Maximum Likelihood Estimation of Misspecified Models

Author(s): Halbert White

Source: *Econometrica*, Jan., 1982, Vol. 50, No. 1 (Jan., 1982), pp. 1-25

Published by: The Econometric Society

Stable URL: https://www.jstor.org/stable/1912526

ECONOMETRICA

VOLUME 50 JANUARY, 1982 NUMBER 1

MAXIMUM LIKELIHOOD ESTIMATION OF MISSPECIFIED MODELS

BY HALBERT WHITE[1]

This paper examines the consequences and detection of model misspecification when using maximum likelihood techniques for estimation and inference. The quasi-maximum likelihood estimator (QMLE) converges to a well defined limit, and may or may not be consistent for particular parameters of interest. Standard tests (Wald, Lagrange Multiplier, or Likelihood Ratio) are invalid in the presence of misspecification, but more general statistics are given which allow inferences to be drawn robustly. The properties of the QMLE and the information matrix are exploited to yield several useful tests for model misspecification.

1. INTRODUCTION

SINCE R. A. FISHER advocated the method of maximum likelihood in his influential papers [**13**, **14**], it has become one of the most important tools for estimation and inference available to statisticians. A fundamental assumption underlying classical results on the properties of the maximum likelihood estimator (e.g., Wald [**32**]; LeCam [**23**]) is that the stochastic law which determines the behavior of the phenomena investigated (the "true" structure) is known to lie within a specified parametric family of probability distributions (the model). In other words, the probability model is assumed to be "correctly specified." In many (if not most) circumstances, one may not have complete confidence that this is so.

If one does not assume that the probability model is correctly specified, it is natural to ask what happens to the properties of the maximum likelihood estimator. Does it still converge to some limit asymptotically, and does this limit have any meaning? If the estimator is somehow consistent, is it also asymptotically normal? Does the estimator have properties which can be used to decide whether or not the the specified family of probability distributions does contain the true structure? This paper provides a unified framework within which specific answers to each of these equations can be given.

The consistency question was apparently first considered independently by Berk [**7**, **8**] and Huber [**20**]. Berk takes a Bayesian approach and mentions in passing the information theoretic interpretation emphasized here. Huber's approach is classical; he provides very general conditions, building on those of Wald [**32**], under which the maximum likelihood estimator converges to a well-defined limit, even when the probability model is not correctly specified.

[1]I am indebted to Jon Wellner, Tom Rothenberg, the referees, and the participants of the Harvard/MIT econometrics workshop for helpful comments and suggestions.

Huber's limit is identical to that of Berk; however, Huber does not explicitly discuss the information theoretic interpretation of this limit. This interpretation has been emphasized by Akaike [3], who has observed that when the true distribution is unknown, the maximum likelihood estimator is a natural estimator for the parameters which minimize the Kullback-Leibler [22] Information Criterion (KLIC). Huber also elegantly treats the asymptotic normality question, and Souza and Gallant [30] definitively treat the related problem of inference, in a general implicit nonlinear simultaneous equations framework.

In Section 2, we provide simple conditions under which the maximum likelihood estimator is a strongly consistent estimator for the parameter vector which minimizes the KLIC. Our conditions are more closely related to the classical treatment of maximum likelihood given by LeCam [23] than to the earlier conditions of Wald [32]. While not as general as Huber's [20] conditions, they are nevertheless sufficiently general to have broad applicability. They are also more easily verified in common situations and have somewhat greater intuitive appeal than do Huber's.

Our treatment of asymptotic normality, given in Section 3, builds on the assumptions used to obtain consistency. While it too is more restrictive than Huber's approach, it does include LeCam's [23] asymptotic normality result as a special case. An interesting feature of this result is that with misspecification, the asymptotic covariance matrix of the QMLE no longer equals the inverse of Fisher's information matrix. Nevertheless, the covariance matrix can be consistently estimated and, as expected, simplifies to the familiar form in the absence of misspecification.

This latter property is exploited in Section 4 to yield a new test for misspecification, applicable to a broad range of problems, including omnibus or directional tests for univariate or multivariate normality, as well as tests for misspecification of linear or nonlinear regression equations.

In Section 5, properties of the QMLE are further exploited to yield specification tests of the Hausman [17] type. A new statistic, based on evaluating the scores for a maintained log-likelihood at an alternative consistent QMLE, is proposed and shown to be asymptotically equivalent to the Hausman statistic. This new statistic is often simpler to compute, since it doesn't require full maximization of the likelihood function.

2. CONSISTENCY

Our first assumption defines the structure which generates the observations.

ASSUMPTION A1: The independent random $1 \times M$ vectors $U_t, t = 1, \ldots, n$, have common joint distribution function G on Ω, a measurable Euclidean space, with measurable Radon-Nikodým density $g = dG/d\nu$.

Since G is unknown a priori, we choose a family of distribution functions which may or may not contain the true structure, G. It is usually easy to choose this family to satisfy the next assumption.

ASSUMPTION A2: The family of distribution functions $F(u,\theta)$ has Radon-Nikodým densities $f(u,\theta) = dF(u,\theta)/d\nu$ which are measurable in u for every θ in Θ, a compact subset of a p-dimensional Euclidean space, and continuous in θ for every u in Ω.

Next, we define the quasi-log-likelihood of the sample as

$$L_n(U,\theta) \equiv n^{-1}\sum_{t=1}^{n} \log f(U_t,\theta),$$

and we define a quasi-maximum likelihood estimator (QMLE) as a parameter vector $\hat{\theta}_n$ which solves the problem

$$\max_{\theta\in\Theta} L_n(U,\theta). \tag{2.1}$$

THEOREM 2.1 (Existence): *Given Assumptions A1 and A2, for all n there exists a measurable* QMLE, $\hat{\theta}_n$.

All proofs are provided in the Mathematical Appendix. Theorem 2.1 ensures that a QMLE always exists, but does not say anything about uniqueness.

Given the existence of a QMLE, we may examine its properties. It is well known that when F contains the true structure G (i.e., $G(u) \equiv F(u,\theta_0)$ for some θ_0 in Θ) the MLE is consistent for θ_0 under suitable regularity conditions (Wald [**32**, Theorem 2]; LeCam [**23**, Theorem 5.a]). Without this restriction Akaike [**3**] has noted that since $L_n(U,\theta)$ is a natural estimator for $E(\log f(U_t,\theta))$, $\hat{\theta}_n$ is a natural estimator for θ_*, the parameter vector which minimizes the Kullback-Leibler [**22**] Information Criterion (KLIC),

$$I(g:f,\theta) \equiv E\big(\log\big[\, g(U_t)/f(U_t,\theta)\,\big]\big).$$

Here, and in what follows, expectations are taken with respect to the true distribution. Hence,

$$I(g:f,\theta) = \int \log g(u)dG(u) - \int \log f(u,\theta)dG(u).$$

The opposite of $I(g:f,\theta)$ is called the entropy of the distribution $G(u)$ with respect to $F(u,\theta)$. Intuitively, $I(f:g,\theta)$ measures our ignorance about the true structure.[2]

To support Akaike's observation that $\hat{\theta}_n$ is a natural estimator for θ_*, we impose the following additional condition.

ASSUMPTION A3: (a) $E(\log g(U_t))$ exists and $|\log f(u,\theta)| \leqq m(u)$ for all θ in Θ, where m is integrable with respect to G; (b) $I(g:f,\theta)$ has a unique minimum at θ_* in Θ.

[2] Akaike [3] provides a useful discussion of the appropriateness of the KLIC for discriminating between models. Rényi [**27**] gives an axiomatic justification for the entropy as an information measure. Important properties of the KLIC which will be used later are $I(g:f,\theta) \geqq 0$ for all θ in Θ, and $I(g:f,\theta_0) = 0$ for some θ_0 in Θ if and only if $g(u) = f(u,\theta_0)$ almost everywhere $-\nu$ (Rao [**26**, Theorem 1e.6(ii)]).

Assumption A3(a) ensures that the KLIC is well-defined; for example, if $M = 1$ and f and g are normal density functions Assumption A3(a) is satisfied whenever the true variance σ_0^2 is finite and Θ does not contain $\sigma^2 = 0$. Assumption A3(b) is the fundamental identification condition (cf. Bowden [**9**]), which, for example, rules out redundant regressors in the linear regression framework and is equivalent to the rank condition in the simultaneous equations framework[3] (cf. Rothenberg [**28**]). When Assumption A3(b) holds, we say that θ_* is *globally identifiable*. We can now state the desired result.

THEOREM 2.2 (Consistency): *Given Assumptions A1–A3,* $\hat{\theta}_n \to \theta_*$ *as* $n \to \infty$ *for almost every sequence* (U_t); *i.e.,* $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta_*$.

In other words, the QMLE is generally a strongly consistent estimator for θ_*, the parameter vector which minimizes the KLIC. This ensures that we minimize our ignorance about the true structure; thus we might call the QMLE the "minimum ignorance" estimator.

If the probability model is correctly specified (i.e., $g(u) = f(u, \theta_0)$ for some θ_0 in Θ), then $I(g:f,\theta)$ attains its unique minimum at $\theta_* = \theta_0$, so that $\hat{\theta}_n$ is consistent for the "true" parameter vector θ_0. The present result with $g(u) = f(u, \theta_0)$ is closely related to the classical MLE consistency result of LeCam [**23**, Theorem 5a]; it is straightforward to verify that given Assumptions A1–A3 and $g(u) = f(u, \theta_0)$, LeCam's [**23**, pp. 303–304] consistency conditions (i)–(iv) are satisfied.

It is important to point out that the correct specification of the probability model is a sufficient, but by no means a necessary condition for the consistent estimation of particular parameters of interest. For example, even when the true distribution is not normal, maximum likelihood carried out under the assumption of normality (i.e., least squares) yields consistent estimates of the mean and variance of distributions for which these quantities are finite. Indeed, it is the consistency of the QMLE for the parameters of interest in a wide range of situations which ensures its usefulness as the basis for the robust estimation techniques (the M-estimators) proposed by Huber [**20**]. This fact also provides the basis for specification tests of the Hausman [**17**] type, which we develop in Section 5.

Although $\hat{\theta}_n$ can fail to be consistent for the parameters of interest as a result of the failure of the distributional assumptions (e.g., many robust estimation techniques require symmetric distributions for consistency), another important reason for this failure is the failure of restrictions implicitly or explicitly imposed on the elements of θ, particularly on the location parameters. As a simple

[3]Strictly speaking, the regularity conditions given here allow our results to be applied only to regression equations with i.i.d. regressors (e.g., as in White [**33**, **36**]). Nevertheless, we expect the conclusions of our theorems to hold under similar conditions for the general implicit nonlinear simultaneous equations model, as the results of Souza and Gallant [**30**] suggest. The present approach is taken to avoid burying the reader in a mass of notation and detail.

example, incorrectly assuming that the mean is zero when one estimates the variance of a population leads to inconsistent estimates. In the linear regression framework, omitting a relevant variable correlated with the included regressors will lead to inconsistent parameter estimates. Similarly, incorrect parametric constraints in the simultaneous equations framework can lead to inconsistent estimates.

3. ASYMPTOTIC NORMALITY

With additional conditions provided in this section, we can show that the QMLE is asymptotically normally distributed. When the partial derivatives exist, we define the matrices

$$A_n(\theta) = \left\{ n^{-1} \sum_{t=1}^{n} \partial^2 \log f(U_t, \theta)/\partial\theta_i \partial\theta_j \right\},$$

$$B_n(\theta) = \left\{ n^{-1} \sum_{t=1}^{n} \partial \log f(U_t, \theta)/\partial\theta_i \cdot \partial \log f(U_t, \theta)/\partial\theta_j \right\}.$$

If expectations also exist, we define the matrices

$$A(\theta) = \left\{ E\left(\partial^2 \log f(U_t, \theta)/\partial\theta_i \partial\theta_j\right) \right\},$$

$$B(\theta) = \left\{ E\left(\partial \log f(U_t, \theta)/\partial\theta_i \cdot \partial \log f(U_t, \theta)/\partial\theta_j\right) \right\}.$$

When the appropriate inverses exist, define

$$C_n(\theta) = A_n(\theta)^{-1} B_n(\theta) A_n(\theta)^{-1},$$

$$C(\theta) = A(\theta)^{-1} B(\theta) A(\theta)^{-1}.$$

ASSUMPTION A4: $\partial \log f(u, \theta)/\partial\theta_i$, $i = 1, \ldots, p$, are measurable functions of u for each θ in Θ and continuously differentiable functions of θ for each u in Ω.

ASSUMPTION A5: $|\partial^2 \log f(u, \theta)/\partial\theta_i \partial\theta_j|$ and $|\partial \log f(u, \theta)/\partial\theta_i \cdot \partial \log f(u, \theta)/\partial\theta_j|$, $i, j = 1, \ldots, p$ are dominated by functions integrable with respect to G for all u in Ω and θ in Θ.

ASSUMPTION A6: (a) θ_* is interior to Θ; (b) $B(\theta_*)$ is nonsingular; (c) θ_* is a regular point of $A(\theta)$.

Assumption 4 ensures that the first two derivatives with respect to θ exist; that these derivatives are measurable in u follows from Assumption A2, since the derivative can be considered as the limit of a sequence of measurable difference quotients. These conditions allow us to apply a mean value theorem for random functions given by Jennrich [**21**, Lemma 3]. Assumption 5 ensures that the derivatives are appropriately dominated by functions integrable with respect to

G, which ensures that $A(\theta)$ and $B(\theta)$ are continuous in θ and that we can apply a uniform law of large numbers (LeCam [**23**, Corollary 4.1]; Jennrich [**21**, Theorem 2]) to $A_n(\theta)$ and $B_n(\theta)$. In Assumption A6(c), we define a *regular point* of the matrix $A(\theta)$ as a value for θ such that $A(\theta)$ has constant rank in some open neighborhood of θ.

Before stating the asymptotic normality result, we give a very general result for the identification problem. Before, we observed that if Assumption A3(b) holds, θ_* is globally identifiable; we say that θ_* is *locally identifiable* if for some open neighborhood $\mathfrak{N} \subset \Theta$, $I(g:f,\theta)$ has a unique minimum at θ_*.

THEOREM 3.1 (Identification): *(i) Given Assumptions A1–A3(a) and Assumptions A4–A6(a), if θ_* is globally (locally) identifiable and if θ_* is a regular point of $A(\theta)$, then $A(\theta_*)$ is negative definite. (ii) Given Assumptions A1–A3(a) and Assumptions A4–A6(a), if $A(\theta_*)$ is negative definite and if θ_* minimizes $I(g:f,\theta)$ in an open neighborhood $\mathfrak{N} \subset \Theta$, then θ_* is locally identifiable.*

This result shows that the identification problem has content even when the model is misspecified. If we further suppose that $g(u) = f(u,\theta_0)$ for θ_0 in Θ and that Assumption A7 below holds, the identification results of Rothenberg [**28**, Theorem 1] and Bowden [**9**, p. 1073] follow as corollaries. Theorem 3.1 shows that a necessary condition for the identifiability of θ_* is the negative definiteness of $A(\theta_*)$. If we find that the sample analog $A_n(\hat{\theta}_n)$ is singular or nearly singular, we have an indication that Assumption A3(b) does not hold.

THEOREM 3.2 (Asymptotic Normality): *Given Assumptions A1–A6*

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \overset{A}{\sim} N(0, C(\theta_*)).$$

Moreover, $C_n(\hat{\theta}_n) \overset{\text{a.s.}}{\to} C(\theta_*)$, element by element.

If we further assume that the model is correctly specified, so that $g(u) = f(u, \theta_0)$ for some θ_0 in Θ, then Theorem 3.2 contains the asymptotic normality result of LeCam [**23**, Theorem 6.(i)]. LeCam also requires

$$\begin{aligned} (3.1) \qquad & \int \partial^2 \log f(u,\theta)/\partial\theta_i \partial\theta_j \cdot f(u,\theta)\, dv \\ & = -\int \partial \log f(u,\theta)/\partial\theta_i \cdot \partial \log f(u,\theta)/\partial\theta_j \cdot f(u,\theta)\, dv \\ & \qquad\qquad (i,j = 1, \ldots, p) \end{aligned}$$

for all θ in Θ. Equation (3.1) is the familiar equality in maximum likelihood theory which ensures the equivalence of the Hessian (left-hand side) and outer product (right-hand side) forms for the information matrix.

In the present case, this equivalence generally won't hold, as an example below demonstrates. However, when the model is correctly specified and the next assumption holds, we obtain an information matrix equivalence result.

ASSUMPTION A7: $|\partial[\partial f(u,\theta)/\partial\theta_i \cdot f(u,\theta)]/\partial\theta_j|$, $i, j = 1, \ldots, p$, are dominated by functions integrable with respect to ν for all θ in Θ, and the minimal support of $f(u,\theta)$ does not depend on θ.[4]

THEOREM 3.3 (Information Matrix Equivalence): *Given Assumptions A1–A7, if* $g(u) = f(u,\theta_0)$ *for* θ_0 *in* Θ, *then* $\theta_* = \theta_0$ *and* $A(\theta_0) = -B(\theta_0)$, *so that* $C(\theta_0) = -A(\theta_0)^{-1} = B(\theta_0)^{-1}$, *where* $-A(\theta_0)$ *is Fisher's information matrix.*

Together, Conditions A1–A7 and $g(u) = f(u,\theta_0)$ for θ_0 in Θ may be thought of as the "usual maximum likelihood regularity conditions," since they ensure that all the familiar results hold.

To see that $A(\theta_*)$ will not generally equal $-B(\theta_*)$, consider the estimation of the mean and variance of i.i.d. random variables U_t, assumed to be distributed as $N(\mu_0, \sigma_0^2)$. The quasi-log-likelihood of an observation is

$$\log f(U_t, \mu, \sigma^2) = -.5\log 2\pi - .5\log\sigma^2 - .5(U_t - \mu)^2/\sigma^2.$$

Provided that U_t has nonzero variance and finite fourth moment, it follows that $\mu_* = \mu_0$ and $\sigma_*^2 = \sigma_0^2$, while

$$A(\mu_0, \sigma_0^2) = \begin{bmatrix} -1/\sigma_0^2 & 0 \\ 0 & -1/2\sigma_0^4 \end{bmatrix} \quad \text{and}$$

$$B(\mu_0, \sigma_0^2) = \begin{bmatrix} 1/\sigma_0^2 & \sqrt{\beta_1}/2\sigma_0^3 \\ \sqrt{\beta_1}/2\sigma_0^3 & (\beta_2 - 1)/4\sigma_0^4 \end{bmatrix}, \quad \text{so that}$$

$$C(\mu_0, \sigma_0^2) = \begin{bmatrix} \sigma_0^2 & \sqrt{\beta_1}\sigma_0^3 \\ \sqrt{\beta_1}\sigma_0^3 & (\beta_2 - 1)\sigma_0^4 \end{bmatrix},$$

where $\sqrt{\beta_1}$ and β_2 are the skewness and kurtosis measures, $\sqrt{\beta_1} = E[(U_t - \mu_0)^3]/\sigma_0^3$ and $\beta_2 = E[(U_t - \mu_0)^4]/\sigma_0^4$. Obviously, a necessary and sufficient condition that $A(\mu_0, \sigma_0^2) = -B(\mu_0, \sigma_0^2)$ is $\sqrt{\beta_1} = 0$ and $\beta_2 = 3$, for which normality is sufficient. With rare exceptions, inferences in the maximum likelihood framework are drawn using estimators for $A(\mu_0, \sigma_0^2)$ or $B(\mu_0, \sigma_0^2)$, taking advantage of the information matrix equivalence. In the present example, the presence of skewness and/or kurtosis can lead to serious errors in inference when standard techniques are applied. Note that inferences about the mean based on an estimator for $A(\mu_0, \sigma_0^2)$ will be correct due to the diagonality of $A(\mu_0, \sigma_0^2)$. However, inferences about σ_0^2 will be affected, as will inferences about either μ_0 or σ_0^2 based on an estimator for $B(\mu_0, \sigma_0^2)$ such as $B_n(\hat{\mu}_n, \hat{\sigma}_n^2)$.

This example makes it clear that care must be taken in drawing inferences in the presence of model misspecification. The fact that $A(\theta_*)$ generally doesn't

[4]This latter condition apparently can be weakened to a requirement that $f(u,\theta)$ vanishes on the boundary of its minimal support.

equal $-B(\theta_*)$ causes the familiar asymptotic equivalence (Silvey [**29**]) of the Lagrange Multiplier (Aitchison and Silvey [**2**]) and Wald [**31**] statistics to the likelihood ratio statistic to break down, as we show below. Nevertheless, Theorem 3.2 can be used to construct appropriate statistics for hypothesis testing when the model is misspecified. In particular, suppose we wish to test the hypothesis $H_0: s(\theta_*) = 0$, where $s: R^p \to \mathbb{R}^r$ is a continuous vector function of θ such that its Jacobian at θ_*, $\nabla s(\theta_*)$, is finite with full row rank r, against the alternative $H_1: s(\theta_*) \neq 0$.

The appropriate form for the Wald statistic is given by the following result.

THEOREM 3.4 (Wald Test): *Given Assumptions A1–A6 and* H_0,

$$\mathfrak{W}_n = n s(\hat{\theta}_n)' \left[\nabla s(\hat{\theta}_n) C_n(\hat{\theta}_n) \nabla s(\hat{\theta}_n)' \right]^{-1} s(\hat{\theta}_n) \overset{A}{\sim} \chi_r^2. \tag{3.2}$$

Note that the usual Wald statistic uses either $-A_n(\hat{\theta}_n)^{-1}$ or $B_n(\hat{\theta}_n)^{-1}$ in place of $C_n(\hat{\theta}_n)$. With model misspecification, $C_n(\hat{\theta}_n)$ *must* be used to ensure that the test has the proper size.

Let $\tilde{\theta}_n$ solve the constrained maximization problem

$$\max_{\theta \in \Theta} L_n(U, \theta) \qquad \text{subject to} \qquad s(\theta) = 0.$$

The proper form for the Lagrange Multiplier statistic is given by the next result.

THEOREM 3.5 (Lagrange Multiplier Test): *Given Assumptions A1–A6 and* H_0,

$$\begin{aligned} \mathfrak{LM}_n = {} & \nabla L_n(U, \tilde{\theta}_n)' A_n(\tilde{\theta}_n)^{-1} \nabla s(\tilde{\theta}_n)' \\ & \times \left[\nabla s(\tilde{\theta}_n) C_n(\tilde{\theta}_n) \Delta s(\tilde{\theta}_n)' \right]^{-1} \nabla s(\tilde{\theta}_n) A_n(\tilde{\theta}_n)^{-1} \nabla L_n(U, \tilde{\theta}_n) \overset{A}{\sim} \chi_r^2. \end{aligned} \tag{3.3}$$

Moreover, $\mathfrak{W}_n - \mathfrak{LM}_n \xrightarrow{p} 0$.

The usual form for the Lagrange Multiplier statistic replaces $C_n(\tilde{\theta}_n)$ with $-A_n(\tilde{\theta}_n)^{-1}$. Again, $C_n(\tilde{\theta}_n)$ must be used in the presence of model misspecification to ensure that the test has proper size.

This result further establishes the asymptotic equivalence of the specification robust versions of the Wald and Lagrange Multiplier statistics. However, the likelihood ratio statistic is not generally equivalent to these in the presence of model misspecification. To see this, we observe that a two term mean value expansion of the likelihood ratio statistic for testing the hypothesis $\theta_* = \theta^0$ yields

$$-2n\left(L_n(U, \theta^0) - L_n(U, \hat{\theta}_n)\right) + n(\hat{\theta}_n - \theta^0)' A(\theta^0)(\hat{\theta}_n - \theta^0) \xrightarrow{p} 0.$$

The term $-n(\hat{\theta}_n - \theta^0)' A(\theta^0)(\hat{\theta}_n - \theta^0)$ is an appropriate Wald statistic when the model is correctly specified. Otherwise, $C(\theta^0)^{-1}$ must replace $-A(\theta^0)$ to obtain a test of the correct size; the likelihood ratio fails to do this and is not asymptotically distributed as χ_r^2 (cf. Souza and Gallant [**30**, Theorem 9], Foutz and Srivastava [**15**]).

4. THE INFORMATION MATRIX TEST FOR MISSPECIFICATION

The information matrix equivalence theorem says essentially that when the model is correctly specified, the information matrix can be expressed in either Hessian form, $-A(\theta_0)$, or outer product form, $B(\theta_0)$. Equivalently, $A(\theta_0) + B(\theta_0) = 0$. When this equality fails, it follows that the model is misspecified, and we saw that this misspecification can have serious consequences when standard inferential techniques are applied.

The failure of information matrix equivalence can also indicate misspecifications which render the QMLE inconsistent for particular parameters of interest. For example, suppose we estimate the variance of random variables U_t assumed to be distributed as $N(0, \sigma_0^2)$, when in fact $E(U_t) = \mu_0 \neq 0$. Then $\hat{\sigma}_n^2 = n^{-1}\sum_{t=1}^n U_t^2$ converges to $\sigma_*^2 = \sigma_0^2 + \mu_0^2$. With $U_t \sim N(\mu_0, \sigma_0^2)$, it follows that $A(\sigma_*^2) = -.5\sigma_*^4$ while $B(\sigma_*^2) = .5\sigma_*^4 - .5\mu_0^4/2\sigma_*^8$. Obviously, $A(\sigma_*^2) = -B(\sigma_*^2)$ if and only if $\mu_0 = 0$, so $A(\sigma_*^2) \neq -B(\sigma_*^2)$ indicates estimator inconsistency. Such examples are easily multiplied, suggesting that $A(\theta_*) + B(\theta_*)$ is a useful indicator of misspecifications which cause either parameter or covariance matrix estimator inconsistency.

The matrix $A(\theta_*) + B(\theta_*)$ is unobservable, but it can be consistently estimated by $A_n(\hat{\theta}_n) + B_n(\hat{\theta}_n)$. To obtain a test statistic, we consider the asymptotic distribution of the elements of $\sqrt{n}(A_n(\hat{\theta}_n) + B_n(\hat{\theta}_n))$, anticipating that under appropriate conditions, these elements will be jointly normally distributed asymptotically, with mean zero in the absence of misspecification. Given a consistent estimator for the asymptotic covariance matrix, we can form an asymptotic χ^2 statistic of the Wald [**31**] type.

It is important to point out that the large sample approach is not the only way of proceeding. For example, when estimating the mean and variance from a sample hypothesized to be normal, the hypothesis that $A(\theta_0) + B(\theta_0) = 0$ is equivalent to the joint hypothesis that $\sqrt{\beta_1} = 0$ and $\beta_2 = 3$. There are presently available several useful approximate finite sample omnibus tests for normality which make joint use of the sample analogs of $\sqrt{\beta_1}$ and β_2, i.e., $\sqrt{b_1}$ and b_2 (e.g., D'Agostino and Pearson [**12**]; Bowman and Shenton [**10**]; Pearson, D'Agostino, and Bowman [**25**]). These are all possible alternatives to our large sample approach. However, since these are specific to the problem of testing normality, and since we wish to give a general procedure, we take a large sample approach.

To simplify the notation which follows, we define

$$d_l(u, \theta) = \partial \log f(u, \theta)/\partial \theta_i \cdot \partial \log f(u, \theta)/\partial \theta_j + \partial^2 \log f(u, \theta)/\partial \theta_i \partial \theta_j,$$

$$(l = 1, \ldots, p(p+1)/2; i = 1, \ldots, p; j = i, \ldots, p).$$

The test will be based on the "indicators" $D_{ln}(\hat{\theta}_n) = n^{-1}\sum_{t=1}^n d_l(u_t, \hat{\theta}_n)$, which are the elements of $A_n(\hat{\theta}_n) + B_n(\hat{\theta}_n)$. In many cases, however, it is inappropriate to base the test on all $p(p+1)/2$ indicators. First, some indicators may be

identically zero, as in the example of estimating the mean and variance of a supposed normal variate. There, $D_{1n}(\hat{\theta}_n) = -1/\hat{\sigma}_n^2 + 1/\hat{\sigma}_n^2 = 0$. Second, some indicators may be linear combinations of others. This occurs in the linear regression framework when the regression contains a constant and polynomial terms in a particular regressor (see White [**35**]). In either case, it is appropriate to ignore such indicators.

Finally, and just as importantly, when one is estimating a moderate number of parameters, it may simply be infeasible to test all indicators jointly. For example, if $p = 10$, there can be as many as 55 indicators. Even with a moderately large sample, say $n = 75$, one may be concerned about the available degrees of freedom. In such situations it is possible to avoid this problem by performing tests on linear combinations of the indicators, or more simply, by considering only a subset. This yields a more directional test, and which subset should be chosen depends on the alternatives against which power is desired. In the normal example, one will obtain tests powerful against skewed or kurtotic alternatives depending on whether $D_{2n}(\hat{\theta}_n)$ or $D_{3n}(\hat{\theta}_n)$ is the basis for a directional test.

Accordingly, define the $q \times 1$ vector $d(u,\theta)$ ($q \leqq p(p+1)/2$) so that $D_n(\hat{\theta}_n) = n^{-1}\sum_{t=1}^{n} d(U_t,\theta_n)$ is the $q \times 1$ vector containing the indicators of interest, and let $D(\theta) = E(d(U_t,\theta))$. When partial derivatives and expectations exist, define the $q \times p$ Jacobian matrices

$$\nabla D_n(\theta) = \left\{ n^{-1} \sum_{t=1}^{n} \partial d_l(U_t,\theta)/\partial\theta_k \right\} \qquad \text{and}$$

$$\nabla D(\theta) = \{ E(\partial d_l(U_t,\theta)/\partial\theta_k) \},$$

where the indexes $l = 1, \ldots, q$ have been reassigned appropriately. We add the following conditions:

ASSUMPTION A8: $\partial d_l(u,\theta)/\partial\theta_k$, $l = 1, \ldots, q$, $k = 1, \ldots, p$, exist and are continuous functions of θ for each u.

ASSUMPTION A9: $|d_l(u,\theta)d_m(u,\theta)|$, $|\partial d_l(u,\theta)/\partial\theta_k|$, and $|d_l(u,\theta) \cdot \partial \log f(u,\theta)/\partial\theta_k|$, $k = 1, \ldots, p$; $l, m = 1, \ldots, q$, are dominated by functions integrable with respect to G for all u and θ in Θ.

These assumptions play roles analogous to Assumptions A4 and A5. Note that Assumption A8 requires continuous third derivatives for the quasi-log-likelihood function. Among other things, Assumption A9 ensures that $\nabla D(\theta)$ is finite for all θ in Θ.

Next, define

$$V(\theta) = E\Big(\big[d(U_t,\theta) - \nabla D(\theta) A(\theta)^{-1} \nabla \log f(U_t,\theta) \big] \cdot \big[d(U_t,\theta) - \nabla D(\theta) A(\theta)^{-1} \nabla \log f(U_t,\theta) \big]' \Big).$$

It turns out that $V(\theta_*)$ is the asymptotic covariance matrix of $\sqrt{n}D_n(\hat{\theta}_n)$, and we make the following assumption:

ASSUMPTION A10: $V(\theta_*)$ is nonsingular.

In practice, Assumption A10 can always be guaranteed by appropriate choice of indicators.

We require a consistent estimator for $V(\theta_*)$; a natural choice is

$$V_n(\hat{\theta}_n) = n^{-1}\sum_{t=1}^{n}\left[d(U_t,\hat{\theta}_n) - \nabla D_n(\hat{\theta}_n)A_n(\hat{\theta}_n)^{-1}\nabla \log f(U_t,\hat{\theta}_n)\right]$$
$$\cdot\left[d(U_t,\hat{\theta}_n) - \nabla D_n(\hat{\theta}_n)A_n(\hat{\theta}_n)^{-1}\nabla \log f(U_t,\hat{\theta}_n)\right]'.$$

Now we can state the desired result.

THEOREM 4.1 (Information Matrix Test): *Given Assumptions A1–A10, if* $g(u) = f(u,\theta_0)$ *for* θ_0 *in* Θ, *then* (i) $\sqrt{n}D_n(\hat{\theta}_n) \overset{A}{\sim} N(0, V(\theta_0))$; (ii) $V_n(\hat{\theta}_n) \overset{\text{a.s.}}{\to} V(\theta_0)$, *and* $V_n(\hat{\theta}_n)$ *is nonsingular almost surely for all n sufficiently large*; (iii) *the information matrix test statistic*

$$\text{(4.1)} \qquad \mathfrak{I}_n = nD_n(\hat{\theta}_n)'V_n(\hat{\theta}_n)^{-1}D_n(\hat{\theta}_n)$$

is distributed asymptotically as χ^2_q.

To carry out the test, one computes (4.1) and compares it to the critical value of the χ^2_q distribution for a given size of test. If (4.1) does not exceed this value, one can't reject the null hypothesis that the model has been correctly specified. Otherwise, one concludes that the model is misspecified, implying the inconsistency of the usual maximum likelihood covariance matrix estimators $-A_n(\hat{\theta}_n)^{-1}$ or $B_n(\hat{\theta}_n)^{-1}$ at the very least, as well as possible inconsistency of the QMLE for parameters of interest. (Whether this latter problem exists can be investigated using the tests of the next section.) When the null hypothesis is rejected, inferences are properly drawn (with respect to θ_*) using the Wald or Lagrange Multiplier statistics of Section 3. Note that even in the absence of parameter estimator inconsistency, a statistically significant value for (4.1) indicates potential efficiency gains to removing the misspecification.

The information matrix test provides a unified framework for specification (goodness of fit) tests for a wide variety of probability laws, uni- or multivariate, continuous or discrete. In addition, it can reasonably be expected to have validity in frameworks much more general than that explicitly considered here. As special cases, it contains the heteroskedasticity test of White [**35**], as well as White's [**36**, Theorem 4.2] specification test for nonlinear regression models. Under appropriate regularity conditions, the statistic (4.1) should also be applicable to general simultaneous equations or limited dependent variables models.

Although we don't formally consider the power of the information matrix test

here, it's reasonable to expect that the test will be consistent (i.e., have unit power asymptotically) against any alternative which renders the usual maximum likelihood inference techniques invalid. Misspecifications which don't affect the usual techniques won't be detected. The loss associated with a type II error in these cases amounts only to the loss in efficiency associated with quasi-maximum likelihood estimation, rather than that resulting from parameter or covariance matrix estimator inconsistency.

Typically, it's straightforward to determine which alternatives will be detected by the information matrix test in specific cases. The information matrix test for normality is sensitive to skewness or kurtosis. Other alternatives are ignored at little cost since these cause neither parameter nor covariance matrix estimator inconsistency when the usual (i.e., least squares) techniques are used. In the linear regression framework, the test is sensitive to forms of heteroskedasticity or model misspecification which result in correlations between the squared regression errors and the second order cross-products of the regressors (see White [**35**, pp. 824–826]).

As a practical matter, computation of (4.1) can be cumbersome due to the presence of $\nabla D_n(\hat{\theta}_n)$, which contains third derivatives. Often it can be shown under the null hypothesis that $\nabla D(\theta_0)$ vanishes, so that $V(\theta_0)$ is consistently estimated by $n^{-1}\sum_{t=1}^{n} d(U_t, \hat{\theta}_n) d(U_t, \hat{\theta}_n)'$. In White's heteroskedasticity and nonlinear regression specification tests [**35**, **36**] $\nabla D(\theta)$ vanishes under the alternative as well. Even when $\nabla D(\theta)$ doesn't vanish, the null hypothesis can be exploited to yield

$$\nabla D(\theta_0) A(\theta_0)^{-1} E(\nabla \log f(U_t, \theta_0) d(U_t, \theta_0)) = \nabla D(\theta_0) C(\theta_0) \nabla D(\theta_0)'$$

so that $V(\theta_0)$ can be consistently estimated by

$$\tilde{V}_n(\hat{\theta}_n) = n^{-1} \sum_{t=1}^{n} d(U_t, \hat{\theta}_n) d(U_t, \hat{\theta}_n)' - \nabla D_n(\hat{\theta}_n) C_n(\hat{\theta}_n) \nabla D_n(\theta_n)'.$$

This estimator is neither consistent nor necessarily positive semi-definite when the null hypothesis fails, so it may be a poor choice in practice.

5. TESTS FOR PARAMETER ESTIMATOR INCONSISTENCY

The information matrix test of the previous section is sensitive to model misspecifications which invalidate the usual maximum likelihood inference procedures. Often, such misspecifications will also cause the QMLE to be inconsistent for particular parameters of interest, but it may sometimes be difficult *a priori* to tell if this is so. In this section, we present two tests sensitive to misspecifications which cause parameter estimator inconsistency.

Given a correctly specified model, not only can we obtain the maximum likelihood estimator (MLE), but it is typically easy to find a QMLE which

contains a subset of estimators consistent for particular parameters of interest. For example, in the normal location estimation problem, the sample mean is the MLE, while the sample median is an alternative consistent QMLE. In the regression framework with a normality assumption, least squares gives the MLE, while weighted least squares gives a consistent QMLE (White [**33**, **34**, **36**].) In the simultaneous equation framework, the three-stage least squares estimator yields the MLE under a normality assumption, while two-stage least squares equation by equation gives a consistent QMLE (e.g., Hausman [**17**]).

Whenever the MLE and an alternative consistent QMLE are available, the distance between them can be used as an indicator of model misspecification since this distance vanishes asymptotically in the absence of misspecification, but generally doesn't vanish otherwise. Although this fact has often been exploited (e.g., by Byron [**11**], Wu [**37**]), Hausman [**17**] was apparently the first to advocate making this fact the basis for a general class of specification tests. Accordingly, we refer to any tests based on a comparison of the MLE to a QMLE as a Hausman test.

For present purposes, we let Θ and Γ be p- and q-dimensional compact subsets of Euclidean spaces such that $\Theta = \mathbb{B} \times \Psi$ and $\Gamma = \mathbb{B} \times \mathbb{A}$, where $\mathbb{B}$ is a compact subset of k-dimensional Euclidean space. A typical element of $\mathbb{B}$ is denoted β. Let $\hat{\theta}_n' = (\hat{\beta}_n', \hat{\psi}_n')$ maximize $n^{-1}\sum_{t=1}^n \log f(U_t, \theta)$ over Θ, while $\tilde{\gamma}_n' = (\tilde{\beta}_n', \tilde{\alpha}_n')$ maximizes $n^{-1}\sum_{t=1}^n \log h(U_t, \gamma)$ over Γ. We require h to be a density function satisfying the following assumption.

ASSUMPTION A11: h satisfies Assumptions A2–A6, and if $g(u) = f(u, \theta_0)$ for any $\theta_0' = (\beta_0', \psi_0')$ in Θ, then $\gamma'_* = (\beta_0', \alpha'_*)$ for γ_* in Γ.

This ensures that $\tilde{\beta}_n$ is a QMLE consistent for β_0, the parameter vector of interest, and that $\sqrt{n}(\tilde{\beta}_n - \beta_0)$ is asymptotically normal, via Theorem 3.2.

The misspecification indicator is $\tilde{\beta}_n - \hat{\beta}_n$, so we investigate the asymptotic distribution of $\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n)$. We anticipate that with appropriate regularity conditions, $\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n)$ will be normally distributed asymptotically with mean zero in the absence of misspecification. Given a consistent estimator for the asymptotic covariance matrix, we can form an asymptotic χ^2 statistic of the Wald [**31**] type.

We adopt the following notation. Let $A^f(\theta) = \{E(\partial^2 \log f(U_t, \theta)/\partial\theta_i \partial\theta_j)\}$ and let $A^h(\gamma) = \{E(\partial^2 \log h(U_t, \gamma)/\partial\gamma_i \partial\gamma_j)\}$, and define $B^f(\theta)$ and $B^h(\gamma)$ similarly. The matrices $A^f(\theta)$ and $B^f(\theta)$ are $p \times p$, while $A^h(\gamma)$ and $B^h(\gamma)$ are $q \times q$. For those values of θ and γ for which inverses exist, we write $A^f(\theta)^{-1}$ and $A^h(\gamma)^{-1}$. We shall often use the $k \times p$ and $k \times q$ submatrices of $A^f(\theta)^{-1}$ and $A^h(\gamma)^{-1}$ obtained by deleting the last $p - k$ and $q - k$ rows respectively. These submatrices will be denoted $A^{f,\,\beta\theta}(\theta)^{-1}$ and $A^{h,\,\beta\gamma}(\gamma)^{-1}$. We define the $p \times q$ matrix

$$R(\theta, \gamma) = \{E(\partial \log f(U_t, \theta)/\partial\theta_i \cdot \partial \log h(U_t, \gamma)/\partial\gamma_j)\},$$
$$(i = 1, \ldots, p;\ j = 1, \ldots, q).$$

Next, when inverses exist we define the $k \times k$ matrix

$$\begin{aligned} S(\theta,\gamma) = {} & A^{h,\,\beta\gamma}(\gamma)^{-1}B^{h}(\gamma)A^{h,\,\beta\gamma}(\gamma)^{-1'} \\ & - A^{h,\,\beta\gamma}(\gamma)^{-1}R(\theta,\gamma)'A^{f,\,\beta\theta}(\theta)^{-1'} \\ & - A^{f,\,\beta\theta}(\theta)^{-1}R(\theta,\gamma)A^{h,\,\beta\gamma}(\gamma)^{-1'} \\ & + A^{f,\,\beta\theta}(\theta)^{-1}B^{f}(\theta)A^{f,\,\beta\theta}(\theta)^{-1}. \end{aligned}$$

$S(\theta_*,\gamma_*)$ turns out to be the asymptotic covariance matrix of $\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n)$, so we require the following assumption.

ASSUMPTION A12: $S(\theta_*,\gamma_*)$ is nonsingular.

In practice, Assumption A12 can always be satisfied by proper choice of $\mathbb{B}$. We need a consistent estimator for $S(\theta_*,\gamma_*)$, and a natural choice is

$$\begin{aligned} S_n(\hat{\theta}_n,\tilde{\gamma}_n) = {} & A_n^{h,\,\beta\gamma}(\tilde{\gamma}_n)^{-1}B_n^{h}(\tilde{\gamma}_n)A_n^{h,\,\beta\gamma}(\tilde{\gamma}_n)^{-1'} \\ & - A_n^{h,\,\beta\gamma}(\tilde{\gamma}_n)^{-1}R_n(\hat{\theta}_n,\tilde{\gamma}_n)A_n^{f,\,\beta\theta}(\hat{\theta}_n)^{-1'} \\ & - A_n^{f,\,\beta\theta}(\hat{\theta}_n)^{-1}R_n(\hat{\theta}_n\tilde{\gamma}_n)A_n^{h,\,\beta\gamma}(\tilde{\gamma}_n)^{-1'} \\ & + A_n^{f,\,\beta\theta}(\hat{\theta}_n)^{-1}B_n^{f}(\hat{\theta}_n)A_n^{f,\,\beta\theta}(\hat{\theta}_n)^{-1'} \end{aligned}$$

where $A_n^{h,\,\beta\gamma}, B_n^{h}, A_n^{f,\,\beta\theta}, B_n^{f}$ are the finite sample analogs of $A^{h,\,\beta\gamma}, B^{h}, A^{f,\,\beta\theta}, B^{f}$, and

$$R_n(\theta,\gamma) = \left\{ n^{-1}\sum_{t=1}^{n} \partial \log f(U_t,\theta)/\partial\theta_i \cdot \partial \log h(U_t,\gamma)/\partial\gamma_j \right\}$$

$$(i = 1,\ldots,p;\ j = 1,\ldots,q).$$

That $R_n(\hat{\theta}_n,\tilde{\gamma}_n)$ converges to the appropriate limit is ensured by Assumptions A5 and A11.

THEOREM 5.1 (Hausman Test): *Given Assumptions A1–A6, A11 and A12, if* $g(u) = f(u,\theta_0)$ *for* θ_0 *in* Θ, *then*

$$\mathcal{H}_n = n(\tilde{\beta}_n - \hat{\beta}_n)'S_n(\hat{\theta}_n,\tilde{\gamma}_n)^{-1}(\tilde{\beta}_n - \hat{\beta}_n) \overset{A}{\sim} \chi_k^2. \tag{5.1}$$

To carry out the Hausman test, one computes (5.1) and compares it to the critical value for the χ_k^2 distribution at a given significance level. If (5.1) exceeds

this value and Assumptions A1–A6, A11 and A12 are maintained, one must reject the hypothesis that the model is correctly specified.

This result contains as special cases the linear and nonlinear regression model specification tests of White [**33, 34**]. It differs from Hausman's result [**17**, Theorem 2.1] in several particulars, but not in spirit. Hausman's result requires $\tilde{\beta}_n$ and $\hat{\beta}_n$ to be jointly consistent uniformly asymptotically normal (JCUAN) estimators for β_0. In particular cases, this can be rather tedious to verify (e.g., using Parzen's [**24**] results). Our conditions automatically ensure that $\tilde{\beta}_n$ and $\hat{\beta}_n$ are JCUAN.

The statistic (5.1) also differs from Hausman's in the choice of the covariance matrix estimator. Hausman's statistic replaces $S_n(\hat{\theta}_n, \tilde{\gamma}_n)$ with the difference of the covariance matrix estimator for $\tilde{\beta}_n$ and that for $\hat{\beta}_n$. This estimator is consistent in the absence of misspecification (Hausman [**17**, Lemma 2.1]), and significantly reduces the required computation. When the model is misspecified, it can fail to be positive semi-definite for any n and is not necessarily consistent. These difficulties are avoided by using $S_n(\hat{\theta}_n, \tilde{\gamma}_n)$.[5]

Holly [**19**] points out that the Hausman test can have low power against particular alternatives. However, as Hausman and Taylor [**18**] show, the Hausman test has optimal power properties against alternatives which result in parameter estimator inconsistency.

Another way to detect the inconsistency of a supposed MLE for the parameters of interest is to observe that when the model is correctly specified, the gradient $\nabla L_n(U, \theta_0)$ has expectation zero. In the absence of misspecification, it is usually easy to find a QMLE consistent for θ_0, say $\tilde{\theta}_n$. Thus, we would expect $\nabla L_n(U, \tilde{\theta}_n)$ to be close to zero in the absence of misspecification, but generally not otherwise, since $\tilde{\theta}_n$ generally won't converge to θ_*.

A very useful result can be obtained by constructing $\tilde{\theta}_n$ in the following way. As before, let $\tilde{\gamma}_n' = (\tilde{\beta}_n', \tilde{\alpha}_n')$ maximize $n^{-1}\sum_{t=1}^n \log h(U_t, \gamma)$ over Γ. Next, let $\tilde{\psi}_n$ maximize $\nabla L_n(U, \tilde{\beta}_n, \psi)$ over Ψ (so that $\nabla_\psi L_n(U, \tilde{\beta}_n, \tilde{\psi}_n) = 0$), and set $\tilde{\theta}_n' = (\tilde{\beta}_n', \tilde{\psi}_n')$. Then $\nabla_\beta L_n(U, \tilde{\theta}_n)$ serves as an indicator of model misspecification and we investigate the asymptotic distribution of $\sqrt{n}\nabla_\beta L_n(U, \tilde{\theta}_n)$. It's reasonable to expect that under appropriate regularity conditions, this will be normally distributed asymptotically with mean zero in the absence of misspecification. Given a consistent estimator of the asymptotic covariance matrix, we can then form an asymptotic χ^2 statistic.

In fact, the necessary regularity conditions have already been given. Let $A_n^{f,\,\beta\beta}(\theta)^{-1}$ be the $k \times k$ submatrix of $A_n^f(\theta)^{-1}$ obtained by deleting the last $p - k$ columns from $A_n^{f,\,\beta\theta}(\theta)^{-1}$. The desired result follows.

[5] A referee points out that the asymptotic slopes criterion of Bahadur [**4**] (see also Geweke [**16**]) applies here, suggesting that the present statistic will dominate Hausman's under some alternatives. This domination is not necessarily uniform over the alternative hypothesis space, indicating that a statistic which dominates both Hausman's and the present statistic in the approximate slopes sense could be obtained as the maximum of the two.

THEOREM 5.2 (Gradient Test): *Given Assumptions A1–A6, A11 and A12, if* $g(u) = f(u, \theta_0)$ *for* θ_0 *in* Θ *then*

$$\mathcal{G}_n = \nabla_\beta L_n(U, \tilde{\theta}_n)' A_n^{f,\beta\beta}(\tilde{\theta}_n)^{-1} S_n(\tilde{\theta}_n, \tilde{\gamma}_n)^{-1} \times A_n^{f,\beta\beta}(\tilde{\theta}_n)^{-1} \nabla_\beta L_n(U, \tilde{\theta}_n) \overset{A}{\sim} \chi_k^2. \tag{5.2}$$

Moreover, $\mathcal{K}_n - \mathcal{G}_n \xrightarrow{p} 0$.

The gradient test is performed by comparing $\mathcal{G}_n$ to the critical value for the χ_k^2 distribution at a given significance level, and rejecting the hypothesis of no misspecification if $\mathcal{G}_n$ exceeds this value.

In effect, the gradient test is a Lagrange-multiplier procedure which tests the hypothesis that $\theta_* = \theta_0$ is consistently estimated by $\tilde{\theta}_n$. Theorem 5.2 establishes that the $\mathcal{G}_n$ statistic is asymptotically equivalent to the $\mathcal{K}_n$ statistic under the null hypothesis, a fact precisely analogous to the asymptotic equivalence of the Wald and Lagrange Multiplier statistics of Section 3. In contrast to the $\mathcal{K}_n$ statistic, the $\mathcal{G}_n$ statistic doesn't require full computation of the MLE $\hat{\theta}_n$, which is often a substantial convenience. Essentially, the $\mathcal{G}_n$ statistic compares $\tilde{\beta}_n$ to the value obtained by taking one step of a Newton-Raphson iteration from $\tilde{\beta}_n$ (which is asymptotically equivalent to the MLE, as LeCam [**23**] shows).

6. SUMMARY AND CONCLUDING REMARKS

In this paper we provide a unified framework for studying the consequences and detection of model misspecification when maximum likelihood techniques are used. Misspecification can cause parameter estimators to be inconsistent for particular parameters of interest, as well as invalidating standard techniques of inference. Specification robust procedures are provided here. The properties of the QMLE are also exploited to yield several useful tests for model misspecification.

Taken together, the specification tests of Sections 4 and 5 have the potential to detect a broad range of model misspecifications. Given the characteristics of the tests, the following sequential procedures may often be convenient. First, apply the information matrix test of Section 4. If the null hypothesis of no misspecification is not rejected, one may have confidence that standard maximum likelihood techniques of estimation and inference are valid. If the null hypothesis is rejected, one can investigate the seriousness of the misspecification using the tests of Section 5. If these don't reject the null hypothesis of no misspecification, one may have confidence that the estimated parameters will be consistent for parameters of interest, although inferences must be based on the specification robust procedures of Section 3. Otherwise, one has an indication that the parameter estimator is inconsistent for the parameters of interest, so that the

model specification must be carefully re-examined.[6] Since the tests are not obviously independent, the actual size of a test for misspecification using this procedure may be difficult to determine. Nevertheless, Bonferroni bounds on the size of such a test are easily found, and this procedure should provide relatively low cost insurance against the improper use of a misspecified model.

Finally, we note that misspecifications which only result in estimator inefficiency (but no parameter or covariance matrix estimator inconsistency) will not be readily detected by the tests of Sections 4 and 5. In some cases, one may be interested in whether such misspecifications remain. Since $I(g:f,\theta_*)=0$ if and only if the probability model is correct (see Footnote 2), the KLIC serves as an indicator for such misspecifications. The KLIC is not observable; however, it can be consistently estimated by

$$\hat{I}_n(g:f,\hat{\theta}_n) = -\hat{H}(g) - n^{-1}\sum_{t=1}^{n}\log f(U_t,\hat{\theta}_n)$$

where $\hat{H}(g)$ is the nonparametric entropy estimator of Ahmad and Lin [**1**]. Tests of the hypothesis that $I(g:f,\theta_*)=0$ might then be based on $\sqrt{n}\hat{I}_n(g:f,\hat{\theta}_n)$. However, establishing the asymptotic distribution of this statistic is a non-trivial problem which we leave to future work.

University of California, San Diego

Manuscript received March, 1980; revision received December, 1980.

MATHEMATICAL APPENDIX

PROOF OF THEOREM 2.1: This follows immediately from Lemma 3 of LeCam [**23**].

PROOF OF THEOREM 2.2: Given Assumptions A1–A3, the conditions of Theorem 2.1 of White [**36**] are satisfied, and the result follows immediately.

PROOF OF THEOREM 3.1: (i) By the mean value theorem, for all θ in Θ', a convex compact subset of Θ,

$$\int\log f(u,\theta)\,g(u)\,dv = \int\log f(u,\theta_*)\,g(u)\,dv + \nabla\int\log f(u,\theta)\,g(u)\,dv|_{\theta^*}\cdot(\theta-\theta_*)$$

$$+ (\theta-\theta_*)'\nabla^2\int\log f(u,\theta)\,g(u)\,dv|_{\bar{\theta}}\cdot(\theta-\theta_*)/2,$$

where $\bar{\theta}$ lies on the segment joining θ and θ_*, since $\int\log f(u,\theta)g(u)dv$ is continuously differentiable of order two given assumption A3(a), A4, and A5. Further, Assumptions A3(a), A4, and A5 ensure that

[6] For an application of this procedure to the nonlinear regression model see White [**36**]. There, it turns out that a good choice for the QMLE of the Hausman test is suggested by the results of the information matrix test.

the derivatives may be taken inside the integrals (Corollary 5.8 of Bartle [5]) so that

$$\int \log f(u,\theta)\, g(u)\, dv - \int \log f\left(u,\theta_*\right) g(u)\, dv = \int \nabla \log f\left(u,\theta_*\right) g(u)\, dv \cdot \left(\theta - \theta_*\right)$$

$$+ \left(\theta - \theta_*\right)' \int \nabla^2 \log f(u,\bar{\theta})\, g(u)\, dv \cdot \left(\theta - \theta_*\right)/2$$

for all θ in Θ'. If θ_* is globally identifiable, then $\theta - \theta_* \neq 0$ implies that

$$\int \log f(u,\theta)\, g(u)\, dv - \int \log f\left(u,\theta_*\right) g(u)\, dv < 0.$$

Further, $\int \nabla \log f(u,\theta_*) g(u)\, dv = 0$ under Assumptions A1–A6(a) (see (A.1) below), so that for all $\theta - \theta_* \neq 0$ ($\theta \in \Theta'$),

$$\left(\theta - \theta_*\right)' A(\bar{\theta}) \left(\theta - \theta_*\right) < 0,$$

which implies that $A(\bar{\theta})$ is negative definite. Now choosing an arbitrary open neighborhood $\mathcal{N}$ appropriately, we can make $\bar{\theta}$ as close to θ_* as we like; since θ_* is a regular point of $A(\theta)$, the rank of $A(\theta)$ is constant in a local neighborhood of θ_*, so that if $A(\bar{\theta})$ is negative definite in that neighborhood, so must be $A(\theta_*)$.

(ii) If $A(\theta_*)$ is negative definite and $A(\theta)$ is continuous in θ (as ensured by Assumptions A4 and A5), then there exists an open neighborhood $\mathcal{N}$ of θ_* such that $A(\theta)$ is negative definite for any θ in $\mathcal{N}$. Since (A.1) holds regardless of whether or not θ_* uniquely minimizes $I(g : f, \theta)$ on $\mathcal{N}$, we again have

$$\int \log f(u,\theta)\, g(u)\, dv - \int \log f\left(u,\theta_*\right) g(u)\, dv = \left(\theta - \theta_*\right)' A(\bar{\theta}) \left(\theta - \theta_*\right)/2 < 0.$$

Since this holds for all $\theta \neq \theta_*$ in $\mathcal{N}$, θ_* is the unique minimizer of $I(g : f, \theta)$ in $\mathcal{N}$, and is therefore locally identifiable.

PROOF OF THEOREM 3.2: Given Assumptions A1–A6, the conditions of Theorem 3.3 of White [36] are satisfied, and the result follows immediately. For later reference we state several useful intermediate results. Given Assumptions A1–A6,

$$E\left(\nabla \log f\left(U_t, \theta_*\right)\right) = 0, \tag{A.1}$$

$$\sqrt{n}\left(\hat{\theta}_n - \theta_*\right) + A\left(\theta_*\right)^{-1} n^{-1/2} \sum_{t=1}^{n} \nabla \log f\left(U_t, \theta_*\right) \xrightarrow{p} 0, \tag{A.2}$$

and

$$A_n(\hat{\theta}_n) \xrightarrow{\text{a.s.}} A\left(\theta_*\right), \qquad B_n(\hat{\theta}_n) \xrightarrow{\text{a.s.}} B\left(\theta_*\right). \tag{A.3}$$

PROOF OF THEOREM 3.3: Given Assumptions A1–A6 and $g(u) = f(u,\theta)$ for any θ interior to Θ, (A.1) implies that for any such θ,

$$\int \nabla \log f(u,\theta)\, f(u,\theta)\, dv = 0.$$

If Assumption A7 holds, then Corollary 5.8 of Bartle [5] allows differentiation to be taken inside the integral above, so that

$$\int \left(\nabla^2 \log f(u,\theta) + \nabla \log f(u,\theta) \nabla \log f(u,\theta)'\right) f(u,\theta)\, dv = 0.$$

Since $I(g : f, \theta_0) = 0$ if and only if $g(u) = f(u,\theta_0)$ for θ_0 in Θ, we have immediately that $\theta_* = \theta_0$. Evaluating the integral above at θ_0, we obtain $A(\theta_0) + B(\theta_0) = 0$, implying $C(\theta_0) = -A(\theta_0)^{-1} = B(\theta_0)^{-1}$.

PROOF OF THEOREM 3.4: The proof is identical to that of Theorem 3.4 of White [**34**], where ∇q_t^0 in the notation of the proof is replaced by $-\nabla \log f(U_t, \theta_*)$ in the present notation.

PROOF OF THEOREM 3.5: First we show that under H_0, the constrained QMLE, $\tilde{\theta}_n$, is consistent for θ_* (which minimizes the KLIC over Θ). By definition, $\tilde{\theta}_n$ solves

$$\max_{\theta \in \Theta_s} L_n(U, \theta)$$

where $\Theta_s = \{\theta \in \Theta : s(\theta) = 0\}$. Since s is continuous and Θ is compact, Θ_s is compact. Given Assumptions A1–A3 (with Θ_s replacing Θ) and H_0 (ensuring $\theta_* \in \Theta_s$), it follows from Theorem 2.2 that $\tilde{\theta}_n \xrightarrow{\text{a.s.}} \theta_*$.

Given Assumption A4 and since $\nabla s(\theta_*)$ has full row rank, the Lagrange Multiplier Theorem (e.g., Bartle [**6**, Theorem 42.9]) ensures the existence of a real $r \times 1$ vector of Lagrange multipliers $\tilde{\lambda}_n$ such that

$$n^{-1} \sum_{t=1}^{n} \nabla \log f(U_t, \tilde{\theta}_n) + \nabla s(\tilde{\theta}_n)' \tilde{\lambda}_n = 0, \tag{A.4}$$

$$s(\tilde{\theta}_n) = 0. \tag{A.5}$$

Given Assumptions A2, A4, and H_0, the mean-value theorem for random functions (Jennrich [**21**, Lemma 3]) allows us to write

$$n^{-1} \sum_{t=1}^{n} \nabla \log f(U_t, \ddot{\theta}_n) = n^{-1} \sum_{t=1}^{n} \nabla \log f(U_t, \theta_*) + A_n(\bar{\theta}_n)(\ddot{\theta}_n - \theta_*) \tag{A.6}$$

where $(\ddot{\theta}_n)$ is a sequence tail-equivalent to $(\tilde{\theta}_n)$ lying in a convex compact neighborhood of θ_* interior to Θ, and $\bar{\theta}_n$ (which varies from row to row of A_n) lies on the segment connecting $\ddot{\theta}_n$ and θ_*. Given H_0, the mean-value theorem applied to $s(\ddot{\theta}_n) = 0$ yields

$$s(\ddot{\theta}_n) = s(\theta_*) + \nabla s(\check{\theta}_n)(\ddot{\theta}_n - \theta_*) = 0 \tag{A.7}$$

where $\check{\theta}_n$ lies on the segment connecting $\ddot{\theta}_n$ to θ_*. Substituting (A.6) and (A.7) into (A.4) and (A.5), setting $s(\theta_*) = 0$ and multiplying by $\sqrt{n}$ yields

$$n^{-1/2} \sum_{t=1}^{n} \nabla \log f(U_t, \theta_*) + A_n(\bar{\theta}_n) \sqrt{n}(\ddot{\theta}_n - \theta_*) + \nabla s(\ddot{\theta}_n)' \sqrt{n} \tilde{\lambda}_n = 0, \tag{A.8}$$

$$\nabla s(\check{\theta}_n) \cdot \sqrt{n}(\ddot{\theta}_n - \theta_*) = 0, \tag{A.9}$$

for all n sufficiently large.

Since $\bar{\theta}_n \xrightarrow{\text{a.s.}} \theta_*$, it follows analogously to (A.3) that $A_n(\bar{\theta}_n) \xrightarrow{\text{a.s.}} A(\theta_*)$. The nonsingularity of $A(\theta_*)$ ensures that $A_n(\bar{\theta}_n)$ is nonsingular almost surely for all n sufficiently large, allowing us to premultiply (A.8) by $\nabla s(\check{\theta}_n) A_n(\bar{\theta}_n)^{-1}$, which yields

$$\nabla s(\check{\theta}_n) A_n(\bar{\theta}_n)^{-1} n^{-1/2} \sum_{t=1}^{n} \nabla \log f(U_t, \theta_*) + \nabla s(\check{\theta}_n) A_n(\bar{\theta}_n)^{-1} \nabla s(\ddot{\theta}_n)' \sqrt{n} \tilde{\lambda}_n = 0, \tag{A.10}$$

upon setting $\nabla s(\check{\theta}_n) \sqrt{n}(\ddot{\theta}_n - \theta_*) = 0$.

Since $\nabla s(\check{\theta}_n) A_n(\bar{\theta}_n)^{-1} \nabla s(\ddot{\theta}_n)'$ is nonsingular almost surely for all n sufficiently large, we have

$$\sqrt{n} \tilde{\lambda}_n = -\left[\nabla s(\check{\theta}_n) A_n(\bar{\theta}_n)^{-1} \nabla s(\ddot{\theta}_n)'\right]^{-1} \nabla s(\check{\theta}_n) A_n(\bar{\theta}_n)^{-1} n^{-1/2} \sum_{t=1}^{n} \nabla \log f(U_t, \theta_*).$$

But by Lemma 3.3 of White [**34**] this has the same asymptotic distribution as

$$\sqrt{n} \lambda_n^* \equiv -\left[\nabla s(\theta_*) A(\theta_*)^{-1} \nabla s(\theta_*)'\right]^{-1} \nabla s(\theta_*) A(\theta_*)^{-1} n^{-1/2} \sum_{t=1}^{n} \nabla \log f(U_t, \theta_*)$$

since $\sqrt{n}\tilde{\lambda}_n - \sqrt{n}\lambda_n^* \xrightarrow{p} 0$ by 2c.4(x.a) of Rao [26]. This latter fact follows from the consistency of $\check{\theta}_n$ and $\ddot{\theta}_n$ for θ_*, $A_n(\ddot{\theta}_n)$ for $A(\theta_*)$ and continuity, together with the convergence in distribution of $n^{-1/2}\sum_{t=1}^n \nabla \log f(U_t, \theta_*)$.

The multivariate Lindeberg-Levy central limit theorem ensures

$$\sqrt{n}\lambda_n^* \overset{A}{\sim} N\big(0,\ Q(\theta_*)\big), \qquad \text{where}$$

$Q(\theta_*) = [\nabla s(\theta_*)A(\theta_*)^{-1}\nabla s(\theta_*)']^{-1}\nabla s(\theta_*)C(\theta_*)\nabla s(\theta_*)'[\nabla s(\theta_*)A(\theta_*)^{-1}\nabla s(\theta_*)']^{-1}$ given Assumptions A1–A6 and H_0. It follows from Lemma 3.3 of White [34] that $\mathfrak{LR}_n = n\tilde{\lambda}_n' Q_n(\tilde{\theta}_n)^{-1}\tilde{\lambda}_n \overset{A}{\sim} \chi_r^2$, where

$$Q_n(\tilde{\theta}_n) = \left[\nabla s(\tilde{\theta}_n)A_n(\tilde{\theta}_n)^{-1}\nabla s(\tilde{\theta}_n)'\right]^{-1}\nabla s(\tilde{\theta}_n)C_n(\tilde{\theta}_n)\nabla s(\tilde{\theta}_n)' \times \left[\nabla s(\tilde{\theta}_n)A_n(\tilde{\theta}_n)^{-1}\nabla s(\tilde{\theta}_n)'\right]^{-1}$$

is consistent for $Q(\theta_*)$ by Lemma 2.2 of White [34], given Assumptions A1–A6 and H_0.

From (A.4) we note that we may write

$$\tilde{\lambda}_n = \left[\nabla s(\tilde{\theta}_n)A_n(\tilde{\theta}_n)^{-1}s(\tilde{\theta}_n)'\right]^{-1}\nabla s(\tilde{\theta}_n)A_n(\tilde{\theta}_n)^{-1}n^{-1}\nabla L_n(U, \tilde{\theta}_n)$$

to obtain the scores form of the Lagrange Multiplier statistic

$$\mathfrak{LR}_n = \nabla L_n(U, \tilde{\theta}_n)'A_n(\tilde{\theta}_n)^{-1}\nabla s(\tilde{\theta}_n)'[\nabla s(\tilde{\theta}_n)C_n(\tilde{\theta}_n)\nabla s(\tilde{\theta}_n)']^{-1} \times \nabla s(\tilde{\theta}_n)A_n(\tilde{\theta}_n)^{-1}\nabla L_n(U, \tilde{\theta}_n) = n\tilde{\lambda}_n' Q_n(\tilde{\theta}_n)^{-1}\tilde{\lambda}_n \overset{A}{\sim} \chi_r^2.$$

The fact that $\mathfrak{LR}_n - \mathfrak{W}_n \xrightarrow{p} 0$ follows from 2c.4(xiv) of Rao [26] since

$$\sqrt{n}s(\hat{\theta}_n) - \nabla s(\theta_*)A(\theta_*)^{-1}\nabla s(\theta_*)'\sqrt{n}\lambda_n^* \xrightarrow{p} 0$$

(see the proof of Theorem 3.5 of White [34]), where $\hat{\theta}_n$ is the unconstrained QMLE.

PROOF OF THEOREM 4.1: By Theorem 2.2, Assumptions A1–A3 ensure that $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_*$; since $f(u, \theta_0) = g(u)$, we have $\theta_* = \theta_0$, so that $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$. Thus, there exists a sequence $(\tilde{\theta}_n)$ tail equivalent to $(\hat{\theta}_n)$ such that each $\tilde{\theta}_n$ takes its values in a convex compact neighborhood of θ_0, which is interior to Θ by Assumption A.6(a). Given Assumptions A4 and A8, Lemma 3 of Jennrich guarantees the existence of measurable Θ-valued functions $\bar{\theta}_n$ such that

$$\sqrt{n}D_n(\tilde{\theta}_n) = \sqrt{n}D_n(\theta_0) + \nabla D_n(\bar{\theta}_n)\sqrt{n}(\tilde{\theta}_n - \theta_0)$$

where each $\bar{\theta}_n$ lies on the sequent joining $\tilde{\theta}_n$ and θ_0. (As before, $\nabla D_n(\bar{\theta}_n)$ is a shorthand notation; each row of ∇D_n depends on a different $\bar{\theta}_n$. This makes no difference asymptotically.)

Since $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is asymptotically normal by Theorem 3.2 under Assumptions A1–A6 and the tail equivalence of $\hat{\theta}_n$ and $\tilde{\theta}_n$, and since $\nabla D_n(\bar{\theta}_n) - \nabla D(\theta_0) \xrightarrow{\text{a.s.}} 0$ by Lemma 3.1 of White [36] given Assumptions A1–A4, A8, and A9, $(\nabla D_n(\bar{\theta}_n) - \nabla D(\theta_0))\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{p} 0$ by 2c.4(x.a) and 2c.4(xiii) of Rao [26]. Since

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) + A(\theta_0)^{-1}n^{-1/2}\sum_{t=1}^n \nabla \log f(U_t, \theta_0) \xrightarrow{p} 0,$$

given Assumptions A1–A6 from the tail equivalence of $\hat{\theta}_n$ and $\tilde{\theta}_n$ and since $\nabla D(\theta_0)$ is finite by Assumption A9,

$$\nabla D(\theta_0)\left(\sqrt{n}(\tilde{\theta}_n - \theta_0) + A(\theta_0)^{-1}n^{-1/2}\sum_{t=1}^n \nabla \log f(U_t, \theta_0)\right) \xrightarrow{p} 0.$$

It follows that

$$\nabla D_n(\bar{\theta}_n)\sqrt{n}(\tilde{\theta}_n - \theta_0) + \nabla D(\theta_0)A(\theta_0)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla \log f(U_t, \theta_0) \xrightarrow{p} 0,$$

so that

$$\sqrt{n}D_n(\tilde{\theta}_n) - \sqrt{n}D_n(\theta_0) + \nabla D(\theta_0)A(\theta_0)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla \log f(U_t, \theta_0) \xrightarrow{p} 0.$$

Let $V(\theta_0)^{-1/2}$ be the symmetric positive definite matrix such that $V(\theta_0)^{-1/2}V(\theta_0)^{-1/2} = V(\theta_0)^{-1}$, which exists and is finite by Assumptions A9 and A10. Then

$$V(\theta_0)^{-1/2}\left(\sqrt{n}D_n(\tilde{\theta}_n) - \sqrt{n}D_n(\theta_0) + \nabla D(\theta_0)A(\theta_0)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla \log f(U_t, \theta_0)\right) \xrightarrow{p} 0.$$

It is easily verified that $V(\theta_0)^{-1/2}\nabla D(\theta_0)A(\theta_0)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla \log f(U_t, \theta_0)$ is distributed asymptotically as $N(0, I_q)$ given Assumptions A1–A6, A8–A10 using the Lindeberg-Levy central limit theorem. By Lemma 3.3 of White [**34**] and the tail equivalence of $\hat{\theta}_n$ and $\tilde{\theta}_n$ it follows that $\sqrt{n}D_n(\hat{\theta}_n) \overset{A}{\sim} N(0, V(\theta_0))$ and that

$$\mathcal{J}_n = nD_n(\hat{\theta}_n)'V_n(\hat{\theta}_n)^{-1}D_n(\hat{\theta}_n) \overset{A}{\sim} \chi^2_q$$

provided that $V_n(\hat{\theta}_n)$ is consistent for $V(\theta_0)$. But this follows from repeated application of Lemma 3.1 of White [**36**] given Assumptions A1–A6, A8, and A9, and the proof is complete.

PROOF OF THEOREM 5.1: Given Assumptions A1–A6, (A.2) implies

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{p} -A^f(\theta_*)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla_\theta \log f(U_t, \theta_*);$$

similarly, given Assumptions A1 and A11, (A.2) implies

$$\sqrt{n}(\tilde{\gamma}_n - \gamma_*) \xrightarrow{p} -A^h(\gamma_*)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla_\gamma \log h(U_t, \gamma_*).$$

Since we are interested only in the $k \times 1$ subvectors β of θ and γ, we specialize the above to

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{p} -A^{f,\,\beta\theta}(\theta_*)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla_\theta \log f(U_t, \theta_*),$$

$$\sqrt{n}(\tilde{\beta}_n - \beta_0) \xrightarrow{p} -A^{h,\,\beta\gamma}(\gamma_*)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla_\gamma \log h(U_t, \gamma_*),$$

where $A^{f,\,\beta\theta}(\theta_*)^{-1}$ and $A^{h,\,\beta\gamma}(\gamma_*)^{-1}$ are the appropriate $k \times p$ and $k \times q$ submatrices of $A^f(\theta_*)^{-1}$ and $A^h(\gamma_*)^{-1}$ and we have used the fact that $\theta_*' = (\beta_0', \psi_0')$, $\gamma_*' = (\beta_0', \alpha_*')$ when $g(u) = f(u, \theta_0)$ for θ_0 in Θ. It follows that

$$\text{(A.11)}\qquad S(\theta_*, \gamma_*)^{-1/2}\left[\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n) - n^{-1/2}\sum_{t=1}^{n}\left(A^{f,\,\beta\theta}(\theta_*)^{-1}\nabla_\theta \log f(U_t, \theta_*) - A^{h,\,\beta\gamma}(\gamma_*)\nabla_\gamma \log h(U_t, \gamma_*)\right)\right] \xrightarrow{p} 0,$$

where $S(\theta_*, \gamma_*)^{-1/2}$ is the symmetric positive definite matrix such that $S(\theta_*, \gamma_*)^{-1/2}S(\theta_*, \gamma_*)^{-1/2} = S(\theta_*, \gamma_*)^{-1}$, which exists and is finite given Assumptions A5, A11, and A12. It is easily verified that

$$S(\theta_*, \gamma_*)^{-1/2}\left[n^{-1/2}\sum_{t=1}^{n}A^{f,\,\beta\theta}(\theta_*)^{-1}\nabla_\theta \log f(U_t, \theta_*) - A^{h,\,\beta\gamma}(\gamma_*)^{-1}\nabla_\gamma \log h(U_t, \gamma_*)\right]$$

is distributed asymptotically as $N(0, I_k)$ given Assumptions A1–A6, A11, and A12 using the Lindeberg-Levy central limit theorem. This fact and (A.11) imply by Lemma 3.3 of White [**34**] that

$$\mathcal{K}_n = n(\tilde{\beta}_n - \hat{\beta}_n)' S_n(\hat{\theta}_n, \hat{\gamma}_n)^{-1}(\tilde{\beta}_n - \hat{\beta}_n) \overset{A}{\sim} \chi^2_k$$

provided that $S_n(\hat{\theta}_n, \tilde{\gamma}_n)$ is consistent for $S(\theta_*, \gamma_*)$. But this follows from repeated application of Lemma 3.1 of White [**36**] given Assumptions A1–A6, and A11, so the proof is complete.

For the next result, partition $A^f(\theta)$ as

$$A^f(\theta) = \begin{bmatrix} A^f_{\beta\beta}(\theta) & A^f_{\beta\psi}(\theta) \\ A^f_{\psi\beta}(\theta) & A^f_{\psi\psi}(\theta) \end{bmatrix}$$

and $A^f(\theta)^{-1}$ as

$$A^f(\theta)^{-1} = \begin{bmatrix} A^{f,\,\beta\beta}(\theta)^{-1} & A^{f,\,\beta\psi}(\theta)^{-1} \\ A^{f,\,\psi\beta}(\theta)^{-1} & A^{f,\,\psi\psi}(\theta)^{-1} \end{bmatrix} = \begin{bmatrix} A^{f,\,\beta\theta}(\theta)^{-1} \\ A^{f,\,\psi\theta}(\theta)^{-1} \end{bmatrix}.$$

PROOF OF THEOREM 5.2: Given Assumption A11, it follows from Theorem 2.2 that $\tilde{\beta}_n$ is a QMLE consistent for β_0, i.e., $\tilde{\beta}_n \overset{\text{a.s.}}{\to} \beta_0$ when $g(u) = f(u, \theta_0)$ for θ_0 in Θ. By definition, $\tilde{\psi}_n$ maximizes $n^{-1}\sum_{t=1}^n \log f(U_t, \tilde{\beta}_n, \psi)$ over Ψ; we need to show that $\tilde{\psi}_n$ is consistent for ψ_0. Given Assumptions A1–A3 $L_n(U, \theta)$ converges to $E(\log f(U_t, \theta))$ uniformly for all θ in Θ and almost every sequence (U_t) by Theorem 2 of Jennrich [**21**]. Choose (U_t) so that this occurs and also $\tilde{\beta}_n \to \beta_0$. Since Ψ is compact, a sequence $(\tilde{\psi}_n)$ has a limit point in Ψ, say ψ^*. Consider a subsequence $(\tilde{\psi}_{n_j})$ which converges to ψ^*. By the triangle inequality

$$\begin{aligned} (A.12) \qquad & |L_{n_j}(U, \tilde{\beta}_{n_j}, \tilde{\psi}_{n_j}) - E(\log f(U_t, \beta_0, \psi^*))| \\ & \leqslant |L_{n_j}(U, \tilde{\beta}_{n_j}, \tilde{\psi}_{n_j}) - E\big(\log f(U_t, \tilde{\beta}_{n_j}, \tilde{\psi}_{n_j})\big)| \\ & + |E\big(\log f(U_t, \tilde{\beta}_{n_j}, \tilde{\psi}_{n_j})\big) - E(\log f(U_t, \beta_0, \psi^*))|. \end{aligned}$$

Since $L_n(U, \theta)$ converges uniformly to $E(\log f(U_t, \theta))$, the first term on the right of (A.12) can be made arbitrarily small for all n_j sufficiently large. The second term is arbitrarily small for all n_j sufficiently large by the uniform continuity in θ of $E(\log f(U_t, \theta))$, since $\tilde{\beta}_{n_j} \to \beta_0$ and $\tilde{\psi}_{n_j} \to \psi^*$. Hence for $\delta > 0$ and all n_j sufficiently large

$$|L_{n_j}(U, \tilde{\beta}_{n_j}, \tilde{\psi}_{n_j}) - E(\log f(U_t, \beta_0, \psi^*))| \leqslant \delta,$$

so that

$$E(\log f(U_t, \beta_0, \psi^*)) \geqslant L_{n_j}(U, \tilde{\beta}_{n_j}, \tilde{\psi}_{n_j}) - \delta.$$

Since $\tilde{\psi}_{n_j}$ maximizes $L_{n_j}(U, \tilde{\beta}_{n_j}, \psi)$,

$$E(\log f(U_t, \beta_0, \psi^*)) \geqslant L_{n_j}(U, \tilde{\beta}_{n_j}, \psi_0) - \delta.$$

Since L_n is uniformly continuous in θ and $\tilde{\beta}_n \to \beta_0$, for all n sufficiently large

$$|L_{n_j}(U, \beta_{n_j}, \psi_0) - L_{n_j}(U, \beta_0, \psi_0)| < \delta,$$

so that

$$E(\log f(U_t, \beta_0, \psi^*)) \geqq L_{n_j}(U, \beta_0, \psi_0) - 2\delta.$$

The uniform convergence of $L_n(U, \theta)$ to $E(\log f(U_t, \theta))$ guarantees that

$$|L_{n_j}(U, \beta_0, \psi_0) - E(\log f(U_t, \beta_0, \psi_0)| < \delta$$

for all n_j sufficiently large, so that

$$E(\log f(U_t, \beta_0, \psi^*)) \geqq E(\log f(U_t, \beta_0, \psi_0)) - 3\delta.$$

Since δ is arbitrary and since Assumption A3 guarantees that ψ_0 uniquely maximizes $E(\log f(U_t, \beta_0, \psi))$ when $g(u) = f(u, \theta_0)$ for θ_0 in Θ, it follows that $\psi^* = \psi_0$, regardless of the subsequence $(\tilde{\psi}_{j_n})$. Hence $(\tilde{\psi}_n)$ converges to ψ_0. Since the uniform convergence of $L_n(U, \theta)$ to $E(\log f(U_t, \theta))$ and $\tilde{\beta}_n \to \beta_0$ fail only on a set of measure zero, $\tilde{\psi}_n \overset{\text{a.s.}}{\to} \psi_0$.

Now consider the asymptotic behavior of $n^{-1/2}\sum_{t=1}^{n} \nabla_\beta \log f(U_t, \tilde{\theta}_n)$. Given $g(u) = f(u, \theta_0)$ for θ_0 in Θ, Assumptions A1, A2, A4, and A6, Lemma 3 of Jennrich [**21**] (the mean-value theorem for random functions) allows us to write

$$\text{(A.13)} \qquad n^{-1/2}\sum_{t=1}^{n} \nabla_\beta \log f\left(U_t, \ddot{\theta}_n\right) = n^{-1/2}\sum_{t=1}^{n} \nabla_\beta \log f(U_t, \theta_0) + n^{-1}\sum_{t=1}^{n} \nabla^2_{\beta\theta} \log f\left(U_t, \bar{\theta}_n\right) \cdot \sqrt{n}\left(\ddot{\theta}_n - \theta_0\right)$$

where $(\ddot{\theta}_n)$ is a sequence tail equivalent to $(\tilde{\theta}_n)$ lying in a convex compact neighborhood of θ_0, and $\bar{\theta}_n$ (which differs from row to row of $\nabla^2_{\beta\theta} \log f$) lies on the segment connecting $\ddot{\theta}_n$ to θ_0. To proceed, we replace $\sqrt{n}(\ddot{\theta}_n - \theta_0)$ with an asymptotically equivalent expression. From (A.2) and the tail equivalence of $\ddot{\beta}_n$ and $\tilde{\beta}_n$,

$$\text{(A.14)} \qquad \sqrt{n}\left(\ddot{\beta}_n - \beta_0\right) + A^{h,\,\beta\theta}\left(\gamma_*\right)^{-1} n^{-1/2}\sum_{t=1}^{n} \nabla_\gamma \log h\left(U_t, \gamma_*\right) \xrightarrow{p} 0.$$

By applying a mean-value expansion to $n^{-1}\sum_{t=1}^{n} \nabla_\psi \log f(U_t, \tilde{\theta}_n) = 0$, it is straightforward to show that, given $g(u) = f(u, \theta_0)$ for θ_0 in Θ, Assumptions A1–A6 and A11,

$$\text{(A.15)} \qquad \sqrt{n}\left(\ddot{\psi}_n - \psi_0\right) + A^f_{\psi\psi}(\theta_0)^{-1}\left[n^{-1/2}\sum_{t=1}^{n} \nabla_\psi \log f(U_t, \theta_0) + A^f_{\psi\beta}(\theta_0)\sqrt{n}\left(\ddot{\beta}_n - \beta_0\right)\right] \xrightarrow{p} 0.$$

Also, given Assumptions A1–A6 and A11, $n^{-1}\sum_{t=1}^{n} \nabla^2_{\beta\theta} \log f(U_t, \bar{\theta}_n) \overset{\text{a.s.}}{\to} A^f_{\beta\theta}(\theta_0)$ when $g(u) = f(u, \theta_0)$ for θ_0 in Θ; applying 2c.4(xiii) of Rao [**26**], we use (A.13), (A.14) and (A.15) to obtain

$$\text{(A.16)} \qquad n^{-1/2}\sum_{t=1}^{n} \nabla_\beta \log f\left(U_t, \ddot{\theta}_n\right) - \left\{n^{-1/2}\sum_{t=1}^{n}\left[I_k, -A^f_{\beta\psi}(\theta_0)A^f_{\psi\psi}(\theta_0)^{-1}\right]\nabla_\theta \log f(U_t, \theta_0) - \left[A^f_{\beta\beta}(\theta_0) - A^f_{\beta\psi}(\theta_0)A^f_{\psi\psi}(\theta_0)^{-1}A^f_{\psi\beta}(\theta_0)\right]A^{h,\,\beta\gamma}\left(\gamma_*\right)^{-1}\nabla_\gamma \log h\left(U_t, \gamma_*\right)\right\} \xrightarrow{p} 0.$$

Next, we observe that the rule for partitioned inversion implies

$$\left[A^f_{\beta\beta}(\theta_0) - A^f_{\beta\psi}(\theta_0)A^f_{\psi\psi}(\theta_0)^{-1}A^f_{\psi\beta}(\theta_0)\right]^{-1}\left[I_k, -A^f_{\beta\psi}(\theta_0)A^f_{\psi\psi}(\theta_0)^{-1}\right] = A^{f,\,\beta\theta}(\theta_0)^{-1},$$

so that (A.14), (A.15), and (A.16) imply

$$A^{f,\,\beta\beta}(\theta_0)^{-1} n^{-1/2}\sum_{t=1}^{n} \nabla_\beta \log f\left(U_t, \ddot{\theta}_n\right) - \sqrt{n}\left(\hat{\beta}_n - \ddot{\beta}_n\right) \xrightarrow{p} 0.$$

Now $A_n^{f,\,\beta\beta}(\tilde{\theta}_n)^{-1} \overset{\text{a.s.}}{\to} A^{f,\,\beta\beta}(\theta_0)^{-1}$ given Assumptions A1–A6 and A11 (cf. (A.3)). The tail equivalence of $\ddot{\theta}_n$ and $\tilde{\theta}_n$ and the convergence in distribution of $n^{-1/2}\sum_{t=1}^{n} \nabla_\beta \log f(U_t, \ddot{\theta}_n)$ ensured by (A.16) via 2c.4(x.a) of Rao [**26**] imply that

$$\left(A_n^{f,\,\beta\beta}(\tilde{\theta}_n)^{-1} - A^{f,\,\beta\beta}(\theta_0)^{-1}\right)n^{-1/2}\sum_{t=1}^{n} \nabla_\beta \log f(U_t, \tilde{\theta}_n) \xrightarrow{p} 0$$

by 2c.4(x.b) of Rao [21] so that

$$\text{(A.17)} \qquad A^{f,\,\beta\beta}(\tilde{\theta})^{-1}n^{-1/2}\sum^{n}\nabla\,\log f(U,\tilde{\theta})-\sqrt{n}(\hat{\beta}-\tilde{\beta}\quad)^{p}$$

Let $S(\theta_0,\gamma_*)^{-1/2}$ be the symmetric positive definite matrix such that $S(\theta_0,\gamma_*)^{-1/2}S(\theta_0,\gamma_*)^{-1/2}=S(\theta_0,\gamma_*)^{-1}$, which exists and is finite given Assumptions A5, A11, and A12 and $g(u)=f(u,\theta_0)$ for θ_0 in Θ. Then

$$S(\theta_0,\gamma_*)^{-1/2}\left[A_n^{f,\,\beta\beta}(\tilde{\theta}_n)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla_\beta\log f(U_t,\tilde{\theta}_n)-\sqrt{n}(\hat{\beta}_n-\tilde{\beta}_n)\right]\xrightarrow{p}0.$$

In view of (A.11) and the argument following (A.11) we have

$$\mathcal{G}_n=\nabla_\beta L_n(U,\tilde{\theta}_n)'A_n^{f,\,\beta\beta}(\tilde{\theta}_n)^{-1}S_n(\tilde{\theta}_n,\tilde{\gamma}_n)^{-1}A_n^{f,\,\beta\beta}(\tilde{\theta}_n)^{-1}\nabla_\beta L_n(U,\tilde{\theta}_n)\overset{A}{\sim}\chi_k^2$$

from Lemma 3.3 of White [34], provided $S_n(\tilde{\theta}_n,\tilde{\gamma}_n)$ is consistent for $S(\theta_0,\gamma_*)$. But this follows from repeated application of Lemma 3.1 of White [36] given Assumptions A1–A6, A11 and $g(u)=f(u,\theta_0)$ for θ_0 in Θ. For all n sufficiently large $S_n(\tilde{\theta}_n,\tilde{\gamma}_n)^{-1/2}$ exists and has bounded elements almost surely, as does $S_n(\hat{\theta}_n,\tilde{\gamma}_n)^{-1/2}$ of Theorem 5.1. Since $S_n(\hat{\theta}_n,\tilde{\gamma}_n)^{-1/2}-S_n(\tilde{\theta}_n,\tilde{\gamma}_n)^{-1/2}\xrightarrow{p}0$ when $g(u)=f(u,\theta_0)$ for θ_0 in Θ, it follows from (A.17) that

$$S_n(\tilde{\theta}_n,\tilde{\gamma}_n)^{-1/2}A_n^{f,\,\beta\beta}(\tilde{\theta}_n)^{-1}n^{-1/2}\sum_{t=1}^{n}\nabla_\beta\log f(U_t,\tilde{\theta}_n)-S_n(\hat{\theta}_n,\tilde{\gamma}_n)^{-1/2}\sqrt{n}(\tilde{\beta}_n-\hat{\beta}_n)\xrightarrow{p}0.$$

It follows immediately from 2c.4(xiv) of Rao [26] that $\mathcal{K}_n-\mathcal{G}_n\xrightarrow{p}0$.

REFERENCES

[1] AHMAD, P., AND I. LIN: "A Nonparametric Estimation of the Entropy for Absolutely Continuous Distributions," *IEEE Transactions on Information Theory*, 22(1976), 372–375.

[2] AITCHISON, J., AND S. D. SILVEY: "Maximum Likelihood Estimation of Parameters Subject to Restraints," *Annals of Mathematical Statistics*, 29(1958), 813–828.

[3] AKAIKE, H.: "Information Theory and an Extension of the Likelihood Principle," in *Proceedings of the Second International Symposium of Information Theory*, ed. B. N. Petrov and F. Csáki. Budapest: Akadémiai Kiado, 1973.

[4] BAHADUR, R.R.: "Stochastic Comparison of Tests," *Annals of Mathematical Statistics*, 31(1960), 276–295.

[5] BARTLE, R.: *The Elements of Integration*. New York: John Wiley and Sons, 1966.

[6] ———: *The Elements of Real Analysis*, Second Edition. New York: John Wiley and Sons, 1976.

[7] BERK, R. H.: "Limiting Behavior of Posterior Distributions When the Model Is Incorrect," *Annals of Mathematical Statistics*, 37(1966), 51–58.

[8] ———: "Consistency a Posteriori," *Annals of Mathematical Statistics*, 41(1970), 894–906.

[9] BOWDEN, R.: "The Theory of Parametric Identification," *Econometrica*, 41(1973), 1069–1074.

[10] BOWMAN, K. O., AND L. R. SHENTON: "Omnibus Contours for Departures from Normality Based on $\sqrt{b_1}$ and b_2," *Biometrika*, 62(1975), 243–250.

[11] BYRON, R. P.: "Testing for Misspecification in Econometric Systems Using Full Information," *International Economic Review*, 13(1972), 745–756.

[12] D'AGOSTINO, R. A., AND E. S. PEARSON: "Tests for Departure from Normality. Empirical Results for the Distributions of b_2 and $\sqrt{b_1}$," *Biometrika*, 60(1973), 613–622. Correction *Biometrika*, 61(1974), 647.

[13] FISHER, R.A.: "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London*, Series A, 222(1922), 309–368.

[14] ———: "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, 22(1925), 700–725.

[15] FOUTZ, R.V., AND R. C. SRIVASTANA: "The Performance of the Likelihood Ratio Test When the Model Is Incorrect," *Annals of Statistics*, 5(1977), 1183–1194.

[16] GEWEKE, J.: "The Approximate Slopes of Some Tests Used in Time Series Analysis," Social Systems Research Institute, University of Wisconsin Discussion Paper No. 7926, Madison, Wisconsin.
[17] HAUSMAN, J. A.: "Specification Tests in Econometrics," *Econometrica*, 46(1978), 1251–1272.
[18] HAUSMAN, J. A., AND W. E. TAYLOR: "Comparing Specification Tests and Classical Tests," M.I.T. Department of Economics Working Paper No. 266, 1980.
[19] HOLLY, A.: "A Remark on Hausman's Specification Test," Harvard Institute of Economic Research, Discussion Paper No. 763, 1980.
[20] HUBER, P. J.: "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*. Berkeley: University of California Press, 1967.
[21] JENNRICH, R. I.: "Asymptotic Properties of Non-Linear Least Squares Estimators," *Annals of Mathematical Statistics*, 40(1969), 633–643.
[22] KULLBACK, S., AND R. A. LEIBLER: "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22(1951), 79–86.
[23] LECAM, L.: "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates," *University of California Publications in Statistics*, 1(1953), 277–330.
[24] PARZEN, E.: "On Uniform Convergence of Families of Sequences of Random Variables," *University of California Publications in Statistics*, 2(1953), 23–53.
[25] PEARSON, E. S., R. A. D'AGOSTINO, AND K. O. BOWMAN: " Tests for Departure from Normality: Comparison of Powers," *Biometrika*, 64(1977), 231–246.
[26] RAO, C. R.: *Linear Statistical Inference and Its Applications*. New York: John Wiley and Sons, 1973.
[27] RÉNYI, A.: "On Measures of Entropy and Information," in *Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics*. Berkeley: University of California Press, 1961.
[28] ROTHENBERG, T.: "Identification in Parametric Models," *Econometrica*, 39(1971), 577–591.
[29] SILVEY, S. D.: "The Lagrangian Multiplier Test," *Annals of Mathematical Statistics*, 30(1959), 389–407.
[30] SOUZA, G., AND A. R. GALLANT: "Statistical Inference Based on *M*-Estimators for the Multivariate Nonlinear Regression Model in Implicit Form," North Carolina State University Institute of Statistics Mimeograph Series No. 1229, 1979.
[31] WALD, A.: "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Transactions of the American Mathematical Society*, 54(1943), 426–482.
[32] ———: "Note on the Consistency of the Maximum Likelihood Estimate," *Annals of Mathematical Statistics*, 60(1949), 595–603.
[33] WHITE, HALBERT: "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21(1980), 149–170.
[34] ———: "Nonlinear Regression on Cross-Section Data," *Econometrica*, 48(1980), 721–746.
[35] ———: "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48(1980), 817–838.
[36] ———: "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association*, 76(1981), 419–433.
[37] WU, DE-MIN: "Alternative Tests of Independence Between Stochastic Regressors and Disturbances," *Econometrica*, 41(1973), 733–750.