Asymptotic Variance of Test Statistics in ML and QML Frameworks*

Anil K. Bera† Osman Doğan‡ Süleyman Taşpınar§

August 15, 2017

Abstract

In this study, we consider test statistics that can be written as the sample averages of data and derive their limiting distribution under the maximum likelihood (ML) and the quasi-maximum likelihood (QML) frameworks. We first generalize the asymptotic variance formula suggested in Pierce (1982) in the ML framework and illustrate its applications through some well-known test statistics: (i) the skewness statistic, (ii) the kurtosis statistic, (iii) the Cox's statistic, (iv) the information matrix test statistic, and (v) the Durbin's h-statistic. We next provide a similar result in the QML setting and illustrate its applications by providing two examples. Illustrations show the simplicity and the effectiveness of our results for the asymptotic variance of test statistics, and therefore, they are recommended for practical applications.

JEL-Classification: C13, C21, C31.
Keywords: Variance, Asymptotic variance, MLE, QMLE, Inference, Test statistics, Skewness statistic, Kurtosis statistic, The Cox's statistic, The information matrix test, the Durbin's h-statistic.

*An earlier version of this paper was presented in a seminar at the Stat-Math Unit (SMU) of the Indian Statistical Institute (ISI), Kolkata, in July 2017. We are grateful to the seminar participants for constructive comments and suggestions. Any remaining shortcomings and errors are, of course, ours.

†Economics Program, University of Illinois, Illinois, United States, email: abera@illinois.edu.

‡Economics Program, University of Illinois, Illinois, United States, email: odogan@illinois.edu.

§Economics Program, Queens College, The City University of New York, United States, email: staspinar@qc.cuny.edu.

1 Introduction

Taking account of the nuisance parameters, particularly in the context of hypotheses testing, is an age-old problem in statistics. Attempts to solve this problem goes back as far as Student (1908) who explored and solved a *finite sample* problem relating to testing for mean with the variance as the nuisance parameter. The problem persists also *asymptotically*; for instance, as investigated in a series of paper by Neyman (1935, 1957, 1959). One of the outcomes of Neyman's research effort was his celebrated $C(\alpha)$ test where the nuisance parameters are replaced with $\sqrt{n}$-consistent estimators without changing the underlying asymptotic distribution of the test statistic.

Similarly, Pierce (1982) considered a "simpler" but immensely important practical problem of investigating the effect of replacing the unknown parameter θ_0 by its *efficient estimator* $\widehat{\theta}$ in a test statistic $T(\mathbf{y}, \theta_0)$, where $\mathbb{E}(T(\mathbf{y}, \theta_0))$ is free of θ_0. He provided an attractive practical solution along with the condition when no adjustment to variance will be necessary for estimation of θ_0. Quite coincidently, many of testing problems in econometrics , both old and recent ones, fall under Pierce (1982) framework. However, we see hardly any reference to Pierce (1982) in the econometric literature. Some exceptions are Newey and McFadden (1994), Bera and Zuo (1996), Tse (2002), Prokhorov and Schmidt (2009), Andreou and Werker (2010) and Gorodnichenko et al. (2012). It appears that econometricians have tackled the problems of nuisance parameters in testing *case by case* through extensive derivations and complex algebra. For instance, take the case of Cox (1961, 1962) statistic for separate families of hypotheses. White (1982b), for the first time, provided rigorous derivation of the asymptotic distribution of Cox statistic. We show the essential part of White (1982b) results can be easily obtained using Pierce (1982). The same can be said about, for example, Durbin (1970) h-test, White (1982a) information matrix test and Jarque and Bera (1987) skewness and kurtosis statistics. Given such ubiquitous occurrence of the same issue, it seems to be very important to bring Pierce's work to the forefront of econometrics.

As we mentioned, Pierce (1982) replaced the unknown parameter θ_0 by an efficient estimator, such as maximum likelihood estimator (MLE). When the true data generating process (DGP) is unknown, the best we can hope for having a quasi MLE (QMLE) after assuming a parametric distribution $F(y, \theta_0)$. White (1982a) suggested a sandwich variance formula for QMLE; the same formula also has been attributed to Eicker (1963, 1967) and Huber (1967). However, its history goes back to Koopmans et al. (1950). They derived the sandwich formula as a part of the large sample properties of the full-information MLE (FMLE) of the parameters of simultaneous equation system. It is fairly safe to say that almost all likelihood based estimators are QMLE, as the true DGP is rarely known. Thus, a natural progression would be to generalize Pierce (1982) when θ_0 in $T(\mathbf{y}, \theta_0)$ is replaced by QMLE rather than MLE. In this study, we provide such a result with practical illustrations. Similar results are also considered in Newey (1985a,b) and Tauchen (1985). In these studies, the authors do not focus on a general formula of the asymptotic variance of test statistics, instead they show how to conduct certain tests through auxiliary regressions. However, our focus is on the general variance formula and its practical implications for well-known test statistics.

The rest of this paper is organized as follows. In Section 2, we revisit Pierce (1982) and

generalize his result for certain type of test statistics in the ML framework. In the following sub-section, Section 2.1, we illustrate the practical aspect of our results by providing examples on the well-known test statistics: the skewness statistic, the kurtosis statistic, the Cox's statistic, the information matrix statistic, and the Durbin's h-statistic. In Section 3, we generalize Pierce (1982) result to the QMLE setting considered by White (1982a). In the following sub-section, Section 3.1, we apply our result to skewness and kurtosis statistics under distributional misspecification. We conclude in Section 4. Some technical results are relegated to an appendix.

2 The Asymptotic Variance of Statistics Based on MLE

In this section, we first state the assumptions needed to characterize the true DGP and define the MLE in a general setting by following White (1982a). We next define the test statistic and state the regularity conditions that are required for its limiting distribution. Our main interest is to generalize Pierce (1982) and to establish the limiting distribution of test statistic formulated with the MLE.

Assumption 1. *The random variables y_i, for $i = 1, \dots n$, are i.i.d with common joint distribution function $F(y, \theta)$, where θ is a $p \times 1$ parameter vector, on a measurable space $(\mathfrak{A}, \mathfrak{F})$, where y_i's assume values in $\mathfrak{A}$ and $\mathfrak{F}$ is the relevant sigma algebra. Let ν be a measure defined on $(\mathfrak{A}, \mathfrak{F})$ such that it dominates F. Given F, there exists a measurable non-negative Radon-Nikodym density $f(y, \theta) = dF(y, \theta)/d\nu$.*

Assumption 2. *(i) Let Θ be a compact subset of $\mathbb{R}^p$. The density function $f(y, \theta)$ is measurable with respect to $y \in \mathfrak{A}$ and continuous with respect to all $\theta \in \Theta$. (ii) $|\log f(y, \theta)| \leq m(y)$ for all $\theta \in \Theta$ and all $y \in \mathfrak{A}$, where m is an integrable function with respect to F. (iii) $\mathbb{E}[\log f(y, \theta)]$ has unique maximum at θ_0.*

Under Assumptions 1 and 2, the log-likelihood function generated by $F(y, \theta)$ is defined by

$$l(\mathbf{y}, \theta) = \sum_{i=1}^{n} \log f(y_i, \theta). \tag{2.1}$$

Then, the MLE is defined by $\widehat{\theta} = \arg\max_{\theta \in \Theta} l(\mathbf{y}, \theta)$. Assumptions 1 and 2 ensure the existence of the MLE. Also under these two assumptions, it can be shown that $\widehat{\theta} = \theta_0 + o_p(1)$. To establish asymptotic normality for the MLE, we need the following additional assumptions.

Assumption 3. *(i) The first order derivatives $\partial \log f(y, \theta)/\partial \theta_j$, for $j = 1, \dots, k$, are $\mathfrak{F}$ measurable for $\theta \in \Theta$, and continuously differentiable function of θ for all $y \in \mathfrak{A}$. (ii) There are integrable functions with respect to F for all $y \in \mathfrak{A}$ and $\theta \in \Theta$, that dominate $\left|\partial^2 \log f(y, \theta)/\partial \theta_i \partial \theta_j\right|$ and $|\partial \log f(y, \theta)/\partial \theta_i \cdot \partial \log f(y, \theta)/\partial \theta_j|$ for $i, j = 1, \dots, p$.*

Assumption 4. *The interior of Θ contains θ_0.*

Using Assumption 3, we can define the following two information matrices:

$$A(\theta) = \left\{ -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i,\theta)}{\partial\theta_i \partial\theta_j} \right\}, \quad \mathcal{A}(\theta) = \left\{ -\mathbb{E}\left(\frac{\partial^2 \log f(y,\theta)}{\partial\theta_i \partial\theta_j} \right) \right\}. \tag{2.2}$$

The asymptotic distribution of $\widehat{\theta}$ is based on the first order Taylor expansion of $\frac{1}{n}\frac{\partial l(\mathbf{y},\widehat{\theta})}{\partial\theta}$ around θ_0, which yields

$$\sqrt{n}\big(\widehat{\theta} - \theta_0\big) = \mathcal{A}^{-1}(\theta_0)\frac{1}{\sqrt{n}}\frac{\partial l(\mathbf{y},\theta_0)}{\partial\theta} + o_p(1). \tag{2.3}$$

Our stated assumptions ensure that $\frac{1}{\sqrt{n}}\frac{\partial l(\mathbf{y},\theta_0)}{\partial\theta} \xrightarrow{d} N\left[0,\ \mathcal{A}(\theta_0)\right]$ and $A(\widehat{\theta}) = \mathcal{A}(\theta_0) + o_p(1)$. Thus, under the assumption that $\mathcal{A}(\theta_0)$ is non-singular, we have $\sqrt{n}\big(\widehat{\theta} - \theta_0\big) \xrightarrow{d} N\left[0,\ \mathcal{A}^{-1}(\theta_0)\right]$.

Following Huber (1967), we consider the test statistics that can be written as the sample averages of data. Specifically, we consider:

$$T(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \rho(y_i,\widehat{\theta}), \tag{2.4}$$

where $\rho(y_i,\theta)$ is a real valued function defined on $\mathfrak{A}\times\Theta$, satisfying the following assumption.

Assumption 5. *(i) $\rho(y,\theta)$ is $\mathfrak{F}$-measurable for all $\theta\in\Theta$ and continuously differentiable for all $\theta\in\Theta$. (ii) $\mathbb{E}(\rho(y,\theta_0)) = \int \rho(y,\theta_0)dF(y,\theta_0) = \bar{\rho}$ is independent of θ_0. (iii) The first order derivatives $\partial\rho(y,\theta)/\partial\theta_j$ for $j = 1,\ldots,p$ are $\mathfrak{F}$ measurable for all $\theta\in\Theta$, and are dominated by integrable function with respect to F.*

Assumption 5 characterizes the statistics that we considered in this paper; in particular, 5(ii) plays a role in simplifying the variance formula, and 5(iii) allows the interchange of the order of differentiation and integration in our analysis, and ensures that the asymptotic variance of test statistic can be consistently estimated.

First, we present a short review of Pierce (1982). Pierce (1982) starts with the assumption that $\sqrt{n}(T(\theta_0) - \bar{\rho})$ and $\sqrt{n}(\widehat{\theta} - \theta_0)$ has a joint asymptotic multivariate normal distribution and derives a general asymptotic variance formula for $\sqrt{n}T(\widehat{\theta})$. Pierce (1982) sets $\bar{\rho} = 0$ in Assumption 5, and starts with the following joint asymptotic multivariate normal distribution:

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_0) \\ \sqrt{n}T(\theta_0) \end{pmatrix} \xrightarrow{d} N\left[0,\ \begin{pmatrix} \mathcal{A}^{-1}(\theta_0) & \mathcal{M}'(\theta_0) \\ \mathcal{M}(\theta_0) & \mathcal{C}(\theta_0) \end{pmatrix}\right], \tag{2.5}$$

where $\mathcal{M}(\theta_0)$ is the asymptotic covariance between $\sqrt{n}T(\theta_0)$ and $\sqrt{n}(\widehat{\theta} - \theta_0)$. A first order Taylor expansion of $T(\widehat{\theta})$ around θ_0 gives

$$\sqrt{n}T(\widehat{\theta}) = \sqrt{n}T(\theta_0) + \mathcal{D}(\theta_0)\sqrt{n}(\widehat{\theta} - \theta_0) + o_p(1), \tag{2.6}$$

where $\mathcal{D}(\theta_0) = \mathbb{E}\left(\frac{\partial \rho(y,\theta_0)}{\partial \theta'}\right)$. Thus, we have

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = \mathcal{C}(\theta_0) + \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{D}'(\theta_0) + \mathcal{M}(\theta_0)\mathcal{D}'(\theta_0) + \mathcal{D}(\theta_0)\mathcal{M}'(\theta_0). \tag{2.7}$$

With the assumption that $\mathbb{E}\left(\rho(y,\theta_0)\right)$ is independent of θ_0, Pierce (1982) demonstrates that

$$\text{Cov}\left(\sqrt{n}T(\theta_0),\, \frac{1}{\sqrt{n}}\frac{\partial l(\mathbf{y},\,\theta_0)}{\partial \theta}\right) = -\mathcal{D}(\theta_0). \tag{2.8}$$

Since from (2.3), $\mathcal{A}^{-1}(\theta_0)\frac{1}{\sqrt{n}}\frac{\partial l(\mathbf{y},\theta_0)}{\partial \theta}$ is asymptotically equivalent to $\sqrt{n}\big(\widehat{\theta} - \theta_0\big)$,

$$\begin{aligned} -\mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0) &= \text{Cov}\left(\sqrt{n}T(\theta_0),\, \mathcal{A}^{-1}(\theta_0)\frac{1}{\sqrt{n}}\frac{\partial l(\mathbf{y},\,\theta_0)}{\partial \theta}\right) \\ &\approx \text{Cov}\left(\sqrt{n}T(\theta_0),\, \sqrt{n}\big(\widehat{\theta} - \theta_0\big)\right) \\ &= \mathcal{M}(\theta_0), \end{aligned} \tag{2.9}$$

leading to

$$\mathcal{M}(\theta_0) \approx -\mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0). \tag{2.10}$$

Using (2.10) in (2.7), we have the asymptotic variance formula stated in Pierce (1982):

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = \mathcal{C}(\theta_0) - \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{D}'(\theta_0). \tag{2.11}$$

Following the asymptotic arguments given in Huber (1967) and White (1980), we generalize Pierce (1982) by *directly* establishing the joint asymptotic distribution of $\sqrt{n}(T(\widehat{\theta}) - \bar{\rho})$ and $\sqrt{n}(\widehat{\theta} - \theta_0)$, as stated in the following proposition.

Proposition 1. *Under Assumptions 1-5, the joint limiting distribution of* $\sqrt{n}(T(\widehat{\theta}) - \bar{\rho})$ *and* $\sqrt{n}(\widehat{\theta} - \theta_0)$ *is given by*

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_0) \\ \sqrt{n}(T(\widehat{\theta}) - \bar{\rho}) \end{pmatrix} \xrightarrow{d} N\left[0,\, \begin{pmatrix} \mathcal{A}^{-1}(\theta_0) & 0 \\ 0 & \mathcal{C}(\theta_0) - \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{D}'(\theta_0) \end{pmatrix}\right], \tag{2.12}$$

where

$$\mathcal{C}(\theta_0) = \mathbb{E}\left((\rho(y,\theta_0) - \bar{\rho}) \times (\rho(y,\theta_0) - \bar{\rho})'\right) \quad \textit{and} \quad \mathcal{D}(\theta_0) = \mathbb{E}\left(\frac{\partial \rho(y,\,\theta_0)}{\partial \theta'}\right). \tag{2.13}$$

Proof. See Appendix A.1. □

Here we should note two important points. First, Proposition 1 indicates that the asymptotic covariance between $\sqrt{n}\left(T(\widehat{\theta}) - \bar{\rho}\right)$ and $\sqrt{n}\left(\widehat{\theta} - \theta_0\right)$ is zero. This result highlights the fact that the MLE $\widehat{\theta}$ is an efficient estimator. For details, see Rao (1973, Section 5a.2). Second, the simple

asymptotic covariance matrix of $\sqrt{n}(T(\widehat{\theta})-\bar{\rho})$ is the same as that in Pierce (1982). As shown in the proof, this simplified asymptotic variance formula can be derived in two alternative ways. Consider the following covariance between the score and test indicator

$$\mathcal{P}(\theta_0) = \mathbb{E}\left(\frac{\partial \log f(y,\theta_0)}{\partial\theta} \times (\rho(y,\theta_0)-\bar{\rho})^{'}\right).$$

In the first approach that we use in our proof in Appendix A.1, the assumption that $\mathbb{E}\left(\rho(y,\,\theta_0)\right)=\bar{\rho}$ is free of θ_0 leads to the result $\mathcal{P}^{'}(\theta_0) = -\mathcal{D}(\theta_0)$, which can be considered as a type of information matrix equality result (Neyman 1959, p.217, Equation (14)). This equality directly implies the asymptotic covariance matrix of $\sqrt{n}(T(\widehat{\theta})-\bar{\rho})$ in (2.12). A second approach can be based on the efficiency argument of the MLE $\widehat{\theta}$. In (A.6), we have the asymptotic covariance matrix between $\sqrt{n}(\widehat{\theta}-\theta_0)$ and $\sqrt{n}(T(\widehat{\theta})-\bar{\rho})$ as

$$\mathcal{V}(\theta_0) = \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0) + \mathcal{P}^{'}(\theta_0)\mathcal{A}^{-1}(\theta_0),$$

which is zero, and that implies the equality $\mathcal{P}^{'}(\theta_0) = -\mathcal{D}(\theta_0)$.

Finally, under our assumptions, consistent estimators for the components of asymptotic variance-covariance matrix in (2.12) can be constructed. We may use the plug-in method when $\mathcal{C}(\theta_0)$ and $\mathcal{D}(\theta_0)$ in (2.13) have closed forms. Moreover, the i.i.d property of data ensures that they can be estimated by their corresponding sample counterparts as in the following proposition.

Proposition 2. *Consider the following sample moments:*

$$D(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial\rho(y_i,\,\theta)}{\partial\theta}\Big|_{\widehat{\theta}}, \tag{2.14}$$

$$C(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left((\rho(y_i,\widehat{\theta})-T(\widehat{\theta})) \times (\rho(y_i,\widehat{\theta})-T(\widehat{\theta}))^{'}\right). \tag{2.15}$$

Then, under our assumptions, we have $D(\widehat{\theta}) = \mathcal{D}(\theta_0) + o_p(1)$ *and* $C(\widehat{\theta}) = \mathcal{C}(\theta_0) + o_p(1)$.

Proof. See Appendix A.2. □

2.1 Illustrations

In this section, we provide examples that illustrate the practical applications of the general result stated in Proposition 1. We use the asymptotic variance formula developed in the previous section to formulate the asymptotic variance of the following well-known statistics: (i) the skewness statistic, (ii) the kurtosis statistic, (iii) the Cox's statistic, (iv) the information matrix statistic, and (v) the Durbin's h-statistic.

2.1.1 The Skewness Statistic

Consider the following data generating process

$$y_i = x_i^{'}\beta_0 + \varepsilon_i, \tag{2.16}$$

where x_i is the $k \times 1$ vector of exogenous variables and ε_i is an i.i.d normal random variable with mean zero and variance σ_0^2. Let $\theta_0 = (\beta_0^{'}, \sigma_0^2)^{'}$ be the true parameter vector and $\theta = (\beta^{'}, \sigma^2)^{'}$ be an arbitrary vale of parameter vector in the parameter space. The log-likelihood function of (2.16) is

$$l(\mathbf{y},\, \theta) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2(\theta), \tag{2.17}$$

where $\varepsilon_i(\theta) = y_i - x_i^{'}\beta$. The first and second order conditions are

$$\begin{aligned}
&\frac{\partial l(\mathbf{y},\theta)}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\varepsilon_i(\theta)x_i, \quad \frac{\partial l(\mathbf{y},\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\varepsilon_i^2(\theta), \\
&\frac{\partial l(\mathbf{y},\theta)}{\partial \beta \partial \beta^{'}} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i x_i^{'}, \quad \frac{\partial l(\mathbf{y},\theta)}{\partial \beta \partial \sigma^2} = -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}\varepsilon_i(\theta)x_i, \\
&\frac{\partial l(\mathbf{y},\theta)}{\partial \sigma^2 \partial \sigma^2} = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}\varepsilon_i^2(\theta), \quad \frac{\partial l(\mathbf{y},\theta)}{\partial \sigma^2 \partial \beta^{'}} = -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}\varepsilon_i(\theta)x_i^{'}.
\end{aligned} \tag{2.18}$$

Let $X = (x_1, \ldots, x_n)^{'}$ be $n \times k$ matrix of exogenous variables. We assume that $Q_x = \lim_{n\to\infty}\frac{1}{n}X^{'}X$ exists and is nonsingular. Then, it can be shown that

$$\mathcal{A}(\theta_0) = \begin{pmatrix} \frac{1}{\sigma_0^2}Q_x & 0_{k\times 1} \\ 0_{k\times 1} & \frac{1}{2\sigma_0^4} \end{pmatrix}. \tag{2.19}$$

Let $\widehat{\theta} = \arg\max_{\theta\in\Theta} l(\mathbf{y},\theta)$ be the ML estimator. Then, the skewness statistic can be expressed as

$$T(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\rho(y_i, \widehat{\theta}), \quad \text{where} \quad \rho(y_i, \widehat{\theta}) = \widehat{\varepsilon}_i^3/\widehat{\sigma}^3. \tag{2.20}$$

Define $\xi_i = \varepsilon_i/\sigma$ as the i.i.d random variable with mean zero and unit variance. Then, the variance of $\sqrt{n}T(\theta_0)$ is

$$\mathcal{C}(\theta_0) = \mathrm{Var}(\xi_i^3) = \mathbb{E}\left[\xi_i^3 - \mathbb{E}(\xi_i^3)\right]^2 = \mu_6\mu_2^{-3} = 15, \tag{2.21}$$

where $\mu_6 = \mathbb{E}(\varepsilon_i^6)$ and $\mu_2 = \mathbb{E}(\varepsilon_i^2) = \sigma_0^2$. Simple calculations show that

$$\mathcal{D}(\theta_0) = \begin{pmatrix} -\frac{3}{\sigma_0}M_x^{'} & 0 \end{pmatrix}. \tag{2.22}$$

where $M_x = \lim_{n\to\infty} \frac{1}{n} X' l_n$ and l_n is the $n \times 1$ vector of ones. Then, using Proposition 1, the asymptotic variance of skewness statistic is

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = 15 - 9M_x' Q_x^{-1} M_x. \tag{2.23}$$

Remark 1. *An estimate of* $M_x' Q_x^{-1} M_x = \left(\lim_{n\to\infty} \frac{1}{n} l_n' X\right) \left(\lim_{n\to\infty} \frac{1}{n} X'X\right)^{-1} \left(\lim_{n\to\infty} \frac{1}{n} X' l_n\right)$ *is* $\frac{1}{n} l_n' X \left(X'X\right)^{-1} X' l_n$. *Consider the regression model* $l_n = X\delta_0 + u$, *where* δ_0 *is* $k \times 1$ *vector of parameters and* u *is an* $n \times 1$ *vector of disturbance term. The OLS estimator is* $\widehat{\delta} = \left(X'X\right)^{-1} X' l_n$. *From the sum of residuals, we get*

$$\begin{aligned} & l_n' \widehat{u} = l_n'(l_n - X\widehat{\delta}) = l_n' \left(l_n - X\left(X'X\right)^{-1} X' l_n\right) = 0 \\ & \quad \Longrightarrow \; l_n' X \left(X'X\right)^{-1} X' l_n = n. \end{aligned} \tag{2.24}$$

Hence, the estimate of $M_x' Q_x^{-1} M_x$ *is simply* 1.

Using Remark 1, the variance formula in (2.23) simplifies to

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = 15 - 9M_x' Q_x^{-1} M_x = 6. \tag{2.25}$$

2.1.2 The Kurtosis Statistic

In this section, we investigate the asymptotic variance of kurtosis statistic under the data generating process described in Section 2.1.1. The kurtosis statistic can be expressed as

$$T(\widehat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \rho(y_i, \widehat{\theta}), \quad \text{where} \quad \rho(y_i, \widehat{\theta}) = \widehat{\varepsilon}_i^4 / \widehat{\sigma}^4. \tag{2.26}$$

Note that $\mathbb{E}\left(\rho(y_i, \theta_0)\right) = \mu_4/\sigma_0^4 = 3$, where $\mu_4 = \mathbb{E}(\varepsilon_i^4)$. The variance of $\sqrt{n}T(\theta_0)$ is

$$\mathcal{C}(\theta_0) = \text{Var}(\xi_i^4) = \mathbb{E}\left[\xi_i^4 - \mathbb{E}(\xi_i^4)\right]^2 = \mu_2^{-4}(\mu_8 - \mu_4^2) = 96. \tag{2.27}$$

Simple calculations show that

$$\mathcal{D}(\theta_0) = \begin{pmatrix} 0 & -\frac{2\mu_4}{\sigma_0^6} \end{pmatrix} = \begin{pmatrix} 0 & -6\mu_2^{-1} \end{pmatrix}. \tag{2.28}$$

Then, Proposition 1 implies the following asymptotic variance

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = 96 - \begin{pmatrix} 0 & -6\mu_2^{-1} \end{pmatrix} \begin{pmatrix} \sigma_0^2 Q_x^{-1} & 0_{k\times 1} \\ 0_{1\times k} & 2\sigma_0^4 \end{pmatrix} \begin{pmatrix} 0 \\ -6\mu_2^{-1} \end{pmatrix} = 24. \tag{2.29}$$

2.1.3 The Cox's Statistic

In this section, we consider the derivation of the asymptotic variance formula for the Cox's test statistic (Cox 1961, 1962). We assume that we have i.i.d observations $y_1, \ldots, y_n$, and we aim to test the null hypothesis H_f that y_i has a density $f(y, \theta)$ for some $\theta \in \Theta$ against the alternative that y_i has a density function $h(y, \gamma)$ for some $\gamma \in \Gamma$, where Γ is a compact parameter space. Assume that θ is $k \times 1$ and γ is $p \times 1$ vectors of parameters. Under H_f, let $\widehat{\theta}$ be the MLE and $\widehat{\gamma}$ be the QMLE. We assume that $\widehat{\theta} = \theta_0 + o_p(1)$ and $\widehat{\gamma} = \gamma_0 + o_p(1)$. Note that the consistency of QMLE $\widehat{\gamma}$ can be verified by adopting regularity conditions listed in Section 3 for $h(y, \gamma)$. Let $\delta = (\theta', \gamma')'$ be the combined parameter vector of dimension $(k + p) \times 1$. Then, the Cox's test statistic is given by

$$T(\widehat{\delta}) = \frac{1}{n} \sum_{i=1}^{n} \rho(y_i, \widehat{\delta}), \tag{2.30}$$

where

$$\rho(y_i, \widehat{\delta}) = \log \left(f(y_i, \widehat{\theta}) / h(y_i, \widehat{\gamma}) \right) - \int \log \left(f(y, \widehat{\theta}) / h(y, \widehat{\gamma}) \right) f(y, \widehat{\theta}) d\nu. \tag{2.31}$$

Note that under H_f, we have $\mathbb{E}(\rho(y, \delta_0)) = 0$, where expectation is taken with respect to $f(y, \theta)$, which implies that

$$\begin{aligned} \mathcal{C}(\delta_0) &= \mathbb{E} \left(\rho(y, \delta_0) \times \rho(y, \delta_0) \right) \\ &= \int \left[\log \left(f(y, \theta_0) / h(y, \gamma_0) \right) \right]^2 f(y, \theta) d\nu - \left[\int \log \left(f(y, \theta_0) / h(y, \gamma_0) \right) f(y, \theta_0) d\nu \right]^2. \end{aligned} \tag{2.32}$$

The gradient of the test statistic is given by

$$\mathcal{D}(\theta_0) = \mathbb{E} \left(\left. \frac{\partial \rho(y, \delta)}{\partial \theta'} \right|_{\delta_0}, \left. \frac{\partial \rho(y, \delta)}{\partial \gamma'} \right|_{\delta_0} \right) = \left(\psi(\delta_0), \phi(\delta_0) \right), \tag{2.33}$$

where $\psi(\delta_0) = \mathbb{E} \left(\left. \frac{\partial \rho(y, \delta)}{\partial \theta'} \right|_{\delta_0} \right)$ and $\phi(\delta_0) = \mathbb{E} \left(\left. \frac{\partial \rho(y, \delta)}{\partial \gamma'} \right|_{\delta_0} \right)$. In order to allow the exchange of the order of differentiation and integration, we adopt the following assumption.

Assumption 6. *$|\partial \log \left(f(y, \theta) / h(y, \gamma) \right) f(y, \theta) / \partial \theta_i|$ and $|\partial \log \left(f(y, \theta) / h(y, \gamma) \right) f(y, \theta) / \partial \gamma_j|$ for $i = 1, \ldots, k$ and $j = 1, \ldots, p$ are dominated for all θ and γ in $\Theta \times \Gamma$ by measurable functions that are integrable with respect to ν.*

Let ∇_δ be gradient operator with respect to δ. Then

$$\begin{aligned} \psi(\delta_0) &= \mathbb{E} \left(\nabla_\theta \log \left(f(y_i, \theta_0) / h(y_i, \gamma_0) \right) - \nabla_\theta \int \log \left(f(y, \theta_0) / h(y, \gamma_0) \right) f(y, \theta_0) d\nu \right) \\ &= \int \nabla_\theta \left(\log f(y, \theta_0) \right) \times \log \left(f(y, \theta_0) / h(y, \gamma_0) \right) f(y, \theta_0) d\nu, \end{aligned} \tag{2.34}$$

where we use the fact that $\nabla_\theta f(y,\,\theta) = \nabla_\theta \left(\log f(y,\,\theta)\right) f(y,\,\theta)$. For $\phi(\delta_0)$, we simply have

$$\phi(\delta_0) = \mathbb{E}\left(\nabla_\gamma \log\left(f(y_i,\theta_0)/h(y_i,\,\gamma_0)\right) - \nabla_\gamma \int \log\left(f(y,\,\theta_0)/h(y,\,\gamma_0)\right) f(y,\,\theta_0) d\nu\right) = 0. \tag{2.35}$$

Hence, the gradient of test statistic is simply given by

$$\mathcal{D}(\theta_0) = \left(\psi(\delta_0),\, 0_{1\times p}\right). \tag{2.36}$$

Under our regularity conditions in Sections 2 and 3 for $f(y,\theta)$ and $h(y,\,\gamma)$, it can be shown that

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_0) \\ \sqrt{n}(\widehat{\gamma} - \gamma_0) \end{pmatrix} \xrightarrow{d} N\left[0_{(k+p)\times 1},\, \begin{pmatrix} \mathcal{A}^{-1}(\theta_0) & \mathcal{C}_{\theta\gamma} \\ \mathcal{C}_{\gamma\theta} & \mathcal{H}(\gamma_0) \end{pmatrix}\right], \tag{2.37}$$

where $\mathcal{C}_{\theta\gamma}$ is the asymptotic covariance between $\sqrt{n}(\widehat{\theta} - \theta_0)$ and $\sqrt{n}(\widehat{\gamma} - \gamma_0)$, and $\mathcal{H}(\gamma_0)$ is the asymptotic covariance of $\sqrt{n}(\widehat{\gamma} - \gamma_0)$. Then, using Proposition 1, we get

$$\begin{aligned} \text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) &= \int \left[\log\left(f(y,\,\theta_0)/h(y,\,\gamma_0)\right)\right]^2 f(y,\,\theta) d\nu - \left[\int \log\left(f(y,\,\theta_0)/h(y,\,\gamma_0)\right) f(y,\,\theta_0) d\nu\right]^2 \\ &\quad - \begin{pmatrix} \psi(\delta_0) & 0_{1\times p} \end{pmatrix} \begin{pmatrix} \mathcal{A}^{-1}(\theta_0) & \mathcal{C}_{\theta\gamma} \\ \mathcal{C}_{\gamma\theta} & \mathcal{H}(\gamma_0) \end{pmatrix} \begin{pmatrix} \psi'(\delta_0) \\ 0_{p\times 1} \end{pmatrix} \\ &= \int \left[\log\left(f(y,\,\theta_0)/h(y,\,\gamma_0)\right)\right]^2 f(y,\,\theta) d\nu - \left[\int \log\left(f(y,\,\theta_0)/h(y,\,\gamma_0)\right) f(y,\,\theta_0) d\nu\right]^2 \\ &\quad + \psi(\delta_0)\mathcal{A}^{-1}(\theta_0)\psi'(\delta_0). \end{aligned} \tag{2.38}$$

The variance formula in (2.38) is the same as with the one stated in White (1982b). In order to get a consistent estimator of (2.38), we need the following condition (See Lemma 2).

Assumption 7. *$\left[log\left(f(y,\,\theta)/h(y,\,\gamma)\right)\right]^2$ is dominated by a measurable function that is integrable with respect to ν for all θ and γ in $\Theta \times \Gamma$.*

Assumptions 6 and 7 and Lemma 2 can be used to show that a consistent estimator of $\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right)$ is given by

$$\begin{aligned} \text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) &= \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(f(y_i,\,\widehat{\theta})/h(y_i,\,\widehat{\gamma})\right)\right]^2 - \left[\int \log\left(f(y,\,\widehat{\theta})/h(y,\,\widehat{\gamma})\right) f(y,\,\widehat{\theta}) d\nu\right]^2 \\ &\quad + \psi(\widehat{\delta})A^{-1}(\widehat{\theta})\psi'(\widehat{\delta}), \end{aligned} \tag{2.39}$$

which the estimator proposed by Cox (1962). To get another consistent estimator of $\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right)$, which avoids evaluation of integrals, we follow White (1982b) and adopt the following assumption.

Assumption 8. *$|\partial \log f(y,\,\theta)/\partial\theta_i \times \log\left(f(y,\,\theta)/h(y,\,\theta)\right)|$ for $i = 1, \ldots, k$ are dominated by measurable functions that are integrable with respect to ν.*

Assumptions 6-8 along with Lemma 2 ensure the following consistent estimator:

$$\begin{aligned}\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(f(y_i,\,\widehat{\theta})/h(y_i,\,\widehat{\gamma})\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\left(f(y_i,\,\widehat{\theta})/h(y_i,\,\widehat{\gamma})\right)\right]^2 \\ + \psi(\widehat{\delta})A^{-1}(\widehat{\theta})\psi^{'}(\widehat{\delta}),\end{aligned} \tag{2.40}$$

where

$$\psi(\widehat{\delta}) = \frac{1}{n}\sum_{i=1}^{n}\nabla_\theta\left(\log f(y_i,\,\widehat{\theta})\right) \times \left[\log\left(f(y_i,\,\widehat{\theta})/h(y_i,\,\widehat{\gamma})\right)\right]. \tag{2.41}$$

2.1.4 The Information Matrix Test

White (1982a) uses the information equivalence to suggest a misspecification test, which is called the information matrix (IM) test. Consider the model specification stated in Section 2. The IM test statistic is given by

$$T(\theta) = \frac{1}{n}\sum_{i=1}^{n}\rho(y_i,\theta), \tag{2.42}$$

where $\rho(y_i,\theta)$ is $q \times 1$ vector containing indicators of interest with a typical element given by

$$\rho_l(y_i,\,\theta) = \frac{\partial \log f(y_i,\theta)}{\partial\theta_i}\cdot\frac{\partial \log f(y_i,\theta)}{\partial\theta_j} + \frac{\partial^2 \log f(y_i,\theta)}{\partial\theta_i\partial\theta_j}. \tag{2.43}$$

Note that under the null hypothesis of no misspecification, we have $\mathbb{E}\left(\rho(y,\theta_0)\right) = 0$, where θ_0 is the true parameter value. Then,

$$\mathcal{C}(\theta_0) = \mathbb{E}\left(\rho(y,\theta_0)\rho^{'}(y,\theta_0)\right). \tag{2.44}$$

We adopt the following assumption for the elements of $\rho(y,\,\theta)$.

Assumption 9. *$\partial\rho_l(y,\,\theta)/\partial\theta_i$ for $l = 1, \ldots, q$ and $i = 1, \ldots, k$ exist and are continuous function of θ for each y.*

Under Assumption 10, the expected value of the gradient of test statistic at the truth is

$$\mathcal{D}(\theta_0) = \left\{\mathbb{E}\left(\nabla_\theta\rho_l(y,\theta_0)\right)\right\}, \quad l = 1, \ldots q. \tag{2.45}$$

Under the null of no misspecification, we have $\sqrt{n}\big(\widehat{\theta}-\theta_0\big) \xrightarrow{d} N\left[0,\,\mathcal{A}^{-1}(\theta_0)\right]$, as shown in Section 2.

Then, applying Proposition 1, we obtain

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = \mathcal{C}(\theta_0) - \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{D}^{'}(\theta_0). \tag{2.46}$$

Now we show that (2.46) is the same as White (1982a, p. 10)'s formula, which is

$$\begin{aligned} V(\theta_0) &= \mathbb{E}\big(\left[\rho(y,\, \theta_0) - \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\nabla_\theta \log f(y,\, \theta_0)\right] \\ &\quad \times \left[\rho(y,\, \theta_0) - \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\nabla_\theta \log f(y,\, \theta_0)\right]^{'} \big) \qquad (2.47) \\ &= \mathbb{E}\left(\rho(y,\, \theta_0)\rho^{'}(y,\, \theta_0)\right) - \mathbb{E}\left(\mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\nabla_\theta \log f(y,\, \theta_0)\rho^{'}(y,\, \theta_0)\right) \\ &\quad - \mathbb{E}\left(\rho(y,\, \theta_0)\nabla_\theta^{'} \log f(y,\, \theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{D}^{'}(\theta_0)\right) \\ &\quad + \mathbb{E}\left(\mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\nabla_\theta \log f(y,\, \theta_0)\nabla_\theta^{'} \log f(y,\, \theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{D}^{'}(\theta_0)\right) \\ &= \mathcal{C}(\theta_0) - \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{D}^{'}(\theta_0), \qquad (2.48) \end{aligned}$$

where we used the fact that $\mathbb{E}\left(\rho(y, \theta_0) \cdot \nabla_\theta \log f(y,\, \theta_0)\right) = -\mathcal{D}(\theta_0)$. A consistent estimator of $V(\theta_0)$ is given by

$$V(\widehat{\theta}) = C(\widehat{\theta}) - D(\widehat{\theta})A^{-1}(\widehat{\theta})D^{'}(\widehat{\theta}), \tag{2.49}$$

where $C(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \rho(y_i, \widehat{\theta})\rho^{'}(y_i, \widehat{\theta})$ and $D(\widehat{\theta}) = \left\{\frac{1}{n}\sum_{i=1}^{n} \partial\rho_l(y_i, \widehat{\theta})/\partial\theta_i\right\}$ for $l = 1, \ldots q$ and $i = 1, \ldots, k$.

2.1.5 The Durbin's h-Statistic

In this section, we consider the Durbin's h-statistic suggested by Durbin (1970) for testing the presence of an autoregressive process in the disturbance terms of a linear regression model that includes lagged dependent variables. The purpose of this illustration is to show that the result in Proposition 1 is general enough and can be applicable to the time series models. Consider the following regression model.

$$y_t = \beta_1 y_{t-1} + \ldots + \beta_r y_{t-r} + \beta_{r+1}x_{1t} + \ldots + \beta_{r+s}x_{st} + u_t, \quad t = 0, 1, \ldots, n, \tag{2.50}$$

$$u_t = \alpha u_{t-1} + \varepsilon_t, \quad t = 1, \ldots, n, \tag{2.51}$$

where $|\alpha| < 1$ is the autoregressive parameter, and ε_t are i.i.d normal random variables with mean zero and variance σ_0^2. As in Durbin (1970), we assume that $y_0, y_{-1} \ldots, y_{-r}$ and $x_{10}, \ldots, x_{s0}$ are known constants, and u_0 is constant but unknown. We consider the null hypothesis $H_0 : \alpha = 0$, The Durbin's h-statistic is based on the following serial correlation statistic:

$$\widehat{\alpha} = \frac{\sum_{t=1}^{n} \widehat{u}_t\widehat{u}_{t-1}}{\sum_{t=1}^{n} \widehat{u}_{t-1}^2}, \tag{2.52}$$

where $\widehat{u}_t = y_t - \widehat{\beta}_1 y_{t-1} - \ldots - \widehat{\beta}_r y_{t-r} - \widehat{\beta}_{r+1} x_{1t} - \ldots - \widehat{\beta}_{r+s} x_{st}$ are least squared residuals. Here, $\sqrt{n}\widehat{\alpha}$ is the statistic of interest and we use Proposition 1 to formulate its asymptotic variance. Then, the asymptotic variance of $\sqrt{n}\alpha$ is given by

$$\text{Var}\left(\sqrt{n}\alpha\right) = 1 - \alpha^2. \tag{2.53}$$

Then, under $H_0 : \alpha = 0$, we have $\text{Var}\left(\sqrt{n}\alpha\right) = 1$. Now, consider

$$\frac{\partial \alpha}{\partial \beta'} = \frac{1-\alpha^2}{n\sigma_0^2} \frac{\partial}{\partial \beta'} \left(\sum_{t=1}^{n} u_t u_{t-1} \right), \tag{2.54}$$

where

$$\frac{\partial}{\partial \beta} \left(\sum_{t=1}^{n} u_t u_{t-1} \right) = \begin{pmatrix} \sum_{t=1}^{n} u_t(-y_{t-2}) + u_{t-1}(-y_{t-1}) \\ \vdots \\ \sum_{t=1}^{n} u_t(-y_{t-r-1}) + u_{t-1}(-y_{t-r}) \\ \sum_{t=1}^{n} u_t(-x_{1t-1}) + u_{t-1}(-x_{1t}) \\ \vdots \\ \sum_{t=1}^{n} u_t(-x_{st-1}) + u_{t-1}(-x_{st}) \\ 0 \end{pmatrix}. \tag{2.55}$$

Using (2.54) and (2.55), we have

$$\mathbb{E}\left(\frac{\partial \alpha}{\partial \beta'} \right) = \begin{pmatrix} 1 & 0 & \ldots & 0 \end{pmatrix}. \tag{2.56}$$

So using Proposition 1, we can write

$$\text{Var}\left(\sqrt{n}\widehat{\alpha}\right) = 1 - n\text{Var}(\widehat{\beta}_1), \tag{2.57}$$

where $\text{Var}(\widehat{\beta}_1)$ is the asymptotic variance of the OLS estimator $\widehat{\beta}_1$. Thus, the Durbin's h-statistic is

$$h = \sqrt{\frac{n\widehat{\alpha}^2}{1 - n\text{Var}(\widehat{\beta}_1)}}, \tag{2.58}$$

which has an asymptotic standard normal distribution.

3 The Asymptotic Variance of Statistics Based on QMLE

In this section, we generalize the Pierce (1982) formula under distributional misspecification, i.e., under the QMLE setting as in White (1982a). We start with the assumption characterizing the true DGP.

Assumption 10. *The random variables y_i, for $i = 1, \ldots n$, are i.i.d with common joint distribution function G on a measurable space $(\mathfrak{A}, \mathfrak{F})$, where y_i's assume values in $\mathfrak{A}$ and $\mathfrak{F}$ is the relevant sigma algebra. Let ν be a measure defined on $(\mathfrak{A}, \mathfrak{F})$ such that it dominates G. Given G, there exists a measurable non-negative Radon-Nikodym density $g = dG/d\nu$.*

Since the true distribution function G is rarely known, we choose to work with a parametric family of distribution functions $\mathcal{F} = \{F(y, \theta)\}$ on $(\mathfrak{A}, \mathfrak{F})$, which may or may not contain G, where θ is a $p \times 1$ vector of parameter. The family $\mathcal{F}$ is correctly specified for y if it contains G, otherwise it is misspecified. $F(y, \theta)$ satisfies the conditions of the following assumption.

Assumption 11. *$F(y, \theta)$ has Radon-Nikodym density $f(y, \theta) = dF(y, \theta)/d\nu$. Let Θ be a compact subset of $\mathbb{R}^p$. The density function $f(y, \theta)$ is measurable with respect to $y \in \mathfrak{A}$ and continuous with respect to all $\theta \in \Theta$.*

Under Assumptions 1 and 2, the quasi-log-likelihood function generated by $F(y, \theta)$ is defined by

$$l(\mathbf{y}, \theta) = \sum_{i=1}^{n} \log f(y_i, \theta). \tag{3.1}$$

Then, the QMLE is defined by $\widehat{\theta} = \arg\max_{\theta \in \Theta} l(\mathbf{y}, \theta)$. If $\mathcal{F}$ contains true distribution function, that is, if $G(y) = F(y, \theta_0)$ for some $\theta_0 \in \Theta$, then the QMLE is just the MLE of θ_0. If $\mathcal{F}$ does not include G, then the QMLE is an estimator of a parameter $\theta_\star$ that minimizes the following Kullback-Leibler Information Criterion (KLIC):

$$\mathbb{I}(g : f, \theta) = \mathbb{E}\left[\log(g(y_i)/f(y_i, \theta)\right] = \int \log(g/f) g d\nu. \tag{3.2}$$

Throughout this section, the expectations are taken with respect to the true distribution function G. The KLIC measures the divergence of f from g, hence the QMLE $\widehat{\theta}$ of $\theta_\star$ minimizes the discrepancy between f and g. To ensure this interpretation for $\widehat{\theta}$ as an estimator of $\theta_\star$, we need the following assumption.

Assumption 12. *(i) $\mathbb{E}(\log g(y))$ exists and $|\log f(y, \theta)| \leq m(y)$ for all $\theta \in \Theta$ and all $y \in \mathfrak{A}$, where m is an integrable function with respect to G. (ii) Identification condition: $\mathbb{I}(g : f, \theta)$ has a unique minimum at pseudo-true value $\theta_\star$ in Θ.*

By Assumption 12, the KLIC is well-defined and $\theta_\star$ is globally identifiable. If the matrix $\mathcal{A}(\theta_\star)$ (see equation (2.2)) is positive definite and if $\theta_\star$ minimizes $\mathbb{I}(g : f, \theta)$ on an open neighborhood $\mathcal{O} \subset \Theta$, then $\theta_\star$ is locally identifiable (White 1982a). This result indicates that if the sample analog $A(\widehat{\theta})$ of $\mathcal{A}(\theta_\star)$ is singular or close to being singular, then we will have an indication for an identification problem.

Now we (re)state Assumption 3 in terms of the true distribution function G.

Assumption 13. *(i) The first order derivatives $\partial \log f(y,\theta)/\partial\theta_j$, for $j = 1,\ldots,k$, are $\mathfrak{F}$ measurable for $\theta \in \Theta$, and continuously differentiable function of θ for all $y \in \mathfrak{A}$. (ii) There are integrable functions with respect to G for all $y \in \mathfrak{A}$ and $\theta \in \Theta$, that dominate $\left|\partial^2 \log f(y,\theta)/\partial\theta_i\partial\theta_j\right|$ and $|\partial \log f(y,\theta)/\partial\theta_i \cdot \partial \log f(y,\theta)/\partial\theta_j|$ for $i,j = 1,\ldots,p$.*

Using Assumption 13, we can define the following matrices:

$$B(\theta) = \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(y,\theta)}{\partial\theta_i} \cdot \frac{\partial \log f(y,\theta)}{\partial\theta_j} \right\}, \quad \mathcal{B}(\theta) = \left\{ \mathbb{E}\left(\frac{\partial \log f(y,\theta)}{\partial\theta_i} \cdot \frac{\partial \log f(y,\theta)}{\partial\theta_j} \right) \right\}. \tag{3.3}$$

Under our stated assumptions, it can be shown that $\widehat{\theta} = \theta_\star + o_p(1)$ (White 1982a). The asymptotic distribution of $\widehat{\theta}$ is based on the first order Taylor expansion of $\frac{1}{n}\frac{\partial l(\mathbf{y},\widehat{\theta})}{\partial\theta}$ around $\theta_\star$, which can be written as

$$\sqrt{n}\big(\widehat{\theta} - \theta_\star\big) = \mathcal{A}^{-1}(\theta_\star)\frac{1}{\sqrt{n}}\frac{\partial l(\mathbf{y},\theta_\star)}{\partial\theta} + o_p(1). \tag{3.4}$$

Our stated assumptions ensure that $\frac{1}{\sqrt{n}}\frac{\partial l(\mathbf{y},\theta_\star)}{\partial\theta} \xrightarrow{d} N\left[0,\ \mathcal{B}(\theta_\star)\right]$, $A(\widehat{\theta}) = \mathcal{A}(\theta_\star) + o_p(1)$ and $B(\widehat{\theta}) = \mathcal{B}(\theta_\star) + o_p(1)$. Thus, (3.4) implies that

$$\sqrt{n}\big(\widehat{\theta} - \theta_\star\big) \xrightarrow{d} N\left[0,\ \mathcal{A}^{-1}(\theta_\star)\mathcal{B}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\right], \tag{3.5}$$

under the assumption that $\mathcal{A}(\theta_\star)$ is non-singular. If the model is correctly specified, that is, $g(y) = f(y,\theta_0)$ for some $\theta_0 \in \Theta$, then $\mathbb{I}(g : f,\theta)$ attains its unique minimum at $\theta_\star = \theta_0$, and thus the QMLE $\widehat{\theta}$ is the consistent estimator of θ_0.

The sandwich formula $\mathcal{A}^{-1}(\theta_\star)\mathcal{B}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)$ is generally attributed to Eicker (1963, 1967) and Huber (1967). For example, Huber (1967, Corollary) derives the sandwich formula for the asymptotic distribution of a consistent estimator under some regularity conditions when data is simply i.i.d., and then establishes the information matrix equivalence for a correctly specified model. However, long before Eicker (1963, 1967) and Huber (1967), Koopmans et al. (1950) derived the sandwich formula while studying the large-sample properties of the MLE of the parameters of the system of structural equations. Koopmans et al. (1950, p. 134) in their Assumption 3.3.1.4, assume joint normality of the disturbances. However, on p.135, they explicitly recognize the possibility that the assumed distribution function has no necessary connection with the distribution of the observations, and wrote: "Nevertheless, we can use the [assumed distribution] function to define parameters by the same maximizing procedure. In these circumstances, we shall call it the quasi-likelihood function, and call the maximizing values of its parameters quasi-maximum-likelihood estimates." Possibly, this is the *first appearance* of the terms "quasi-likelihood function" and "quasi-maximum-likelihood estimates" in the statistics and econometrics literature. In Section 3.3.10, Koopmans et al. (1950) provide "asymptotic sampling variances and covariances of the maximum likelihood estimates," and on page p.150, they derive the sandwich form of the covariance matrix as given in

our equation (3.5).

In the QML framework, our test statistic defined in (2.4) satisfies the conditions of the following assumptions, which is a counterpart of our earlier Assumption 5 under the true distribution function G.

Assumption 14. *(i) $\rho(y,\theta)$ is $\mathfrak{F}$-measurable for all $\theta \in \Theta$ and continuously differentiable for all $\theta \in \Theta$. (ii) $\mathbb{E}(\rho(y,\theta_0)) = \int \rho(y,\theta_\star)g(y)d\nu = \rho_\star$ is independent of $\theta_\star$. (iii) The first order derivatives $\partial\rho(y,\theta)/\partial\theta_j$ for $j = 1,\ldots,p$ are $\mathfrak{F}$ measurable for all $\theta \in \Theta$, and are dominated by integrable function with respect to G.*

Following the asymptotic argument used in Proposition 1, we establish the joint asymptotic distribution of $\sqrt{n}(\widehat{\theta} - \theta_\star)$ and $\sqrt{n}(T(\widehat{\theta}) - \rho_\star)$ in the QML framework, as stated in the following proposition.

Proposition 3. *Under our regularity conditions, the joint limiting distribution of $\sqrt{n}(T(\widehat{\theta}) - \rho_\star)$ and $\sqrt{n}(\widehat{\theta} - \theta_\star)$ is given by*

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_\star) \\ \sqrt{n}(T(\widehat{\theta}) - \rho_\star) \end{pmatrix} \xrightarrow{d} N\left[0, \begin{pmatrix} \mathcal{A}^{-1}(\theta_\star)\mathcal{B}(\theta_\star)\mathcal{A}^{-1}(\theta_\star) & \mathcal{V}^{'}(\theta_\star) \\ \mathcal{V}(\theta_\star) & \mathcal{S}(\theta_\star) \end{pmatrix}\right], \tag{3.6}$$

$$\mathcal{V}(\theta_\star) = \mathcal{D}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{B}(\theta_\star)\mathcal{A}^{-1}(\theta_\star) + \mathcal{P}^{'}(\theta_\star)\mathcal{A}^{-1}(\theta_\star), \tag{3.7}$$

$$\begin{aligned}\mathcal{S}(\theta_\star) = \mathcal{C}(\theta_\star) &+ \mathcal{D}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{B}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{D}^{'}(\theta_\star) + \mathcal{P}^{'}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{D}^{'}(\theta_\star) \\ &+ \mathcal{D}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{P}(\theta_\star).\end{aligned} \tag{3.8}$$

where $\mathcal{D}(\theta_\star) = \mathbb{E}\left(\frac{\partial\rho(y,\theta_\star)}{\partial\theta'}\right)$, $\mathcal{P}(\theta_\star) = \mathbb{E}\left(\frac{\partial \log f(y,\theta_\star)}{\partial\theta} \times (\rho(y,\theta_\star) - \rho_\star)^{'}\right)$, *and* $\mathcal{C}(\theta_\star) = \mathbb{E}\left((\rho(y,\theta_\star) - \rho_\star) \times (\rho(y,\theta_\star) - \rho_\star)^{'}\right)$.

Proof. See Appendix A.3. □

Comparing Propositions 1 and 3, we first note that the off-diagonal block $\mathcal{V}(\theta_\star)$ is not a null matrix. Since there is no information matrix equality under distributional misspecification, $\mathcal{V}(\theta_\star)$ is not a null matrix even if $\mathcal{P}^{'}(\theta_\star) = -\mathcal{D}(\theta_\star)$ in (3.7). Second, when there is correct specification, that is, $G(y) = F(y,\theta_0)$ for some $\theta_0 \in \Theta$, Proposition 3 reduces to Proposition 1. Thus, the simplification in the asymptotic variance $\mathcal{S}(\theta_\star)$ derived in Proposition 1 is possible. Under the distributional misspecification, the argument given for the simplification in Proposition 1 is not applicable, since there is no parametric density function f such that $g(y) = f(y,\theta_0)$. Finally, as shown in Proposition 2, the pertinent sample product moments or plug-in estimators can be formulated to estimate the asymptotic variance in (3.8) in the QML framework.

3.1 Illustrations

In this section, we illustrate the application of the formula stated in (3.8) within the context of two important example: (i) the skewness statistic and (ii) the Kurtosis statistic.

3.1.1 The Skewness Statistic

We consider the regression model stated in Section 2.1.1 under the assumption that ε_i is an i.i.d random variable with mean zero and variance σ_0^2. For notational simplification, we denote the true parameter vector with $\theta_0 = (\beta_0^{'}, \sigma_0^2)^{'}$ even the model is misspecified. Using the first order conditions stated in Section 2.1.1, it can easily be shown that

$$\mathcal{B}(\theta_0) = \begin{pmatrix} \frac{1}{\sigma_0^2} Q_x & \frac{\mu_3}{2\sigma_0^6} M_x \\ \frac{\mu_3}{2\sigma_0^6} M_x^{'} & \frac{\mu_4 - \sigma^4}{4\sigma_0^8} \end{pmatrix}, \tag{3.9}$$

where $\mu_3 = \mathbb{E}(\varepsilon_i^3)$ and $\mu_4 = \mathbb{E}(\varepsilon_i^4)$. Note that under the normal distribution assumption, we have $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$, which lead to $\mathcal{B}(\theta_0) = \mathcal{A}(\theta_0)$. Using (2.19) and (3.9), it can be shown that

$$\mathcal{A}^{-1}(\theta_0)\mathcal{B}(\theta_0)\mathcal{A}^{-1}(\theta_0) = \begin{pmatrix} \sigma_0^2 Q_x^{-1} & \mu_3 Q_x^{-1} M_x \\ \mu_3 M_x^{'} Q_x^{-1} & \mu_4 - \sigma_0^4 \end{pmatrix}. \tag{3.10}$$

Remark 2. *Note that the diagonality of $\mathcal{A}(\theta_0)$ indicates that inference about β_0 based on $\mathcal{A}(\theta_0)$ will be correct even when there is distributional misspecification in the model. However, this is not the case for σ_0^2. Under the distributional misspecification, the correct asymptotic variance of $\sqrt{n}(\widehat{\sigma}^2 - \sigma_0^2)$ is $\mu_4 - \sigma_0^4$, not $2\sigma_0^4$. In addition, the result in* (3.10) *indicates that inference about both β_0 and σ_0^2 based on $\mathcal{B}(\theta_0)$ will be affected and be incorrect under distributional misspecification.*

Let $\widehat{\theta} = (\widehat{\beta}^{'}, \widehat{\sigma}^2)^{'}$ be the QMLE of θ_0. We consider the following skewness statistic.

$$T(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \rho(y_i, \widehat{\theta}), \quad \text{where} \quad \rho(y_i, \widehat{\theta}) = \widehat{\varepsilon}_i^3 / \widehat{\sigma}^3 - \mu_3 \mu_2^{-3/2}. \tag{3.11}$$

Note that $\mathbb{E}\left(\rho(y_i, \theta_0)\right) = \mathbb{E}(\varepsilon_i^3 / \sigma_0^3 - \mu_3 \mu_2^{-3/2}) = 0$. The variance of unfeasible version $\sqrt{n}T(\theta_0)$ is

$$\mathcal{C}(\theta_0) = \mathbb{E}\left((\varepsilon^3 / \sigma_0^3 - \mu_3 \mu_2^{-3/2})^2\right) = \mu_2^{-3}(\mu_6 - \mu_3^2), \tag{3.12}$$

where $\mu_6 = \mathbb{E}(\varepsilon_i^6)$. Simple calculations show that

$$\mathcal{D}(\theta_0) = \begin{pmatrix} -\frac{3}{\sigma_0} M_x^{'} & -\frac{3\mu_3}{2\sigma_0^5} \end{pmatrix}. \tag{3.13}$$

Next, we will find $\mathcal{P}(\theta_0)$. Using the first order conditions stated in Section 2.1.1, we can easily

calculate that

$$\mathcal{P}(\theta_0) = \begin{pmatrix} \frac{\mu_4}{\sigma_0^5} M_x \\ -\frac{\mu_3}{2\sigma_0^5} + \frac{\mu_5}{2\sigma_0^7} \end{pmatrix}. \tag{3.14}$$

Remark 3. *Note that unlike in Illustration 2.1.1, here $\mathcal{D}(\theta_0) \neq -\mathcal{P}'(\theta_0)$, in general. Under the normality of disturbance term, i.e., when there is no distributional misspecification, we have $\mu_4 = 3\sigma_0^4$ and $\mu_3 = \mu_5 = 0$, and we obtain the IM equality of Section 2: $\mathcal{D}(\theta_0) = -\mathcal{P}'(\theta_0)$.*

Using Proposition 3, we obtain

$$\begin{aligned} \text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = \frac{1}{\sigma_0^6}(\mu_6 - \mu_3^2) + 9M_x' Q_x^{-1} M_x + 9\frac{\mu_3^2}{\sigma_0^6} M_x' Q_x^{-1} M_x + \frac{9}{4}\frac{\mu_3^2}{\sigma_0^{10}}(\mu_4 - \sigma_0^4) \\ - 2\left(\frac{3\mu_4}{\sigma_0^4} M_x' Q_x^{-1} M_x - \frac{3\mu_3^2}{2\sigma_0^6} + \frac{3\mu_3\mu_5}{2\sigma_0^8}\right). \end{aligned} \tag{3.15}$$

Using Remark 1, $\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right)$ in (3.15) simplifies to

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = \frac{1}{\sigma_0^6}(\mu_6 - \mu_3^2) + 9 + 9\frac{\mu_3^2}{\sigma_0^6} + \frac{9}{4}\frac{\mu_3^2}{\sigma_0^{10}}(\mu_4 - \sigma_0^4) - \frac{6\mu_4}{\sigma_0^4} + \frac{3\mu_3^2}{\sigma_0^6} - \frac{3\mu_3\mu_5}{\sigma_0^8}. \tag{3.16}$$

Under the null hypothesis of no skewness, i.e., $\mu_3 = \mu_5 = 0$, the above expression further simplifies to

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = \frac{\mu_6}{\sigma_0^6} - \frac{6\mu_4}{\sigma_0^4} + 9. \tag{3.17}$$

Remark 4. *Under no misspecification, i.e., under the normality assumption of disturbance term, we have $\mu_6 = 15\sigma_0^6$, and $\mu_4 = 3\sigma_0^4$. Then, from (3.17), we get $\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = 6$ as shown in Section 2.1.1.*

3.1.2 The Kurtosis Statistic

In this section, we investigate the asymptotic variance of kurtosis statistic when the disturbance terms of the regression model stated in Section 2.1.1 are simply i.i.d with mean zero and variance σ_0^2. We consider the following kurtosis statistic:

$$T(\widehat{\theta}) = \frac{1}{n}\sum_{i=1} \rho(y_i, \widehat{\theta}), \quad \text{where} \quad \rho(y_i, \widehat{\theta}) = \widehat{\varepsilon}_i^4/\widehat{\sigma}^4 - \mu_4/\sigma_0^4, \tag{3.18}$$

where $\widehat{\theta} = (\widehat{\beta}', \widehat{\sigma}^2)'$ is the QMLE of θ_0. Then, the variance of $\sqrt{n}T(\theta_0)$ is

$$\mathcal{C}(\theta_0) = \mathbb{E}\left(\left(\varepsilon^4/\sigma_0^4 - \mu_4/\sigma_0^4\right)^2\right) = \frac{(\mu_8 - \mu_4^2)}{\sigma_0^8}. \tag{3.19}$$

Simple calculations show that

$$\mathcal{D}(\theta_0) = \left(-\frac{4\mu_3}{\sigma_0^4} M_x' \quad -\frac{2\mu_4}{\sigma_0^6} \right). \tag{3.20}$$

Using the first order conditions in Section 2.1.1, we obtain

$$\mathcal{P}(\theta_0) = \begin{pmatrix} \frac{\mu_5}{\sigma_0^6} M_x \\ -\frac{\mu_4}{2\sigma_0^6} + \frac{\mu_6}{2\sigma_0^8} \end{pmatrix}. \tag{3.21}$$

Note that $\mathcal{P}(\theta_0) \neq -\mathcal{D}'(\theta_0)$, in general, and only under symmetry, i.e., when there is no distributional misspecification, we will get the IM type-equality of Section 2. An application of Proposition 3 gives

$$\begin{aligned} \text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = \frac{1}{\sigma_0^8}(\mu_8 - \mu_4^2) + \frac{16\mu_3^2}{\sigma_0^6} M_x' Q_x^{-1} M_x + \frac{16\mu_4\mu_3^2}{\sigma_0^{10}} M_x' Q_x^{-1} M_x + \frac{4\mu_4^2}{\sigma_0^{12}}(\mu_4 - \sigma_0^4) \\ - 2\left(\frac{4\mu_3\mu_5}{\sigma_0^8} M_x' Q_x^{-1} M_x - \frac{2\mu_4^2}{\sigma_0^8} + \frac{2\mu_4\mu_6}{\sigma_0^{10}} \right). \end{aligned} \tag{3.22}$$

Then, by Remark 1, we obtain

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = \frac{1}{\sigma_0^8}(\mu_8 - \mu_4^2) + \frac{16\mu_3^2}{\sigma_0^6} + \frac{16\mu_4\mu_3^2}{\sigma_0^{10}} + \frac{4\mu_4^2}{\sigma_0^{12}}(\mu_4 - \sigma_0^4) - \frac{8\mu_3\mu_5}{\sigma_0^8} + \frac{4\mu_4^2}{\sigma_0^8} - \frac{4\mu_4\mu_6}{\sigma_0^{10}}. \tag{3.23}$$

Under the null hypothesis of no excess kurtosis, i.e., $\mu_4 = 3\sigma_0^4$, $\mu_6 = 15\sigma_0^6$ and $\mu_8 = 105\sigma_0^8$, the above expression simplifies to

$$\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = 64\frac{\mu_3^2}{\sigma_0^6} - 8\frac{\mu_3\mu_5}{\sigma_0^8} + 24. \tag{3.24}$$

Remark 5. *Under no misspecification, i.e., under the normality assumption of disturbance term, we have* $\mu_3 = \mu_5 = 0$. *Then, from* (3.24), *we get* $\text{Var}\left(\sqrt{n}T(\widehat{\theta})\right) = 24$ *as shown in Section 2.1.2.*

4 Conclusion

In this study, we provide the variance formulas for the asymptotic variance of test statistics that can be written as the sample averages of data. We first generalize the Pierce formula in the ML setting and illustrate the practical applications of the formula within the context of some well-known test statistics. We next drive a similar formula in the QML setting for the same type of test statistics and provide two illustrations. We show that the asymptotic covariance between the MLE and the test statistic equals to the expectation of gradient of the test statistic. This information matrix type-equality allows us to simplify the asymptotic variance formula in the ML setting. Since there is no such equality-relations in the QML setting, it is not possible to simplify the asymptotic

variance formula. Our examples clearly indicate the usefulness of asymptotic variance formulas in hypothesis testing.

Appendix

A Some Useful Lemmas

Lemma 1. *Let $g(x,\theta)$ be a continuous function of θ for each x and a measurable function of x for each θ on $\mathcal{X}\times\Theta$, where $\mathcal{X}$ is a Euclidean space and Θ is a compact subset of a Euclidean space. Assume that*

(i) $|g(x,\theta)|\leq h(x)$ for all x and θ, where h is integrable with respect to a probability distribution function F on $\mathcal{X}$.

(ii) $x_1, x_2, \ldots x_n$ is a random sample from F.

Then:

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)\to\int g(x,\theta)dF(x)=\mathbb{E}\big(g(x,\theta)\big),$$

almost everywhere uniformly for all θ in Θ.

Proof. See Jennrich (1969, Theorem 2). □

Lemma 2. *Let Z_i for $i=1,\ldots,n$ be i.i.d random variables assuming values in some set Ψ endowed with a sigma-field $\mathcal{A}$. Let $q:\Psi\times\Theta\to\mathbb{R}$, where $\Theta\subset\mathbb{R}^k$ is compact. Let $Q_n(z,\theta)=\frac{1}{n}\sum_{i=1}^{n} q(z_i,\theta)$ be a measurable function for all $\theta\in\Theta$, and a continuous function of θ for all $z\in\Psi$. Assume that*

(i) $\left|Q_n(z,\theta)-\overline{Q}_n(\theta)\right|\to 0$ a.e uniformly for all $\theta\in\Theta$, where $\overline{Q}_n(\theta)=\mathbb{E}\left(q(z,\theta)\right)$,

(ii) $\widehat{\theta}\to\theta_0$ a.e.,

Then:

$$\left|Q_n(z,\widehat{\theta})-\overline{Q}_n(\theta_0)\right|\to 0,\quad a.e.$$

Proof. This lemma is a simple modification of White (1980, Lemma 2.6). □

A.1 Proof of Proposition 1

The claim can be proved by following the asymptotic argument given in Huber (1967, Corollary, p.231). Define the following vector

$$\xi(y,\theta,\bar{\rho})=\begin{pmatrix}\frac{\partial\log f(y,\theta)}{\partial\theta}\\ \rho(y,\theta)-\bar{\rho}\end{pmatrix}. \tag{A.1}$$

Note that $\operatorname{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\rho(y_i,\widehat{\theta})=\bar{\rho}$ by Lemmas 1 and 2. Thus, under our regularity conditions, we have $\operatorname{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\xi(y_i,\widehat{\theta},\bar{\rho})=0$. Note that

$$\mathbb{E}\left(\nabla_{\theta\bar{\rho}}\xi(y,\theta_0,\bar{\rho})\right)=\mathbb{E}\begin{pmatrix}\frac{\partial^2\log f(y,\theta_0)}{\partial\theta\partial\theta'} & 0\\ \frac{\partial\rho(y,\theta_0)}{\partial\theta'} & -I\end{pmatrix}=\begin{pmatrix}-\mathcal{A}(\theta_0) & 0\\ \mathcal{D}(\theta_0) & -I\end{pmatrix}, \tag{A.2}$$

where $\mathcal{D}(\theta_0) = \mathbb{E}\left(\frac{\partial \rho(y,\theta_0)}{\partial \theta'}\right)$. Also note that

$$\mathbb{E}\left(\xi(y,\theta_0,\bar{\rho}) \times \xi^{'}(y,\theta_0,\bar{\rho})\right) = \begin{pmatrix} \mathcal{A}(\theta_0) & \mathcal{P}(\theta_0) \\ \mathcal{P}^{'}(\theta_0) & \mathcal{C}(\theta_0) \end{pmatrix}, \tag{A.3}$$

where $\mathcal{P}(\theta_0) = \mathbb{E}\left(\frac{\partial \log f(y,\theta_0)}{\partial \theta} \times (\rho(y,\theta_0) - \bar{\rho})^{'}\right)$ and $\mathcal{C}(\theta_0) = \mathbb{E}\left((\rho(y,\theta_0) - \bar{\rho}) \times (\rho(y,\theta_0) - \bar{\rho})^{'}\right)$. Then, an application of Huber (1967, Corollary, p.231) yields:

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_0) \\ \sqrt{n}(T(\widehat{\theta}) - \bar{\rho}) \end{pmatrix} \xrightarrow{d} N\left[0, \begin{pmatrix} \mathcal{A}^{-1}(\theta_0) & \mathcal{V}^{'}(\theta_0) \\ \mathcal{V}(\theta_0) & \mathcal{S}(\theta_0) \end{pmatrix}\right], \tag{A.4}$$

where

$$\begin{pmatrix} \mathcal{A}^{-1}(\theta_0) & \mathcal{V}^{'}(\theta_0) \\ \mathcal{V}(\theta_0) & \mathcal{S}(\theta_0) \end{pmatrix} = \begin{pmatrix} -\mathcal{A}(\theta_0) & 0 \\ \mathcal{D}(\theta_0) & -I \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{A}(\theta_0) & \mathcal{P}(\theta_0) \\ \mathcal{P}^{'}(\theta_0) & \mathcal{C}(\theta_0) \end{pmatrix} \begin{pmatrix} -\mathcal{A}(\theta_0) & \mathcal{D}^{'}(\theta_0) \\ 0 & -I \end{pmatrix}^{-1}.$$

Using the inverse partitioned matrix formula,

$$\begin{pmatrix} -\mathcal{A}(\theta_0) & 0 \\ \mathcal{D}(\theta_0) & -I \end{pmatrix}^{-1} = \begin{pmatrix} -\mathcal{A}^{-1}(\theta_0) & 0 \\ -\mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0) & -I \end{pmatrix}, \tag{A.5}$$

$\mathcal{V}(\theta_0)$ and $\mathcal{S}(\theta_0)$ in (A.4) can be expressed as

$$\mathcal{V}(\theta_0) = \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0) + \mathcal{P}^{'}(\theta_0)\mathcal{A}^{-1}(\theta_0), \tag{A.6}$$

$$\mathcal{S}(\theta_0) = \mathcal{C}(\theta_0) + \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{D}^{'}(\theta_0) + \mathcal{P}^{'}(\theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{D}^{'}(\theta_0) + \mathcal{D}(\theta_0)\mathcal{A}^{-1}(\theta_0)\mathcal{P}(\theta_0). \tag{A.7}$$

The assumption that $\mathbb{E}(\rho(y,\theta_0))$ is independent of θ_0 implies that

$$\begin{aligned} &\frac{\partial \mathbb{E}\left(T(\theta)\right)}{\partial \theta'}\Big|_{\theta_0} = \frac{\partial \mathbb{E}(\rho(y,\theta))}{\partial \theta'}\Big|_{\theta_0} = \frac{\partial}{\partial \theta'}\int \rho(y,\,\theta_0) \times f(y,\theta_0)d\nu \\ &= \int \frac{\partial \rho(y,\,\theta)}{\partial \theta'}\bigg|_{\theta_0} \times f(y,\theta_0)d\nu + \int \sqrt{n}\rho(y,\,\theta_0)\left(\frac{1}{\sqrt{n}}\frac{\partial \log f(y,\,\theta)}{\partial \theta}\bigg|_{\theta_0}\right)^{'} f(y,\theta_0)d\nu = 0. \end{aligned} \tag{A.8}$$

Since $\mathbb{E}\left(\frac{\partial \log f(y,\theta)}{\partial \theta}\big|_{\theta_0}\right) = 0$, (A.8) can be expressed as

$$\int \frac{\partial \rho(y,\,\theta)}{\partial \theta}\bigg|_{\theta_0} \times f(y,\theta_0)d\nu + \int \sqrt{n}\left(\rho(y,\,\theta_0) - \bar{\rho}\right)\left(\frac{1}{\sqrt{n}}\frac{\partial \log f(y,\,\theta)}{\partial \theta}\bigg|_{\theta_0}\right)^{'} f(y,\theta_0)d\nu = 0, \tag{A.9}$$

which implies that

$$\mathcal{P}^{'}(\theta_0) = -\mathcal{D}(\theta_0). \tag{A.10}$$

Using (A.10) in (A.6) and (A.7) for $\mathcal{V}(\theta_0)$ and $\mathcal{S}(\theta_0)$, respectively, yields the desired results.

A.2 Proof of Proposition 2

Using Lemma 2, the estimators for the components of $\mathcal{S}(\theta_0)$ and $\mathcal{V}(\theta_0)$ can be formulated from the sample counterparts. These sample counterparts are

$$D(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \rho(y_i,\, \theta)}{\partial \theta}\Big|_{\widehat{\theta}}, \tag{A.11}$$

$$C(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \left((\rho(y_i, \widehat{\theta}) - T(\widehat{\theta})) \times (\rho(y_i, \widehat{\theta}) - T(\widehat{\theta}))' \right). \tag{A.12}$$

Then, Lemmas 1 and 2 ensure that $D(\widehat{\theta}) = \mathcal{D}(\theta_0) + o_p(1)$ and $C(\widehat{\theta}) = \mathcal{C}(\theta_0) + o_p(1)$.

A.3 Proof of Proposition 3

The proof is similar to that of Proposition 1. Again, define the following vector

$$\vartheta(y, \widehat{\theta}, \rho_\star) = \begin{pmatrix} \frac{\partial \log f(y,\widehat{\theta})}{\partial \theta} \\ \rho(y, \widehat{\theta}) - \rho_\star \end{pmatrix}. \tag{A.13}$$

Note that $\operatorname{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \rho(y_i, \widehat{\theta}) = \rho_\star$ by Lemmas 1 and 2. Thus, under our regularity conditions, we have $\operatorname{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \vartheta(y_i, \widehat{\theta}, \rho_\star) = 0$. Also note that

$$\mathbb{E}\left(\nabla_{\theta\rho_\star} \vartheta(y, \theta_\star, \rho_\star)\right) = \mathbb{E}\begin{pmatrix} \frac{\partial^2 \log f(y,\theta_\star)}{\partial\theta\partial\theta'} & 0 \\ \frac{\partial \rho(y,\theta_\star)}{\partial \theta'} & -I \end{pmatrix} = \begin{pmatrix} -\mathcal{A}(\theta_\star) & 0 \\ \mathcal{D}(\theta_\star) & -I \end{pmatrix}, \tag{A.14}$$

$$\mathbb{E}\left(\vartheta(y, \theta_\star, \rho_\star) \times \vartheta'(y, \theta_\star, \rho_\star)\right) = \begin{pmatrix} \mathcal{B}(\theta_\star) & \mathcal{P}(\theta_\star) \\ \mathcal{P}'(\theta_\star) & \mathcal{C}(\theta_\star) \end{pmatrix}, \tag{A.15}$$

where $\mathcal{D}(\theta_\star) = \mathbb{E}\left(\frac{\partial \rho(y,\theta_\star)}{\partial \theta'}\right)$, $\mathcal{P}(\theta_\star) = \mathbb{E}\left(\frac{\partial \log f(y,\theta_\star)}{\partial \theta}\left(\rho(y,\theta_\star) - \rho_\star\right)\right)$, and $\mathcal{C}(\theta_\star) = \mathbb{E}\left((\rho(y,\theta_\star) - \rho_\star) \times (\rho(y,\theta_\star) - \rho_\star)'\right)$. Then, an application of Huber (1967, Corollary, p.231) yields:

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_\star) \\ \sqrt{n}(T(\widehat{\theta}) - \rho_\star) \end{pmatrix} \xrightarrow{d} N\left[0, \begin{pmatrix} \mathcal{A}^{-1}(\theta_\star)\mathcal{B}(\theta_\star)\mathcal{A}^{-1}(\theta_\star) & \mathcal{V}'(\theta_\star) \\ \mathcal{V}(\theta_\star) & \mathcal{S}(\theta_\star) \end{pmatrix}\right], \tag{A.16}$$

where

$$\begin{pmatrix} \mathcal{A}^{-1}(\theta_\star)\mathcal{B}(\theta_\star)\mathcal{A}^{-1}(\theta_\star) & \mathcal{V}'(\theta_\star) \\ \mathcal{V}(\theta_\star) & \mathcal{S}(\theta_\star) \end{pmatrix} = \begin{pmatrix} -\mathcal{A}(\theta_\star) & 0 \\ \mathcal{D}(\theta_\star) & -I \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{B}(\theta_\star) & \mathcal{P}(\theta_\star) \\ \mathcal{P}'(\theta_\star) & \mathcal{C}(\theta_\star) \end{pmatrix} \begin{pmatrix} -\mathcal{A}(\theta_\star) & \mathcal{D}'(\theta_\star) \\ 0 & -I \end{pmatrix}^{-1}.$$

Using the inverse partitioned matrix formula (see (A.5)), it can be shown that

$$\mathcal{V}(\theta_\star) = \mathcal{D}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{B}(\theta_\star)\mathcal{A}^{-1}(\theta_\star) + \mathcal{P}^{'}(\theta_\star)\mathcal{A}^{-1}(\theta_\star), \tag{A.17}$$

$$\begin{aligned}\mathcal{S}(\theta_\star) = \mathcal{C}(\theta_\star) &+ \mathcal{D}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{B}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{D}^{'}(\theta_\star) + \mathcal{P}^{'}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{D}^{'}(\theta_\star) \\ &+ \mathcal{D}(\theta_\star)\mathcal{A}^{-1}(\theta_\star)\mathcal{P}(\theta_\star).\end{aligned} \tag{A.18}$$

Since the information matrix type equality $\mathcal{D}(\theta_0) = -\mathcal{P}^{'}(\theta_0)$ does not hold under QML setting, expressions in (A.17) and (A.18) cannot be further simplified.

References

Andreou, Elena and Bas J. M. Werker (2010). "An Alternative Asymptotic Analysis of Residual-Based Statistics". In: *The Review of Economics and Statistics* 94.1, pp. 88–99.

Bera, Anil K. and Xiao-Lei Zuo (1996). "Specification test for a linear regression model with ARCH process". In: *Journal of Statistical Planning and Inference* 50.2, pp. 283 –308.

Cox, D. R. (1961). "Tests of Separate Families of Hypotheses". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, pp. 105–123.

— (1962). "Further Results on Tests of Separate Families of Hypotheses". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 24.2, pp. 406–424.

Durbin, J. (1970). "Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables". In: *Econometrica* 38.3, pp. 410–421.

Eicker, Friedhelm (1963). "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions". In: *The Annals of Mathematical Statistics* 34.2, pp. 447–456.

— (1967). "Limit theorems for regressions with unequal and dependent errors". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Ed. by L.M. LeCam and J. Neyman. University of California Press, pp. 59–82.

Gorodnichenko, Yuriy, Anna Mikusheva, and Serena Ng (2012). "Estimators for persistent and possibly nonstationary data with classical properties". In: *Econometric Theory* 28.5, pp. 1003–1036.

Huber, Peter J. (1967). "The behavior of maximum likelihood estimates under nonstandard conditions". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, pp. 221–233.

Jarque, Carlos M. and Anil K. Bera (1987). "A Test for Normality of Observations and Regression Residuals". In: *International Statistical Review / Revue Internationale de Statistique* 55.2, pp. 163–172.

Jennrich, Robert I. (1969). "Asymptotic Properties of Non-Linear Least Squares Estimators". In: *Ann. Math. Statist.* 40.2, pp. 633–643.

Koopmans, T.C., H. Rubin, and R.B Leipnik (1950). "Measuring the Equation Systems of Dynamic Economics". In: *Statistical Inference in Dynamic Economic Models by Cowles Commission Monograph no10*. Ed. by T.C. Koopmans. John Wiley and Sons, Inc., pp. 53–237.

Newey, Whitney K. (1985a). "Generalized method of moments specification testing". In: *Journal of Econometrics* 29.3, pp. 229 –256.

— (1985b). "Maximum Likelihood Specification Testing and Conditional Moment Tests". In: *Econometrica* 53.5, pp. 1047–1070.

Newey, Whitney K. and Daniel McFadden (1994). "Chapter 36 Large sample estimation and hypothesis testing". In: ed. by Robert F. Engle and Daniel L. McFadden. Vol. 4. Handbook of Econometrics. Elsevier, pp. 2111 –2245.

Neyman, Jerzy (1935). "Sur la vérification des hypothèses statistiques composées." French. In: *Bulletin de la Société Mathématique de France* 63, pp. 246–266.

— (1957). *Current problems of mathematical statistics*. Statistical Laboratory, University of California.

— (1959). "Optimal asymptotic tests of composite statistical hypotheses". In: *Probability and Statistics, the Harald Cramer Volume*. Ed. by U. Grenander. Wiley, New York, pp. 416–444.

Pierce, Donald A. (1982). "The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics". In: *The Annals of Statistics* 10.2, pp. 475–478.

Prokhorov, Artem and Peter Schmidt (2009). "GMM redundancy results for general missing data problems". In: *Journal of Econometrics* 151.1, pp. 47 –55.

Rao, C. Radhakrishna (1973). *Linear statistical inference and its applications*. 2nd Edition. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. John Wiley & Sons, Inc.

Student (1908). "The Probable Error of a Mean". In: *Biometrika* 6.1, pp. 1–25.

Tauchen, George (1985). "Diagnostic testing and evaluation of maximum likelihood models". In: *Journal of Econometrics* 30.1, pp. 415 –443.

Tse, Y. K. (2002). "Residual-based diagnostics for conditional heteroscedasticity models". In: *The Econometrics Journal* 5.2, pp. 358–373.

White, Halbert (1980). "Nonlinear Regression on Cross-Section Data". In: *Econometrica* 48.3, pp. 721–746.

— (1982a). "Maximum Likelihood Estimation of Misspecified Models". In: *Econometrica* 50.1, pp. 1–25.

— (1982b). "Regularity conditions for cox's test of non-nested hypotheses". In: *Journal of Econometrics* 19.2, pp. 301 –318.