsymbol{z}_t\boldsymbol{y}_t$ obey a suitable LLN with the corresponding limits $\boldsymbol{M}_{zx}$ and $\boldsymbol{m}_{zy}$, we immediately have

$$\hat{\boldsymbol{\beta}}_{T,\mathrm{IV}} \to \boldsymbol{M}_{zx}^{-1}\boldsymbol{m}_{zy},$$

almost surely (or in probability).

By construction,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\boldsymbol{z}_t y_t) = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{z}_t\boldsymbol{x}_t')\boldsymbol{\beta}_o.$$

Passing to the limit we have $\boldsymbol{\beta}_o = \boldsymbol{M}_{zx}^{-1}\boldsymbol{m}_{zy}$. Hence, $\hat{\boldsymbol{\beta}}_{T,\mathrm{IV}}$ is consistent for $\boldsymbol{\beta}_o$, where the variables $\boldsymbol{z}_t$ employed in estimation are mainly instrumental for "recovering" consistency. As such, the estimator $\hat{\boldsymbol{\beta}}_{T,\mathrm{IV}}$ is known as the *instrumental variable* (IV) estimator, with the *instruments* $\boldsymbol{z}_t$. Note that this estimator is also a *method of moment* estimator, because it is obtained by solving the sample counterpart of the moment conditions: $\mathbb{E}[\boldsymbol{z}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_o)] = \mathbf{0}$. This method breaks down when more than k instruments are available, however.

It is not too difficult to see that, when $T^{-1/2}\sum_{t=1}^{T}\boldsymbol{z}_t\epsilon_t$ obeys a suitable CLT and converges to $\mathcal{N}(\mathbf{0}, \boldsymbol{V}_o)$ with

$$\boldsymbol{V}_o = \lim_{T\to\infty}\mathrm{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{z}_t\epsilon_t\right),$$

the normalized IV estimator is also asymptotically normally distributed:

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_{T,\mathrm{IV}} - \boldsymbol{\beta}_o) = \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{x}_t'\right)^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{z}_t \epsilon_t\right) \xrightarrow{D} \mathcal{N}(\boldsymbol{0},\, \boldsymbol{D}_o),$$

where $\boldsymbol{D}_o = \boldsymbol{M}_{zx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{zx}^{-1}$. Similar to OLS estimation, we can compute a consistent estimator $\widehat{\boldsymbol{V}}_T$ for $\boldsymbol{V}_o$, so that

$$\widehat{\boldsymbol{V}}_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_{T,\mathrm{IV}} - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k).$$

A χ^2 test is then readily computed from this asymptotic normality property.

6.6 Asymptotic Properties of the GLS and FGLS Estimators

In this section we will digress from the OLS estimator and investigates the asymptotic properties of the GLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ and the FGLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$. We consider the case that $\boldsymbol{X}$ is stochastic and does not include lagged dependent variables. Assuming that $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ and $\mathrm{var}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{\Sigma}_o$, we have $\mathbb{E}(\hat{\boldsymbol{\beta}}_T) = \boldsymbol{\beta}_o$ and

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_T) = \mathbb{E}\big[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\big].$$

The GLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ is also unbiased and

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = \mathbb{E}\big(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X}\big)^{-1}.$$

As in Section 4.1, $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - (\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}$ is positive semi-definite with probability one, so that $\mathrm{var}(\hat{\boldsymbol{\beta}}_T) - \mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}})$ is a positive semi-definite matrix. The GLS estimator thus remains a more efficient estimator.

Analyzing the asymptotic properties of the GLS estimator is not straightforward. Recall that the GLS estimator can be computed as the OLS estimator of the transformed specification:

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

where $\tilde{\boldsymbol{y}} = \boldsymbol{\Sigma}_o^{-1/2}\boldsymbol{y}$, $\tilde{\boldsymbol{X}} = \boldsymbol{\Sigma}_o^{-1/2}\boldsymbol{X}$, and $\tilde{\boldsymbol{e}} = \boldsymbol{\Sigma}_o^{-1/2}\boldsymbol{e}$. Note that each element of $\tilde{\boldsymbol{y}}$, $\tilde{y}_t$, is a linear combination of all y_t with weights taken from $\boldsymbol{\Sigma}_o^{-1/2}$. Similarly, the $t^{\,\mathrm{th}}$ column of $\tilde{\boldsymbol{X}}'$, $\tilde{\boldsymbol{x}}_t$, is a linear combination of all $\boldsymbol{x}_t$. As such, even when y_t ($\boldsymbol{x}_t$) are independent across t, $\tilde{y}_t$ ($\tilde{\boldsymbol{x}}_t$) are highly correlated and may not obey a LLN and a CLT. It is therefore difficult to analyze the behavior of the GLS estimator, let alone the FGLS estimator.

Typically, $\boldsymbol{\Sigma}_o$ depends on a p-dimensional parameter vector $\boldsymbol{\alpha}_o$ and can be written as $\boldsymbol{\Sigma}(\boldsymbol{\alpha}_o)$. For simplicity, we shall consider only the case that $\boldsymbol{\Sigma}_o$ is a diagonal matrix

with the t^{th} diagonal element $\sigma_t^2(\boldsymbol{\alpha}_o)$. The transformed data are: $\tilde{y}_t = y_t/\sigma_t(\boldsymbol{\alpha}_o)$ and $\tilde{\boldsymbol{x}}_t = \boldsymbol{x}_t/\sigma_t(\boldsymbol{\alpha}_o)$; the GLS estimator is

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left(\sum_{t=1}^T \frac{\boldsymbol{x}_t\boldsymbol{x}_t'}{\sigma_t^2(\boldsymbol{\alpha}_o)}\right)^{-1}\left(\sum_{t=1}^T \frac{\boldsymbol{x}_t y_t'}{\sigma_t^2(\boldsymbol{\alpha}_o)}\right).$$

Under suitable conditions on y_t/σ_t and $\boldsymbol{x}_t/\sigma_t$, we are still able to show that $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is strongly (weakly) consistent for $\boldsymbol{\beta}_o$, and

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0},\, \tilde{\boldsymbol{M}}_{xx}^{-1}).$$

where $\tilde{\boldsymbol{M}}_{xx} = \lim_{T\to\infty} T^{-1}\sum_{t=1}^T \mathbb{E}[(\boldsymbol{x}_t\boldsymbol{x}_t')/\sigma_t^2(\boldsymbol{\alpha}_o)]$. Note that when $\sigma_t^2 = \sigma_o^2$ for all t, this asymptotic normality result is the same as that of the OLS estimator.

To compute the FGLS estimator, $\boldsymbol{\Sigma}_o$ is estimated by substituting an estimator $\hat{\boldsymbol{\alpha}}_T$ for $\boldsymbol{\alpha}_o$, where $\hat{\boldsymbol{\alpha}}_T$ is typically computed from the OLS results; see Section 4.2 and Section 4.3 for examples. The resulting estimator of $\boldsymbol{\Sigma}_o$ is $\hat{\boldsymbol{\Sigma}}_T = \boldsymbol{\Sigma}(\hat{\boldsymbol{\alpha}}_T)$ with the t^{th} diagonal element $\sigma_t^2(\hat{\boldsymbol{\alpha}}_T)$. The FGLS estimator is then

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = \left(\sum_{t=1}^T \frac{\boldsymbol{x}_t\boldsymbol{x}_t'}{\sigma_t^2(\hat{\boldsymbol{\alpha}}_T)}\right)^{-1}\left(\sum_{t=1}^T \frac{\boldsymbol{x}_t y_t'}{\sigma_t^2(\hat{\boldsymbol{\alpha}}_T)}\right).$$

Provided that $\hat{\boldsymbol{\alpha}}_T$ is consistent for $\boldsymbol{\alpha}_o$ and $\sigma_t^2(\cdot)$ is continuous at $\boldsymbol{\alpha}_o$, the FGLS estimator is asymptotically equivalent to the GLS estimator. Consequently,

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_{\text{FGLS}} - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0},\, \tilde{\boldsymbol{M}}_{xx}^{-1}).$$

Example 6.15 Consider the case that $\boldsymbol{y}$ exhibits groupwise heteroskedasticity:

$$\boldsymbol{\Sigma}_o = \begin{bmatrix} \sigma_1^2\boldsymbol{I}_{T_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2\boldsymbol{I}_{T_2} \end{bmatrix},$$

as discussed in Section 4.2. In the light of Exercise 6.6, we expect that the OLS variance estimator $\hat{\sigma}_1^2$ obtained from the first $T_1 = [Tm]$ observations is consistent for σ_1 and that $\hat{\sigma}_2^2$ obtained from the last $T - [Tm]$ observations is consistent for σ_2, where $0 < m < 1$. Under suitable conditions on y_t and $\boldsymbol{x}_t$,

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = \left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\hat{\sigma}_1^2} + \frac{\boldsymbol{X}_2'\boldsymbol{X}_2}{\hat{\sigma}_2^2}\right)^{-1}\left(\frac{\boldsymbol{X}_1'\boldsymbol{y}_1}{\hat{\sigma}_1^2} + \frac{\boldsymbol{X}_2'\boldsymbol{y}_2}{\hat{\sigma}_2^2}\right) \xrightarrow{\text{a.s.}} \boldsymbol{\beta}_o,$$

and

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_{\text{FGLS}} - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}\left(\mathbf{0},\, \left(\frac{m}{\sigma_1^2} + \frac{1-m}{\sigma_2^2}\right)^{-1}\boldsymbol{M}^{-1}\right),$$

where $\boldsymbol{M} = \lim_{T\to\infty} \boldsymbol{X}_1'\boldsymbol{X}_1/[Tm] = \lim_{T\to\infty} \boldsymbol{X}_2'\boldsymbol{X}_2/(T-[Tm])$. □

Exercises

6.1 Consider a linear specification with $\boldsymbol{x}_t = (1 \;\; d_t)'$, where d_t is a one-time dummy: $d_t = 1$ if $t = t^*$, a pre-specified time, and $d_t = 0$ otherwise. What is

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t')?$$

Does the OLS estimator have a finite limit?

6.2 Consider the specification $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$, where $\boldsymbol{x}_t$ is $k \times 1$. Suppose that

$$\mathbb{E}\big(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t\big) = \boldsymbol{z}_t'\boldsymbol{\gamma}_o,$$

where $\boldsymbol{z}_t$ is an $m \times 1$ vector with some elements different from $\boldsymbol{x}_t$. Assuming suitable strong laws for $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$, what is the almost sure limit of the OLS estimator of $\boldsymbol{\beta}$?

6.3 Consider the specification $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{z}_t'\boldsymbol{\gamma} + e_t$, where $\boldsymbol{x}_t$ is $k_1 \times 1$ and $\boldsymbol{z}_t$ is $k_2 \times 1$. Suppose that

$$\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o.$$

Assuming suitable strong laws for $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$, what are the almost sure limits of the OLS estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$?

6.4 Given the binary dependent variable $y_t = 1$ or 0 and random explanatory variables $\boldsymbol{x}_t$, suppose that a linear specification is

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t.$$

This is the linear probability model of Section 4.4 in the context that $\boldsymbol{x}_t$ are random. Let $F(\boldsymbol{x}_t'\boldsymbol{\theta}_o) = \mathbb{P}(y_t = 1 \mid \boldsymbol{x}_t)$ for some $\boldsymbol{\theta}_o$ and assume that $\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ and $\{\boldsymbol{x}_t F(\boldsymbol{x}_t'\boldsymbol{\theta}_o)\}$ obey a suitable SLLN (WLLN). What is the almost sure (probability) limit of $\hat{\boldsymbol{\beta}}_T$?

6.5 Given $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_o + \epsilon_t$, suppose that $\{\epsilon_t\}$ is a martingale difference sequence with respect to $\{\mathcal{Y}^{t-1}, \mathcal{W}^t\}$. Show that $\mathbb{E}(\epsilon_t) = 0$ and $\mathbb{E}(\epsilon_t\epsilon_\tau) = 0$ for all $t \neq \tau$. Is $\{\epsilon_t\}$ a white noise? Why or why not?

6.6 Given $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_o + \epsilon_t$, suppose that $\{\epsilon_t\}$ is a martingale difference sequence with respect to $\{\mathcal{Y}^{t-1}, \mathcal{W}^t\}$. State the conditions under which the OLS variance estimator $\hat{\sigma}_T^2$ is strongly consistent for σ_o^2.

6.7 State the conditions under which the OLS estimators of seemingly unrelated regressions are consistent and asymptotically normally distributed.

6.8 Suppose that $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ is the linear projection of y_t, where y_t are observable variables, but $\boldsymbol{x}_t$ can only be observed with random errors $\boldsymbol{u}_t$:

$$\boldsymbol{w}_t = \boldsymbol{x}_t + \boldsymbol{u}_t,$$

with $\mathbb{E}(\boldsymbol{u}_t) = \boldsymbol{0}$, $\mathrm{var}(\boldsymbol{u}_t) = \boldsymbol{\Sigma}_u$, and $\mathbb{E}(\boldsymbol{x}_t\boldsymbol{u}_t') = \boldsymbol{0}$, and $\mathbb{E}(y_t\boldsymbol{u}_t) = \boldsymbol{0}$. The linear specification $y_t = \boldsymbol{w}_t'\boldsymbol{\beta} + e_t$, together with these conditions, is known as a model with *measurement errors*. When this specification is evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$, we write $y_t = \boldsymbol{w}_t'\boldsymbol{\beta}_o + v_t$.

(a) Is $\boldsymbol{w}_t'\boldsymbol{\beta}_o$ also a linear projection of y_t?

(b) Assume that all the variables are well behaved in the sense that they obey some SLLN. Is $\hat{\boldsymbol{\beta}}_T$ strongly consistent for $\boldsymbol{\beta}_o$? If yes, explain why; if no, find the almost sure limit of $\hat{\boldsymbol{\beta}}_T$.

6.9 Given the specification: $y_t = \alpha y_{t-1} + e_t$, let $\hat{\alpha}_T$ denote the OLS estimator of α. Suppose that y_t are weakly stationary and generated according to $y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + u_t$, where u_t are i.i.d. with mean zero and variance σ_u^2.

(a) What is the almost sure (probability) limit α^* of $\hat{\alpha}_T$?

(b) What is the limiting distribution of $\sqrt{T}(\hat{\alpha}_T - \alpha^*)$?

6.10 Given the specification

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + e_t,$$

let $\hat{\alpha}_{1T}$ and $\hat{\alpha}_{2T}$ denote the OLS estimators of α_1 and α_2. Suppose that y_t are generated according to $y_t = \psi_1 y_{t-1} + u_t$ with $|\psi_1| < 1$, where u_t are i.i.d. with mean zero and variance σ_u^2.

(a) What are the almost sure (probability) limits of $\hat{\alpha}_{1T}$ and $\hat{\alpha}_{2T}$? Let α_1^* and α_2^* denote these limits.

(b) State the asymptotic normality results of the normalized OLS estimators.

6.11 Consider the log-likelihood function:

$$L_T(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{T}\sum_{t=1}^{T}\frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}.$$

(a) What is the LR test of $\mathbb{R}\boldsymbol{\beta}_o = r$ when $\sigma^2 = \sigma_o^2$ is known? Let $\mathcal{LR}_T(\sigma_o^2)$ denote this LR test. Given an intuitive explanation of $\mathcal{LR}_T(\sigma_o^2)$.

(b) When σ^2 is unknown, show that $\mathcal{W}_T = \mathcal{LR}_T(\tilde{\sigma}_T^2)$, where $\mathcal{W}_T$ is the Wald test (6.14) with $\hat{\boldsymbol{V}}_T = \tilde{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$, and $\tilde{\sigma}_T^2$ is the unconstrained MLE of σ^2.

(c) Show that

$$\mathcal{LR}_T(\tilde{\sigma}_T^2) = -2T\big[L_T(\tilde{\beta}_T^r, \tilde{\sigma}_T^2) - L_T(\tilde{\beta}_T, \tilde{\sigma}_T^2)\big],$$

where $\tilde{\beta}_T^r$ maximizes $L_T(\beta, \tilde{\sigma}_T^2)$ subject to the constraint $R\beta = r$. Use this fact to prove that $\mathcal{W}_T - \mathcal{LR}_T \geq 0$.

6.12 Consider the same framework as Exercise 6.11.

(a) When σ^2 is unknown, show that $\mathcal{LM}_T = \mathcal{LR}_T(\ddot{\sigma}_T^2)$, where $\mathcal{LM}_T$ is the LM test (6.18) with $\ddot{\boldsymbol{V}}_T = \ddot{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$, and $\ddot{\sigma}_T^2$ is the constrained MLE of σ^2.

(b) Show that

$$\mathcal{LR}_T(\ddot{\sigma}_T^2) = -2T\big[L_T(\ddot{\beta}_T, \ddot{\sigma}_T^2) - L_T(\ddot{\beta}_T^u, \ddot{\sigma}_T^2)\big],$$

where $\ddot{\beta}_T^u$ maximizes $L_T(\beta, \ddot{\sigma}_T^2)$. Use this fact to prove that $\mathcal{LR}_T - \mathcal{LM}_T \geq 0$.

References

Andrews, Donald W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, **59**, 817–858.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors, in L. M. LeCam and J. Neyman (eds.), *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 59–82, University of California, Berkeley.

Gallant, A. Ronald (1987). *Nonlinear Statistical Models*, New York, NY: Wiley.

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*, New York, NY: Cambridge University Press.

Newey, Whitney K. and Kenneth West (1987). A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, **55**, 703–708.

Ng, Serena and Pierre Perron (1996). The exact error in estimating the spectral density at the origin, *Journal of Time Series Analysis*, **17**, 379–408.

White, Halbert (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, **48**, 817–838.

White, Halbert (2001). *Asymptotic Theory for Econometricians*, revised edition, Orlando, FL: Academic Press.