Journal of Econometrics 30 (1985) 415–443. North-Holland

DIAGNOSTIC TESTING AND EVALUATION OF MAXIMUM LIKELIHOOD MODELS*

George TAUCHEN

Duke University, Durham, NC 27706, USA

The paper develops a unified theory of maximum likelihood specification testing based on M-estimators of auxiliary parameters. The theory is sufficiently general to encompass a wide class of specification tests including moment-based tests, Pearson-type goodness of fit tests, the information matrix test, and the Cox test. The paper also presents a framework based on Frechet differentiation for determining the effects of misspecification on the almost sure limits of parameter estimates and specification test statistics.

1. Introduction

This paper develops the asymptotic distribution theory for a class of specification tests for the non-linear maximum likelihood model. The ideas that motivate consideration of this class of specification tests have their origins in Hausman's (1978) paper. Hausman suggested that in general, i.e., not only for the ML model, a useful specification test can be based upon the difference between two estimates of the vector of parameters of interest. This idea, however, is somewhat difficult to apply in a multivariate context when the likelihood function depends upon the parameters in a highly non-linear fashion. The difficulty lies in finding a computationally tractable form for the second 'specification-robust' estimate of the parameter vector that is required to implement Hausman's test. White (1982) suggests a different but related approach. Specifically, White derives a test that is based not upon difference between two estimates of the parameters of direct interest, but instead is based upon the difference between the two natural estimates of the expected information matrix. This paper extends White's work further by deriving the asymptotic properties of an entire class of specification tests that includes as a special case the information matrix test, and other specification tests, e.g., the Cox test [Aguirre-Torres and Gallant (1983)] and the Lagrange multiplier test [Engle

*Helpful comments were obtained from Ronald Gallant, James Heckman, James Mackinnon, Richard Robb, Robin Sickles, Donald Waldman, Dudley Wallace, Adonis Yatchew, and two referees. Earlier versions of this paper were presented at seminars at the University of Chicago, the University of Pennsylvania, Queen's University, the University of Toronto, the Triangle Area Econometrics Seminar, and the 1984 Austin Conference on Model Selection.

(1982)]. The asymptotic theory developed here is sufficiently general to include in the class of allowable tests those that are based upon non-differentiable and even discontinuous functions of the data and the parameter vector. In particular, the class of tests includes Pearson-type goodness of fit tests with random cell boundaries [Moore and Spruill (1975)].

This paper also develops a framework based on Frechet differentiation for characterizing the non-null behavior of these various specification tests. Within this framework, 'directions' of misspecification are identified against which the various specification tests can be expected to have maximum or minimum power.

Before describing the class of specification tests in more detail, it is helpful to review briefly the asymptotic distribution theory of the quasi-maximum likelihood estimator (the ML estimator with an incorrect likelihood function). Assume the observed data $Y_1, Y_2, \ldots, Y_n$ are mutually independent and identically distributed $m \times 1$ random vectors with common unknown distribution function G and density function g, both defined on R^m. Let $\{F(y,\theta)\colon y \in R^m, \theta \in \theta \subset R^p\}$ be a family of distribution functions on R^m that is the basis for the estimation. For each fixed parameter vector θ the function $F(y,\theta)$ is a probability distribution on R^m with density function denoted by $f(y,\theta)$. Together the elements of the family of distribution functions $\{F(y,\theta)\}$, or equivalently the family of density functions $\{f(y,\theta)\}$, comprise a probability model for the observed data. The quasi-maximum likelihood estimator $\hat{\theta}_n$ is the value of the parameter that maximizes the sample quasi-loglikelihood function

$$L_n(\theta) = \frac{1}{n}\sum_{1}^{n} l(Y_i,\theta), \tag{1}$$

where $l(y,\theta) = \log(f(y,\theta))$ is the log-density function. Burguette, Gallant and Souza (1982), Huber (1967) and White (1982) have shown that under a variety of regularity conditions the QML estimator $\hat{\theta}_n$ converges almost surely to the value $\bar{\theta}$ at which the expected log-density function,

$$L(\theta) = \mathrm{E}[l(Y_i,\theta)] = \int l(y,\theta)\,\mathrm{d}G(y), \tag{2}$$

achieves its maximum. Now, if the underlying model is correctly specified, then there exists a θ_0 such that the density $f(y,\theta_0)$ is a version of the true density $g(y)$. In this case the maximizing $\bar{\theta}$ for L in (2) equals θ_0 and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normally distributed with mean zero and variance–covariance matrix equal to the inverse of the information matrix. On the other hand, if the model is misspecified, then of course no such θ_0 exists; but the maximizing $\bar{\theta}$ for the expected quasi-loglikelihood function still exists and $\sqrt{n}(\hat{\theta}_n - \bar{\theta})$ has a well-defined asymptotic distribution. One interpretation for $\bar{\theta}$ is that it is the

'true' parameter value that is induced directly by the estimation procedure itself.

The class of specification tests considered in this paper consists of those tests based on the magnitude of the statistic

$$\hat{\tau}_n = \frac{1}{n}\sum_{1}^{n} c(Y_i, \hat{\theta}_n), \tag{3}$$

where $\hat{\theta}_n$ is the QML estimator and where the vector-valued function c satisfies

$$\int c(y, \theta)\, \mathrm{d}F(y, \theta) = 0, \tag{4}$$

for all θ. The condition (4) says that the function $c(y, \theta)$ has mean zero with respect to each distribution function in the probability model. A function that satisfies this condition will be called an *auxiliary criterion function*. As will become clearer below, for any given family of distribution functions $\{F(y, \theta)\}$ there are many auxiliary criterion functions. In practice, the better auxiliary criterion functions will be those for which the magnitude of the elements of the vector $\hat{\tau}_n$ in (3) provide useful diagnostic information about the specification of the model. A strategy for getting an informative $\hat{\tau}_n$ is to construct the auxiliary criterion function in such a way that the components of $\hat{\tau}_n$ equal the differences between two estimates of some statistical quantities of interest.

The statistic $\hat{\tau}_n$ is useful for specification testing because it converges almost surely to zero when the model is correctly specified and it converges to a non-zero quantity when the model is incorrectly specified. This result is proved in section 2, but it is intuitively clear from inspection of the expressions (3) and (4). In the former case when θ_0 exists,

$$\hat{\tau}_n \xrightarrow{\text{a.s.}} \mathrm{E}[c(Y_i, \theta_0)] = \int c(y, \theta_0)\, \mathrm{d}F(y, \theta_0),$$

which is zero by construction of the auxiliary criterion function. In the latter case,

$$\hat{\tau}_n \xrightarrow{\text{a.s.}} \bar{\tau} = \mathrm{E}[c(Y_i, \bar{\theta})] = \int c(y, \bar{\theta})\, \mathrm{d}G(y),$$

which in general is non-zero. As shown in section 3, the statistic $\hat{\tau}_n$ also has a well defined asymptotic distribution in either case. Its asymptotic variance–covariance matrix can be expressed as the sum of two parts, one of which corresponds to the variability in $(1/n)\sum_1^n c(Y_i, \bar{\theta})$ about $\bar{\tau}$ and the other to the variability in $\hat{\theta}_n$ about $\bar{\theta}$.

The following three examples help to illustrate the practical applications of the general results in this paper:

Example 1 (low-order moments)

For simplicity in exposition take y as scaler though θ may be multi-dimensional. Define

$$\mu_j(\theta) = \int y^j \mathrm{d}F(y, \theta),$$

for integer j. Thus $\mu_j(\hat{\theta}_n)$ is the predicted jth non-central moment from the estimated probability model. Let j be fixed at some integer and define

$$c(y, \theta) = y^j - \mu_j(\theta).$$

This function is a legitimate auxiliary criterion function since it satisfies the condition (4). Moreover, the statistic

$$\hat{\tau}_n = \frac{1}{n}\sum_{1}^{n} c(Y_i, \hat{\theta}_n)$$

is simply the difference between sample jth non-central moment and the predicted moment from the probability model. A large value for $|\hat{\tau}_n|$ would tend to indicate that the probability model does a poor job of 'matching' the jth moment of the distribution of the data. As shown in section 5 of this paper, there is a regression-based procedure for testing for whether the magnitude $|\hat{\tau}_n|$ is too large to be accounted for by sampling fluctuations: One regresses the values $\hat{c}_i = c(Y_i, \hat{\theta}_n)$ on the scores $\hat{h}_i = \partial l(Y_i, \hat{\theta}_n)/\partial\theta$ and performs a t test for a non-zero intercept. If the t statistic is large from a statistical point of view the model may need to be reformulated or else an explanation given as to why the difference in moments is too small to be of practical importance. Of course in some cases the estimation procedure may force some of the sample and predicted moments to be equal and no such test is possible. For instance, if the underlying model is the univariate normal distribution, then the first two sample and predicted moments must be equal. Diagnostic tests in this case would then have to be based on moments higher than the second. For the asymptotic theory to provide a good approximation to distribution of $\hat{\tau}_n$, the order of the moments above two should be kept reasonably small.

The extension of this to other unconditional moments is straightforward. For central moments in the scaler case let the auxiliary criterion function be $[y - \mu_1(\theta)]^j$ minus the expected value of this quantity with respect to $F(y, \theta)$. For central moments in the multivariate case, the auxiliary criterion function

would be the distinct elements of the j-fold Kronecker product of the vector $y - \mu_1(\theta)$ with itself minus the expectation with respect to $F(y, \theta)$.

As noted by Newey (1984) in an independent paper, moment conditions can be used to form useful auxiliary criterion functions when the data vector is partitioned as $y' = (w', x')$ and the probability model is $f_1(w|x, \theta)$. Here w is a vector of jointly dependent variables and x is a vector of exogenous variables. The marginal density $f_2(x)$ for x is not specified by the model. A function of the form

$$c(w, x, \theta) = \left(\frac{\partial}{\partial \theta} \log f_1(w|x, \theta) \right) a(x, \theta),$$

where $a(x, \theta)$ depends only on x and θ, satisfies (4). A test based on this auxiliary criterion function is an 'instrumented score test'. Newey examines in detail the statistical properties of such tests and presents useful applications for regression models and limited dependent variable models.

Example 2 (tail areas)

In some applied work it is important to have information on how well the probability model predicts tail areas. An auxiliary criterion function that provides such information can be constructed along the following lines. Assume for simplicity in exposition that y is scalar though θ may be multi-dimensional. Let $\mu(\theta)$ and $\sigma(\theta)$ denote the mean and standard deviation of the distribution $F(y, \theta)$. Fix α as a small probability and let z_α satisfy

$$\text{prob}_F[y - \mu(\theta) \geq \sigma(\theta) z_\alpha] = \alpha,$$

where the subscript F is self-explanatory. Now put

$$c(y, \theta) = I[y - \mu(\theta) \geq \sigma(\theta) z_\alpha] - \alpha,$$

where $I[\cdot]$ is the 0–1 indicator function. Then the statistic

$$\hat{\tau}_n = \frac{1}{n} \sum_1^n c(Y_i, \hat{\theta}_n)$$

is the difference between the observed and the predicted frequency with which right-hand extreme values occur.

As illustrated in section 5, an asymptotically valid test for no difference in the frequencies can be computed by regressing the values $\hat{c}_i = (Y_i, \hat{\theta}_n)$ on the 'scores', i.e., the gradients $\partial l(Y_i, \hat{\theta}_n)/\partial \theta$ of the log-density function, and then performing the usual t test for no intercept. Interestingly, the square of this t

statistic is asymptotically a chi-square variate with one degree of freedom, but the t^2 does not equal the classical Pearson statistic. The reason is that this t^2 statistic properly accounts for the randomness in $\hat{\theta}_n$, where the classical statistic does not. The classical Pearson procedure implicitly assumes that the asymptotic variance is $\alpha(1-\alpha)$ which exceeds the true variance. [When the model is $\phi(y-\mu)$, where ϕ is the standard normal pdf, then the variance is $\alpha(1-\alpha)-\phi(z_\alpha)^2$.] Put another way, the classical procedure ignores the randomness in $\hat{\theta}_n$ and treats $(1/n)\sum_1^n c(Y_i, \hat{\theta}_n)$ as if it has the same asymptotic distribution as $(1/n)\sum_1^n c(Y_i, \theta_0)$, which is a 'Durbin' problem that leads to the incorrect expression for the asymptotic variance.

A more general chi-square goodness of fit test is as follows. Suppose the data vector is of the form $y' = (w', x')$ where, in a notation consistent with that used at the end of Example 1, the vector w contains the jointly dependent variables and x the exogenous variables. The probability model is the conditional density $f_1(w|x,\theta)$ of the dependent variables given x, with the marginal density $f_2(x)$ not specified. Let the components of the $K \times 1$ auxiliary criterion function be

$$c_k(y,\theta) = c_k(w,x,\theta) = I\left[w \in R_k(x,\theta)\right] - \pi_{0k}, \qquad k = 1,2,\ldots,K,$$

where $I[\ ]$ is the 0–1 indicator function, the π_{0k} are fixed probabilities such that $\sum_{k=1}^K \pi_{0k} = 1$, and the regions $R_k(x,\theta)$ are chosen so that

$$\int I\left[w \in R_k(x,\theta)\right] f_1(w|x,\theta)\,\mathrm{d}w = \pi_{0k},$$

for each $k = 1,2,\ldots,K$. Then the $K \times 1$ vector

$$\hat{\tau}_n = \frac{1}{n}\sum_{i=1}^n c(w_i, x_i, \hat{\theta}_n)$$

contains the differences between the observed and expected frequencies. The regression-based method described in section 5 can be used to construct an asymptotically valid chi-square statistic based on $\hat{\tau}_n$. This test is based on random cell boundaries [Moore and Spruill (1975)] and it accounts for covariates x. It differs from Heckman's (1984) test because here the regions $R(x,\theta)$ depend not only on x but also on θ. More specifically, here the probabilities are viewed as fixed and the regions then determined, whereas Heckman views the regions $R(x)$ as given independently of θ, and then the probabilities $\pi_k(x,\theta) = \int I[w \in R(x)] f(x,\theta)\,\mathrm{d}w$ are determined. The asymptotic theory of this paper is general enough to cover the case when the test is set up in Heckman's manner, but there may be advantages to setting it up the other way. First, with a priori fixed probabilities the test outcome could be

easier to interpret and provide better diagnostic information. Second, with this setup the user can choose the probabilities so that $\pi_{0k} = 1/K$, i.e., so that the regions are equiprobable, which is a method that has been shown to have optimum properties [Kendall and Stewart (1973, ch. 30)] in the case with no covariates.

Example 3 (White's information matrix test)

To include White's test in this setup, take as the auxiliary criterion function $c(y, \theta)$ the vector function comprised of the distinct elements of the symmetric matrix function

$$h(y, \theta)h(y, \theta)' + \frac{\partial h}{\partial \theta'}(y, \theta),$$

where $h(y, \theta) = \partial l(y, \theta)/\partial \theta$ is the gradient of the log-density function. With the function c defined in this fashion the vector

$$\hat{\tau}_n = \frac{1}{n}\sum_{1}^{n} c(Y_i, \hat{\theta}_n)$$

contains all of the differences between the distinct elements of the two natural estimates of the information matrix. White derives an estimator for the asymptotic variance–covariance matrix of this $\hat{\tau}_n$ that requires the user to calculate analytical third-order partial derivatives of the log-density function. In section 3 it is shown that there is an extension of the classical information equality which, as also noted by Chesher (1983) and Lancaster (1984), eliminates the need for third partials and leads to regression based procedures for conducting White's test.

The remainder of this paper is organized as follows. Sections 2 and 3 present the consistency and asymptotic normality results. Section 4 examines some measures of the performance of the specification tests. Section 5 presents the regression-based procedure for conducting the specification tests discussed here. Section 6 contains some concluding remarks.

For the sake of completeness, the various assumptions which were either implicit or explicit in this introduction are now listed in one place.

Assumption 1

(i) The observed data $Y_1, Y_2, \ldots, Y_n$ are iid $m \times 1$ random vectors with distribution function G on R^m.

(ii) The probability model is the family of distribution functions $\{F(y, \theta): y \in R^m, \theta \in \Theta \subset R^p\}$, where the parameter space Θ is a compact convex subset of R^p with a non-empty interior.

(iii) Both $G(y)$ and $F(y,\theta)$ are absolutely continuous with respect to some measure $\mu(y)$ on R^m with generalized (Radon–Nikodym) densities denoted by $g(y) = \mathrm{d}G(y)/\mathrm{d}\mu(y)$ and $f(y,\theta) = \mathrm{d}F(y,\theta)/\mathrm{d}\mu(y)$.

(iv) The auxiliary criterion function satisfies $\int c(y,\theta)\,\mathrm{d}F(y,\theta) = 0$ for each $\theta \in \Theta$.

2. Consistency

As in the introduction let l: $R^m \times \Theta \to R^1$ be the log-density function and let c: $R^m \times \Theta \to R^s$ be the auxiliary criterion function. The QML estimator $\hat{\theta}_n$ and the statistic $\hat{\tau}_n$ are defined by

$$L_n(\hat{\theta}_n) = \max_{\theta \in \Theta} L_n(\theta), \tag{5}$$

$$\hat{\tau}_n = \psi_n(\hat{\theta}_n), \tag{6}$$

where ψ_n and L_n are the functions

$$L_n(\theta) = \frac{1}{n}\sum_1^n l(Y_i,\theta), \qquad \psi_n(\theta) = \frac{1}{n}\sum_1^n c(Y_i,\theta).$$

The key step in proving the consistency results is to establish the almost sure convergence of $L_n(\theta)$ and $\psi_n(\theta)$ to their expectations uniformly in the parameter θ. The almost sure convergence of $\hat{\theta}_n$ and $\hat{\tau}_n$ to well defined limits will then follow from assumptions guaranteeing that the almost sure limit of the function L_n has a unique maximum.

It proves useful to identify a large class of vector-valued functions on $R^m \times \Theta$ for which uniform almost sure convergence will hold.

Definition 1. A function ϕ: $R^m \times \Theta \to R^k$ is said to be *regular* if

(i) $\phi(y,\theta)$ is measurable in y for each $\theta \in \Theta$,

(ii) ϕ is separable [see Huber (1967, p. 222)],

(iii) ϕ is dominated, $|\phi(y,\theta)| \leq b(y)$, where the function b is integrable with respect to G,

(iv) ϕ is almost surely continuous in the sense that for each fixed θ the set $\{y: \lim_{\gamma \to \theta}\phi(y,\gamma) = \phi(y,\theta)\}$ has probability 1($\mathrm{d}G$). The null set may depend on θ.

The measurability and separability conditions (i) and (ii) are weak and essentially non-restrictive side conditions. The domination assumption (iii) ensures that the expectation

$$\lambda(\theta) = \int \phi(y,\theta)\,\mathrm{d}G(y) \tag{7}$$

exists, while the almost sure continuity condition (iv) implies by dominated convergence that λ is a continuous function of θ. As the following lemma indicates, sample averages of $\phi(Y_i, \theta)$ have the requisite convergence properties if ϕ is regular.

Lemma 1. If ϕ is regular, then the function

$$\lambda_n(\theta) = \frac{1}{n}\sum_1^n \phi(Y_i, \theta)$$

converges uniformly almost surely to function λ in (7). (Proof: Appendix.)

The next two assumptions contain the conditions for consistency of $\hat{\theta}_n$ and $\hat{\tau}_n$:

Assumption 2. The auxiliary criterion function c is regular and the log-density function l satisfies (i)–(iii) of Definition 1 and a stronger version of (iv), namely, $l(y, \theta)$ is continuous in θ for all y.

The stronger continuity assumption is needed for the log-density function in order to ensure that the maximizing $\hat{\theta}_n$ for L_n in (5) exists for all n. The weaker continuity condition for the auxiliary criterion function c suffices because the existence of $\hat{\tau}_n$ in (6) is guaranteed once the existence of $\hat{\theta}_n$ is established.

Define the functions

$$L(\theta) = \mathrm{E}[l(Y_i, \theta)] = \int l(y, \theta)\, \mathrm{d}G(y), \tag{8a}$$

$$\psi(\theta) = \mathrm{E}[c(Y_i, \theta)] = \int c(y, \theta)\, \mathrm{d}G(y), \tag{8b}$$

both of which exist and are continuous by Assumption 2. From Lemma 1 it is known that $L_n(\theta) \overset{\text{a.s.}}{\to} L(\theta)$ and $\psi_n(\theta) = (1/n)\sum_1^n c(Y_i, \theta) \overset{\text{a.s.}}{\to} \psi(\theta)$ uniformly in θ. By continuity and compactness the limiting function L achieves its maximum at least once in the parameter space Θ. For the limit of the estimator $\hat{\theta}_n$ to be well defined, it is necessary to assume that there is only one such maximum.

Assumption 3. The limiting quasi-loglikelihood function L achieves its maximum uniquely at $\bar{\theta}$ in the interior of the parameter space.

The basic consistency result is:

Theorem 1. $\hat{\theta}_n \xrightarrow{\text{a.s.}} \bar{\theta}$ *and* $\hat{\tau}_n \xrightarrow{\text{a.s.}} \bar{\tau}$, *where*

$$\bar{\tau} = \psi(\bar{\theta}) = \int c(y, \bar{\theta})\, \mathrm{d}G(y). \tag{9}$$

Proof. The convergence of $\hat{\theta}_n \xrightarrow{\text{a.s.}} \bar{\theta}$ follows from arguments similar to those in Burguette, Gallant and Souza (1982). Since $\psi_n(\theta) \xrightarrow{\text{a.s.}} \psi(\theta)$ uniformly in θ and since ψ is continuous, then by standard arguments $\psi_n(\hat{\theta}_n) \xrightarrow{\text{a.s.}} \psi(\bar{\theta}) = \bar{\tau}$ as given in (9).

Note that this theorem covers both the null case in which the model is correctly specified and there exists $\theta_0 \in \Theta$ such that $f(y, \theta_0)$ is a version of $g(y)$, and it covers the non-null case in which no such θ_0 exist. In the null case, $\bar{\theta} = \theta_0$ and the almost sure limit of $\hat{\tau}_n$ is

$$\psi(\theta_0) = \int c(y, \theta_0)\, \mathrm{d}F(y, \theta_0) = 0,$$

by the construction of the auxiliary criterion function. In the non-null case the almost sure limit of $\hat{\tau}_n$ is $\psi(\bar{\theta})$, which is in general non-zero.

3. Asymptotic normality

3.1. The joint asymptotic distribution of $\hat{\theta}_n$ *and* $\hat{\tau}_n$

In order to allow for a large class of auxiliary criterion functions – in particular, those based on frequency counts or absolute moments – the conditions for asymptotic normality that are placed on the auxiliary criterion function $c(y, \theta)$ do not require differentiability with respect to θ. Instead, the conditions only require c to satisfy certain Huber-type Lipshitz conditions and $\psi(\theta) = \int c(y, \theta)\, \mathrm{d}G(y)$ to be a continuously differentiable function of θ. One of the costs, however, of not imposing differentiability on c is that a strategy for proving asymptotic normality that is based on Taylor approximations does not work. Specifically, it is not possible to adopt methods of proof similar to those commonly used in non-linear econometrics, because the difference $\sqrt{n}\,(\hat{\tau}_n - \psi_n(\bar{\theta}))$ cannot be approximated by $(\partial\psi_n/\partial\theta')(\bar{\theta})\sqrt{n}\,(\hat{\theta}_n - \bar{\theta})$ and then the asymptotic normality of $\sqrt{n}\,(\hat{\theta}_n - \bar{\theta})$ exploited. The alternative strategy adopted here is to embed the determination of $\hat{\theta}_n$ and $\hat{\tau}_n$ into a larger M-estimation problem which gives the joint asymptotic distribution of $\hat{\theta}_n$ and $\hat{\tau}_n$.

Joint asymptotic normality of $\hat{\theta}_n$ and $\hat{\tau}_n$ will be proved under the following conditions for l and c. In the statement of the conditions, the vector-valued function ϕ is

$$\phi(y,\theta)=\begin{bmatrix}h(y,\theta)\\ c(y,\theta)\end{bmatrix},$$

where h is the gradient of the log-density function and c is the auxiliary criterion function; the function u in (iii) below is

$$u(y,\theta,d)=\sup_{|\gamma-\theta|\le d}|\phi(y,\gamma)-\phi(y,\theta)|.$$

Assumption 4

(i) $l(y,\theta)$ is continuously differentiable in θ for all y with gradient denoted by

$$h(y,\theta)=\frac{\partial}{\partial\theta}l(y,\theta).$$

(ii) $|\phi(y,\theta)|\le b(y)$, where the function b is square integrable with respect to G.

(iii) There exist positive constants β_1 and β_2 such that for all θ

$$\limsup_{d\downarrow 0}\frac{\mathrm{E}[u(Y_i,\theta,d)]}{d}\le\beta_1,\qquad \limsup_{d\downarrow 0}\frac{\mathrm{E}[u(Y_i,\theta,d)^2]}{d}\le\beta_2.$$

(iv) The components of $\lambda(\theta)=\mathrm{E}[\phi(Y_i,\theta)]$ are continuously differentiable in θ and the matrix $\partial\mathrm{E}[h(Y_i,\theta)]/\partial\theta'|_{\bar{\theta}}$, which is the upper left $p\times p$ submatrix of $\partial\lambda(\bar{\theta})/\partial\theta'$, is non-singular.

A sketch of the asymptotic normality proof is as follows. The almost sure convergence of $\hat{\theta}_n$ and $\hat{\tau}_n$ to $\bar{\theta}$ and $\bar{\tau}$ was established in the previous section. By assumption, the limit $\bar{\theta}$ lies in the interior of the parameter space, and so ultimately the maximizing $\hat{\theta}_n$ must remain in the interior of the parameter space. Thus, ultimately the first-order condition

$$\frac{1}{n}\sum_{1}^{n}h(Y_i,\theta)=0$$

must be satisfied at $\theta=\hat{\theta}_n$. Now let T be a non-trivial closed ball about $\bar{\tau}$, and

define the vector-valued function η on $R^m \times \Theta \times T$ by

$$\eta(y,\theta,\tau) = \begin{bmatrix} h(y,\theta) \\ c(y,\theta) - \tau \end{bmatrix}. \tag{10}$$

It is seen immediately that $\hat{\theta}_n$ and $\hat{\tau}_n$ will ultimately solve the expanded system of equations

$$\frac{1}{n}\sum_{1}^{n} \eta(Y_i,\theta,\tau) = 0. \tag{11}$$

Therefore Huber's (1967, p. 231; 1981, p. 133) results for M-estimators determined by solving a system of implicit equations can thus be applied to $\hat{\theta}_n$ and $\hat{\tau}_n$.

The main asymptotic normality result is:

Theorem 2. The random vector

$$\sqrt{n}\begin{bmatrix} \hat{\theta}_n - \bar{\theta} \\ \hat{\tau}_n - \bar{\tau} \end{bmatrix}$$

converges in distribution to a multivariate normal with mean zero and with the variance–covariance matrix given by

$$\begin{bmatrix} \Sigma_\theta & \Sigma_{\theta\tau} \\ \Sigma_{\tau\theta} & \Sigma_\tau \end{bmatrix} = \begin{bmatrix} K_h & 0 \\ K_c & -I \end{bmatrix}^{-1} \begin{bmatrix} J_{hh} & J_{hc} \\ J_{ch} & J_{cc} \end{bmatrix} \begin{bmatrix} K_h & K_c' \\ 0 & -I \end{bmatrix}^{-1}, \tag{12}$$

where the submatrices on the right-hand side are

$$K_h = \frac{\partial}{\partial \theta'} \mathrm{E}\big[h(Y_i,\theta)\big] \quad \textit{at} \quad \theta = \bar{\theta}, \tag{13a}$$

$$K_c = \frac{\partial}{\partial \theta'} \mathrm{E}\big[c(Y_i,\theta)\big] \quad \textit{at} \quad \theta = \bar{\theta}, \tag{13b}$$

$$J_{hh} = \mathrm{E}\big[h(Y_i,\bar{\theta})h(Y_i,\bar{\theta})'\big], \tag{14a}$$

$$J_{hc} = \mathrm{E}\big[h(Y_i,\bar{\theta})(c(Y_i,\bar{\theta}) - \bar{\tau})'\big], \tag{14b}$$

$$J_{ch} = J_{hc}', \tag{14c}$$

$$J_{cc} = \mathrm{E}\big[(c(Y_i,\bar{\theta}) - \bar{\tau})(c(Y_i,\bar{\theta}) - \bar{\tau})'\big]. \tag{14d}$$

Proof. The main argument was given in remarks preceding the statement of the theorem. The expression for the joint asymptotic variance–covariance matrix of $\hat{\theta}_n$ and $\hat{\tau}_n$ follows from applying Huber's (1967, p. 231) corollary to his Theorem 3.

By way of interpretation, note that the rows of the matrices K_h and K_c are the gradients with respect to θ' of the components of $\mathrm{E}[h(Y_i, \theta)]$ and $\mathrm{E}[C(Y_i, \theta)]$, while the matrices J_{hh}, J_{hc} and J_{cc} are simply the variance–covariance matrices of the random variables $\tilde{h}_i = h(Y_i, \bar{\theta})$ and $\tilde{c}_i = c(Y_i, \bar{\theta})$. Note also that Theorem 2 gives

$$\Sigma_\theta = K_h^{-1} J_{hh} K_h^{-1}.$$

For the marginal asymptotic distribution of $\bar{\theta}$ which is the familiar form for the asymptotic variance–covariance matrix of the QML estimator.

For the purposes of formal testing and calculating confidence intervals, estimates are needed of the various K's and J's that appear in (12) through (14). The J's can be consistently estimated in the natural way by forming the corresponding sample product moment matrices:

Theorem 3. The estimates

$$\hat{J}_{hh} = \frac{1}{n}\sum_1^n h(Y_i, \hat{\theta}_n) h(Y_i, \hat{\theta}_n)',$$

$$\hat{J}_{hc} = \frac{1}{n}\sum_1^n h(Y_i, \hat{\theta}_n)(c(Y_i, \hat{\theta}_n) - \hat{\tau}_n)',$$

$$\hat{J}_{cc} = \frac{1}{n}\sum_1^n (c(Y_i, \hat{\theta}_n) - \hat{\tau}_n)(c(Y_i, \hat{\theta}_n) - \hat{\tau}_n)',$$

of the matrices J_{hh}, J_{hc} and J_{cc} in (14) are consistent in the sense of element-wise almost sure convergence.

Proof. By Assumption 2 and items (i) and (ii) of Assumption 4 each of the columns of the matrix $\phi(y, \theta)\phi(y, \theta)'$, where $\phi(y, \theta)' = [h(y, \theta)' c(y, \theta)']$, is a regular function in the sense of Definition 1; apply Lemma 1 to each column. This, plus the result $\hat{\tau}_n \xrightarrow{\text{a.s.}} \bar{\tau}$ from Theorem 1, establishes the conclusion.

These estimates of the J's are 'specification-robust' in the sense that they are valid even if the underlying probability model is misspecified. If the components of the gradient of the log-density function $h(y, \theta)$ and the auxiliary

criterion function $c(y,\theta)$ are continuously differentiable in θ, then there are similar natural specification-robust estimates of the K's.

Theorem 4. Assume that the components of $h(y,\theta)$ and $c(y,\theta)$ are continuously differentiable in θ for all y and put $\tilde{K}_h(y,\theta) = \partial h(y,\theta)/\partial\theta'$ and $\tilde{K}_c(y,\theta) = \partial c(y,\theta)/\partial\theta'$. If the columns of the matrix functions $\tilde{K}_h$ and $\tilde{K}_c$ are regular as defined in Definition 1 (note that continuity of the columns in θ is presupposed in the hypotheses of this theorem), then the random matrices

$$\hat{K}_h = \frac{1}{n}\sum_{1}^{n} \tilde{K}_h(Y_i, \hat{\theta}_n), \tag{15a}$$

$$\hat{K}_c = \frac{1}{n}\sum_{1}^{n} \tilde{K}_c(Y_i, \hat{\theta}_n), \tag{15b}$$

are consistent estimates of K_h and K_c in the sense of element-wise almost sure convergence.

Proof. The proof is entirely analogous to that for Theorem 3.

In most applied work the log-density function $l(y,\theta)$ satisfies the differentiability condition in the hypotheses of Theorem 4, and so the natural estimate $\hat{K}_h$ in (15a) is nearly always available. If the auxiliary criterion is reasonably smooth – as would usually be the case if the model evaluation is based on the difference between predicted and sample moments – then the estimate $\hat{K}_c$ in (15b) is also available. In these cases, then, Theorems 3 and 4 lead to a specification-robust estimate $\hat{\Sigma}_\tau$.

3.2. *The generalized information equality and the estimation of Σ_τ under correct specification*

An estimate of Σ_τ that is valid under the maintained hypothesis that the probability model is correctly specified turns out to be very easy to compute, even if the auxiliary criterion function is not differentiable in θ. The reduction in computational burden is brought about by the availability of an extension of the classical information equality. This equality says that under suitable regularity conditions the expected information matrix equals minus the expected Hessian matrix. In the notation of Theorem 4, the information equality can be expressed as

$$K_h^o + J_{hh}^o = 0, \tag{16}$$

where the superscript o means that these are the J_{hh} and K_h matrices in (13a) and (13b) when the model is correctly specified and θ_0 exists.

To motivate the generalization of the information equality, consider

$$\int c(y,\theta) f(y,\theta)\,\mathrm{d}\mu = 0,$$

and note that the equality holds identically in θ. Now, if differentiation could be brought freely in and out of the integration, then the usual Cramer calculus gives

$$\int c_{\theta'}(y,\theta_0) f(y,\theta_0)\,\mathrm{d}\mu + \int c(y,\theta_0) f_{\theta'}(y,\theta_0)\,\mathrm{d}\mu = 0,$$

or

$$\frac{\partial}{\partial \theta'} \int c(y,\theta) f(y,\theta_0)\,\mathrm{d}\mu \Big|_{\theta_0} + \int c(y,\theta_0) h(y,\theta_0)' f(y,\theta_0)\,\mathrm{d}\mu = 0,$$

where the subscript θ' on c and f in the first equality denotes partial differentiation. The last equality is more compactly written

$$K_c^o + J_{ch}^o = 0, \tag{17}$$

where as in (16) the superscript o means that these are the corresponding K_h and J_{ch} matrices whenever θ_0 exists. The next theorem states that both the basic information equality and its generalization in (17) are valid even if $c(y,\theta)$ is not differentiable in θ and the differentiation cannot in general be brought inside the integration.

Theorem 5 (generalized information equality). Assume (i): θ_0 exists such that $f(y,\theta_0)$ is a version of $g(y)$, and (ii): the function

$$q(y,\theta) = f(y,\theta)/f(y,\theta_0) \tag{18}$$

satisfies $|\partial q(y,\theta)/\partial\theta| \leq b(y)$, where b is square integrable with respect to $F(y,\theta_0)$. Then both equalities (16) and (17) hold. (Proof: Appendix.)

The following corollary gives a simple method for getting null-consistent estimates of the asymptotic variance–covariance matrix of $\hat{\tau}_n$.

Corollary 5.1. If the hypotheses of Theorem 5 hold, then

$$\Sigma_\tau^o = J_{cc}^o - J_{ch}^o \left(J_{hh}^o\right)^{-1} J_{hc}^o \tag{19}$$

is the asymptotic variance–covariance matrix of $\sqrt{n}\,\hat{\tau}_n$. *Moreover, the natural estimator*

$$\hat{\Sigma}_\tau^o = \hat{J}_{cc} - \hat{J}_{ch}\hat{J}_{hh}^{-1}\hat{J}_{hc}, \tag{20}$$

with the $\hat{J}$*'s as in Theorem 3, is consistent in the sense of element-wise almost sure convergence.*

Proof. Apply the two equalities (16) and (17) to the expression (12) and then select off the lower right-hand corner of the joint variance–covariance matrix that corresponds to $\hat{\tau}_n$; the convergence of (20) follows from Theorem 3.

Interestingly, the estimate $\hat{\Sigma}_\tau^o$ in (20) is simply the usual estimate of the residual variance–covariance matrix from a seemingly unrelated regression of the components of $\hat{c}_i = c(Y_i, \hat{\theta}_n)$ on the 'scores' $\hat{h}_i = h(Y_i, \hat{\theta}_n)$.

4. The local behavior of $\bar{\tau}$ under misspecification

In the previous sections it was established that the estimator $\hat{\theta}_n$ and the statistic $\hat{\tau}_n$ converge almost surely to

$$\bar{\theta} = \arg\min_\theta \left\{ \int l(y,\theta)\,\mathrm{d}G(y) \right\}, \tag{21}$$

$$\bar{\tau} = \int c(y,\bar{\theta})\,\mathrm{d}G(y), \tag{22}$$

and that $\sqrt{n}(\hat{\theta}_n - \bar{\theta})$ and $\sqrt{n}(\hat{\tau}_n - \bar{\tau})$ have a joint asymptotic normal distribution with variance and covariance matrices $\Sigma_\theta, \Sigma_\tau, \Sigma_{\theta\tau}$, as given in Theorem 2 above. When the model is correctly specified and $G(y) = F(y,\theta_0)$ for some θ_0, then $\bar{\theta} = \theta_0$ and $\bar{\tau} = 0$. Under misspecification however, $\bar{\theta}$ need not equal θ_0, and likewise $\bar{\tau}$ will be non-zero, which is where the specification tests gets its power.

In this section we will investigate the ability of the test to detect misspecification by examining the local behavior of the $\bar{\tau}$ for small deviations of the idealized model from the true model. These deviations are generated in the following manner. Consider alternative true distributions G_v given by

$$\mathrm{d}G_v(y) = [1 + v(y)]\,\mathrm{d}F_0(y), \tag{23}$$

where v is a function on R^m and $\mathrm{d}F_0(y) = \mathrm{d}F(y,\theta_0)$. The parameter value θ_0 is fixed throughout, while v will vary over a class of functions on R^m, with

each v giving rise to an alternative true model or distribution of the data. The particular class of functions are the elements of the following space.

Notation. Let V denote the set of functions v: $R^m \to R^1$ such that

(i) $\quad 1 + v(y) \geq 0, \quad$ for all y,

(ii) $\quad \int v(y)\, \mathrm{d}F_0(y) = 0,$

(iii) $\quad \int v(y)^2\, \mathrm{d}F_0(y) < \infty.$

The first two of these conditions simply ensure that G_v is a bona fide distribution function. The third condition ensures that a random variable of the form $v(Y)$, with $Y \sim F_0$, has finite variance, which proves to be convenient below. The norm of the space V is taken to be the natural $L_2(F_0)$ norm, $\|v\|_2 = (\int v(y)^2 \mathrm{d}F_0(y))^{1/2}$.

In this setup, then, for each $v \in V$ there is a true distribution G_v given by (23). For any non-zero v in V the probability model is misspecified since there will in general be no θ in Θ such that $G_v(y) = F(y, \theta)$. However, at $v = 0$ the model is correctly specified with $G_v(y) = F_0(y) = F(y, \theta_0)$, by construction. Although this setup does restrict the true distribution to be absolutely continuous with respect to the distribution F_0, it does nonetheless generate a very wide class of alternative models. Furthermore one might argue that absolute continuity is no restriction at all, since any region in R^m over which the true distribution puts positive probability mass will ultimately be discovered in large samples anyways.

Under suitable regularly conditions both $\hat{\theta}_n$ and $\hat{\tau}_n$ will, for each G_v with v fixed, have almost sure limits

$$\hat{\theta}_n \overset{\text{a.s.}}{\to} \bar{\theta}(v) = \arg\min_{\theta} \left\{ \int l(y, \theta)\, \mathrm{d}G_v(y) \right\}, \tag{24}$$

$$\bar{\tau}_n \overset{\text{a.s.}}{\to} \bar{\tau}(v) = \int c(y, \bar{\theta}(v))\, \mathrm{d}G_v(y). \tag{25}$$

The almost surely here means with probability one with respect to the joint distribution of $\{Y_i\}$, where the Y_i are independent and have common distribution function G_v.

Both $\bar{\theta}(v)$ and $\bar{\tau}(v)$ are functionals on V that are highly non-linear in v, which makes a complete analysis of them difficult to obtain. However, an

analysis of their local behavior near $v = 0$, i.e., near a correctly specified model, is tractable and gives several interesting insights into the characteristics of the estimator and the specification test under misspecification. The analysis proceeds by examining the Frechet derivatives [Wouk (1979, ch. 12)] of $\bar{\theta}(v)$ and $\bar{\tau}(v)$ at $v = 0$. These derivatives are by definition linear functionals, $D\bar{\theta}$ and $D\bar{\tau}$, on V such that

$$\bar{\theta}(v) = \theta_0 + D\bar{\theta}[v] + o(\|v\|_2), \quad \bar{\tau}(v) = D\bar{\tau}[v] + o(\|v\|_2).$$

In other words, $\bar{\theta}(v)$ and $\bar{\tau}(v)$ can each be approximated by a linear function of v up to an approximation error that is $o(\|v\|_2)$; the analogy is that a first-term Taylor approximation to an ordinary function. Strictly speaking, V itself is not a linear space for the domains of the linear operators $D\bar{\theta}$ and $D\bar{\tau}$. But this is not a problem because the derivatives are actually defined for all $v \in L_2(F_0)$, while here we are only restricting their domains to those values of v for which the G_v is a genuine distribution function.

Three recent papers that have also undertaken local specification analysis are Kiefer and Skoog (1984), Newey (1984) and Davidson and Mackinnon (1984a). The Kiefer–Skoog paper analyzes the effects of misspecification on the limiting parameter value $\bar{\theta}$ for a set of finite-dimensional 'directions' of misspecification generated by incorrect parametric restrictions. Here we study the effects of misspecification on both $\bar{\theta}$ and $\bar{\tau}$ for the more extensive infinite-dimensional set of directions generated by incorrect distributional assumptions. The Newey and Davidson–Mackinnon papers also use an infinite-dimensional set of directions, but the emphasis in these papers is more of studying the behavior of the limiting chi-square non-centrality parameter, while here we focus more on $\bar{\theta}$ and $\bar{\tau}$. Also, the mathematical methods of these other two papers are different than those used here. These papers use concepts of differentiation similar to the Gateaux derivative, whereas here we use the Frechet derivative [see Wouk (1979, ch. 12)]. The requirements for the existence of the Frechet derivative are more stringent than those for the Gateaux derivative – e.g., a Cobb–Douglas production function has Gateaux derivatives at the origin in all directions in the positive orthant but does not have a Frechet derivative there – and establishing the existence of the Frechet derivative entails more detailed arguments. However, because the Frechet derivative, or more precisely the corresponding linear functional, is independent of the direction at which it is evaluated, the qualitative predictions based on it are stronger. For instance, suppose it is found that a specification test locally has zero power in two different directions. Then, if the appropriate Frechet derivatives exist, the test will also be guaranteed to have zero power locally in all directions that are linear combinations of these two directions. A similar guarantee is not available if only the Gateaux derivatives exist and thus conclusions based on the

weaker concept of differentiation could potentially be misleading in some cases.

The main results for the Frechet derivatives are in Theorem 6 below, which is proceeded by a technical lemma:

Lemma 2. Let Assumptions 1–4 hold for each G_v of the form $\mathrm{d}G_v = [1 + v(y)]\,\mathrm{d}F_0(y)$ and put

$$\alpha(\theta, \tau, v) = \begin{bmatrix} \lambda_1(\theta, v) \\ \lambda_2(\theta, v) - \tau \end{bmatrix},$$

where

$$\lambda_1(\theta, v) = \int h(y, \theta)\,\mathrm{d}G_v(y), \qquad \lambda_2(\theta, v) = \int c(y, \theta)\,\mathrm{d}G_v(y).$$

Then the Frechet derivative of α at $(\theta, \tau, 0) \in \Theta \times T \times V$ exists and is given by

$$\mathrm{D}\alpha[\Delta\theta, \Delta\tau, v] = \begin{bmatrix} \frac{\partial \lambda_1}{\partial \theta'}(\theta, 0) & 0 \\ \frac{\partial \lambda_2}{\partial \theta'}(\theta, 0) & -I \end{bmatrix} \begin{bmatrix} \Delta\theta \\ \Delta\tau \end{bmatrix} + \begin{bmatrix} \int h(y, \theta) v(y)\,\mathrm{d}F_0(y) \\ \int c(y, \theta) v(y)\,\mathrm{d}F_0(y) \end{bmatrix},$$

where $(\Delta\theta, \Delta\tau, v) \in \Theta \times T \times V$, with the norm on $\Theta \times T \times V$ being $|\Delta\theta| + |\Delta\tau| + \|v\|_2$. In addition, $\mathrm{D}\alpha$ is continuous in (θ, τ). (Proof: Appendix.)

Theorem 6. Let Assumptions 1–4 hold for each G_v, let condition (ii) of Theorem 5 hold at θ_0, and let $\bar{\theta}(v)$ and $\bar{\tau}(v)$ be the implied almost sure limits of $\hat{\theta}_n$ and $\hat{\tau}_n$. Then $\bar{\theta}(v)$ and $\bar{\tau}(v)$ are Frechet differentiable in v at $v = 0$, with the derivatives given by

$$\mathrm{D}\bar{\theta}[v] = (J^o_{hh})^{-1} \int h(y, \theta_0) v(y)\,\mathrm{d}F_0(y),$$

$$\mathrm{D}\bar{\tau}[v] = \int c(y, \theta_0) v(y)\,\mathrm{d}F_0(y) - J^o_{ch}\,\mathrm{D}\bar{\theta}[v],$$

where, in the notation of the previous section,

$$J^o_{hh} = \int h(y, \theta_0) h(y, \theta_0)'\,\mathrm{d}F_0(y), \quad J^o_{ch} = \int c(y, \theta_0) h(y, \theta_0)'\,\mathrm{d}F_0(y).$$

Proof. Use Lemma 2 to apply the implicit function theorems for Frechet derivatives [Wouk (1979, Theorem 12,4.1, p. 294; Corollary 1, p. 296)] and the extended information equalities (16) and (17).

Note that implicit in the hypotheses of this theorem are the assumed existence and uniqueness of $\bar{\theta}(v)$ and $\bar{\tau}(v)$. Clarke (1983) provides a set of regularity conditions, albeit stronger than those assumed here, that guarantee existence and uniqueness.

An interesting way to express these derivatives is to put $\bar{\theta}_v = \mathrm{D}\bar{\theta}[v]$ and $\bar{\tau}_v = \mathrm{D}\bar{\theta}[v]$, and write them as

$$\bar{\theta}_v = \left(J^o_{hh}\right)^{-1} J^o_{hv}, \tag{26}$$

$$\bar{\tau}_v = J^o_{cv} - J^o_{ch}\bar{\theta}_v, \tag{27}$$

where

$$J^o_{hv} = \int h(y,\theta_0)v(y)\,\mathrm{d}F_0(y), \quad J^o_{cv} = \int c(y,\theta_0)v(y)\,\mathrm{d}F_0(y).$$

By analogy with least squares, the derivative $\bar{\theta}_v$ is the vector of regression coefficients in a regression of a random variable $v = v(Y)$ on a random variable $h = h(Y,\theta_0)$, with $Y \sim F(y,\theta_0)$. Thus, misspecification of the model leads to an inconsistent estimate of θ_0, except when the misspecification is in a direction v that is orthogonal to the gradient of the log-density function in the sense that $\mathrm{cov}(v(Y), h(Y,\theta_0)) = 0$ under $F(y,\theta_0)$. The special set of directions in which orthogonality holds corresponds to estimation situations in which the distributional assumptions implicit in the hypothesized model for the data are incorrect but the estimator $\hat{\theta}_n$ is still consistent (i.e., its genuine quasi-maximum likelihood). One well-known example of this in econometrics is FIML applied to a linear simultaneous equations system under the assumption of normally distributed errors when in fact the errors are not normal. Another is Phillips's (1982) example consisting of a two-equation non-linear simultaneous system and a family of non-normal error distributions. Phillips shows that FIML estimation of his system under a normality assumption gives consistent estimates of the parameters despite the failure of the distributional assumptions. The essential feature of Phillips's example is that each true model generated by an error distribution in the family of allowable distributions lies in a direction v that is uncorrelated with h under a normality assumption. Such an example, however, is clearly special and misspecification in general can lie in directions that are not orthogonal to h, so that the limiting value $\bar{\theta}$ is directly affected by misspecification. An example of this more serious type of misspecification would be an omitted variable from a simultaneous equations system.

Considering the derivative $\bar{\tau}_v$ in (27), we see that under misspecification the limit $\bar{\tau}$ is perturbed away from zero in ways. The first is through J^o_{cv}, which is the covariance under F_0 between $c(Y, \theta_0)$ and $v(Y)$; the second is through $\bar{\theta}_v$, i.e., through the effect that the misspecification has on the limit of the estimated parameters. In the special case when the model is wrong but the derivative v is uncorrelated with h under F_0 and $\bar{\theta}_v = 0$, the specification test will detect the failure of the distributional assumptions so long as the auxiliary criterion function has some covariance with v. An example of this use would be applying White's information matrix test under the maintained model $Y \sim N(\mu, \sigma^2)$ when the true distribution is symmetric about μ but with tails 'thicker' than those of the normal. In this example the ML estimators $\hat{\mu}$ and $\hat{\sigma}^2$ are consistent, but White's test would detect the departure from normality via the failure of the fourther moment about the mean to equal three times the square of the second moment. Again, though, this type of example is clearly special and in general misspecification will perturb $\bar{\theta}$ away from θ_0 and the second term in (27) will be non-zero.

The derivative $\bar{\tau}_v$ also has interesting interpretations based on a least squares analogy. Substitution of (26) into (27) gives

$$\bar{\tau}_v = J^o_{cv} - J^o_{ch}(J^o_{hh})^{-1} J^o_{hv} \quad \text{or} \quad \bar{\tau}_v = \int [c - B_{ch} h] v \, \mathrm{d}F_0, \tag{28}$$

where $B_{ch} = J^o_{ch}(J^o_{hh})^{-1}$. The matrix B_{ch} is simply the matrix of regression coefficients in a regression of the random vector $c = c(Y, \theta_0)$ on the vector $h = h(Y, \theta_0)$, with $Y \sim F_0(y)$. That is, $B_{ch} = \mathrm{E}_0[ch'](\mathrm{E}_0[hh'])^{-1}$, where the expectation $\mathrm{E}_0[\cdot]$ is under $\mathrm{d}F_0$. Thus $\bar{\tau}_v$ is the covariance between v and the auxiliary criterion function c, after the linear effects of h have been removed from c. Intuitively, the reason that the direction h is 'parsed out' of c, so to speak, is that the condition $\mathrm{E}_0[h(Y, \theta_0)] = 0$ was imposed directly in the estimation, and thus no specification test can be based on this direction.

5. Procedures for application

5.1. Outline

The following three-step algorithm for diagnostic testing and model evaluation summarizes the method for applying Theorems 1–5:

Step 1: Calculate $\hat{\theta}$ by maximum likelihood and retain for subsequent use the 'scores' $\hat{h}_i = h(Y_i, \hat{\theta})$, which are the gradients of the log-density function evaluated at the ML estimate.

Step 2: Choose a set of k auxiliarv criterion functions, $c_k(y, \theta)$, for $k = 1, 2, \ldots, K$. Each of the $c_k(y, \theta)$ should have the property that a large absolute

value for $\hat{\tau}_k = (1/n)\sum_1^n c_k(Y_i, \hat{\theta})$ would tend to cast some doubt on the assumptions underlying the likelihood model.

Step 3: Form $\hat{c}_{ki} = c_k(Y_i, \hat{\theta})$ and regress the $\hat{c}$'s on the scores. Specifically, estimate for each k the parameters of the equation

$$\hat{c}_{ki} = \beta_{k0} + \hat{h}_i'\beta_k + u_i, \qquad i = 1, 2, \ldots, n, \tag{29}$$

where u_i represents the error, and β_{k0} and β_k are the constant and the slope coefficients. The estimates $\hat{\beta}_{0k}$ of the intercepts in the regression equations will be the $\hat{\tau}_k$. Furthermore, individual t tests for non-zero intercepts in the regressions using the printed standard errors will be asymptotically valid tests for whether or not the corresponding $\hat{\tau}_k$ are significantly different from zero. Finally, the statistic $\hat{\tau}'(\hat{\Omega}/n)^{-1}\hat{\tau}$, where $\hat{\tau}$ is the vector of $\hat{\tau}_k$ and $\hat{\Omega}$ is the $K \times K$ cross-equation residual covariance matrix, is an appropriate chi-square statistic for testing whether or not all of the intercepts are jointly statistically significantly different from zero.

This regression-based procedure is the multivariate extension of a method proposed by Cox (1962, p. 411) for calculating the test statistic for his test for separate families. The procedure differs somewhat from that proposed by White (1982), Chesher (1983), Davidson and Mackinnon (1984b) and others for specification tests that are special cases of those considered here. In the other procedure the user calculates a single chi-square statistic equal to nR^2 (uncentered) for a regression of an $n \times 1$ column of ones on the $n \times (K + p)$ matrix whose ith row is $(\hat{c}_i', \hat{h}_i')$, where $\hat{c}_i' = (\hat{c}_{1i}, \hat{c}_{2i}, \ldots, \hat{c}_{ki})$ and $\hat{h}_i'$ is the transposed score vector. A major advantage of the procedure outlined above is that the individual t tests on the intercepts provide the user with detailed information on the statistical significance of each of the components $\hat{\tau}_k$. Using the other procedure the user obtains information only on the joint significance of all components taken together. A disadvantage, however, is that to get the chi-square statistic $\hat{\tau}'(\hat{\Omega}/n)^{-1}\hat{\tau}$ for joint significance the user must calculate a quadratic form in the residual covariance matrix for K separate regressions, while for the other procedure the user only calculates nR^2 for a single regression. Interestingly, the two chi-square statistics from the procedures are only asymptotically equivalent but not computationally equivalent in finite samples. If no degree of adjustment is used in calculating $\hat{\Omega}$, then least squares algebra shows that the statistics satisfy $\hat{\tau}'(\hat{\Omega}/n)^{-1}\hat{\tau} \geq nR^2$. With a degree of freedom adjustment, however, the inequality can go in either direction for finite n. Clearly, further work on the small sample properties of the chi-square statistics would be useful.

5.2. Empirical example

The potential uses of the type of specification tests considered in this paper can be illustrated with an application from the study of price behavior on speculative markets. One of the major stylized facts that has emerged from extensive research into the characteristics of short-term price movements on futures and equity markets is that the price changes generally have mean zero and are independent of one another, but their probability distribution is not a normal distribution. In particular, the pdf of the price changes has thick tails or is leptokurtic relative to the normal pdf. It is interesting, then, to see if the types of specification tests discussed here do in fact detect departure from normality in speculative price changes.

Suppose that the maintained model for the daily price change is $\Delta P_i \sim \mathrm{N}(0, \sigma^2)$. Given n observations on ΔP_i, the ML estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{1}^{n} \Delta P_i^2.$$

Under weak conditions that do not require normality the estimater $\hat{\sigma}^2$ is consistent. However, the model $\Delta P_i \sim \mathrm{N}(0, \sigma^2)$ when viewed as a probability model for the data is misspecified if the ΔP_i are not normally distributed, and may give biased and misleading predictions. Consider the following four auxiliary criterion functions for detecting misspecification of the probability model:

$$c_1(\Delta P, \sigma) = \Delta P^4 - 3\sigma^4,$$

$$c_2(\Delta P, \sigma) = |\Delta P| - \sqrt{2/\pi}\,\sigma,$$

$$c_3(\Delta P, \sigma) = I[|\Delta P/\sigma| \geq z_{0.40}] - 0.80,$$

$$c_4(\Delta P, \sigma) = I[|\Delta P/\sigma| \geq z_{0.005}] - 0.01,$$

where $I[\cdot]$ denotes the 0–1 indicator function, and where $z_{0.40}$ and $z_{0.005}$ are upper critical points of the normal distribution. Elementary calculations show that each of the $c_k(\Delta P, \sigma)$ integrates to zero under $\Delta P \sim \mathrm{N}(0, \sigma^2)$. Now put

$$\hat{c}_{ki} = c_k(\Delta P_i, \hat{\sigma}), \qquad i = 1, \ldots, n,$$

$$\hat{\tau}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{c}_{ki}, \qquad k = 1, \ldots, 4.$$

The random variable $\hat{\tau}_1$ is the difference between the observed and predicted fourth moments; $\hat{\tau}_2$ is the difference between the observed and predicted absolute first moments; $\hat{\tau}_3$ is the difference between the observed and the predicted fraction of the observations for which the magnitude $|\Delta P_i|$ lies more

than $z_{0.40}$ standard deviations above zero; and $\hat{\tau}_4$ is analogously defined for $z_{0.005}$.

These quantities were evaluated using 876 observations on the daily price change for the T-bills futures markets, 1976–79, from a data set described more fully in Tauchen and Pitts (1983). The results are as follows:

		Observed	Predicted	Difference ($\hat{\tau}_k$)
(1)	Fourth moment	0.0124	0.0063	0.0061
(2)	Abs. first moment	0.153	0.171	−0.017
(3)	Outer 80%	0.728	0.800	−0.072
(4)	Outer 1%	0.027	0.010	0.017

The first two rows of this display suggest that the actual fourth moment is somewhat larger than what is predicted by the probability model, while the absolute first moment is somewhat smaller. The last two rows indicate that there are too few observations above $\hat{\sigma}z_{0.40}$ in magnitude and too many above $\hat{\sigma}z_{0.005}$ than would be expected on the basis of the normal distribution.

To determine the statistical significance of each of the four $\hat{\tau}_k$, we simply test for a non-zero intercept in the appropriate auxiliary regression

$$\hat{c}_{ik} = \beta_{0k} + \beta_{1k}\hat{h}_i + \text{error}, \qquad i = 1, 2, \ldots, n,$$

where $\hat{h}_i = \partial l(\Delta P_i, \hat{\sigma}^2)/\partial\sigma^2$ is the gradient of the log-density function. Specifically,

$$\hat{h}_i = -\tfrac{1}{2}\left(\hat{\sigma}^{-2} - \Delta P_i^2\hat{\sigma}^{-4}\right), \qquad i = 1, 2, \ldots, 876,$$

with $\hat{\sigma} = 0.214$. The results are:

		$\hat{\beta}_{0k}$(s.d.)	$\hat{\beta}_{1k}$(s.d)
(1)	Fourth moment	0.0061 (0.0011)	0.0026 (0.000047)
(2)	Abs. first moment	−0.0170 (0.0020)	0.0056 (0.000089)
(3)	Outer 80%	−0.0717 (0.0145)	0.0050 (0.00060)
(4)	Outer 1%	0.0174 (0.0033)	0.0055 (0.00014)

The magnitudes of each of the four estimated intercepts exceeds twice their standard errors, which suggests that the four $\hat{\tau}_k$ are statistically significantly different from zero. The normal distribution thus appears to be inadequate as an approximation of the pdf of the ΔP_i.

6. Conclusions: suggestions for future research

This paper has developed the general asymptotic distribution theory for specification tests that are based on M-estimates of auxiliary parameters. Though this class of specification tests is quite large, there are some tests not included within it. For instance, the Kolmogorov–Smirnov test cannot be interpreted as being based on an M-estimator of some auxiliary statistic. Of course the large-sample theory for the K–S test statistic is well known, even for the case in which the distribution function being tested involves an estimated parameter. However, by embedding the K–S test statistic and others like it into the general theory of L- or U-estimates, one should be able to develop a new and very wide class of specification tests that is analogous to the class of tests developed here. This work is deferred to another paper.

Appendix

Proof of Lemma 1 (based on the ideas in Huber (1967, pp. 224–226)). Let $\varepsilon > 0$ be given. Define

$$u(y,\theta,d) = \sup_{|\gamma-\theta|\le d} |\phi(y,\gamma)-\phi(y,\theta)|.$$

By almost sure continuity, $\lim u(y,\theta,d)=0$ as $d\to 0$, with θ fixed, almost surely dG. Thus by dominated convergence, $\mathrm{E}[u(Y,\theta,d)]\le\varepsilon$ whenever $d\le \bar{d}(\theta)$. Let $B(\theta)$ denote an open ball of radius $\bar{d}(\theta)$ about θ. Together the $B(\theta)$ cover Θ. By compactness the $B(\theta)$ can be reduced to a finite open covering $B_k = B(\theta_k)$, $k=1,\ldots,K$. Put $d_k=\bar{d}(\theta_k)$ and $\mu_k=\mathrm{E}[u(Y,\theta_k,d_k)]$, and note that if $\theta\in B_k$, then $\mu_k\le\varepsilon$, and $|\lambda(\theta)-\lambda(\theta_k)|\le\varepsilon$. Now let $\theta\in B_k$ and consider

$$\begin{aligned}
&\left|\frac{1}{n}\sum_1^n\phi(Y_i,\theta)-\lambda(\theta)\right| \\
&\le \frac{1}{n}\sum_1^n|\phi(Y_i,\theta)-\phi(Y_i,\theta_k)| + \left|\frac{1}{n}\sum_1^n\phi(Y_i,\theta_k)-\lambda(\theta_k)\right| \\
&\quad + |\lambda(\theta_k)-\lambda(\theta)| \\
&\le \left(\frac{1}{n}\sum_1^n u(Y_i,\theta_k,d_k)-\mu_k\right)+\mu_k+\left|\frac{1}{n}\sum_1^n\phi(Y_i,\theta_k)-\lambda(\theta_k)\right|+\varepsilon \\
&\le 4\varepsilon,
\end{aligned}$$

whenever $n \geq N_k(\varepsilon)$ almost surely, by applying twice the strong law of large numbers and using $\mu_k \leq \varepsilon$. Thus

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{1}^{n} \phi(Y_i, \theta) - \lambda(\theta) \right| \leq 4\varepsilon,$$

whenever $n \geq \max_k N_k(\varepsilon)$ almost surely, which proves the result.

Proof of Theorem 5. Let Δ by any non-zero vector in R^p such that $|\Delta|$ is small enough that $\gamma = \theta_0 + \Delta \in \Theta$, and consider

$$|(J_{ch}^o + K_c^o)(\Delta/|\Delta|)|. \tag{A.1}$$

Since $\int c(y, \theta) f(y, \theta) \, d\mu = 0$ for all θ in Θ, then (A.1) can be written as

$$\Big| (1/|\Delta|) \int [c(y, \gamma) f(y, \gamma) - c(y, \theta_0) f(y, \theta_0)] \, d\mu$$

$$- (J_{ch}^o + K_c^o)(\Delta/|\Delta|) \Big|,$$

which is dominated by

$$(1/|\Delta|) \left| \int [c(y, \gamma) - c(y, \theta_0)] f(y, \theta_0) \, d\mu - K_c^o \Delta \right|$$

$$+ \left| (1/|\Delta|) \int c(y, \gamma) [f(y, \gamma) - f(y, \theta_0)] \, d\mu - J_{ch}^o (\Delta/|\Delta|) \right|. \tag{A.2}$$

The first of the two terms in (A.2) tends to zero as $|\Delta| \to 0$ since K_c^o is, by definition, the matrix of partial derivatives with respect to θ' of $\int c(y, \theta) f(y, \theta_0) \, d\mu$ evaluated at $\theta = \theta_0$. Now write

$$f(y, \gamma) - f(y, \theta_0) = f(y, \dot{\theta}) h(y, \dot{\theta})' \Delta,$$

where $\dot{\theta}$ (which depends upon y and Δ) is on the line segment between θ_0 and γ. By definition $J_{ch}^o = \int c(y, \theta_0) h(y, \theta_0)' f(y, \theta_0) \, d\mu$ and so the second term in (A.2) is dominated by the sum of

$$\left| \int c(y, \gamma) \left[\frac{f(y, \dot{\theta})}{f(y, \theta_0)} h(y, \dot{\theta}) - h(y, \theta_0) \right]' \frac{\Delta}{|\Delta|} f(y, \theta_0) \, d\mu \right|, \tag{A.3}$$

and

$$\left|\int [c(y,\gamma)-c(y,\theta_0)]\,h(y,\theta_0)'\frac{\Delta}{|\Delta|}f(y,\theta_0)\,\mathrm{d}\mu\right|. \tag{A.4}$$

From (ii) of Assumption 4 the expected value of $|c(y,\gamma)|^2$ with respect to $f(y,\theta_0)$ is uniformly bounded in γ, and so by the Schwarz inequality the square of the term (A.3) is dominated by a constant times

$$\int\left|\frac{f(y,\dot\theta)}{f(y,\theta_0)}h(y,\dot\theta)-h(y,\theta_0)\right|^2 f(y,\theta_0)\,\mathrm{d}\mu.$$

The second hypothesis of the theorem and (ii) of Assumption 4 imply that the squared term in this integral is dominated by a function integrable with respect to $f(y,\theta_0)\,\mathrm{d}\mu$. Thus by dominated convergence this integral, and hence (A.3), tends to zero as $|\Delta|\to 0$. Finally, since c satisfies the Lipshitz conditions in Assumption 4 and $|h|^2$ is dominated, the term (A.4) is $O(|\Delta|^{1/2})$.

In summary, the matrix $J_{ch}^o + K_c^o$ satisfies for any non-zero $\Delta \in R^p$

$$\lim_{|\Delta|\to 0}\left|(J_{ch}^o + K_c^o)(\Delta/|\Delta|)\right| = 0,$$

irrespective of how $|\Delta|\to 0$; hence it equals zero because it maps every vector of unit length into zero.

Proof of Lemma 2. Put $\delta = |\Delta\theta| + |\Delta\tau| + \|v\|_2$, and consider

$$\begin{aligned}
&\delta^{-1}\left|\alpha(\theta+\Delta\theta,\tau+\Delta\tau,v)-\alpha(\theta,\tau,0)-\mathrm{D}\alpha[\Delta\theta,\Delta\tau,v]\right| \\
&\le \delta^{-1}\Big|\lambda_1(\theta+\Delta\theta,v)-\lambda_1(\theta,0)-(\partial\lambda_1/\partial\theta')(\theta,0)\Delta\theta \\
&\quad -\int h(y,\theta)v(y)\,\mathrm{d}F_0(y)\Big| \\
&\quad +\delta^{-1}\Big|\lambda_2(\theta+\Delta\theta,v)-\lambda_2(\theta,0)-(\partial\lambda_2/\partial\theta')(\theta,0)\Delta\theta \\
&\quad -\int c(y,\theta)v(y)\,\mathrm{d}F_0(y)\Big|.
\end{aligned} \tag{A.5}$$

We must show that the right-hand side of this inequality tends to zero as $\delta\to 0$. Now the first term on the right-hand side of the inequality (A.5) cannot

exceed the sum of

$$\delta^{-1}|\lambda_1(\theta+\Delta\theta,0)-\lambda_1(\theta,0)-(\partial\lambda_1/\partial\theta')(\theta,0)\,\Delta\theta| \tag{A.6}$$

and

$$\delta^{-1}\left|\int[h(y,\theta+\Delta\theta)-h(y,\theta)]\,v(y)\,\mathrm{d}F_0(y)\right|. \tag{A.7}$$

From the definition of $\partial\lambda_1/\partial\theta'$, the expression (A.6) must tend to zero with δ. On the other hand, the expression (A.7) is dominated by

$$\delta^{-1}\|v\|_2\left[\int|h(y,\theta+\Delta\theta)-h(y,\theta)|^2\,\mathrm{d}F_0(y)\right]^{1/2}, \tag{A.8}$$

and by the second Lipshitz condition (iii) in Assumption 4 there is a constant β such that (A.8) is of the form

$$\delta^{-1}\|v\|_2\beta|\Delta\theta|^{1/2}\leq\beta|\Delta\theta|^{1/2},$$

which tends to zero with δ. In an exactly analogous manner the second term on the right-hand side of the initial inequality (A.5) (the term corresponding to λ_2) tends to zero with δ, and so the existence of the Frechet derivative of α with respect to θ,τ,v at $(\theta,\tau,0)$ has been established. The continuity of the derivative in (θ,τ) is presupposed in (iv) of Assumption 4.

References

Aguirre-Torres, V. and A.R. Gallant, 1983, The null and non-null asymptotic distribution of the Cox test for multivariate nonlinear regression, Journal of Econometrics 21, 1–33.

Burguette, J.F., A.R. Gallant and G. Souza, 1982, On unification of the asymptotic theory of nonlinear econometric models, Econometric Review 1, 151–190.

Chesher, Andrew, 1983, The information matrix test: Simplified calculation via a score interpretation, Economics Letters 13, 45–48.

Clarke, Brenton R., 1983, Uniqueness and Frechet differentiability of functional solutions to maximum likelihood type equations, Annals of Statistics 11, 1196–1205.

Cox, D.R., 1962, Further results on tests of separate families of hypotheses, Journal of the Royal Statistical Society B 24, 406–424.

Davidson, R. and J. Mackinnon, 1984a, Implicit alternatives and the local power of test statistics, Queen's University discussion paper no. 556.

Davidson, R. and J. Mackinnon, 1984b, Model specification tests based on artificial linear regressions, International Economic Review 25, 485–502.

Engle, Robert F., 1982, A general approach to Lagrange multiplier model diagnostics, Journal of Econometrics 20, 83–104.

Hausman, J., 1978, Specification tests in econometrics, Econometrica 46, 1251–1271.

Heckman, J., 1984, The χ^2 goodness of fit for models with parameters estimated from microdata, Econometrica 52, 1543–1548.

Huber, Peter J., 1967, The behavior of maximum likelihood estimates under nonstandard conditions, in: Lucien M. LeCam and Jerzy Neyman, eds., Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1 (University of California Press, Berkeley, CA) 221–234.
Huber, Peter J., 1981, Robust statistics (Wiley, New York).
Kendall, M.G. and A. Stuart, 1973, The advanced theory of statistics, 3rd ed. (Griffin, London).
Kiefer, N. and G. Skoog, Local asymptotic specification error analysis, Econometrica 52, 873–886.
Lancaster, T., 1984, The covariance matrix of the information matrix text, Econometrica 52, 1051–1054.
Moore, D.S. and M.C. Spruill, 1975, Unified large-sample of general chi-squared statistics for tests of fit, Annals of Statistics 3, 599–616.
Newey, W.K., 1984, Maximum likelihood specification testing and instrumented score tests, Princeton University mimeo.
Phillips, P.C.B., 1982, On the consistency of FIML, Econometrica 50, 1307–1324.
Tauchen, George and M. Pitts, 1983, The price variability–volume relationship on speculative markets, Econometrica 51, 485–505.
White, H., 1982, Maximum likelihood estimation of misspecified models, Econometrica 50, 1–26.
Wouk, A., 1979, A course of applied functional analysis (Wiley, New York).