Chapter 9

The Quasi-Maximum Likelihood Method: Theory

As discussed in preceding chapters, estimating linear and nonlinear regressions by the least squares method results in an approximation to the conditional mean function of the dependent variable. While this approach is important and common in practice, its scope is still limited and does not fit the purposes of various empirical studies. First, this approach is not suitable for analyzing the dependent variables with special features, such as binary and truncated (censored) variables, because it is unable to take data characteristics into account. Second, it has little room to characterize other conditional moments, such as conditional variance, of the dependent variable. As far as a complete description of the conditional behavior of a dependent variable is concerned, it is desirable to specify a density function that admits specifications of different conditional moments and other distribution characteristics. This motivates researchers to consider the method of *quasi-maximum likelihood* (QML).

The QML method is essentially the same as the ML method usually seen in statistics and econometrics textbooks. It is conceivable that specifying a density 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 traditional ML method assumes that the specified density function is the true density function, so that specification errors are "assumed away." Therefore, 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, (1994).

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 when p is small but less important when p is large. For example, when $p = 0.99$, the message that A will occur is not much informative, but when $p = 0.01$, such a message is very helpful. Hence, the information content of the message that A will occur ought to be a decreasing function of 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

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

It is also common to interpret $\iota(p)$ as the "surprise" resulted from knowing that the event A will occur when $\mathbb{P}(A) = p$. The expected information I is thus also a weighted average of "surprises" and 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 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 knowing g is in fact 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 the Lebesgue measure zero).*

Proof: Using the fact that $\log(1+x) < x$ for all $x > -1$, we have

$$\log\Big(\frac{g}{f}\Big) = -\log\Big(1 + \frac{f-g}{g}\Big) > 1 - \frac{f}{g}.$$

It follows that

$$\int \log\Big(\frac{g(\zeta)}{f(\zeta)}\Big) g(\zeta)\,\mathrm{d}\zeta > \int \Big(1 - \frac{f(\zeta)}{g(\zeta)}\Big) g(\zeta)\,\mathrm{d}\zeta = 0.$$

Clearly, if $g = f$ almost everywhere, $\mathbb{I}(g:f) = 0$. Conversely, given $\mathbb{I}(g:f) = 0$, suppose without loss of generality that $g = f$ except that $g > f$ on the set B that has a non-zero Lebesgue measure. Then,

$$\int_{\mathbb{R}} \log\Big(\frac{g(\zeta)}{f(\zeta)}\Big) g(\zeta)\,\mathrm{d}\zeta = \int_B \log\Big(\frac{g(\zeta)}{f(\zeta)}\Big) g(\zeta)\,\mathrm{d}\zeta > \int_B (\log 1) g(\zeta)\,\mathrm{d}\zeta = 0,$$

contradicting $\mathbb{I}(g:f) = 0$. Thus, g must be the same as f almost everywhere. □

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 crude 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 $(\Omega, \mathcal{F}, \mathbb{P}_o)$. Without causing any confusion, we shall use the same notations for random vectors and their realizations. Let

$$\boldsymbol{z}^T = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_T)$$

denote the the information set generated by all $\boldsymbol{z}_t$. The vector $\boldsymbol{z}_t$ is divided into two parts y_t and $\boldsymbol{w}_t$, where y_t is a scalar and $\boldsymbol{w}_t$ is $(\nu-1)\times 1$. Corresponding to $\mathbb{P}_o$, there exists a joint density function g^T for $\boldsymbol{z}^T$:

$$g^T(\boldsymbol{z}^T) = g(\boldsymbol{z}_1) \prod_{t=2}^{T} \frac{g^t(\boldsymbol{z}^t)}{g^{t-1}(\boldsymbol{z}^{t-1})} = g(\boldsymbol{z}_1) \prod_{t=2}^{T} g_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1}),$$

where g^t denote the joint density of t random variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t$, and g_t is the density function of $\boldsymbol{z}_t$ conditional on past information $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{t-1}$. The joint density function

g^T is the random mechanism governing the behavior of $\boldsymbol{z}^T$ and will be referred to as the *data generation process* (DGP) of $\boldsymbol{z}^T$.

As g^T is unknown, we may postulate a conditional density function $f_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$, and approximate $g^T(\boldsymbol{z}^T)$ by

$$f^T(\boldsymbol{z}^T; \boldsymbol{\theta}) = f(\boldsymbol{z}_1) \prod_{t=2}^{T} f_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1}; \boldsymbol{\theta}).$$

The function f^T is referred to as the *quasi-likelihood function*, in the sense that f^T need not agree with g^T. For notation convenience, we set the unconditional density $g(\boldsymbol{z}_1)$ (or $f(\boldsymbol{z}_1)$) as a conditional density and write g^T (or f^T) as the product of all conditional density functions. Clearly, the postulated density f^T would be useful if it is close to the DGP g^T. It is therefore natural to consider minimizing the KLIC of g^T relative to f^T:

$$\mathbb{I}(g^T : f^T; \boldsymbol{\theta}) = \int_{\mathbb{R}^T} \log \left(\frac{g^T(\boldsymbol{\zeta}^T)}{f^T(\boldsymbol{\zeta}^T; \boldsymbol{\theta})} \right) g^T(\boldsymbol{\zeta}^T) \, \mathrm{d}(\boldsymbol{\zeta}^T). \tag{9.1}$$

This amounts to minimizing the "surprise" level resulted from specifying an f^T for the DGP g^T. As g^T does not involve $\boldsymbol{\theta}$, minimizing the KLIC (9.1) with respect to $\boldsymbol{\theta}$ is equivalent to maximizing

$$\int_{\mathbb{R}^T} \log(f^T(\boldsymbol{\zeta}^T; \boldsymbol{\theta})) g^T(\boldsymbol{\zeta}^T) \, \mathrm{d}(\boldsymbol{\zeta}^T) = \mathbb{E}[\log f^T(\boldsymbol{z}^T; \boldsymbol{\theta})],$$

where $\mathbb{E}$ is the expectation operator with respect to the DGP $g^T(\boldsymbol{z}^T)$. This is, in turn, equivalent to maximizing the average of $\log f_t$:

$$\bar{L}_T(\boldsymbol{\theta}) = \frac{1}{T} \mathbb{E}[\log f^T(\boldsymbol{z}^T; \boldsymbol{\theta})] = \frac{1}{T} \mathbb{E} \left(\sum_{t=1}^{T} \log f_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1}; \boldsymbol{\theta}) \right). \tag{9.2}$$

Let $\boldsymbol{\theta}^*$ be the maximizer of $\bar{L}_T(\boldsymbol{\theta})$. Then, $\boldsymbol{\theta}^*$ is also the minimizer of the KLIC (9.1). If there exists a unique $\boldsymbol{\theta}_o \in \Theta$ such that $f^T(\boldsymbol{z}^T; \boldsymbol{\theta}_o) = g^T(\boldsymbol{z}^T)$ for all T, we say that $\{f_t\}$ is *specified correctly in its entirety* for $\{\boldsymbol{z}_t\}$. In this case, the resulting KLIC $\mathbb{I}(g^T : f^T; \boldsymbol{\theta}_o) = 0$ and reaches the minimum. This shows that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_o$ when $\{f_t\}$ is specified correctly in its entirety.

Maximizing $\bar{L}_T$ is, however, not a readily solvable problem because the objective function (9.2) involves the expectation operator and hence depends on the unknown DGP g^T. It is therefore natural to consider maximizing the sample counterpart of $\bar{L}_T(\boldsymbol{\theta})$:

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

which is known as the *quasi-log-likelihood function*. The maximizer of $L_T(\boldsymbol{z}^T;\boldsymbol{\theta})$, $\tilde{\boldsymbol{\theta}}_T$, is known as the *quasi-maximum likelihood estimator* (QMLE) of $\boldsymbol{\theta}$. The prefix "quasi" is used to indicate that this solution may be obtained from a misspecified log-likelihood function. When $\{f_t\}$ is specified correctly in its entirety for $\{\boldsymbol{z}_t\}$, the QMLE is understood as the MLE, as in standard statistics textbooks.

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). Instead, econometricians are typically interested in modeling a variable of interest, say, y_t, conditional on a set of "pre-determined" variables, say, $\boldsymbol{x}_t$, where $\boldsymbol{x}_t$ includes some elements of $\boldsymbol{w}_t$ and $\boldsymbol{z}^{t-1}$. This is a relatively simple job because only the conditional behavior of y_t need to be considered. As $\boldsymbol{w}_t$ are not explicitly modeled, the conditional density $g_t(y_t \mid \boldsymbol{x}_t)$ provides only a partial description of $\{\boldsymbol{z}_t\}$. We then find a quasi-likelihood function $f_t(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta})$ to approximate $g_t(y_t \mid \boldsymbol{x}_t)$. Analogous to (9.1), the resulting average KLIC of g_t relative to f_t is

$$\bar{\mathbb{I}}_T(\{g_t : f_t\};\boldsymbol{\theta}) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}(g_t : f_t;\boldsymbol{\theta}). \tag{9.4}$$

Let $y^T = (y_1,\ldots,y_T)$ and $\boldsymbol{x}^T = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)$. Minimizing $\bar{\mathbb{I}}_T(\{g_t : f_t\};\boldsymbol{\theta})$ in (9.4) is thus equivalent to maximizing

$$\bar{L}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\log f_t(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta})]; \tag{9.5}$$

cf. (9.2). The parameter $\boldsymbol{\theta}^*$ that maximizes (9.5) must also minimize the average KLIC (9.4). 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, $\bar{\mathbb{I}}_T(\{g_t : f_t\};\boldsymbol{\theta}_o)$ is zero, and $\boldsymbol{\theta}^* = \boldsymbol{\theta}_o$.

Similar as before, $\bar{L}_T(\boldsymbol{\theta})$ is not directly observable, we instead maximize its sample counterpart:

$$L_T(y^T,\boldsymbol{x}^T;\boldsymbol{\theta}) := \frac{1}{T}\sum_{t=1}^{T} \log f_t(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta}), \tag{9.6}$$

which will also be referred to as the quasi-log-likelihood function; cf. (9.3). The resulting solution is the QMLE $\tilde{\boldsymbol{\theta}}_T$. When $\{f_t\}$ is correctly specified for $\{y_t \mid \boldsymbol{x}_t\}$, the QMLE $\tilde{\boldsymbol{\theta}}_T$ is also understood as the usual MLE.

As a special case, 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:

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

note that conditional variance is not explicitly modeled. Setting $\boldsymbol{\theta} = (\boldsymbol{\beta}'\ \sigma^2)'$, it is easy to see that the maximizer of the quasi-log-likelihood function $T^{-1}\sum_{t=1}^T \log f_t(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta})$ is also the solution to

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

The resulting QMLE of $\boldsymbol{\beta}$ is thus the NLS estimator. Therefore, the NLS estimator can be viewed as a QMLE under the assumption of conditional normality with conditional homoskedasticity. We say that $\{\mu_t\}$ is correctly specified for the conditional mean $\mathbb{E}(y_t \mid \boldsymbol{x}_t)$ if there exists a $\boldsymbol{\theta}_o$ such that $\mu_t(\boldsymbol{x}_t;\boldsymbol{\theta}_o) = \mathbb{E}(y_t \mid \boldsymbol{x}_t)$. A more flexible specification, such as

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

would allow us to characterize conditional variance as well.

9.2 Asymptotic Properties of the QMLE

The quasi-log-likelihood function is, in general, a nonlinear function in $\boldsymbol{\theta}$, so that the QMLE thus must be computed numerically using a nonlinear optimization algorithm. Given that maximizing L_T is equivalent to minimizing $-L_T$, the algorithms discussed in Section 8.2.2 are readily applied. We shall not repeat these methods here but proceed to the discussion of 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 the 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) or (9.4).

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 $L_T(\boldsymbol{z}^T;\boldsymbol{\theta})$ (or $L_T(\boldsymbol{y}^T,\boldsymbol{x}^T;\boldsymbol{\theta})$) tends to $\bar{L}_T(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta}\in\Theta$, i.e., $L_T(\boldsymbol{z}^T;\boldsymbol{\theta})$ obeys a WULLN, then

$\tilde{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}^*$ in probability. This shows that the QMLE is a weakly consistent estimator for the minimizer of KLIC $\mathbb{I}(\mathrm{g}^{\mathrm{T}} : \mathrm{f}^{\mathrm{T}}; \boldsymbol{\theta})$ (or the average KLIC $\bar{\mathbb{I}}_T(\{g_t : f_t\}; \boldsymbol{\theta})$), whether the specification is correct or not. When $\{f_t\}$ is specified correctly in its entirety (or 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 conditions required to ensure QMLE consistency are also those for a WULLN; we omit the details.

9.2.2 Asymptotic Normality

Consider first the specification of the quasi-log-likelihood function for $\boldsymbol{z}^T$: $L_T(\boldsymbol{z}^T; \boldsymbol{\theta})$. When $\boldsymbol{\theta}^*$ is in the interior of Θ, the mean-value expansion of $\nabla L_T(\boldsymbol{z}^T; \tilde{\boldsymbol{\theta}}_T)$ about $\boldsymbol{\theta}^*$ is

$$\nabla L_T(\boldsymbol{z}^T; \tilde{\boldsymbol{\theta}}_T) = \nabla L_T(\boldsymbol{z}^T; \boldsymbol{\theta}^*) + \nabla^2 L_T(\boldsymbol{z}^T; \boldsymbol{\theta}_T^\dagger)(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*), \tag{9.7}$$

where $\boldsymbol{\theta}_T^\dagger$ is between $\tilde{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}^*$. Note that requiring $\boldsymbol{\theta}^*$ in the interior of Θ is to ensure that the mean-value expansion and subsequent asymptotics are valid in the parameter space.

The left-hand side of (9.7) is zero because the QMLE $\tilde{\boldsymbol{\theta}}_T$ solves the first order condition $\nabla L_T(\boldsymbol{z}^T; \boldsymbol{\theta}) = \mathbf{0}$. Then, as long as $\nabla^2 L_T(\boldsymbol{z}^T; \boldsymbol{\theta}_T^\dagger)$ is invertible with probability one, (9.7) can be written as

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\nabla^2 L_T(\boldsymbol{z}^T; \boldsymbol{\theta}_T^\dagger)^{-1} \sqrt{T}\, \nabla L_T(\boldsymbol{z}^T; \boldsymbol{\theta}^*).$$

Note that the invertibility of $\nabla^2 L_T(\boldsymbol{z}^T; \boldsymbol{\theta}_T^\dagger)$ amounts to requiring quasi-log-likelihood function L_T being locally quadratic at $\boldsymbol{\theta}^*$. Let $\mathbb{E}[\nabla^2 L_T(\boldsymbol{z}^T; \boldsymbol{\theta})] = \boldsymbol{H}_T(\boldsymbol{\theta})$ be the expected Hessian matrix. When $\nabla^2 L_T(\boldsymbol{z}^T; \boldsymbol{\theta}_T^\dagger)$ obeys a WULLN, we have

$$\nabla^2 L_T(\boldsymbol{z}^T; \boldsymbol{\theta}) - \boldsymbol{H}_T(\boldsymbol{\theta}) \xrightarrow{\mathbb{P}} \mathbf{0},$$

uniformly in Θ. As $\tilde{\boldsymbol{\theta}}_T$ is weakly consistent for $\boldsymbol{\theta}^*$, so is $\boldsymbol{\theta}_T^\dagger$. The assumed WULLN then implies that

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

It follows that

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1} \sqrt{T}\, \nabla L_T(\boldsymbol{z}^T; \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1). \tag{9.8}$$

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: $\sqrt{T}\, \nabla L_T(\boldsymbol{z}^T; \boldsymbol{\theta}^*)$.

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

$$\boldsymbol{B}_T(\boldsymbol{\theta}) = \operatorname{var}\big(\sqrt{T}\nabla L_T(\boldsymbol{z}^T;\boldsymbol{\theta})\big),$$

which will also be referred to as the *information matrix*. Then provided that $\log f_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1})$ obeys a CLT, we have

$$\boldsymbol{B}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\big(\nabla L_T(\boldsymbol{z}^T;\boldsymbol{\theta}^*) - \mathbb{E}[\nabla L_T(\boldsymbol{z}^T;\boldsymbol{\theta}^*)]\big) \xrightarrow{D} \mathcal{N}(0,\boldsymbol{I}_k). \tag{9.9}$$

When differentiation and integration can be interchanged,

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

where the right-hand side is the first order derivative of (9.2). As $\boldsymbol{\theta}^*$ is the KLIC minimizer, $\nabla\bar{L}_T(\boldsymbol{\theta}^*) = \mathbf{0}$ so that $\mathbb{E}[\nabla L_T(\boldsymbol{z}^T;\boldsymbol{\theta}^*)] = \mathbf{0}$. By (9.8) and (9.9),

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

which has an asymptotic normal distribution. This immediately leads to the following result.

Theorem 9.2 *When* (9.7), (9.8) *and* (9.9) *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 $\boldsymbol{H}_T(\boldsymbol{\theta}^*) = \mathbb{E}[\nabla^2 L_T(\boldsymbol{z}^T;\boldsymbol{\theta}^*)]$ *and* $\boldsymbol{B}_T(\boldsymbol{\theta}^*) = \operatorname{var}\big(\sqrt{T}\nabla L_T(\boldsymbol{z}^T;\boldsymbol{\theta}^*)\big)$.

Remark: For the specification of $\{y_t \mid \boldsymbol{x}_t\}$, the QMLE is obtained from the quasi-log-likelihood function $L_T(y^T,\boldsymbol{x}^T;\boldsymbol{\theta})$, and its asymptotic normality holds similarly as in Theorem 9.2. That is,

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

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 $\boldsymbol{B}_T(\boldsymbol{\theta}^*) = \operatorname{var}\big(\sqrt{T}\nabla L_T(y^T,\boldsymbol{x}^T;\boldsymbol{\theta}^*)\big)$.

9.3 Information Matrix Equality

A useful result in the quasi-maximum likelihood theory is the *information matrix equality*. This equality shows that, under certain conditions, the information matrix $\boldsymbol{B}_T(\boldsymbol{\theta})$ is the same as the negative of the expected Hessian matrix $-\boldsymbol{H}_T(\boldsymbol{\theta})$ so that the covariance matrix $\boldsymbol{C}_T(\boldsymbol{\theta})$ can be simplified. In such a case, the estimation of $\boldsymbol{C}_T(\boldsymbol{\theta})$ would be much simpler.

Define the following score functions: $\boldsymbol{s}_t(\boldsymbol{z}^t;\boldsymbol{\theta}) = \nabla \log f_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1};\boldsymbol{\theta})$. The average of these scores is

$$\boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta}) = \nabla L_T(\boldsymbol{z}^T;\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{s}_t(\boldsymbol{z}^t;\boldsymbol{\theta}).$$

Clearly, $\boldsymbol{s}_t(\boldsymbol{z}^t;\boldsymbol{\theta}) f_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1};\boldsymbol{\theta}) = \nabla f_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1};\boldsymbol{\theta})$. It follows that

$$\begin{aligned}
\boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta}) f^T(\boldsymbol{z}^T;\boldsymbol{\theta}) &= \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{s}_t(\boldsymbol{z}^t;\boldsymbol{\theta})\right)\left(\prod_{t=1}^{T} f^\tau(\boldsymbol{z}_\tau \mid \boldsymbol{z}^{\tau-1};\boldsymbol{\theta})\right) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\nabla f_t(\boldsymbol{z}_t|\boldsymbol{z}^{t-1};\boldsymbol{\theta}) \prod_{\tau\neq t} f^\tau(\boldsymbol{z}_\tau \mid \boldsymbol{z}^{\tau-1};\boldsymbol{\theta})\right) \\
&= \frac{1}{T}\nabla f^T(\boldsymbol{z}^T;\boldsymbol{\theta}).
\end{aligned}$$

As $\int f^T(\boldsymbol{z}^T;\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z}^T = 1$, its derivative must be zero. By permitting interexchange of differentiation and integration we have

$$\mathbf{0} = \frac{1}{T}\int \nabla f^T(\boldsymbol{z}^T;\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z}^T = \int \boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta}) f^T(\boldsymbol{z}^T;\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z}^T,$$

and

$$\begin{aligned}
\mathbf{0} &= \frac{1}{T}\int \nabla^2 f^T(\boldsymbol{z}^T;\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z}^T \\
&= \int \left[\nabla \boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta}) + T\boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta})\boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta})'\right] f^T(\boldsymbol{z}^T;\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z}^T.
\end{aligned}$$

The equalities are simply the expectations with respect to $f^T(\boldsymbol{z}^T;\boldsymbol{\theta})$, yet they need not hold when the integrator is not $f^T(\boldsymbol{z}^T;\boldsymbol{\theta})$.

If $\{f_t\}$ is correctly specified in its entirety for $\{\boldsymbol{z}_t\}$, the results above imply $\mathbb{E}[\boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta}_o)] = \mathbf{0}$ and

$$\mathbb{E}[\nabla \boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta}_o)] + T\,\mathbb{E}[\boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta}_o)\boldsymbol{s}^T(\boldsymbol{z}^T;\boldsymbol{\theta}_o)'] = \mathbf{0},$$

where the expectations are taken with respect to the true density $g^T(\boldsymbol{z}^T) = f^T(\boldsymbol{z}^T;\boldsymbol{\theta}_o)$. This is the information matrix equality stated below.

Theorem 9.3 *Suppose that there exists a $\boldsymbol{\theta}_o$ such that $f_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1}; \boldsymbol{\theta}_o)\} = g_t(\boldsymbol{z}_t \mid \boldsymbol{z}^{t-1})$. Then,*

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

where $\boldsymbol{H}_T(\boldsymbol{\theta}_o) = \mathbb{E}[\nabla^2 L_T(\boldsymbol{z}^T; \boldsymbol{\theta}_o)]$ and $\boldsymbol{B}_T(\boldsymbol{\theta}_o) = \operatorname{var}\big(\sqrt{T}\nabla L_T(\boldsymbol{z}^T; \boldsymbol{\theta}_o)\big)$.

When this equality holds, the covariance matrix $\boldsymbol{C}_T$ in Theorem 9.2 simplifies to

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

That is, the QMLE achieves the Cramér-Rao lower bound asymptotically.

On the other hand, when f^T is not a correct specification for $\boldsymbol{z}^T$, the score function is not related to the true density g^T. Hence, there is no guarantee that the mean score is zero, i.e.,

$$\mathbb{E}[\boldsymbol{s}^T(\boldsymbol{z}^T; \boldsymbol{\theta})] = \int \boldsymbol{s}^T(\boldsymbol{z}^T; \boldsymbol{\theta}) g^T(\boldsymbol{z}^T)\, \mathrm{d}\boldsymbol{z}^T \neq \mathbf{0},$$

even when this expectation is evaluated at $\boldsymbol{\theta}^*$. By the same reason,

$$\mathbb{E}[\nabla \boldsymbol{s}^T(\boldsymbol{z}^T; \boldsymbol{\theta})] + T\,\mathbb{E}[\boldsymbol{s}^T(\boldsymbol{z}^T; \boldsymbol{\theta}) \boldsymbol{s}^T(\boldsymbol{z}^T; \boldsymbol{\theta})'] \neq \mathbf{0},$$

even when it is evaluated at $\boldsymbol{\theta}^*$.

For the specification of $\{y_t \mid \boldsymbol{x}_t\}$, we have $\nabla f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta}) = \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}) f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})$, and $\int f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})\, \mathrm{d}y_t = 1$. Similar as above,

$$\int \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}) f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})\, \mathrm{d}y_t = \mathbf{0},$$

and

$$\int [\nabla \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}) + \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}) \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta})'] f_t(y_t \mid \boldsymbol{x}_t; \boldsymbol{\theta})\, \mathrm{d}y_t = \mathbf{0}.$$

If $\{f_t\}$ is correctly specified for $\{y_t \mid \boldsymbol{x}_t\}$, we have $\mathbb{E}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o) \mid \boldsymbol{x}_t] = \mathbf{0}$. Then by the law of iterated expectations, $\mathbb{E}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)] = \mathbf{0}$, so that the mean score is still zero under correct specification. Moreover,

$$\mathbb{E}[\nabla \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o) \mid \boldsymbol{x}_t] + \mathbb{E}[\boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o) \boldsymbol{s}_t(y_t, \boldsymbol{x}_t; \boldsymbol{\theta}_o)' \mid \boldsymbol{x}_t] = \mathbf{0},$$

which implies

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

The equality above is not necessarily equivalent to the information matrix equality, however.

To see this, consider the specifications $\{f_t(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta})\}$, which are correct for $\{y_t \mid \boldsymbol{x}_t\}$. These specifications are said to have *dynamic misspecification* if they are not correctly specified for $\{y_t \mid \boldsymbol{w}_t, \boldsymbol{z}^{t-1}\}$. That is, there does not exist any $\boldsymbol{\theta}_o$ such that $f_t(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta}_o) = g_t(y_t \mid \boldsymbol{w}_t, \boldsymbol{z}^{t-1})$. Thus, the information contained in $\boldsymbol{w}_t$ and $\boldsymbol{z}^{t-1}$ cannot be fully represented by $\boldsymbol{x}_t$. On the other hand, when dynamic misspecification is absent, it is easily seen that

$$\mathbb{E}[\boldsymbol{s}_t(y_t,\boldsymbol{x}_t;\boldsymbol{\theta}_o) \mid \boldsymbol{x}_t] = \mathbb{E}[\boldsymbol{s}_t(y_t,\boldsymbol{x}_t;\boldsymbol{\theta}_o) \mid \boldsymbol{w}_t, \boldsymbol{z}^{t-1}]. \tag{9.10}$$

It is then easily verified that

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

where the last equality holds by (9.10) and the law of iterated expectations. When there is dynamic misspecification, the last equality fails because the covariances of scores do not vanish. This shows that the average of individual information matrices, i.e., $\mathrm{var}(\boldsymbol{s}_t(y_t,\boldsymbol{x}_t;\boldsymbol{\theta}_o))$, need not be the information matrix.

Theorem 9.4 *Suppose that there exists a $\boldsymbol{\theta}_o$ such that $f_t(y_t \mid \boldsymbol{x}_t;\boldsymbol{\theta}_o) = g_t(y_t \mid \boldsymbol{x}_t)$ and there is no dynamic misspecification. Then,*

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

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

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

When Theorem 9.4 holds, the covariance matrix needed to normalize $\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_o)$ again simplifies to $\boldsymbol{B}_T(\boldsymbol{\theta}_o)^{-1} = -\boldsymbol{H}_T(\boldsymbol{\theta}_o)^{-1}$, and the QMLE achieves the Cramér-Rao lower bound asymptotically.

Example 9.5 Consider the following specification: $y_t \mid \boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$ for all t. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}' \ \sigma^2)'$, then

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

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 $\boldsymbol{\beta}$ to obtain the QMLE. It is easily verified that the QMLE of $\boldsymbol{\beta}$ is the OLS estimator $\hat{\boldsymbol{\beta}}_T$ and 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 $\{y_t \mid \boldsymbol{x}_t\}$, 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

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

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

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

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

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

If there is no dynamic misspecification,

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

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

$$\mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^3] = -3\sigma_o^2\,\mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})] - \mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})^3], \tag{9.11}$$

which is zero when evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$, and that

$$\mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^4] = 3(\sigma_o^2)^2 + 6\sigma_o^2\,\mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})^2] + \mathbb{E}[(\boldsymbol{x}_t'\boldsymbol{\beta}_o - \boldsymbol{x}_t'\boldsymbol{\beta})^3], \tag{9.12}$$

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

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

This shows that the information matrix equality holds.

A typical consistent estimator of $\boldsymbol{H}_T(\boldsymbol{\theta}_o)$ is

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

Due to the information matrix equality, a consistent estimator of $\boldsymbol{B}_T(\boldsymbol{\theta}_o)$ is $\boldsymbol{B}_T(\tilde{\boldsymbol{\theta}}_T) = -\boldsymbol{H}_T(\tilde{\boldsymbol{\theta}}_T)$. It can be seen that the upper-left block of $\boldsymbol{H}_T(\tilde{\boldsymbol{\theta}}_T)$ is the standard estimator for the covariance matrix of $\hat{\boldsymbol{\beta}}_T$. On the other hand, when there is dynamic missepcification, $\boldsymbol{B}_T(\boldsymbol{\theta})$ is not the same as given above so that the information matrix equality fails; see Exercise 9.3. □

The information matrix equality may also fail in different circumstances. The example below shows that, even when the specification for $\mathbb{E}(y_t \mid \boldsymbol{x}_t)$ 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 9.6 Consider the specification as in Example 9.5: $y_t \mid \boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$ for all t. Suppose that the DGP is

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

where the conditional variance $h(\boldsymbol{x}_t'\boldsymbol{\gamma}_o)$ varies with $\boldsymbol{x}_t$. Then, this specification includes a correct specification for the conditional mean but itself is incorrect for $\{y_t \mid \boldsymbol{x}_t\}$ because it ignores conditional heteroskedasticity. From Example 9.5 we see that the upper-left block of $\boldsymbol{H}_T(\boldsymbol{\theta})$ is

$$-\frac{1}{\sigma^2 T}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')\sigma^2,$$

and that the corresponding block of $\boldsymbol{B}_T(\boldsymbol{\theta})$ is

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{E}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2\boldsymbol{x}_t\boldsymbol{x}_t']}{(\sigma^2)^2}.$$

As the specification is correct for the conditional mean but not the conditional variance, the KLIC minimzer is $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_o', (\sigma^*)^2)'$. Evaluating the submatrix of $\boldsymbol{H}_T(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$ yields

$$-\frac{1}{(\sigma^*)^2 T}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'),$$

which differs from the corresponding submatrix of $\boldsymbol{B}_T(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}^*$:

$$\frac{1}{(\sigma^*)^4 T}\sum_{t=1}^{T} \mathbb{E}[h(\boldsymbol{x}_t'\boldsymbol{\gamma}_o)^2\boldsymbol{x}_t\boldsymbol{x}_t'].$$

Thus, the information matrix equality breaks down, despite that the conditional mean specification is correct.

A consistent estimator of $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ is

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

yet a consistent estimator of $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$ is

$$\boldsymbol{B}_T(\tilde{\boldsymbol{\theta}}_T) = \begin{bmatrix} \frac{\sum_{t=1}^{T}\hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'}{T(\tilde{\sigma}_T^2)^2} & \mathbf{0} \\ \mathbf{0} & -\frac{1}{4(\tilde{\sigma}_T^2)^2} + \frac{\sum_{t=1}^{T}\hat{e}_t^4}{T\,4(\tilde{\sigma}_T^2)^4} \end{bmatrix};$$

see Exercise 9.4. Due to the block diagonality of $\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_T)$ and $\boldsymbol{B}(\tilde{\boldsymbol{\theta}}_T)$, it is easy to verify that the upper-left block of $\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_T)^{-1}\boldsymbol{B}(\tilde{\boldsymbol{\theta}}_T)\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_T)^{-1}$ is

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

This is precisely the the Eicker-White estimator (6.10) for the covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}}_T$, as shown in Section 6.3.1. □

9.4 Hypothesis Testing

In this section we discuss three classical large sample tests (Wald, LM, and likelihood ratio tests), the Hausman (1978) test, and the information matrix test of White (1982, 1987) for the null hypothesis $\boldsymbol{R}\boldsymbol{\theta}^* = \boldsymbol{r}$, where $\boldsymbol{R}$ is $q \times k$ matrix with full row rank. Failure to reject the null hypothesis suggests that there is no empirical evidence against the specified model under the null hypothesis.

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 (9.8),

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

This shows that $\boldsymbol{R}\boldsymbol{C}_T(\boldsymbol{\theta}^*)\boldsymbol{R}' = \boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{B}_T(\boldsymbol{\theta}^*)\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'$ can be used to normalize $\sqrt{T}\boldsymbol{R}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*)$ to obtain asymptotic normality. Under the null hypothesis, $\boldsymbol{R}\boldsymbol{\theta}^* = \boldsymbol{r}$ so that

$$[\boldsymbol{R}\boldsymbol{C}_T(\boldsymbol{\theta}^*)\boldsymbol{R}']^{-1/2}\sqrt{T}(\boldsymbol{R}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{r}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q). \tag{9.13}$$

This is the key distribution result for the Wald test.

For notation simplicity, we let $\tilde{\boldsymbol{H}}_T = \boldsymbol{H}_T(\tilde{\boldsymbol{\theta}}_T)$ denote a consistent estimator for $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ and $\tilde{\boldsymbol{B}}_T = \boldsymbol{B}_T(\tilde{\boldsymbol{\theta}}_T)$ denote a consistent estimator for $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$. It follows that a consistent estimator for $\boldsymbol{C}_T(\boldsymbol{\theta}^*)$ is

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

Substituting $\tilde{\boldsymbol{C}}_T$ for $\boldsymbol{C}_T(\boldsymbol{\theta})$ in (9.13) we have

$$[\boldsymbol{R} \qquad tildeb C_T(\boldsymbol{\theta}^*)\boldsymbol{R}']^{-1/2}\sqrt{T}(\boldsymbol{R}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{r}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q). \tag{9.14}$$

The Wald test statistic is the inner product of the left-hand side of (9.14):

$$\mathcal{W}_T = T(\boldsymbol{R}\tilde{\boldsymbol{\theta}}_T - \boldsymbol{r})'(\boldsymbol{R}\tilde{\boldsymbol{C}}_T\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\theta}}_T - \boldsymbol{r}). \tag{9.15}$$

The limiting distribution of the Wald test now follows from (9.14) and the continuous mapping theorem.

Theorem 9.7 *Suppose that Theorem 9.2 for the QMLE $\tilde{\boldsymbol{\theta}}_T$ holds. Then under the null hypothesis,*

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

where $\mathcal{W}_T$ is defined in (9.15) *and q is the number of rows of $\boldsymbol{R}$.*

Example 9.8 Consider the quasi-log-likelihood function specified in Example 9.5. 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 (9.15):

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

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

As shown in Example 9.5, when the information matrix equality holds, $\tilde{\boldsymbol{C}}_T = -\tilde{\boldsymbol{H}}_T^{-1}$ is block diagnoal so that

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

The Wald test then becomes

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

In this case, the Wald test is just s times the standard F statistic which is readily available from most of econometric packages. □

9.4.2 Lagrange Multiplier Test

Consider now the problem of maximizing $L_T(\boldsymbol{\theta})$ subject to the constraint $\boldsymbol{R}\boldsymbol{\theta} = \boldsymbol{r}$. 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 and 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:

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

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

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

where $\boldsymbol{\theta}_T^\dagger$ is the mean value between $\ddot{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}^*$. It has been shown in (9.8) that

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

Hence,

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

Using the WULLN result: $\nabla^2 L_T(\boldsymbol{\theta}_T^\dagger) - \boldsymbol{H}_T(\boldsymbol{\theta}^*) \xrightarrow{\mathbb{P}} \boldsymbol{0}$, we obtain

$$\sqrt{T}(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = \sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) - \boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'\sqrt{T}\ddot{\boldsymbol{\lambda}}_T + o_{\mathbb{P}}(1). \tag{9.16}$$

This establishes a relationship between the constrained and unconstrained QMLEs.

Pre-multiplying both sides of (9.16) by $\boldsymbol{R}$ and noting that the constrained estimator $\ddot{\boldsymbol{\theta}}_T$ must satisfy the constraint so that $\boldsymbol{R}(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = \boldsymbol{0}$, we have

$$\sqrt{T}\ddot{\boldsymbol{\lambda}}_T = [\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1), \tag{9.17}$$

which relates the Lagrangian multiplier and the unconstrained QMLE $\tilde{\boldsymbol{\theta}}_T$. When Theorem 9.2 holds for the normalized $\tilde{\boldsymbol{\theta}}_T$, we obtain the following asymptotic normality result for the normalized Lagrangian multiplier:

$$\boldsymbol{\Lambda}_T^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T = \boldsymbol{\Lambda}_T^{-1/2}[\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q), \tag{9.18}$$

where

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

Let $\ddot{\boldsymbol{H}}_T = \boldsymbol{H}_T(\ddot{\boldsymbol{\theta}}_T)$ denote a consistent estimator for $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ and $\ddot{\boldsymbol{C}}_T = \boldsymbol{C}_T(\ddot{\boldsymbol{\theta}}_T)$ 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 (9.18) that

$$\ddot{\boldsymbol{\Lambda}}_T^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T = \ddot{\boldsymbol{\Lambda}}_T^{-1/2}(\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q), \tag{9.19}$$

The LM test statistic is the inner product of the left-hand side of (9.19):

$$\mathcal{LM}_T = T\ddot{\boldsymbol{\lambda}}_T'\ddot{\boldsymbol{\Lambda}}_T^{-1}\ddot{\boldsymbol{\lambda}}_T = T\ddot{\boldsymbol{\lambda}}_T'\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}'(\boldsymbol{R}\ddot{\boldsymbol{C}}_T\boldsymbol{R}')^{-1}\boldsymbol{R}\ddot{\boldsymbol{H}}_T^{-1}\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T. \tag{9.20}$$

The limiting distribution of the LM test now follows easily from (9.19) and the continuous mapping theorem.

Theorem 9.9 *Suppose that Theorem 9.2 for the QMLE $\tilde{\boldsymbol{\theta}}_T$ holds. Then under the null hypothesis,*

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

where $\mathcal{LM}_T$ is defined in (9.20) *and q is the number of rows of $\boldsymbol{R}$.*

Remark: When the information matrix equality holds, the LM statistic (9.20) becomes

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

which mainly involves the averages of scores: $\nabla L_T(\ddot{\boldsymbol{\theta}}_T)$. The LM test is thus a test that checks if the average of scores is sufficiently close to zero and hence also known as the *score* test.

Example 9.10 Consider the quasi-log-likelihood function specified in Example 9.5. 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

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

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 (9.20):

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

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 matters.

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

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

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

Example 9.11 (Breusch-Pagan) Suppose that the specification is

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

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 $y_t \mid \boldsymbol{x}_t, \boldsymbol{\zeta}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$, where $\sigma^2 = h(\alpha_0)$. This is 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. The score vector corresponding to $\boldsymbol{\alpha}$ is:

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

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

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

The $(p+1)\times(p+1)$ block of the Hessian matrix corresponding to $\boldsymbol{\alpha}$ is

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

Evaluating the expectation of this block at $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_o\ \alpha_0^*\ \mathbf{0}')'$ and noting that $(\sigma^*)^2 = h(\alpha_0^*)$ we have

$$\left(\frac{c^2}{2[(\sigma^*)^2]^2}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t')\right),$$

which, apart from the constant c, can be 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 the results above when the information matrix equality holds.

Setting $d_t = \hat{\epsilon}_t^2/\ddot{\sigma}_T^2 - 1$, the LM statistic is

$$\begin{aligned}\mathcal{LM}_T &= \left(\sum_{t=1}^{T} d_t\boldsymbol{\zeta}_t'\right)\left(\sum_{t=1}^{T}\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t'\right)^{-1}\left(\sum_{t=1}^{T}\boldsymbol{\zeta}_t d_t\right)/2 \\ &= \boldsymbol{d}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{d}/2 \\ &\xrightarrow{D} \chi^2(p),\end{aligned}$$

where $\boldsymbol{d}$ is $T\times 1$ with the t^{th} element d_t, and $\boldsymbol{Z}$ is the $T\times(p+1)$ matrix with the t^{th} row $\boldsymbol{\zeta}_t'$. It can be seen that the numerator of the LM statistic is the (centered) regression sum of squares (RSS) of regressing d_t on $\boldsymbol{\zeta}_t$. This shows that the Breusch-Pagan test can also be 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 capable of explaining the square of the (standardized) OLS residuals. It is also interesting to see that the value of c and the functional form of h do not matter in deriving the

statistic. The latter feature makes the Breusch-Pagan test a general test for conditional heteroskedasticity.

Koenker (1981) noted that under conditional normality, $\sum_{t=1}^T d_t^2/T \xrightarrow{\mathbb{P}} 2$. Thus, a test that is more robust to non-normality and asymptotically equivalent to the Breusch-Pagan test is to replace the denominator 2 with $\sum_{t=1}^T d_t^2/T$. This robust version can be expressed as

$$\mathcal{LM}_T = T[\boldsymbol{d}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{d}/\boldsymbol{d}'\boldsymbol{d}] = TR^2,$$

where R^2 is the (centered) R^2 from regressing d_t on $\boldsymbol{\zeta}_t$. This is also equivalent to the centered R^2 from regressing $\hat{\epsilon}_t^2$ 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.

Example 9.12 (Breusch-Godfrey) Given the specification $y_t \mid \boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$, suppose that one would like to check if the errors are serially correlated. Consider first the AR(1) errors: $y_t - \boldsymbol{x}_t'\boldsymbol{\beta} = \rho(y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta}) + u_t$ with $|\rho| < 1$ and $\{u_t\}$ a white noise. The null hypothesis is $\rho^* = 0$, i.e., no serial correlation. It can be seen that a general specification that allows for serial correlations is

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

The constrained specification is just the standard linear regression model $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}$. Testing the null hypothesis that $\rho^* = 0$ is thus equivalent to testing whether an additional variable $y_{t-1} - \boldsymbol{x}_{t-1}\boldsymbol{\beta}$ should be included in the mean specification. In the light of Example 9.10, if the information matrix equality holds, the LM test can be computed as TR^2, where R^2 is obtained from regressing the OLS residuals $\hat{\epsilon}_t$ on $\boldsymbol{x}_t$ and

$\hat{\epsilon}_{t-1} = y_{t-1} - \boldsymbol{x}_{t-1}\hat{\boldsymbol{\beta}}_T$. This is precisely the Breusch (1978) and Godfrey (1978) test for AR(1) errors with the limiting $\chi^2(1)$ distribution.

The test above can be extended straightforwardly to check AR(p) errors. By regressing $\hat{\epsilon}_t$ on $\boldsymbol{x}_t$ and $\hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-p}$, the resulting TR^2 is the LM test when the information matrix equality holds and has a limiting $\chi^2(p)$ distribution. Such tests are known as the Breusch-Godfrey test. 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

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

The null hypothesis is $\alpha^* = 0$. It is readily seen that the resulting LM test is identical to that for AR(1) errors. Thus, the Breusch-Godfrey test for MA(p) errors is the same as that for AR(p) errors. □

Remarks:

1. The Breusch-Godfrey tests discussed above are obtained under the condition that the information matrix equality holds. We have seen that the information matrix equality may fail when, for example, there is neglected conditional heteroskedasticity. Thus, the Breusch-Godfrey tests are not valid when conditional heteroskedasticity is present.

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 compares the performance of the constrained and unconstrained specifications based on their likelihood ratio:

$$\mathcal{LR}_T = -2T[L_T(\ddot{\boldsymbol{\theta}}_T) - L_T(\tilde{\boldsymbol{\theta}}_T)]. \tag{9.21}$$

Utilizing the relationship between the constrained and unconstrained QMLEs (9.16) and the relationship between the Lagrangian multiplier and unconstrained QMLE (9.17), we can write

$$\begin{aligned}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \ddot{\boldsymbol{\theta}}_T) &= \boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'\sqrt{T}\ddot{\boldsymbol{\lambda}}_T + o_{\mathbb{P}}(1) \\ &= \boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'\end{aligned} \tag{9.22}$$

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

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

where the second equality follows because $\nabla L_T(\tilde{\boldsymbol{\theta}}_T) = \mathbf{0}$. It can be seen that the right-hand side is essentially the Wald statistic when $-\boldsymbol{R}\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{R}'$ is the normalizing variance-covariance matrix. This leads to the following distribution result for the LR test.

Theorem 9.13 *Suppose that Theorem 9.2 for the QMLE $\tilde{\boldsymbol{\theta}}_T$ and the information matrix equality both hold. Then under the null hypothesis,*

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

where $\mathcal{LR}_T$ is defined in (9.21) *and q is the number of rows of $\boldsymbol{R}$.*

Theorem 9.13 differs from Theorem 9.7 and Theorem 9.9 in that it also requires the validity of the information matrix equality. When the information matrix equality does not 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. This result clearly indicates that, when L_T is constructed by specifying density functions for $\{y_t \mid x_t\}$, the LR test is not robust to dynamic misspecification and misspecifications of other conditional attributes, such as neglected conditional heteroskedasticity. By contrast, the Wald and LM tests can be made robust by employing a proper normalizing variance-covariance matrix.

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 Prove equations (9.11) and (9.12) in Example 9.5.

9.3 In Example 9.5, suppose there is dynamic misspecification. What is $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$?

9.4 In Example 9.6, what is $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$? Show that

$$\boldsymbol{B}_T(\tilde{\boldsymbol{\theta}}_T) = \begin{bmatrix} \frac{\sum_{t=1}^T \hat{e}_t^2 \boldsymbol{x}_t \boldsymbol{x}_t'}{T(\tilde{\sigma}_T^2)^2} & \mathbf{0} \\ \mathbf{0} & -\frac{1}{4(\tilde{\sigma}_T^2)^2} + \frac{\sum_{t=1}^T \hat{e}_t^4}{T 4(\tilde{\sigma}_T^2)^4} \end{bmatrix}$$

is a consistent estimator for $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$.

9.5 Consider the specification $y_t \mid \boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}, h(\zeta_t'\boldsymbol{\alpha}))$. What condition would ensure that $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ and $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$ are block diagonal?

9.6 Consider the specification $y_t \mid \boldsymbol{x}_t, y_{t-1} \sim \mathcal{N}(\gamma y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta}, \sigma^2)$ 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.

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.

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.

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.

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.