Chapter 9

Quasi-Maximum Likelihood Theory

As discussed in preceding chapters, postulating a (non-)linear specification and estimating its unknown parameters by the least squares method amounts to approximating the conditional mean function of the dependent variable. This approach is practically useful, yet its scope is quite limited. First, it leaves no room for modeling other conditional moments, such as conditional variance, of the dependent variable. Second, it fails to accommodate certain characteristics of the dependent variable, such as binary response and data truncation. To provide a more complete description of the conditional behavior of a dependent variable, it is desirable to formulate a model that admits specifications of different conditional moments and/or other distribution characteristics. To this end, the method of *quasi-maximum likelihood* (QML) is to be preferred.

The QML method is essentially the same as the ML method usually seen in statistics and econometrics textbooks. A key difference between these two methods is that the former allows for possible misspecification of the likelihood function. It is conceivable that specifying a likelihood function, while being more general and more flexible than specifying a function for conditional mean, is more likely to result in specification errors. How to draw statistical inferences under potential model misspecification is thus a major concern of the QML method. By contrast, the conventional ML method assumes that the postulated likelihood function is specified correct, so that specification errors are "assumed away." As such, the results in the ML method are just special cases of the QML method. Our discussion below is primarily based on White (1994); for related discussion we also refer to White (1982), Amemiya (1985), Godfrey (1988), and Gourieroux and Monfort (1995).

9.1 Kullback-Leibler Information Criterion

We first discuss the concept of *information*. In a random experiment, suppose that the event A occurs with probability p. The message that A will surely occur would be more valuable (or more surprising) when p is small, but it is less informative (or less surprising) when p is large. Hence, the information content of the message that A will occur ought to be a decreasing function of the true event probability p.

A common choice of the *information function* is

$$\iota(p) = \log(1/p),$$

which decreases from positive infinity ($p \approx 0$) to zero ($p = 1$). It should be clear that $\iota(1-p)$, the information that A will not occur, is not the same as $\iota(p)$, unless $p = 0.5$. The expected information of these two messages is

$$I = p\,\iota(p) + (1-p)\,\iota(1-p) = p \log\Big(\frac{1}{p}\Big) + (1-p) \log\Big(\frac{1}{1-p}\Big).$$

The expected information I is also known as the *entropy* of the event A.

Similarly, the information that the probability of the event A changes from p to q would be useful when p and q are very different, but it is not of much value when p and q are close. The resulting information content is then the difference between these two pieces of information:

$$\iota(p) - \iota(q) = \log(q/p),$$

which is positive (negative) when $q > p$ ($q < p$). Given n mutually exclusive events $A_1, \ldots, A_n$, each with an information value $\log(q_i/p_i)$, the expected information value is then

$$I = \sum_{i=1}^{n} q_i \log\Big(\frac{q_i}{p_i}\Big).$$

This idea is readily generalized to discuss the information content of density functions, as discussed below.

Let g be the density function of the random variable z and f be another density function. Define the *Kullback-Leibler Information Criterion* (KLIC) of g relative to f as

$$\mathbb{I}(g\!:\!f) = \int_{\mathbb{R}} \log\Big(\frac{g(\zeta)}{f(\zeta)}\Big) g(\zeta)\, \mathrm{d}\zeta.$$

When f is used to describe z, the value $\mathbb{I}(\mathrm{g}:\mathrm{f})$ is the expected "surprise" resulted from the fact that g is the true density of z. The following result shows that the KLIC of g relative to f is non-negative.

Theorem 9.1 *Let g be the density function of the random variable z and f be another density function. Then $\mathbb{I}(g:f) \geq 0$, with the equality holding if, and only if, $g = f$ almost everywhere (i.e., $g = f$ except on a set with Lebesgue measure zero).*

Note, however, that the KLIC is not a metric because it is not reflexive in general, i.e., $\mathbb{I}(g:f) \neq \mathbb{I}(f:g)$, and it does not obey the triangle inequality; see Exercise 9.1. Hence, the KLIC is only a rough measure of the closeness between f and g.

Let $\{\boldsymbol{z}_t\}$ be a sequence of $\mathbb{R}^\nu$-valued random variables defined on the probability space $(\Omega, \mathcal{F}, \mathrm{IP}_o)$ and $\boldsymbol{z}^t$ be the collection of $(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_t)$. Given a sample of T observations, specifying a complete probability model for $\boldsymbol{z}^T$ may be a formidable task in practice because it involves too many random variables (T random vectors $\boldsymbol{z}_t$, each with ν random variables). Let y_t denote an element of $\boldsymbol{z}_t$ whose behavior is of particular interest to us. Writing $\boldsymbol{z}_t$ as $(y_t\ \boldsymbol{w}_t')'$, it is practically more convenient to specify a probability model for $g_t(y_t \mid \boldsymbol{x}_t)$, the density of y_t conditional on the information generated by a set of "pre-determined" variables, $\boldsymbol{x}_t$, which include some elements of $\boldsymbol{w}_t$ and $\boldsymbol{z}^{t-1}$. Similar to $\boldsymbol{z}^t$, we also write $y^t = (y_1, \ldots, y_t)$ and $\boldsymbol{x}^t = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)$.

To approximate $g_t(y_t \mid \boldsymbol{x}_t)$, we may specify a *quasi-likelihood* function $f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. Note that the prefix "quasi" is used to indicate that the likelihood is possibly misspecified. The KLIC of g_t relative to f_t is

$$\mathbb{I}(g_t : f_t; \boldsymbol{\theta}) = \int_{\mathbb{R}} \log\left(\frac{g_t(y_t \mid \boldsymbol{x}_t)}{f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})}\right) g_t(y_t \mid \boldsymbol{x}_t)\, \mathrm{d}y_t.$$

For a sample of T observations, we consider the average of T individual KLICs:

$$\begin{aligned} \bar{\mathbb{I}}_T(\{g_t : f_t\}; \boldsymbol{\theta}) &:= \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}(g_t : f_t; \boldsymbol{\theta}) \\ &= \frac{1}{T}\sum_{t=1}^{T} \big(\mathrm{IE}[\log g_t(y_t \mid \boldsymbol{x}_t)] - \mathrm{IE}[\log f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})]\big). \end{aligned} \tag{9.1}$$

It is clear that minimizing $\bar{\mathbb{I}}_T(\{g_t : f_t\}; \boldsymbol{\theta})$ in (9.1) is equivalent to maximizing

$$\bar{L}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{IE}[\log f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})]. \tag{9.2}$$

The maximizer of (9.2), $\boldsymbol{\theta}^*$, is thus the minimizer of the average KLIC (9.1). If there exists a $\boldsymbol{\theta}_o \in \Theta$ such that $f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}_o) = g_t(y_t \mid \boldsymbol{x}_t)$ for all t, we say that $\{f_t\}$ is correctly specified for $\{y_t \mid \boldsymbol{x}_t\}$. In this case, $\mathbb{I}(g_t : f_t; \boldsymbol{\theta}_o) = 0$, so that $\bar{\mathbb{I}}_T(\{g_t : f_t\}; \boldsymbol{\theta})$ is minimized at $\boldsymbol{\theta}^* = \boldsymbol{\theta}_o$.

Clearly, $\bar{L}_T(\boldsymbol{\theta})$ is not directly observable because it involves the expectation operator which depends on the joint density of $\boldsymbol{z}^T$. We may then maximize its sample counterpart:

$$L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}) := \frac{1}{T} \sum_{t=1}^{T} \log f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}), \tag{9.3}$$

the average of the individual quasi-log-likelihood functions. The resulting solution, $\tilde{\boldsymbol{\theta}}_T$, is known as the *quasi-maximum likelihood estimator* (QMLE) of $\boldsymbol{\theta}$. When $\{f_t\}$ is specified correctly for $\{y_t \mid \boldsymbol{x}_t\}$, the QMLE is understood as the standard MLE, as in standard statistics and econometrics textbooks.

In practice, one may concentrate on certain conditional attribute of y_t and postulate a specification $\mu_t(\boldsymbol{x}_t; \boldsymbol{\theta})$ for this attribute. A leading example is the following specification of conditional normality with $\mu_t(\boldsymbol{x}_t; \boldsymbol{\theta})$ as the specification of its mean:

$$y_t \mid \boldsymbol{x}_t \sim \mathcal{N}\big(\mu_t(\boldsymbol{x}_t; \boldsymbol{\beta}), \sigma^2\big),$$

or more generally, with $\mu_t(\boldsymbol{x}_t; \boldsymbol{\theta})$ and $h(\boldsymbol{x}_t; \boldsymbol{\alpha})$ as the specifications of its respective mean and variance:

$$y_t \mid \boldsymbol{x}_t \sim \mathcal{N}\big(\mu_t(\boldsymbol{x}_t; \boldsymbol{\beta}), h(\boldsymbol{x}_t; \boldsymbol{\alpha})\big).$$

Note that the functional forms for conditional mean and conditional variance may have specification errors; even the specification of normality may also be incorrect.

For example, consider the specification $y_t \mid \boldsymbol{x}_t \sim \mathcal{N}\big(\mu_t(\boldsymbol{x}_t; \boldsymbol{\beta}), \sigma^2\big)$. Setting $\boldsymbol{\theta} = (\boldsymbol{\beta}' \; \sigma^2)'$, it is easy to see that the maximizer of $T^{-1} \sum_{t=1}^{T} \log f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$ leads to the solution to

$$\min_{\boldsymbol{\beta}} \frac{1}{T} \sum_{t=1}^{T} [y_t - \mu_t(\boldsymbol{x}_t; \boldsymbol{\beta})]' \, [y_t - \mu_t(\boldsymbol{x}_t; \boldsymbol{\beta})] \, .$$

That is, the QMLE of $\boldsymbol{\beta}$ is also the NLS estimator of $\boldsymbol{\beta}$. As such, the NLS estimator can be viewed as a QMLE under the specification of conditional normality with conditional homoskedasticity. Even when $\{\mu_t\}$ is correctly specified for the conditional mean in the sense that there exists a $\boldsymbol{\theta}_o$ such that $\mu_t(\boldsymbol{x}_t; \boldsymbol{\theta}_o) = \mathbb{E}(y_t \mid \boldsymbol{x}_t)$, there is no guarantee that the specifications of σ^2 is correct.

9.2 Asymptotic Properties of the QMLE

The quasi-log-likelihood function is, in general, a nonlinear function in $\boldsymbol{\theta}$. The QMLE must be computed numerically using a nonlinear optimization algorithm, so that the algorithms discussed in Section 8.2.2 are readily applied. We do not repeat these methods here but shall proceed to discuss the asymptotic properties of the QMLE. For our subsequent analysis, we always assume that the specified quasi-log-likelihood function is twice continuously differentiable on a compact parameter space Θ with probability one and that integration and differentiation can be interchanged. Moreover, we maintain the following identification condition.

[ID-2] There exists a unique $\boldsymbol{\theta}^*$ that minimizes the KLIC: (9.1).

9.2.1 Consistency

We sketch the idea of establishing the consistency of the QMLE. In the light of the uniform law of large numbers discussed in Section 8.3.1, we know that if $L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta})$ tends to $\bar{L}_T(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta} \in \Theta$, i.e., $L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta})$ obeys a WULLN, then $\tilde{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}^*$ in probability, where $\boldsymbol{\theta}^*$ is the minimizer of the average KLIC, $\bar{\mathbb{I}}_T(\{g_t : f_t\}; \boldsymbol{\theta})$. When $\{f_t\}$ is specified correctly for $\{y_t \mid \boldsymbol{x}_t\}$, the KLIC minimizer $\boldsymbol{\theta}^*$ is also the true parameter $\boldsymbol{\theta}_o$. In this case, the QMLE is weakly consistent for $\boldsymbol{\theta}_o$. Therefore, the regularity conditions that ensure QMLE consistency are basically those ensuring a WULLN of the quasi-log-likelihood function; we will not pursue the technical details here.

9.2.2 Asymptotic Normality

Given that $\tilde{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}^*$ in probability, the asymptotic normality of $\tilde{\boldsymbol{\theta}}_T$ can be established by analyzing the local behavior of the quasi-log-likelihood function L_T around $\boldsymbol{\theta}^*$. When $\boldsymbol{\theta}^*$ is in the interior of Θ, the mean-value expansion of $\nabla L_T(y^T, \boldsymbol{x}^T; \tilde{\boldsymbol{\theta}}_T)$ about $\boldsymbol{\theta}^*$ is

$$\nabla L_T(y^T, \boldsymbol{x}^T; \tilde{\boldsymbol{\theta}}_T) = \nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}^*) + \nabla^2 L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}_T^\dagger)(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*), \tag{9.4}$$

where $\boldsymbol{\theta}_T^\dagger$ is between $\tilde{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}^*$, and the left-hand side of (9.4) is zero because the QMLE $\tilde{\boldsymbol{\theta}}_T$ solves the first order condition $\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}) = \mathbf{0}$. The asymptotic normality of $\tilde{\boldsymbol{\theta}}_T$ is then determined by the right-hand side of (9.4).

Let $\boldsymbol{H}_T(\boldsymbol{\theta}) = \mathbb{E}[\nabla^2 L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta})]$ be the expected value of the Hessian matrix of the specified quasi-log-likelihood function. As $\tilde{\boldsymbol{\theta}}_T$ is weakly consistent for $\boldsymbol{\theta}^*$, so is $\boldsymbol{\theta}_T^\dagger$. When $\nabla^2 L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}_T^\dagger)$ obeys a WULLN, we have

$$\nabla^2 L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}_T^\dagger) - \boldsymbol{H}_T(\boldsymbol{\theta}^*) \xrightarrow{\mathbb{P}} 0.$$

Then, provided that $\boldsymbol{H}_T(\boldsymbol{\theta})$ is invertible, (9.4) can be written as

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\sqrt{T}\,\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1). \tag{9.5}$$

This shows that the asymptotic distribution of $\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*)$ is essentially determined by the asymptotic distribution of the normalized score: $\sqrt{T}\,\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}^*)$. Note that the invertibility of $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ ensures that $\nabla^2 L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}_T^\dagger)$ is also invertible when T is sufficiently large. This in turn implies that the quasi-log-likelihood function L_T must be locally quadratic at $\boldsymbol{\theta}^*$, at least asymptotically.

Let $\boldsymbol{B}_T$ denote the variance-covariance matrix of $\sqrt{T}\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta})$:

$$\boldsymbol{B}_T(\boldsymbol{\theta}) = \mathrm{var}\big(\sqrt{T}\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta})\big) = \mathrm{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \nabla \log f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})\right),$$

which will also be referred to as the *information matrix*. Then provided that $\nabla \log f_t(y_t \mid \boldsymbol{x}_t)$ obeys a CLT, we have

$$\boldsymbol{B}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\big(\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}^*) - \mathbb{E}[\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}^*)]\big) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{I}_k). \tag{9.6}$$

When differentiation and integration can be interchanged,

$$\mathbb{E}[\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta})] = \nabla\,\mathbb{E}[L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta})] = \nabla \bar{L}_T(\boldsymbol{\theta}),$$

where the right-hand side is the first order derivative of (9.2). As $\boldsymbol{\theta}^*$ is the KLIC minimizer, $\nabla \bar{L}_T(\boldsymbol{\theta}^*) = \boldsymbol{0}$ so that $\mathbb{E}[\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}^*)] = \boldsymbol{0}$. By (9.5) and (9.6),

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\boldsymbol{H}_T(\boldsymbol{\theta}^*)\boldsymbol{B}_T(\boldsymbol{\theta}^*)^{1/2}\big[\boldsymbol{B}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\,\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}^*)\big] + o_{\mathbb{P}}(1),$$

which has an asymptotic normal distribution, as shown in the following result.

Theorem 9.2 *When* (9.4), (9.5) *and* (9.6) *hold,*

$$\boldsymbol{C}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{I}_k),$$

where

$$\boldsymbol{C}_T(\boldsymbol{\theta}^*) = \boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{B}_T(\boldsymbol{\theta}^*)\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1},$$

with $\boldsymbol{H}_T(\boldsymbol{\theta}^*) = \mathbb{E}[\nabla^2 L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}^*)]$ *and* $\boldsymbol{B}_T(\boldsymbol{\theta}^*) = \mathrm{var}\big(\sqrt{T}\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}^*)\big)$.

9.3 Information Matrix Equality

A useful result in the quasi-maximum likelihood theory is the *information matrix equality*. This equality shows that when the specification is correct up to certain extent, the information matrix $\boldsymbol{B}_T(\boldsymbol{\theta})$ is the same as the negative of the expected Hessian matrix $-\boldsymbol{H}_T(\boldsymbol{\theta})$ when evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_o$, i.e.,

$$\boldsymbol{H}_T(\boldsymbol{\theta}_o) + \boldsymbol{B}_T(\boldsymbol{\theta}_o) = \mathbf{0}.$$

In this case, the asymptotic covariance matrix $\boldsymbol{C}_T(\boldsymbol{\theta}_o)$ can be simplified to $-\boldsymbol{H}_T(\boldsymbol{\theta}_o)^{-1}$ or $\boldsymbol{B}_T(\boldsymbol{\theta}_o)^{-1}$ which renders its estimation simpler.

For the specification of $\{y_t|\boldsymbol{x}_t\}$, define the following score functions:

$$\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}) = \nabla \log f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}) = f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta})^{-1} \nabla f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}),$$

so that $\nabla f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}) = \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}) f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta})$. By permitting interexchange of differentiation and integration we have

$$\int_{\mathbb{R}} \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}) f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}) \, \mathrm{d}y_t = \nabla \int_{\mathbb{R}} f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}) \, \mathrm{d}y_t = \mathbf{0}.$$

If $\{f_t\}$ is correctly specified for $\{y_t|\boldsymbol{x}_t\}$, we have $\mathrm{IE}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)|\boldsymbol{x}_t] = \mathbf{0}$, where the conditional expectation is taken with respect to $g_t(y_t|\boldsymbol{x}_t) = f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}_o)$. By the law of iterated expectations, we have $\mathrm{IE}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)] = \mathbf{0}$, so that the mean score is zero under correct specification. Similarly,

$$\begin{aligned}
&\int_{\mathbb{R}} [\nabla \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}) + \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}) \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta})'] \, f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}) \, \mathrm{d}y_t \\
&\qquad = \int_{\mathbb{R}} \nabla \big(\nabla f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta})\big) \, \mathrm{d}y_t \\
&\qquad = \nabla^2 \int_{\mathbb{R}} f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta})\big) \, \mathrm{d}y_t \\
&\qquad = \mathbf{0}.
\end{aligned}$$

It follows that

$$\mathrm{IE}[\nabla \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)|\boldsymbol{x}_t] + \mathrm{IE}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o) \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)'|\boldsymbol{x}_t] = \mathbf{0}.$$

Consequently,

$$\begin{aligned}
&\frac{1}{T} \sum_{t=1}^{T} \mathrm{IE}[\nabla \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)] + \frac{1}{T} \sum_{t=1}^{T} \mathrm{IE}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o) \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)'] \\
&\qquad = \boldsymbol{H}_T(\boldsymbol{\theta}_o) + \frac{1}{T} \sum_{t=1}^{T} \mathrm{IE}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o) \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)'] \qquad (9.7) \\
&\qquad = \mathbf{0}.
\end{aligned}$$

This shows that the expected Hessian matrix is the negative of the averaged of individual information matrices, i.e., $\mathrm{var}(\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o))$. Note, however, that the latter need not be the information matrix $\boldsymbol{B}_T(\theta_o)$.

To see this, recall that

$$\begin{aligned}\boldsymbol{B}_T(\boldsymbol{\theta}_o) &= \frac{1}{T}\,\mathbb{E}\left[\left(\sum_{t=1}^{T} \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)\right)\left(\sum_{t=1}^{T} \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)'\right)\right] \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)'] \\ &\qquad + \frac{1}{T}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T} \mathbb{E}[\boldsymbol{s}_{t-\tau}(y_{t-\tau}, \boldsymbol{x}_{t-\tau}; \boldsymbol{\theta}_o)\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)'] \\ &\qquad + \frac{1}{T}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T} \mathbb{E}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)\boldsymbol{s}_{t+\tau}(y_{t+\tau}, \boldsymbol{x}_{t+\tau}; \boldsymbol{\theta}_o)'].\end{aligned}$$

A specification of $\{y_t|\boldsymbol{x}_t\}$ is said to have *dynamic misspecification* if it is not correctly specified for $\{y_t|\boldsymbol{w}_t, \boldsymbol{z}^{t-1}\}$; that is, there does not exist any $\boldsymbol{\theta}_o$ such that $f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}_o) = g_t(y_t|\boldsymbol{w}_t, \boldsymbol{z}^{t-1})$. Thus, the information contained in $\boldsymbol{w}_t$ and $\boldsymbol{z}^{t-1}$ cannot be fully captured by $\boldsymbol{x}_t$, and some important variables are omitted in the conditioning set, such as remote lagged y_t. On the other hand, when $f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}_o) = g_t(y_t|\boldsymbol{w}_t, \boldsymbol{z}^{t-1})$, we have

$$\mathbb{E}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)|\boldsymbol{x}_t] = \mathbb{E}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)|\boldsymbol{w}_t, \boldsymbol{z}^{t-1}] = \mathbf{0}, \tag{9.8}$$

so that by the law of iterated expectations,

$$\begin{aligned}&\mathbb{E}\big[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)\boldsymbol{s}_{t+\tau}(y_{t+\tau}, \boldsymbol{x}_{t+\tau}; \boldsymbol{\theta}_o)'\big] \\ &\qquad = \mathbb{E}\big[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)\,\mathbb{E}[\boldsymbol{s}_{t+\tau}(y_{t+\tau}, \boldsymbol{x}_{t+\tau}; \boldsymbol{\theta}_o)'|w_{t+\tau}, \boldsymbol{z}^{t+\tau-1}]\big] = \mathbf{0},\end{aligned}$$

for $\tau \geq 1$. We thus have

$$\boldsymbol{B}_T(\boldsymbol{\theta}_o) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)'\big],$$

so that (9.7) is understood as the information matrix equality. On the other hand, when dynamic misspecification is present, (9.7) remains valid, yet it is not the information matrix equality.

Theorem 9.3 *Suppose that there exists a $\boldsymbol{\theta}_o$ such that $f_t(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}_o) = g_t(y_t|\boldsymbol{x}_t)$ and there is no dynamic misspecification. Then,*

$$\boldsymbol{H}_T(\boldsymbol{\theta}_o) + \boldsymbol{B}_T(\boldsymbol{\theta}_o) = \mathbf{0},$$

where $\boldsymbol{H}_T(\boldsymbol{\theta}_o) = T^{-1}\sum_{t=1}^{T} \mathrm{IE}[\nabla \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)]$ *and*

$$\boldsymbol{B}_T(\boldsymbol{\theta}_o) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{IE}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)'].$$

When Theorem 9.3 holds, the covariance matrix needed to normalize $\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_o)$ simplifies to $\boldsymbol{B}_T(\boldsymbol{\theta}_o)^{-1} = -\boldsymbol{H}_T(\boldsymbol{\theta}_o)^{-1}$; that is,

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

This shows that the QMLE achieves the Cramér-Rao lower bound asymptotically. Note that it is typically easy to estimate $\boldsymbol{H}_T(\boldsymbol{\theta}_o)$ consistently. When the information matrix equality does not hold, one must also consistently estimate $\boldsymbol{B}_T(\boldsymbol{\theta}_o)$ and then compute $\boldsymbol{C}(\boldsymbol{\theta}_o)$. When dynamic misspecification is present, a consistent estimator of $\boldsymbol{B}_T(\boldsymbol{\theta}_o)$ may be obtained using a Newey-West type estimator.

Example 9.4 Consider the following specification: $y_t|\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$ for all t. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}' \ \sigma^2)'$, then

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

and

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

Straightforward calculation yields

$$\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \frac{\boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})}{\sigma^2} \\ -\frac{1}{2\sigma^2} + \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2(\sigma^2)^2} \end{bmatrix},$$

$$\nabla^2 L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} -\frac{\boldsymbol{x}_t\boldsymbol{x}_t'}{\sigma^2} & -\frac{\boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})}{(\sigma^2)^2} \\ -\frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})\boldsymbol{x}_t'}{(\sigma^2)^2} & \frac{1}{2(\sigma^2)^2} - \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{(\sigma^2)^3} \end{bmatrix}.$$

Setting $\nabla L_T(y^T, \boldsymbol{x}^T; \boldsymbol{\theta}) = \boldsymbol{0}$ we can solve for $\boldsymbol{\beta}$ from the first set of first order conditions to obtain the QMLE $\tilde{\boldsymbol{\beta}}_T$, which is nothing but the OLS estimator $\hat{\boldsymbol{\beta}}_T$. It can also be seen that the QMLE of σ^2 is the average of the OLS residuals: $\tilde{\sigma}_T^2 = T^{-1}\sum_{t=1}^{T}(y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T)^2$.

If the specification above is correct for $\{y_t|\boldsymbol{x}_t\}$, there exists $\boldsymbol{\theta}_o = (\boldsymbol{\beta}_o' \ \sigma_o^2)'$ such that the conditional distribution of y_t given $\boldsymbol{x}_t$ is $\mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}_o, \sigma_o^2)$. Taking expectation with respect to the true distribution function, we have

$$\mathrm{IE}[\boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})] = \mathrm{IE}[\boldsymbol{x}_t(\mathrm{IE}(y_t|\boldsymbol{x}_t) - \boldsymbol{x}_t'\boldsymbol{\beta})] = \mathrm{IE}(\boldsymbol{x}_t\boldsymbol{x}_t')(\boldsymbol{\beta}_o - \boldsymbol{\beta}),$$

which is zero when evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$. Similarly,

$$\begin{aligned}
\mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2] &= \mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_o + \boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})^2] \\
&= \mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_o)^2] + \mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})^2] \\
&= \sigma_o^2 + \mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})^2],
\end{aligned}$$

where the second term on the right-hand side is zero if it is evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$. These results together show that

$$\begin{aligned}
\boldsymbol{H}_T(\boldsymbol{\theta}) &= \mathbb{E}[\nabla^2 L_T(\boldsymbol{\theta})] \\
&= \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} -\frac{\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')}{\sigma^2} & -\frac{\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')(\boldsymbol{\beta}_o-\boldsymbol{\beta})}{(\sigma^2)^2} \\ -\frac{(\boldsymbol{\beta}_o-\boldsymbol{\beta})'\,\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')}{(\sigma^2)^2} & \frac{1}{2(\sigma^2)^2} - \frac{\sigma_o^2+\mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o-\boldsymbol{x}_t'\boldsymbol{\beta})^2]}{(\sigma^2)^3} \end{bmatrix}.
\end{aligned} \tag{9.9}$$

When $H_T(\boldsymbol{\theta})$ is evaluated at $\boldsymbol{\theta}_o = (\boldsymbol{\beta}_o' \ \sigma_o^2)'$, we have

$$\boldsymbol{H}_T(\boldsymbol{\theta}_o) = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} -\frac{\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')}{\sigma_o^2} & \boldsymbol{0} \\ \boldsymbol{0}' & -\frac{1}{2(\sigma_o^2)^2} \end{bmatrix}.$$

If there is no dynamic misspecification, it is straightforward to show that the information matrix is

$$\boldsymbol{B}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E} \begin{bmatrix} \frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^2\boldsymbol{x}_t\boldsymbol{x}_t'}{(\sigma^2)^2} & -\frac{\boldsymbol{x}_t(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})}{2(\sigma^2)^2} + \frac{\boldsymbol{x}_t(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^3}{2(\sigma^2)^3} \\ -\frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})\boldsymbol{x}_t'}{2(\sigma^2)^2} + \frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^3\boldsymbol{x}_t'}{2(\sigma^2)^3} & \frac{1}{4(\sigma^2)^2} - \frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^2}{2(\sigma^2)^3} + \frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^4}{4(\sigma^2)^4} \end{bmatrix}. \tag{9.10}$$

Given that y_t is conditionally normally distributed, its conditional third and fourth central moments are zero and $3(\sigma_o^2)^2$, respectively. It can then be verified that

$$\mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^3] = 3\sigma_o^2\,\mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})] + \mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})^3],$$

which is zero when evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$. Similarly,

$$\mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^4] = 3(\sigma_o^2)^2 + 6\sigma_o^2\,\mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})^2] + \mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})^4],$$

which is $3(\sigma_o^2)^2$ when evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$; see Exercise 9.2. Consequently,

$$\boldsymbol{B}_T(\boldsymbol{\theta}_o) = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} \frac{\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')}{\sigma_o^2} & \boldsymbol{0} \\ \boldsymbol{0}' & \frac{1}{2(\sigma_o^2)^2} \end{bmatrix}.$$

It is now easily seen that the information matrix equality holds. Yet, when there is dynamic missepcification, $\boldsymbol{B}_T(\boldsymbol{\theta})$ would not be the same as the form given above so that the information matrix equality breaks down; see Exercise 9.3.

A typical consistent estimator of $\boldsymbol{H}_T(\boldsymbol{\theta}_o)$ is the sample counterpart of $\boldsymbol{H}_T(\boldsymbol{\theta}_o)$ with σ_o^2 replaced by its QMLE $\tilde{\sigma}_T^2$:

$$\widetilde{\boldsymbol{H}}_T = \begin{bmatrix} -\frac{\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'}{T\tilde{\sigma}_T^2} & \boldsymbol{0} \\ \boldsymbol{0}' & -\frac{1}{2(\tilde{\sigma}_T^2)^2} \end{bmatrix}.$$

When the information matrix equality holds, a consistent estimator of $\boldsymbol{C}_T(\boldsymbol{\theta}_o)$ is $-\widetilde{\boldsymbol{H}}_T^{-1}$. It can be seen that the upper-left block of $-\widetilde{\boldsymbol{H}}_T^{-1}$ is $\tilde{\sigma}_T^2(\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'/T)^{-1}$, which is the standard OLS estimator for the asymptotic covariance matrix of $T^{1/2}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)$. □

The example below shows that, even when the specification for $\mathbb{E}(y_t|\boldsymbol{x}_t)$ is correct and there is no dynamic misspecification, the information matrix equality may still fail to hold if there is misspecification of other conditional moments, such as neglected conditional heteroskedasticity.

Example 9.5 Suppose that the true conditional density is

$$y_t|\boldsymbol{x}_t \sim \mathcal{N}\big(\boldsymbol{x}_t'\boldsymbol{\beta}_o, h(\boldsymbol{x}_t, \boldsymbol{\alpha}_o)\big),$$

where the conditional variance $h(\boldsymbol{x}_t, \boldsymbol{\alpha}_o)$ varies with $\boldsymbol{x}_t$. Yet, our specification is still that in Example 9.4: $y_t|\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$. Then, this specification includes a correct specification for the conditional mean but a misspecified conditional variance. Assume that there is no dynamic misspecification. Due to misspecification, the KLIC mimimizer is $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_o', (\sigma^*)^2)'$. From Example 9.4 we have from (9.9) that

$$\boldsymbol{H}_T(\boldsymbol{\theta}^*) = \frac{1}{T}\sum_{t=1}^T \begin{bmatrix} -\frac{\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')}{(\sigma^*)^2} & \boldsymbol{0} \\ \boldsymbol{0}' & -\frac{1}{2(\sigma^*)^4} + \frac{\mathbb{E}[h(\boldsymbol{x}_t,\boldsymbol{\alpha}_o)]}{(\sigma^*)^6} \end{bmatrix},$$

and by (9.10),

$$\boldsymbol{B}_T(\boldsymbol{\theta}^*) = \frac{1}{T}\sum_{t=1}^T \begin{bmatrix} \frac{\mathbb{E}[h(\boldsymbol{x}_t,\boldsymbol{\alpha}_o)\boldsymbol{x}_t\boldsymbol{x}_t']}{(\sigma^*)^4} & \boldsymbol{0} \\ \boldsymbol{0}' & \frac{1}{4(\sigma^*)^4} - \frac{\mathbb{E}[h(\boldsymbol{x}_t,\boldsymbol{\alpha}_o)]}{2(\sigma^*)^6} + \frac{3\,\mathbb{E}[h(\boldsymbol{x}_t,\boldsymbol{\alpha}_o)^2]}{4(\sigma^*)^8} \end{bmatrix}.$$

It is easy to verify that the information matrix equality does not hold, despite that the conditional mean function is specified correctly.

In this case, it can be seen that the upper-left block of the estimator $\widetilde{\boldsymbol{H}}_T$ given in Example 9.4, $-\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'/(T\tilde{\sigma}_T^2)$, remains a consistent estimator of the corresponding block in $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$. The information matrix $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$ can be consistently estimated by a block-diagonal matrix with the upper-left block:

$$\frac{\sum_{t=1}^T \hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'}{T(\tilde{\sigma}_T^2)^2};$$

see Section 6.3.1. Then, as far as the estimation of the asymptotic covariance matrix of $T^{1/2}(\tilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)$ is concerned, only the upper-left block of $\widetilde{\boldsymbol{C}}_T = \widetilde{\boldsymbol{H}}_T^{-1}\widetilde{\boldsymbol{B}}_T\widetilde{\boldsymbol{H}}_T^{-1}$ is needed, which reads

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

This is precisely the the Eicker-White estimator (6.8) given in Section 6.3.1, which is consistent when heteroskedasticity is present of unknown form. □

9.4 Hypothesis Testing: Nested Models

In this section we again consider the null hypothesis $\boldsymbol{R}\boldsymbol{\theta}^* = \boldsymbol{r}$, where $\boldsymbol{R}$ is $q \times k$ matrix with full row rank, and discuss three classical large sample tests (Wald, LM, and likelihood ratio tests), and the information matrix test of White (1982, 1987). The postulated hypothesis suggests that the model under the null is obtained by restricting the parameters in a more general model under the alternative. Such models are therefore known as nested models, in the sense that one model is nested in another.

9.4.1 Wald Test

Similar to the Wald test for linear regression discussed in Section 6.4.1, the Wald test under the QMLE framework is based on the difference $\boldsymbol{R}\tilde{\boldsymbol{\theta}}_T - \boldsymbol{r}$. From (9.5),

$$\sqrt{T}\boldsymbol{R}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{B}_T(\boldsymbol{\theta}^*)^{1/2}\left[\boldsymbol{B}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\,\nabla L_T(\boldsymbol{z}^T;\boldsymbol{\theta}^*)\right] + o_{\mathbb{P}}(1).$$

It is then clear that $\boldsymbol{R}\boldsymbol{C}_T(\boldsymbol{\theta}^*)\boldsymbol{R}' = \boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{B}_T(\boldsymbol{\theta}^*)\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'$ is needed to normalize $\sqrt{T}\boldsymbol{R}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*)$. Under the null hypothesis, $\boldsymbol{R}\boldsymbol{\theta}^* = \boldsymbol{r}$, and hence

$$[\boldsymbol{R}\boldsymbol{C}_T(\boldsymbol{\theta}^*)\boldsymbol{R}']^{-1/2}\sqrt{T}(\boldsymbol{R}\tilde{\boldsymbol{\theta}}_T - \boldsymbol{r}) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q). \tag{9.11}$$

This is the key distribution result for the Wald test.

Let $\widetilde{\boldsymbol{H}}_T$ denote a consistent estimator for $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ and $\widetilde{\boldsymbol{B}}_T$ denote a consistent estimator for $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$. It follows that a consistent estimator for $\boldsymbol{C}_T(\boldsymbol{\theta}^*)$ is

$$\widetilde{\boldsymbol{C}}_T = \widetilde{\boldsymbol{H}}_T^{-1} \widetilde{\boldsymbol{B}}_T \widetilde{\boldsymbol{H}}_T^{-1}.$$

Substituting $\widetilde{\boldsymbol{C}}_T$ for $\boldsymbol{C}_T(\boldsymbol{\theta}^*)$ in (9.11) we have

$$[\boldsymbol{R}\widetilde{\boldsymbol{C}}_T\boldsymbol{R}']^{-1/2}\sqrt{T}(\boldsymbol{R}\tilde{\boldsymbol{\theta}}_T - \boldsymbol{r}) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q). \tag{9.12}$$

The Wald test statistic is the inner product of the left-hand side of (9.12):

$$\mathcal{W}_T = T(\boldsymbol{R}\tilde{\boldsymbol{\theta}}_T - \boldsymbol{r})'(\boldsymbol{R}\widetilde{\boldsymbol{C}}_T\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\theta}}_T - \boldsymbol{r}), \tag{9.13}$$

and its asymptotic distribution follows easily from (9.12) and the continuous mapping theorem.

Theorem 9.6 *Suppose that Theorem 9.2 holds for the QMLE $\tilde{\boldsymbol{\theta}}_T$. Then under the null hypothesis,*

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

where $\mathcal{W}_T$ is defined in (9.13) *and q is the number of rows of $\boldsymbol{R}$.*

Example 9.7 Consider the quasi-log-likelihood function specified in Example 9.4. We write $\boldsymbol{\theta} = (\sigma^2\ \boldsymbol{\beta}')'$ and $\boldsymbol{\beta} = (\boldsymbol{b}_1'\ \boldsymbol{b}_2')'$, where $\boldsymbol{b}_1$ is $(k-s) \times 1$, and $\boldsymbol{b}_2$ is $s \times 1$. We are interested in the null hypothesis that $\boldsymbol{b}_2^* = \boldsymbol{R}\boldsymbol{\theta}^* = \boldsymbol{0}$, where $\boldsymbol{R} = [\boldsymbol{0}\ \boldsymbol{R}_1]$ is $s \times (k+1)$ and $\boldsymbol{R}_1 = [\boldsymbol{0}\ \boldsymbol{I}_s]$ is $s \times k$. The Wald test can be computed according to (9.13):

$$\mathcal{W}_T = T\tilde{\boldsymbol{\beta}}_{2,T}'(\boldsymbol{R}\widetilde{\boldsymbol{C}}_T\boldsymbol{R}')^{-1}\tilde{\boldsymbol{\beta}}_{2,T},$$

where $\tilde{\boldsymbol{\beta}}_{2,T} = \boldsymbol{R}\tilde{\boldsymbol{\theta}}_T$ is the estimator of $\boldsymbol{b}_2$.

As shown in Example 9.4, when the information matrix equality holds, $\widetilde{\boldsymbol{C}}_T = -\widetilde{\boldsymbol{H}}_T^{-1}$ is block diagnoal so that

$$\boldsymbol{R}\widetilde{\boldsymbol{C}}_T\boldsymbol{R}' = -\boldsymbol{R}\widetilde{\boldsymbol{H}}_T^{-1}\boldsymbol{R}' = \tilde{\sigma}_T^2\boldsymbol{R}_1(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}_1'.$$

The Wald test then becomes

$$\mathcal{W}_T = T\tilde{\boldsymbol{\beta}}_{2,T}'[\boldsymbol{R}_1(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}_1']^{-1}\tilde{\boldsymbol{\beta}}_{2,T}/\tilde{\sigma}_T^2,$$

which is just s times the standard F statistic, as also shown in Example 6.12. It should be stressed that this simpler version of the Wald test would not be valid if the information matrix equality fails. □

9.4.2 Lagrange Multiplier Test

To derive the LM test, consider the problem of maximizing $L_T(\boldsymbol{\theta})$ subject to the constraint $\boldsymbol{R\theta} = \boldsymbol{r}$. The Lagrangian is

$$L_T(\boldsymbol{\theta}) + \boldsymbol{\theta}'\boldsymbol{R}'\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. The maximizers of the Lagrangian are denoted as $\ddot{\boldsymbol{\theta}}_T$ and $\ddot{\boldsymbol{\lambda}}_T$, where $\ddot{\boldsymbol{\theta}}_T$ is the constrained QMLE of $\boldsymbol{\theta}$. Analogous to Section 6.4.2, the LM test under the QML framework also checks whether $\ddot{\boldsymbol{\lambda}}_T$ is sufficiently close to zero.

First note that $\ddot{\boldsymbol{\theta}}_T$ and $\ddot{\boldsymbol{\lambda}}_T$ satisfy the saddle-point condition:

$$\nabla L_T(\ddot{\boldsymbol{\theta}}_T) + \boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T = \mathbf{0}.$$

The mean-value expansion of $\nabla L_T(\ddot{\boldsymbol{\theta}}_T)$ about $\boldsymbol{\theta}^*$ yields

$$\nabla L_T(\boldsymbol{\theta}^*) + \nabla^2 L_T(\boldsymbol{\theta}_T^\dagger)(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + \boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T = \mathbf{0},$$

where $\boldsymbol{\theta}_T^\dagger$ is the mean value between $\ddot{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}^*$. It has been shown in (9.5) that

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\sqrt{T}\nabla L_T(\boldsymbol{\theta}^*) + o_{\mathrm{IP}}(1).$$

Hence,

$$-\boldsymbol{H}_T(\boldsymbol{\theta}^*)\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + \nabla^2 L_T(\boldsymbol{\theta}_T^\dagger)\sqrt{T}(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + \boldsymbol{R}'\sqrt{T}\ddot{\boldsymbol{\lambda}}_T = o_{\mathrm{IP}}(1).$$

Basing on the WULLN result: $\nabla^2 L_T(\boldsymbol{\theta}_T^\dagger) - \boldsymbol{H}_T(\boldsymbol{\theta}^*) \xrightarrow{\mathrm{IP}} \mathbf{0}$, we obtain

$$\sqrt{T}(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = \sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) - \boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'\sqrt{T}\ddot{\boldsymbol{\lambda}}_T + o_{\mathrm{IP}}(1). \tag{9.14}$$

This establishes a relationship between the constrained and unconstrained QMLEs.

Pre-multiplying both sides of (9.14) by $\boldsymbol{R}$ and noting that the constrained estimator $\ddot{\boldsymbol{\theta}}_T$ must satisfy the constraint so that $\boldsymbol{R}(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = \mathbf{0}$, we have

$$\sqrt{T}\ddot{\boldsymbol{\lambda}}_T = [\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + o_{\mathrm{IP}}(1), \tag{9.15}$$

which relates the Lagrangian multiplier and the unconstrained QMLE $\tilde{\boldsymbol{\theta}}_T$. Let

$$\boldsymbol{\Lambda}_T(\boldsymbol{\theta}^*) = [\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}\boldsymbol{C}_T(\boldsymbol{\theta}^*)\boldsymbol{R}'[\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}']^{-1}.$$

When Theorem 9.2 holds for the normalized $\tilde{\boldsymbol{\theta}}_T$, the following asymptotic normality result for the normalized Lagrangian multiplier is immediate:

$$\begin{aligned}\boldsymbol{\Lambda}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T &= \boldsymbol{\Lambda}_T(\boldsymbol{\theta}^*)^{-1/2}[\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^*)\\ &\xrightarrow{D} \mathcal{N}(\mathbf{0},\boldsymbol{I}_q).\end{aligned} \tag{9.16}$$

Let $\ddot{\boldsymbol{H}}_T$ denote a consistent estimator for $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ and $\ddot{\boldsymbol{C}}_T$ denote a consistent estimator for $\boldsymbol{C}_T(\boldsymbol{\theta}^*)$, both based on the constrained QMLE $\ddot{\boldsymbol{\theta}}_T$. Then,

$$\ddot{\boldsymbol{\Lambda}}_T = (\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\ddot{\boldsymbol{C}}_T\boldsymbol{R}'(\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}')^{-1}$$

is consistent for $\boldsymbol{\Lambda}_T(\boldsymbol{\theta}^*)$. It follows from (9.16) that

$$\ddot{\boldsymbol{\Lambda}}_T^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} \mathcal{N}(\mathbf{0},\boldsymbol{I}_q). \tag{9.17}$$

The LM test statistic is the inner product of the left-hand side of (9.17):

$$\mathcal{LM}_T = T\ddot{\boldsymbol{\lambda}}_T'\ddot{\boldsymbol{\Lambda}}_T^{-1}\ddot{\boldsymbol{\lambda}}_T = T\ddot{\boldsymbol{\lambda}}_T'\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}'(\boldsymbol{R}\ddot{\boldsymbol{C}}_T\boldsymbol{R}')^{-1}\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T. \tag{9.18}$$

The limiting distribution of the LM test now follows easily from (9.17) and the continuous mapping theorem.

Theorem 9.8 *Suppose that Theorem 9.2 holds for the QMLE $\tilde{\boldsymbol{\theta}}_T$. Then under the null hypothesis,*

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

where $\mathcal{LM}_T$ is defined in (9.18) *and q is the number of rows of $\boldsymbol{R}$.*

Remark: From (9.18) we can also write

$$\mathcal{LM}_T = T\nabla L_T(\ddot{\boldsymbol{\theta}}_T)'\ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}'(\boldsymbol{R}\ddot{\boldsymbol{C}}_T\boldsymbol{R}')^{-1}\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\nabla L_T(\ddot{\boldsymbol{\theta}}_T),$$

which is mainly based on the score function ∇L_T evaluated at $\ddot{\boldsymbol{\theta}}_T$. When the information matrix equality holds, the LM statistic further simplifies to

$$\mathcal{LM}_T = -T\ddot{\boldsymbol{\lambda}}_T'\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T = -T\nabla L_T(\ddot{\boldsymbol{\theta}}_T)'\ddot{\boldsymbol{H}}_T^{-1}\nabla L_T(\ddot{\boldsymbol{\theta}}_T).$$

The LM test is thus a test that checks if the average of individual scores is sufficiently close to zero and hence also known as the *score* test.

Example 9.9 Consider the quasi-log-likelihood function specified in Example 9.4. We write $\boldsymbol{\theta} = (\sigma^2\ \boldsymbol{\beta}')'$ and $\boldsymbol{\beta} = (\boldsymbol{b}_1'\ \boldsymbol{b}_2')'$, where $\boldsymbol{b}_1$ is $(k-s)\times 1$, and $\boldsymbol{b}_2$ is $s \times 1$. We are interested in the null hypothesis that $\boldsymbol{b}_2^* = \boldsymbol{R}\boldsymbol{\theta}^* = \boldsymbol{0}$, where $\boldsymbol{R} = [\boldsymbol{0}\ \boldsymbol{R}_1]$ is $s \times (k+1)$ and $\boldsymbol{R}_1 = [\boldsymbol{0}\ \boldsymbol{I}_s]$ is $s \times k$. From the saddle-point condition,

$$\nabla L_T(\ddot{\boldsymbol{\theta}}_T) = -\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T,$$

which can be partitioned as

$$\nabla L_T(\ddot{\boldsymbol{\theta}}_T) = \begin{bmatrix} \nabla_{\sigma^2} L_T(\ddot{\boldsymbol{\theta}}_T) \\ \nabla_{\boldsymbol{b}_1} L_T(\ddot{\boldsymbol{\theta}}_T) \\ \nabla_{\boldsymbol{b}_2} L_T(\ddot{\boldsymbol{\theta}}_T) \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{0} \\ -\ddot{\boldsymbol{\lambda}}_T \end{bmatrix} = -\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T.$$

Thus, the LM test is mainly based on $\nabla_{b_2} L_T(\ddot{\boldsymbol{\theta}}_T)$. Partitioning $\boldsymbol{x}_t$ accordingly as $(\boldsymbol{x}_{1t}'\ \boldsymbol{x}_{2t}')'$, we have

$$\nabla_{\boldsymbol{b}_2} L_T(\ddot{\boldsymbol{\theta}}_T) = \frac{1}{T\,\ddot{\sigma}_T^2}\sum_{t=1}^{T} \boldsymbol{x}_{2t}\ddot{\epsilon}_t = X_2'\ddot{\boldsymbol{\epsilon}}/(T\,\ddot{\sigma}_T^2).$$

where $\ddot{\sigma}_T^2 = \ddot{\boldsymbol{\epsilon}}'\ddot{\boldsymbol{\epsilon}}/T$, and $\ddot{\boldsymbol{\epsilon}}$ is the vector of constrained residuals obtained from regressing y_t on $\boldsymbol{x}_{1t}$ and $\boldsymbol{X}_2$ is the $T \times s$ matrix whose t th row is $\boldsymbol{x}_{2t}'$. The LM test can be computed according to (9.18):

$$\mathcal{LM}_T = T\begin{bmatrix} 0 \\ \boldsymbol{0} \\ X_2'\ddot{\boldsymbol{\epsilon}}/(T\,\ddot{\sigma}_T^2) \end{bmatrix}' \ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}'(\boldsymbol{R}\ddot{\boldsymbol{C}}_T\boldsymbol{R}')^{-1}\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\begin{bmatrix} 0 \\ \boldsymbol{0} \\ X_2'\ddot{\boldsymbol{\epsilon}}/(T\,\ddot{\sigma}_T^2) \end{bmatrix},$$

which converges in distribution to $\chi^2(s)$ under the null hypothesis. Note that we do not have to evaluate the complete score vector for computing the LM test; only the subvector of the score that corresponds to the constraint really matters.

When the information matrix equality holds, the LM statistic has a simpler form:

$$\begin{aligned} \mathcal{LM}_T &= -T\nabla L_T(\ddot{\boldsymbol{\theta}}_T)'\ddot{\boldsymbol{H}}_T^{-1}\nabla L_T(\ddot{\boldsymbol{\theta}}_T) \\ &= T[\boldsymbol{0}'\ \ddot{\boldsymbol{\epsilon}}'\boldsymbol{X}_2/T](\boldsymbol{X}'\boldsymbol{X}/T)^{-1}[\boldsymbol{0}'\ \ddot{\boldsymbol{\epsilon}}'\boldsymbol{X}_2/T]'/\ddot{\sigma}_T^2 \\ &= T[\ddot{\boldsymbol{\epsilon}}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ddot{\boldsymbol{\epsilon}}/\ddot{\boldsymbol{\epsilon}}'\ddot{\boldsymbol{\epsilon}}] \\ &= TR^2, \end{aligned}$$

where R^2 is the non-centered coefficient of determination obtained from the auxiliary regression of the constrained residuals $\ddot{\epsilon}_t$ on $\boldsymbol{x}_{1t}$ and $\boldsymbol{x}_{2t}$. This is also the result obtained in Example 6.14. □

Example 9.10 (Breusch-Pagan) Suppose that the specification is

$$y_t|\boldsymbol{x}_t, \boldsymbol{\zeta}_t \sim \mathcal{N}\big(\boldsymbol{x}_t'\boldsymbol{\beta}, h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})\big),$$

where $h\colon \mathbb{R} \to (0,\infty)$ is a differentiable function, and $\boldsymbol{\zeta}_t'\boldsymbol{\alpha} = \alpha_0 + \sum_{i=1}^{p} \zeta_{ti}\alpha_i$. The null hypothesis is conditional homoskedasticity, i.e., $\alpha_1 = \cdots = \alpha_p = 0$ so that $h(\alpha_0) = \sigma_0^2$. Breusch and Pagan (1979) derived the LM test for this hypothesis under the assumption that the information matrix equality holds. This test is now usually referred to as the Breusch-Pagan test.

Note that the constrained specification is $y_t|\boldsymbol{x}_t, \boldsymbol{\zeta}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$, where $\sigma^2 = h(\alpha_0)$. This leads to the standard linear regression model without heteroskedasticity. The constrained QMLEs for $\boldsymbol{\beta}$ and σ^2 are, respectively, the OLS estimators $\hat{\boldsymbol{\beta}}_T$ and $\ddot{\sigma}_T^2 = \sum_{t=1}^{T} \hat{e}_t^2/T$, where $\hat{e}_t$ are the OLS residuals. As in Example 9.9, we evaluate the score vector corresponding to $\boldsymbol{\alpha}$:

$$\nabla_\alpha L_T(y_t, \boldsymbol{x}_t, \boldsymbol{\zeta}_t; \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{h_1(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})\boldsymbol{\zeta}_t}{2h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})}\left(\frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})} - 1\right)\right],$$

where $h_1(\eta) = \mathrm{d}h(\eta)/\,\mathrm{d}\eta$. Under the null hypothesis, $h_1(\boldsymbol{\zeta}_t'\boldsymbol{\alpha}) = h_1(\alpha_0)$ is just a constant and will be denoted as c. The score vector above evaluated at the constrained QMLEs is

$$\nabla_\alpha L_T(y_t, \boldsymbol{x}_t, \boldsymbol{\zeta}_t; \ddot{\boldsymbol{\theta}}_T) = \frac{c}{T}\sum_{t=1}^{T}\left[\frac{\boldsymbol{\zeta}_t}{2\ddot{\sigma}_T^2}\left(\frac{\hat{e}_t^2}{\ddot{\sigma}_T^2} - 1\right)\right].$$

It can be shown that the $(p+1)\times(p+1)$ block of the Hessian matrix corresponding to $\boldsymbol{\alpha}$ is

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}&\left[\frac{-(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{h^3(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})} + \frac{1}{2h^2(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})}\right][h_1(\boldsymbol{\zeta}'\boldsymbol{\alpha})]^2\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t' \\ &+ \left[\frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2h^2(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})} - \frac{1}{2h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})}\right]h_2(\boldsymbol{\zeta}'\boldsymbol{\alpha})\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t',\end{aligned}$$

where $h_2(\eta) = \,\mathrm{d}h_1(\eta)/\,\mathrm{d}\eta$. Evaluating the expectation of this block at $\boldsymbol{\theta}_o = (\boldsymbol{\beta}_o'\ \alpha_0\ \mathbf{0}')'$ and noting that $\sigma_o^2 = h(\alpha_0)$ we have

$$-\left(\frac{c^2}{2[(\sigma_o^2)^2}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{I\!E}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t')\right),$$

which can be consistently estimated by

$$-\left(\frac{c^2}{2(\ddot{\sigma}_T^2)^2}\right)\left(\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t')\right).$$

The LM test is now readily derived from these results when the information matrix equality holds.

Setting $d_t = \hat{e}_t^2/\ddot{\sigma}_T^2 - 1$, the LM statistic can be expressed as

$$\mathcal{LM}_T = \frac{1}{2}\left(\sum_{t=1}^{T} d_t \boldsymbol{\zeta}_t'\right)\left(\sum_{t=1}^{T} \boldsymbol{\zeta}_t \boldsymbol{\zeta}_t'\right)^{-1}\left(\sum_{t=1}^{T} \boldsymbol{\zeta}_t d_t\right) \xrightarrow{D} \chi^2(p).$$

A novel feature of this statistic is that neither c nor the functional form of h matters in the statistic. Hence, the Breusch-Pagan test is capable of testing for general conditional heteroskedasticity. In terms of test computation, it is easily seen that the numerator of this statistic is the (centered) regression sum of squares (RSS) of regressing d_t on $\boldsymbol{\zeta}_t$. As such, the Breusch-Pagan test can be conveniently computed by running an auxiliary regression and using the resulting RSS/2 as the statistic. Intuitively, this amounts to checking whether the variables in $\boldsymbol{\zeta}_t$ are able to explain the square of the (standardized) OLS residuals.

Given conditional normality, Koenker (1981) noted that $T^{-1}\sum_{t=1}^{T} \hat{e}_t^4 \xrightarrow{\mathbb{P}} 3(\sigma_o^2)^2$ under the null hypothesis, so that

$$\frac{1}{T}\sum_{t=1}^{T} d_t^2 = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{e}_t^4}{(\ddot{\sigma}_T^2)^2} - 1 \xrightarrow{\mathbb{P}} 2.$$

Thus, a test that is asymptotically equivalent to the original Breusch-Pagan test is to replace the denominator 2 in the statistic with $\sum_{t=1}^{T} d_t^2/T$. That is,

$$\mathcal{LM}_T = T\left(\sum_{t=1}^{T} d_t \boldsymbol{\zeta}_t'\right)\left(\sum_{t=1}^{T} \boldsymbol{\zeta}_t \boldsymbol{\zeta}_t'\right)^{-1}\left(\sum_{t=1}^{T} \boldsymbol{\zeta}_t d_t\right) \Big/ \sum_{t=1}^{T} d_t^2,$$

which can be computed as TR^2, with R^2 obtained from the regression of d_t on $\boldsymbol{\zeta}_t$. As $\sum_{i=1}^{T} d_i = 0$, the centered and non-centered R^2 are in fact equivalent. This test is also equivalent to TR^2 with the centered R^2 computed from regressing $\hat{e}_t^2$ on $\boldsymbol{\zeta}_t$. □

Remarks:

1. To compute the Breusch-Pagan test, one must specify a vector $\boldsymbol{\zeta}_t$ that determines the conditional variance. Here, $\boldsymbol{\zeta}_t$ may contain some or all the variables in $\boldsymbol{x}_t$. If $\boldsymbol{\zeta}_t$ is chosen to include all elements of $\boldsymbol{x}_t$, their squares and pairwise products, the resulting TR^2 is also the White (1980) test for (conditional) heteroskedasticity of unknown form. The White test can also be interpreted as an "information matrix test" discussed below.

2. The Breusch-Pagan test is obtained under the condition that the information matrix equality holds. We have seen that the information matrix equality may fail when there is dynamic misspecification. Thus, the Breusch-Pagan test is not valid when, e.g., the errors are serially correlated. For the LM test for conditional heteroskedasticity under dynamic misspecification, see Exercise 9.6.

Example 9.11 (Breusch-Godfrey) Given the specification $y_t|\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$, suppose that one would like to check if the errors $y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$ are serially correlated. Consider first the AR(1) error: $y_t - \boldsymbol{x}_t'\boldsymbol{\beta} = \rho(y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta}) + u_t$ with $|\rho| < 1$ and $\{u_t\}$ a white noise. The null hypothesis is $\rho^* = 0$, i.e., no serial correlation. Instead of deriving the LM test formally, we treat this as a specification with possibly omitted variables, as in Example 9.9. To this end, consider a general specification that allows for serial correlations:

$$y_t|y_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t-1} \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta} + \rho(y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta}), \sigma_u^2),$$

which reduces to the original specification when $\rho = 0$. Thus, the constrained specification is the standard linear regression model $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}$, and the constrained QMLE of $\boldsymbol{\beta}$ is the OLS estimator $\hat{\boldsymbol{\beta}}_T$. Testing the null hypothesis that $\rho^* = 0$ now amounts to testing whether an additional variable $y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta}$ should be included in the mean specification.

When the information matrix equality holds, an LM test can be obtained from an auxiliary regression of the OLS residuals $\hat{e}_t = y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T$ on $\boldsymbol{x}_t$ and $y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta}$. Replacing $\boldsymbol{\beta}$ with its constrained estimator $\hat{\boldsymbol{\beta}}_T$, an LM test is TR^2, with R^2 computed from the regression of $\hat{e}_t$ on $\boldsymbol{x}_t$ and $\hat{e}_{t-1}$, and has the limiting $\chi^2(1)$ distribution under the null hypothesis. This is precisely the Breusch (1978) and Godfrey (1978) test for AR(1) errors. More formally, we can derive the Breusch-Godfrey test along the line discussed in this section; see Exercise 9.7. The Breusch-Godfrey test can be extended straightforwardly to check if the errors follow an AR(p) process. By regressing $\hat{e}_t$ on $\boldsymbol{x}_t$ and $\hat{e}_{t-1}, \ldots, \hat{e}_{t-p}$, the resulting TR^2 is the LM test when the information matrix equality holds and has a limiting $\chi^2(p)$ distribution.

Moreover, if the specification is $y_t - \boldsymbol{x}_t'\boldsymbol{\beta} = u_t + \alpha u_{t-1}$, i.e., the errors follow an MA(1) process, we can write

$$y_t|\boldsymbol{x}_t, u_{t-1} \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta} + \alpha u_{t-1}, \sigma_u^2).$$

The null hypothesis is $\alpha^* = 0$. Again, the constrained specification is the standard linear regression model $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}$, and the constrained QMLE of $\boldsymbol{\beta}$ is still the OLS estimator

$\hat{\boldsymbol{\beta}}_T$. It follows that the LM test of $\alpha^* = 0$ can be computed as TR^2 with R^2 obtained from the regression of $\hat{u}_t = y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T$ on $\boldsymbol{x}_t$ and $\hat{u}_{t-1}$. This is identical to the LM test for AR(1) errors. Similarly, the Breusch-Godfrey test for MA(p) errors is also the same as that for AR(p) errors. □

Remarks:

1. The Breusch-Godfrey tests are obtained under the condition that the information matrix equality holds. If there is neglected conditional heteroskedasticity, the information matrix equality would fail, so that the Breusch-Godfrey tests no longer have a limiting χ^2 distribution.

2. It can be shown that the square of Durbin's h test is also an LM test. While Durbin's h test may not be feasible in practice, the Breusch-Godfrey test can always be computed.

9.4.3 Likelihood Ratio Test

As discussed in Section 6.4.3, the LR test is based on the comparison between the constrained and unconstrained specifications in terms of their log-likelihoods:

$$\mathcal{LR}_T = -2T[L_T(\ddot{\boldsymbol{\theta}}_T) - L_T(\tilde{\boldsymbol{\theta}}_T)]. \qquad (9.19)$$

Clearly, this test depends on the estimation results of both constrained and unconstrained specification.

Recall that (9.14) gives a relation between the constrained and unconstrained QMLEs which implies

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \ddot{\boldsymbol{\theta}}_T) = \boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'\sqrt{T}\ddot{\boldsymbol{\lambda}}_T + o_{\mathbb{P}}(1).$$

By (9.15), we obtain a relation between the Lagrangian multiplier and unconstrained QMLE:

$$\sqrt{T}\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T = \boldsymbol{R}'[\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1).$$

It follows that

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \ddot{\boldsymbol{\theta}}_T) = \boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'[\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1). \qquad (9.20)$$

By Taylor expansion of $L_T(\ddot{\boldsymbol{\theta}}_T)$ about $\tilde{\boldsymbol{\theta}}_T$, we have

$$\begin{aligned}
&-2T\big[L_T(\ddot{\boldsymbol{\theta}}_T) - L_T(\tilde{\boldsymbol{\theta}}_T)\big] \\
&\qquad = -2T\nabla L_T(\tilde{\boldsymbol{\theta}}_T)(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T) - T(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T)'\boldsymbol{H}_T(\tilde{\boldsymbol{\theta}}_T)(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T) + o_{\mathbb{P}}(1) \\
&\qquad = -T(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T)'\boldsymbol{H}_T(\boldsymbol{\theta}^*)(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T) + o_{\mathbb{P}}(1),
\end{aligned}$$

because $\nabla L_T(\tilde{\boldsymbol{\theta}}_T) = \mathbf{0}$. Using (9.20) we have

$$-2T[L_T(\ddot{\boldsymbol{\theta}}_T) - L_T(\tilde{\boldsymbol{\theta}}_T)] = -T(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*)'\boldsymbol{R}'[\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1),$$

where the right-hand side is essentially the Wald statistic with the normalizing variance-covariance matrix $-\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'$. This shows that the LR test would have a limiting χ^2 distribution when the information matrix equality holds.

Theorem 9.12 *Given that Theorem 9.2 holds for the QMLE $\tilde{\boldsymbol{\theta}}_T$, suppose that the information matrix equality also holds. Then under the null hypothesis,*

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

where $\mathcal{LR}_T$ is defined in (9.19) *and q is the number of rows of $\boldsymbol{R}$.*

Theorem 9.12 differs from Theorem 9.6 and Theorem 9.8 in that it also requires the validity of the information matrix equality. When the information matrix equality fails to hold, $-\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'$ is not a proper normalizing matrix so that $\mathcal{LR}_T$ does not have a limiting χ^2 distribution. In other words, the LR test given in (9.19) can not be made robust to the failure of the information matrix equality. This should not be too surprising because L_T is constructed by specifying likelihood functions for $\{y_t|x_t\}$ without considering possible dynamic relations between y_t and the past information not contained in $\boldsymbol{x}_t$. By contrast, the Wald and LM tests can be made robust by employing a proper estimator of the asymptotic variance-covariance matrix.

9.5 Hypothesis Testing: Non-Nested Models

In this section we consider the problem of testing non-nested specifications under the null and alternative hypotheses:

$$\begin{aligned}
&H_0\colon\ y_t|\boldsymbol{x}_t,\boldsymbol{\xi}_t \sim f(y_t|\boldsymbol{x}_t;\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p, \\
&H_1\colon\ y_t|\boldsymbol{x}_t,\boldsymbol{\xi}_t \sim \varphi(y_t|\boldsymbol{\xi}_t;\boldsymbol{\psi}), \quad \boldsymbol{\psi} \in \Psi \subseteq \mathbb{R}^q,
\end{aligned}$$

where $\boldsymbol{x}_t$ and $\boldsymbol{\xi}_t$ are two sets of variables. These two specifications are non-nested in the sense that one can not be derived from the other by imposing restrictions on the parameters. Thus, the tests discussed in the preceding section do not apply. A leading approach to testing non-nested specification is based on the *encompassing principle* of Mizon (1984) and Mizon and Richard (1986). This principle asserts that, if the model under the null hypothesis is true, it should encompass the model under the alternative, such that a statistic of the alternative model should be close to its *pseudo-true value*, the probability limit evaluated under the null model. An *encompassing test* for non-nested hypotheses is then based on the difference between a chosen statistic and the sample counterpart of its pseudo-true value.

9.5.1 Wald Encompassing Test

When the chosen statistic is the QMLE of the alternative specification, the resulting encompassing test is known as the *parameter encompassing test.* This test is analogous to the Wald test for nested models and also known as the *Wald encompassing test* (WET).

To illustrate the WET, we specialize on the following non-nested specifications of the conditional mean function:

$$H_0\colon\ y_t|\boldsymbol{x}_t,\boldsymbol{\xi}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta},\sigma^2), \quad \boldsymbol{\beta}\in\mathcal{B}\subseteq\mathbb{R}^k,$$
$$H_1\colon\ y_t|\boldsymbol{x}_t,\boldsymbol{\xi}_t \sim \mathcal{N}(\boldsymbol{\xi}_t'\boldsymbol{\delta},\sigma^2), \quad \boldsymbol{\delta}\in\mathcal{D}\subseteq\mathbb{R}^r,$$

where $\boldsymbol{x}_t$ and $\boldsymbol{\xi}_t$ do not have elements in common. Let $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{\delta}}_T$ denote the QMLEs of the parameters in the null and alternative models, respectively. Taking $\hat{\boldsymbol{\delta}}_T$ as the statistic for an encompassing test, we need to evaluate its probability limit under the null hypothesis. When the null hypothesis is a correct specification, $y_t|\boldsymbol{x}_t,\boldsymbol{\xi}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}_o,\sigma_o^2)$, so that $\mathbb{E}(\boldsymbol{\xi}_t y_t) = \mathbb{E}(\boldsymbol{\xi}_t\boldsymbol{x}_t')\boldsymbol{\beta}_o$. Hence, with a suitable LLN,

$$\hat{\boldsymbol{\delta}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\boldsymbol{\xi}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t y_t\right) \xrightarrow{\mathbb{P}} \boldsymbol{M}_{\xi\xi}^{-1}\boldsymbol{M}_{\xi x}\boldsymbol{\beta}_o,$$

where $\boldsymbol{M}_{\xi\xi} = \lim_{T\to\infty}\sum_{t=1}^{T}\mathbb{E}(\boldsymbol{\xi}_t\boldsymbol{\xi}_t')/T$ and $\boldsymbol{M}_{\xi x} = \lim_{T\to\infty}\sum_{t=1}^{T}\mathbb{E}(\boldsymbol{\xi}_t\boldsymbol{x}_t')/T$. This probability limit, denoted as $\boldsymbol{\delta}(\boldsymbol{\beta}_o)$, is the pseudo-true value of $\hat{\boldsymbol{\delta}}_T$ and usually referred to as the *pseudo-true parameter.* It is clear that $\boldsymbol{\delta}(\boldsymbol{\beta}_o)$ would not be the probability limit of $\hat{\boldsymbol{\delta}}_T$ if $\boldsymbol{x}_t'\boldsymbol{\beta}$ is an incorrect specification of the conditional mean. Thus, whether $\hat{\boldsymbol{\delta}}_T - \boldsymbol{\delta}(\boldsymbol{\beta}_o)$ is sufficiently close to zero constitutes an evidence for or against the null hypothesis. Note, however, that neither $\boldsymbol{\delta}(\boldsymbol{\beta}_o)$ nor $\boldsymbol{\delta}(\hat{\boldsymbol{\beta}}_T)$ is observable.

The discussion above suggests that an encompassing test can be based on the difference between $\hat{\boldsymbol{\delta}}_T$ and the sample counterpart of $\boldsymbol{\delta}(\hat{\boldsymbol{\beta}}_T)$:

$$\hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T) = \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t \boldsymbol{\xi}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t \boldsymbol{x}_t'\right) \hat{\boldsymbol{\beta}}_T.$$

In particular,

$$\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T) = \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t \boldsymbol{\xi}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t (y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T)\right).$$

Let $\epsilon_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_o$ and $\hat{e}_t = y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T$. Therefore, a test based on $\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)$ can also be interpreted as a test of the correlation between the errors of the null model, ϵ_t, and the regressors of the alternative model, $\boldsymbol{\xi}_t$. A large value of such correlation would indicate that there is still information in $\boldsymbol{\xi}_t$ that is not captured by the null model.

To find the limiting distribution of $\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)$, observe that

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{\xi}_t \hat{e}_t &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{\xi}_t \epsilon_t - \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t \boldsymbol{x}_t'\right) \sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) \\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{\xi}_t \epsilon_t - \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t \boldsymbol{x}_t'\right) \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t'\right)^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{x}_t \epsilon_t\right).\end{aligned}$$

Setting $\hat{\boldsymbol{\xi}}_t = \boldsymbol{M}_{\xi x}\boldsymbol{M}_{xx}^{-1}\boldsymbol{x}_t$, we can write

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{\xi}_t \hat{e}_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} (\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_t)\epsilon_t + o_{\mathbb{P}}(1).$$

With suitable LLN and CLT, we have under the null hypothesis that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{\xi}_t \hat{e}_t \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_o),$$

where

$$\boldsymbol{\Sigma}_o = \sigma_o^2 \left(\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_t)(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_t)'\big]\right) = \sigma_o^2\big(\boldsymbol{M}_{\xi\xi} - \boldsymbol{M}_{\xi x}\boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{x\xi}\big).$$

Consequently, $T^{1/2}[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)] \xrightarrow{D} \mathcal{N}\big(\mathbf{0}, \boldsymbol{M}_{\xi\xi}^{-1}\boldsymbol{\Sigma}_o\boldsymbol{M}_{\xi\xi}^{-1}\big)$, and hence

$$T\big[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\big]'\boldsymbol{M}_{\xi\xi}\boldsymbol{\Sigma}_o^{-1}\boldsymbol{M}_{\xi\xi}\big[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\big] \xrightarrow{D} \chi^2(r).$$

Replacing $\boldsymbol{M}_{\xi\xi}$, $\boldsymbol{M}_{\xi x}$ and $\boldsymbol{M}_{xx}$ with their sample counterparts and replacing σ_o^2 with $\hat{\sigma}_T^2 = \sum_{t=1}^T \hat{e}_t^2/T$, we obtain a consistent estimator for $\boldsymbol{\Sigma}_o$:

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

The WET statistic reads

$$\mathcal{WE}_T = T\big[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\big]' \left(\frac{1}{T} \sum_{t=1}^T \boldsymbol{\xi}_t \boldsymbol{\xi}_t' \right) \widehat{\boldsymbol{\Sigma}}_T^{-1} \left(\frac{1}{T} \sum_{t=1}^T \boldsymbol{\xi}_t \boldsymbol{\xi}_t' \right) \big[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\big],$$

which has the limiting $\chi^2(r)$ distribution under the null hypothesis.

When $\boldsymbol{x}_t$ and $\boldsymbol{\xi}_t$ have s ($s < r$) elements in common, $\sum_{t=1}^T \boldsymbol{\xi}_t \hat{e}_t$ must have s elements that are identically zero. Hence, $\operatorname{rank}(\boldsymbol{\Sigma}_o) = r^* \leq r - s$. In this case,

$$T\big[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\big]' \boldsymbol{M}_{\xi\xi} \boldsymbol{\Sigma}_o^{-} \boldsymbol{M}_{\xi\xi} \big[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\big] \xrightarrow{D} \chi^2(r^*),$$

where $\boldsymbol{\Sigma}_o^{-}$ denotes the generalized inverse of $\boldsymbol{\Sigma}_o$, and the WET can be computed as above but with the generalized inverse of $\widehat{\boldsymbol{\Sigma}}_T$.

Under H_1: $y_t|\boldsymbol{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}(\boldsymbol{\xi}_t'\boldsymbol{\delta}, \sigma^2)$, the score function evaluated at the pseudo-true parameter $\boldsymbol{\delta}(\boldsymbol{\beta}_o)$ is (apart from a constant σ^{-2})

$$\frac{1}{T} \sum_{t=1}^T \boldsymbol{\xi}_t [y_t - \boldsymbol{\xi}_t' \boldsymbol{\delta}(\boldsymbol{\beta}_o)] = \frac{1}{T} \sum_{t=1}^T \boldsymbol{\xi}_t \big[y_t - \boldsymbol{\xi}_t' (\boldsymbol{M}_{\xi\xi}^{-1} \boldsymbol{M}_{\xi x} \boldsymbol{\beta}_o)\big].$$

When the pseudo-true parameter is replaced by its estimator $\hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)$, the score function becomes

$$\frac{1}{T} \sum_{t=1}^T \boldsymbol{\xi}_t (y_t - \boldsymbol{x}_t \hat{\boldsymbol{\beta}}_T) = \frac{1}{T} \sum_{t=1}^T \boldsymbol{\xi}_t \hat{e}_t.$$

The so-called score encompassing test (SET) is based on the normalized score and also amounts to checking the correlation between the error of the null model and the regressor of the alternative model; see Exercise 9.10. Note that the SET and WET are based on the same ingredient and require estimation of the pseud-true parameter.

On the other hand, the applicability of the WET and SET may be quite limited in practice. From the discussion above, it is clear that a crucial step in deriving a WET or a SET is to evaluate the pseudo-true value of a parameter estimator. This is straightforward in the present example because the estimator $\hat{\delta}_T$ has an analytic form. There are, however, numerous examples in which a QMLE must be solved numerically. As examples, consider the following specifications: (1) conditional normality with nonlinear specifications of the conditional mean function:

$$\begin{aligned} &H_0\colon\ y_t|\boldsymbol{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}\big(m(\boldsymbol{x}_t, \boldsymbol{\beta}), \sigma^2\big), \quad \boldsymbol{\beta} \in \mathcal{B} \subseteq \mathbb{R}^k, \\ &H_1\colon\ y_t|\boldsymbol{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}\big(\mu(\boldsymbol{\xi}_t, \boldsymbol{\delta}), \sigma^2\big), \quad \boldsymbol{\delta} \in \mathcal{D} \subseteq \mathbb{R}^r; \end{aligned}$$

(2) conditional normality with specifications for both the conditional mean and conditional variance functions:

$$\begin{aligned} &H_0\colon\ y_t|\boldsymbol{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}\big(\boldsymbol{x}_t'\boldsymbol{\beta}, h(\boldsymbol{x}_t, \boldsymbol{\alpha})\big), \quad \boldsymbol{\theta} = (\boldsymbol{\beta}'\ \boldsymbol{\alpha}')' \in \Theta \subseteq \mathbb{R}^p, \\ &H_1\colon\ y_t|\boldsymbol{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}\big(\boldsymbol{\xi}_t'\boldsymbol{\delta}, \kappa(\boldsymbol{\xi}_t, \boldsymbol{\gamma})\big), \quad \boldsymbol{\psi} = (\boldsymbol{\delta}'\ \boldsymbol{\gamma}')' \in \Psi \subseteq \mathbb{R}^q. \end{aligned}$$

In these cases, the QMLEs do not have closed forms, and it would be practically difficult, if not impossible, to evaluate the pseudo-true parameter. Consequently, computing a WET or a SET may not be feasible.

9.5.2 Pseudo-True Score Encompassing Test

Instead of testing parameter encompassing, Chen and Kuan (2002) considered the pseudo-true value of the score function under the alternative hypothesis and proposed the pseudo-true score encompassing (PSE) test. An advantage of basing an encompassing test on the score function is that, while a QMLE may not have a closed form, the score function usually does. It is therefore easier to derive and estimate the pseudo-true score function in practice.

Let $s_{f,t}(\boldsymbol{\theta}) = \nabla \log f(y_t|\boldsymbol{x}_t; \boldsymbol{\theta})$ and $s_{\varphi,t}(\boldsymbol{\psi}) = \nabla \log \varphi(y_t|\boldsymbol{\xi}_t; \boldsymbol{\psi})$ be individual score functions under the null and alternative hypotheses, respectively. Also let

$$\nabla L_{f,T}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} s_{f,t}(\boldsymbol{\theta}), \qquad \nabla L_{\varphi,T}(\boldsymbol{\psi}) = \frac{1}{T}\sum_{t=1}^{T} s_{\varphi,t}(\boldsymbol{\psi}).$$

The pseudo-true score function of $\nabla L_{\varphi,T}(\boldsymbol{\psi})$ is

$$J_\varphi(\boldsymbol{\theta}, \boldsymbol{\psi}) = \lim_{T\to\infty} \mathbb{E}_{f(\boldsymbol{\theta})}\big[\nabla L_{\varphi,T}(\boldsymbol{\psi})\big],$$

where $\mathbb{E}_{f(\boldsymbol{\theta})}$ denote the expectation that takes into account the null hypothesis. Given the specification of φ, the pseudo-true parameter $\boldsymbol{\psi}(\boldsymbol{\theta}_o)$ is the KLIC minimizer when the null hypothesis is specified correctly, and it must solve $J_\varphi(\boldsymbol{\theta}_o, \boldsymbol{\psi}) = \mathbf{0}$, i.e.,

$$J_\varphi(\boldsymbol{\theta}_o, \boldsymbol{\psi}(\boldsymbol{\theta}_o)) = \mathbf{0}.$$

Thus, whether $J_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T)$ is close to zero constitutes an evidence for or against the null hypothesis. The PSE test is based on the sample counterpart of $J_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T)$. While the

SET relies on the score function evaluated at the pseudo-true parameter and is in effect a test of parameter encompassing, the PSE test is truly a test of score encompassing because the null model has been directly taken into account to derive the pseudo-true score function.

Following Wooldridge (1990), we can incorporate the null model into the the score function under the alternative hypothesis and write

$$\nabla L_{\varphi,T}(\boldsymbol{\psi}) = \frac{1}{T}\sum_{t=1}^{T} d_{1,t}(\boldsymbol{\theta},\boldsymbol{\psi}) + \frac{1}{T}\sum_{t=1}^{T} d_{2,t}(\boldsymbol{\theta},\boldsymbol{\psi}) c_t(\boldsymbol{\theta}),$$

where $d_{1,t}$ and $d_{2,t}$ both depend on $\boldsymbol{x}_t$ and $\boldsymbol{\xi}_t$ and $\mathbb{E}_{f(\boldsymbol{\theta})}[c_t(\boldsymbol{\theta})|\boldsymbol{x}_t,\boldsymbol{\xi}_t] = \mathbf{0}$. As such,

$$J_\varphi(\boldsymbol{\theta},\boldsymbol{\psi}) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{f(\boldsymbol{\theta})}\left[d_{1,t}(\boldsymbol{\theta},\boldsymbol{\psi})\right],$$

and its sample counterpart is

$$\widehat{J}_\varphi(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T) = \frac{1}{T}\sum_{t=1}^{T} d_{1,t}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T) = -\frac{1}{T}\sum_{t=1}^{T} d_{2,t}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T) c_t(\hat{\boldsymbol{\theta}}_T),$$

where the second equality follows because $\nabla L_{\varphi,T}(\hat{\boldsymbol{\psi}}_T) = \mathbf{0}$. Thus, the PSE test depends on the QMLEs under the null and alternative hypotheses but not on the sample counterpart of the pseudo-true parameter.

As an example, consider the non-nested specifications in Section 9.5.1. By incorporating the null model into the score function under the alternative hypothesis we obtain

$$\nabla_{\boldsymbol{\delta}} L_{\varphi,T}(\boldsymbol{\psi}) = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t(\boldsymbol{x}_t'\boldsymbol{\beta} - \boldsymbol{\xi}_t'\boldsymbol{\delta}) + \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\varepsilon_t$$

with $d_{1,t}(\boldsymbol{\theta},\boldsymbol{\psi}) = \boldsymbol{\xi}_t(\boldsymbol{x}_t'\boldsymbol{\beta} - \boldsymbol{\xi}_t'\boldsymbol{\delta})$, $d_{2,t}(\boldsymbol{\theta},\boldsymbol{\psi}) = \boldsymbol{\xi}_t$, and $c_t(\boldsymbol{\theta}) = \varepsilon_t$. In this case, the PSE test is based on

$$\widehat{J}_\varphi(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T) = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t(\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T - \boldsymbol{\xi}_t'\hat{\boldsymbol{\delta}}_T) = -\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\hat{e}_t,$$

as in the SET discussed earlier. On the other hand, when the conditional mean specifications are nonlinear with $m(\boldsymbol{x}_t,\boldsymbol{\beta})$ under the null and $\mu(\boldsymbol{\xi}_t,\boldsymbol{\delta})$ under the alternative, the SET is not readily available because of the difficulty of deriving the pseudo-true parameter. Yet, analogous to the result for linear specifications, the PSE test can be based on

$$\frac{1}{T}\sum_{t=1}^{T}\nabla_{\boldsymbol{\delta}}\mu(\boldsymbol{\xi}_t,\hat{\boldsymbol{\delta}}_T)\big[m(\boldsymbol{x}_t,\hat{\boldsymbol{\beta}}_T)-\mu(\boldsymbol{\xi}_t,\hat{\boldsymbol{\delta}}_T)\big]=-\frac{1}{T}\sum_{t=1}^{T}\nabla_{\boldsymbol{\delta}}\mu(\boldsymbol{\xi}_t,\hat{\boldsymbol{\delta}}_T)\hat{e}_t,$$

with $\hat{e}_t = y_t - m(\boldsymbol{x}_t,\hat{\boldsymbol{\beta}}_T)$ the nonlinear OLS residuals under the null model.

It can be verified that the linear expansion of $T^{1/2}\widehat{J}_{\varphi}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T)$ about $(\boldsymbol{\theta}_o,\boldsymbol{\psi}(\boldsymbol{\theta}_o))$ is

$$\sqrt{T}\widehat{J}_{\varphi}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T)=-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}d_{2,t}(\boldsymbol{\theta}_o,\boldsymbol{\psi}(\boldsymbol{\theta}_o))c_t(\boldsymbol{\theta}_o)-\boldsymbol{A}_o\sqrt{T}(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_o)+o_{\mathbb{P}}(1),$$

where $\boldsymbol{A}_o = \lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathbb{E}_{f(\boldsymbol{\theta}_o)}\big[d_{2,t}(\boldsymbol{\theta}_o,\boldsymbol{\psi}(\boldsymbol{\theta}_o))\nabla_{\boldsymbol{\theta}}c_t(\boldsymbol{\theta}_o)\big]$. Note that the other terms in the expansion that involve c_t would vanish in the limit because they have zero mean. Recall also that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_o)=-H_T(\boldsymbol{\theta}_o)^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{s}_{f,t}(\boldsymbol{\theta}_o)+o_{\mathbb{P}}(1),$$

where $\mathbb{E}_{f(\boldsymbol{\theta}_o)}[s_{f,t}(\boldsymbol{\theta}_o)|\boldsymbol{x}_t,\boldsymbol{\xi}_t]=\mathbf{0}$. Collecting terms we have

$$\sqrt{T}\widehat{J}_{\varphi}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T)=-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{b}_t(\boldsymbol{\theta}_o,\boldsymbol{\psi}(\boldsymbol{\theta}_o))+o_{\mathbb{P}}(1),$$

where $\boldsymbol{b}_t(\boldsymbol{\theta}_o,\boldsymbol{\psi}(\boldsymbol{\theta}_o)) = d_{2,t}(\boldsymbol{\theta}_o,\boldsymbol{\psi}(\boldsymbol{\theta}_o))c_t(\boldsymbol{\theta}_o) - \boldsymbol{A}_o\boldsymbol{H}_T(\boldsymbol{\theta}_o)^{-1}\boldsymbol{s}_{f,t}(\boldsymbol{\theta}_o)$ is such that

$$\mathbb{E}_{f(\boldsymbol{\theta}_o)}[\boldsymbol{b}_t(\boldsymbol{\theta}_o,\boldsymbol{\psi}(\boldsymbol{\theta}_o))|\boldsymbol{x}_t,\boldsymbol{\xi}_t]=\mathbf{0}.$$

By invoking a suitable CLT, $T^{1/2}\widehat{J}_{\varphi}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T)$ has a limiting normal distribution with the asymptotic covariance matrix:

$$\boldsymbol{\Sigma}_o=\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{f(\boldsymbol{\theta}_o)}\big[\boldsymbol{b}_t(\boldsymbol{\theta}_o,\boldsymbol{\psi}(\boldsymbol{\theta}_o))\boldsymbol{b}_t(\boldsymbol{\theta}_o,\boldsymbol{\psi}(\boldsymbol{\theta}_o))'\big],$$

which can be consistently estimated by

$$\widehat{\boldsymbol{\Sigma}}_T=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{b}_t(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T)\boldsymbol{b}_t(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T)'.$$

It follows that the PSE test is

$$\mathcal{PSE}_T=T\widehat{J}_{\varphi}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T)'\widehat{\boldsymbol{\Sigma}}_T^{-}\widehat{J}_{\varphi}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T)\xrightarrow{D}\chi^2(k),$$

where k is the rank of $\widehat{\boldsymbol{\Sigma}}_T$. Wooldridge (1990) proposed the conditional mean encompassing test which is analogous to the PSE test for conditional mean specifications. Yet the Wooldridge test is not applicable to testing other conditional moments.

Example 9.13 Consider the following non-nested specifications of conditional variance:

$$\begin{aligned} &H_0\colon\ y_t|\boldsymbol{x}_t,\boldsymbol{\xi}_t \sim \mathcal{N}\big(0, h(\boldsymbol{x}_t,\boldsymbol{\alpha})\big), \\ &H_1\colon\ y_t|\boldsymbol{x}_t,\boldsymbol{\xi}_t \sim \mathcal{N}\big(0, \kappa(\boldsymbol{\xi}_t,\boldsymbol{\gamma})\big). \end{aligned}$$

For notation simplicity, we shall let h_t denote $h(\boldsymbol{x}_t,\boldsymbol{\alpha})$ and κ_t denote $\kappa(\boldsymbol{\xi}_t,\boldsymbol{\gamma})$. When h_t and κ_t are evaluated at the respective QMLEs $\tilde{\boldsymbol{\alpha}}_T$ and $\tilde{\boldsymbol{\gamma}}_T$, we write $\hat{h}_t$ and $\hat{\kappa}_t$. It can be verified that

$$\begin{aligned} \boldsymbol{s}_{h,t}(\boldsymbol{\alpha}) &= \frac{\nabla_{\boldsymbol{\alpha}} h_t}{2h_t^2}(y_t^2 - h_t), \\ \boldsymbol{s}_{\kappa,t}(\boldsymbol{\gamma}) &= \frac{\nabla_{\boldsymbol{\gamma}} \kappa_t}{2\kappa_t^2}(y_t^2 - \kappa_t) = \underbrace{\frac{\nabla_{\boldsymbol{\gamma}} \kappa_t}{2\kappa_t^2}(h_t - \kappa_t)}_{d_{1,t}} + \underbrace{\frac{\nabla_{\boldsymbol{\gamma}} \kappa_t}{2\kappa_t^2}}_{d_{2,t}} \underbrace{(y_t^2 - h_t)}_{c_t}, \end{aligned}$$

where $\mathbb{E}_{f(\boldsymbol{\theta})}(y_t^2 - h_t|\boldsymbol{x}_t,\boldsymbol{\xi}_t) = 0$. The sample counterpart of the pseudo-true score function is thus

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\nabla_{\boldsymbol{\gamma}} \hat{\kappa}_t}{2\hat{\kappa}_t^2}(y_t^2 - \hat{h}_t).$$

Thus, the PSE test amounts to checking whether $\nabla_{\boldsymbol{\gamma}}\kappa_t/(2\kappa_t^2)$ are correlated with the "generalized" errors $(y_t^2 - h_t)$. The PSE statistic and $\widehat{\boldsymbol{\Sigma}}_T$, the estimate of the asymptotic covariance matrix, are left to Exercise 9.11. □

Exercises

9.1 Let g and f be two density functions. Show that the KLIC $\mathbb{I}(g:f)$ does not obey the triangle inequality, i.e., $\mathbb{I}(g:f) \not\leq \mathbb{I}(g:h) + \mathbb{I}(h:f)$ for any other density function h.

9.2 When $\mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$ is a correct specification of $y_t|\boldsymbol{x}_t$, show that $\mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^3] = 0$ and $\mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^4] = 3(\sigma_o^2)^2$ when $\boldsymbol{\beta}$ is evaluated at $\boldsymbol{\beta}_o$.

9.3 In Example 9.4, the upper-left block of $\boldsymbol{B}_T(\boldsymbol{\theta}_o)$ is $\sum_{t=1}^T \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')/(T\sigma_o^2)$. What would this block be if there is dynamic misspecification? Would $\boldsymbol{B}_T(\boldsymbol{\theta}_o)$ be a block-diagonal matrix if there is dynamic misspecification?

9.4 Suppose that $\mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha}))$ is a correct specification of $y_t|(\boldsymbol{x}_t, \boldsymbol{\zeta}_t)$. Derive $\nabla L_T(\boldsymbol{\theta})$, $\nabla^2 L_T(\boldsymbol{\theta})$, $\boldsymbol{H}_T(\boldsymbol{\theta})$, and $\boldsymbol{B}_T(\boldsymbol{\theta})$.

9.5 Consider the specification $y_t|\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, h(\zeta_t'\boldsymbol{\alpha}))$. What conditions are needed to ensure block diagonality of $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ and $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$?

9.6 Suppose that $\mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha}))$ is a correct specification of $y_t|(\boldsymbol{x}_t, \boldsymbol{\zeta}_t)$, where $\boldsymbol{\zeta}_t'\boldsymbol{\alpha} = \alpha_0 + \sum_{i=1}^p \zeta_{ti}\alpha_i$. Derive the LM test for $\alpha_1 = \cdots = \alpha_p = 0$ when there is dynamic misspecification.

9.7 In the context of Example 9.11, derive the Breusch-Godfrey test for AR(1) errors.

9.8 Consider the specification $y_t|\boldsymbol{x}_t, y_{t-1} \sim \mathcal{N}(\gamma y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$ and the AR(1) errors:

$$y_t - \alpha y_{t-1} - \boldsymbol{x}_t'\boldsymbol{\beta} = \rho(y_{t-1} - \alpha y_{t-2} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta}) + u_t,$$

with $|\rho| < 1$ and $\{u_t\}$ a white noise. Derive the LM test for the null hypothesis $\rho^* = 0$ and show its square root is Durbin's h test; see Section 4.3.3.

9.9 Suppose that the specification is

$$y_t|\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, y_{t-1} \sim \mathcal{N}\big(\boldsymbol{x}_t'\boldsymbol{\beta},\ \alpha_0 + \alpha_1(y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta})^2\big).$$

Letting $e_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$, this specification postulates that the conditional variance of y_t is $\alpha_0 + \alpha_1 e_{t-1}^2$. This is an ARCH (AutoRegressive Conditional Heteroskedasticity) model of order one introduced in Engle (1982). Derive the LM test for the null hypothesis $\alpha_1 = 0$, i.e., no ARCH effect. An ARCH(p) model postulates that the conditional variance of y_t is $\alpha_0 + \alpha_1 e_{t-1}^2 + \cdots + \alpha_p e_{t-p}^2$. What is the LM test of $\alpha_1 = \cdots = \alpha_p = 0$?

9.10 Given the non-nested specifications in Section 9.5.1, construct the SET based on $T^{-1/2}\sum_{t=1}^{T}\boldsymbol{\xi}_t\hat{e}_t$, where $\hat{e}_t$ are the OLS residuals under H_0, and derive the asymptotic distribution of the SET.

9.11 Derive the required asymptotic covariance matrix for the PSE test in Example 9.13 and state the PSE statistic.

References

Amemiya, Takeshi (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models, *Australian Economic Papers*, **17**, 334–355.

Breusch, T. S. and A. R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation, *Econometrica*, **47**, 1287–1294.

Chen, Yi-Ting and Chung-Ming Kuan (2002). The pseudo-true score encompassing test for non-nested hypotheses, *Journal of Econometrics*, **106**, 271–295.

Chen, Yi-Ting and Chung-Ming Kuan (2007). Corrigendum, *Journal of Econometrics*, forthcoming.

Engle, Robert F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation, *Econometrica*, **50**, 987–1007.

Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables, *Econometrica*, **46**, 1293–1301.

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

Gourieroux, Christian and Alain Monfort (1995). *Statistics and Econometric Models*, Volume 1 and 2, Cambridge: Cambridge University Press.

Hamilton, James D. (1994). *Time Series Analysis*, Princeton: Princeton University Press.

Hausman, Jerry A. (1978). Specification tests in econometrics, *Econometrica*, **46**, 1251–1272.

Koenker, Roger (1981). A note on studentizing a test for heteroscedasticity, *Journal of Econometrics*, **17**, 107–112.

Mizon, Grayham E. (1984). The encompassing approach in econometrics, in: D. F. Hendry and K. F. Wallis (eds.), *Econometrics and Quantitative Economics*, pp. 135–172, Oxford: Basil Blackwell.

Mizon, Grayham E. and Jean-Francois. Richard (1986). The encompassing principle and its application to testing non-nested hypotheses, *Econometrica*, **54**, 657–678.

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

White, Halbert (1982). Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1–25.

White, Halbert (1994). *Estimation, Inference, and Specification Analysis*, New York: Cambridge University Press.

Wooldridge, J. M. (1990). An encompassing approach to conditional mean tests with applications to testing non-nested hypotheses, *Journal of Econometrics*, **45**, 331–350.