CHI-SQUARE DIAGNOSTIC TESTS FOR ECONOMETRIC MODELS: THEORY

BY DONALD W. K. ANDREWS

This paper extends the Pearson chi-square testing method to nondynamic parametric econometric models, in particular, to models with covariates. The paper establishes the asymptotic distribution of the test statistic under the null and local alternatives when the test statistic is based on data-dependent random cells of a general form and on an arbitrary asymptotically normal estimator. These results are attained by extending recent probabilistic results for the weak convergence of empirical processes indexed by sets. The chi-square test that is introduced can be used to test goodness-of-fit of a parametric model, as well as to test particular aspects of the parametric model that are of interest.

KEYWORDS: Brownian bridge, chi-square statistic, diagnostic test, empirical process, goodness-of-fit test, Pearson chi-square test, random cells, Vapnik-Cervonenkis class, weak convergence

1 INTRODUCTION

This paper extends the Pearson chi-square testing method to nondynamic parametric models with covariates. By allowing covariates, an extremely wide range of cross-sectional econometric and statistical models can be investigated using chi-square tests. The extension allows for data-dependent random cells, flexible choice of cell shapes, and estimation of unknown parameters by general methods. These features enable one to test the classical goodness-of-fit hypothesis that the parametric model is correctly specified, as well as to test specific aspects of the model.

The literature on the Pearson (1900) chi-square test and its extensions is enormous. In consequence, we mention here a select few papers that are particularly pertinent to the present results. The data-dependent random cells used in this paper are specified quite generally following the approach of Pollard (1979). His results, in turn, build upon those of Watson (1957), Chibisov (1970), and Moore (1971). For an alternative approach, see Moore and Spruill (1975) and Tauchen (1985). The estimation procedure used by the test need not be the peculiar multinomial maximum likelihood (ML) estimator required by the Pearson chi-square test, but can be any asymptotically normal estimator, as in Nikulin (1973) and Rao and Robson (1974).

The extension of Pearson's chi-square test to models with covariates was initiated by McFadden (1974) for a special case of the multinomial logit model and extended to a larger class of models by Heckman (1984) and Horowitz (1985). Also see Kwei (1987). The present paper extends it further to most parametric cross-sectional and panel data models used in econometrics. It pre-

1 I would like to thank Chris Monahan for excellent research assistance and David Pollard for his helpful comments and suggestions. I also thank several referees, the co-editor Gary Chamberlain, and the participants of the econometrics workshops at Berkeley, Stanford, Harvard/MIT, Chicago, and Princeton for their very helpful comments and suggestions. The financial support of the National Science Foundation through Grant No. SES-8419789 is acknowledged gratefully.
sents results that are the most general available with respect to the choice of cells and their shapes, the applicable estimation procedures, and the models covered. The generality of the results given here allows one to construct a wide variety of goodness-of-fit tests not previously available in the literature.

The paper is organized as follows: Section 2 defines the test statistic, discusses the choice of cells, and sets out the assumptions and regularity conditions used to obtain the asymptotic results. Section 3 describes the use of the weak convergence of the conditional empirical process indexed by partitions to derive the asymptotic distribution of the test statistic under the null. Simplified computation of the test statistic also is discussed. Section 4 presents local power, consistency, and asymptotic optimality results. Section 5 presents Monte Carlo comparisons of chi-square tests with other specification tests in censored regression models. An Appendix contains proofs of results given in Sections 3 and 4.

A less technical description of the testing procedures described in this paper is available in Andrews (1988). In addition, the latter discusses the choice of cells for particular testing problems in greater detail than is given here.

2 DEFINITIONS AND ASSUMPTIONS

This section defines the chi-square test statistic $X^2_\lambda(\hat{\Lambda}, \hat{\beta})$. It also presents assumptions on the model, estimator, and random cells that are used below to obtain the asymptotic distribution of $X^2_\lambda(\hat{\Lambda}, \hat{\beta})$ under the null hypothesis of correct specification.

2.1. Definition of the Chi-Square Test Statistic

The observed sample consists of the first $n$ terms of the sequence of random vectors $((Y_i, X_i): i = 1, 2, \ldots)$. $Y_i$ and $X_i$ are vectors of response variables and covariates that take values in $Y \subset \mathbb{R}^k$ and $X \subset \mathbb{R}^k$, respectively. $P$ denotes the distribution of $((Y_i, X_i): i = 1, 2, \ldots)$ under the null hypothesis. The model is assumed to be nondynamic when correctly specified:

**Assumption M1**: $((Y_i, X_i): i = 1, 2, \ldots)$ are independent and identically distributed under $P$.

The parametric models considered here consist of parametric families of conditional distributions of response variables given covariates. The marginal distributions of the covariates are left unrestricted, as is usually the case in practice. Hence, the null hypothesis of correct specification is:

$H_0$: The conditional distribution of $Y_i$ given $X_i$ is in the parametric family $(f(y|x, \theta): \theta \in \Theta)$, where $f(y|x, \theta)$ is a density with respect to some $\sigma$-finite measure $\mu$ and $\Theta \subset \mathbb{R}^l$ is a parameter space.

Since $\mu$ need not be Lebesgue measure, $Y_i$ may be discrete, continuous, or mixed.
Let $P_X$ denote the distribution of $X_i$ under $P$. Since $P_X$ is not restricted by the null hypothesis, $X_i$ also may be discrete, continuous, or mixed. Let $\theta_0$ denote the true parameter value when the null hypothesis is true.

The proposed chi-square test statistic is constructed by partitioning into disjoint cells the region $Y \times X$ in which the response variables and covariates lie. The test statistic is given by a quadratic form based on the differences between the observed and conditionally expected numbers of outcomes in each cell with the latter being calculated using the parametric model. If the parametric model is correct, then the differences are due solely to random fluctuations. On the other hand, if the parametric model is incorrect, both random and systematic components contribute to the differences and the quadratic form takes on larger values.

Following the approach of Pollard (1979) (who considers models with no covariates), the cells are chosen from a class $C$ of measurable sets in $Y \times X$. Let $J$ denote the number of cells used in constructing the test statistic. $J$ is assumed to be fixed for all $n$. Let $D$ be a class of partitions of $Y \times X$, each partition being comprised of $J$ sets from $C$. That is,

\begin{equation}
D = \left\{ \gamma \in C^J : \bigcup_{j=1}^{J} \gamma_j = Y \times X, \gamma_j \cap \gamma_k = \emptyset, \forall j \neq k \right\},
\end{equation}

where $\gamma_j$ and $\gamma_k$ denote elements of the partition $\gamma$. For each sample size $n$, the $J$ cells used to construct the test statistic are given by a random element of $D$, denoted $\bar{\Gamma}$ (where $\bar{\Gamma}$ depends on $n$ in general).

Next, we define a stochastic process, called the conditional empirical process, that is the basis of the chi-square test statistic. Let $P_n(\cdot)$ denote the empirical measure of the sample $\{(Y_i, X_i), i = 1, \ldots, n\}$ indexed by elements $\gamma$ in $D$. That is,

\begin{equation}
P_n(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \gamma(Y_i, X_i),
\end{equation}

where $\gamma(Y_i, X_i)$ denotes the vector of indicator functions of $(Y_i, X_i) \in \gamma_j$ for $j = 1, \ldots, J$. Let $F_n(\cdot, \theta)$ denote the conditional empirical measure constructed using the parametric conditional distribution of $Y_i$ given $X_i$. That is,

\begin{equation}
F_n(\gamma, \theta) = \frac{1}{n} \sum_{i=1}^{n} \int_{\gamma} \gamma(Y_i, X_i)f(Y|X_i, \theta) \, d\mu(Y) = \frac{1}{n} \sum_{i=1}^{n} F(\gamma, X_i, \theta).
\end{equation}

**Definition:** The conditional empirical process $\nu_n(\cdot, \theta)$ indexed by elements $\gamma$ of $D$ is defined as

\begin{equation}
\nu_n(\cdot, \theta) = \sqrt{n} \left( P_n(\cdot) - F_n(\cdot, \theta) \right).
\end{equation}

Let $\hat{\theta}$ be some estimator of $\theta_0$. Then, the random vector $\nu_n(\bar{\Gamma}, \hat{\theta})$ is proportional to the differences between the observed and (estimated) conditionally expected cell frequencies. This vector is the basis of the test statistic. Under the
assumptions introduced below, it has an asymptotic normal distribution with
covariance matrix $\Sigma_0$ (defined in Section 3 below).

Let $\hat{W}$ be a consistent estimator of some generalized inverse of $\Sigma_0$. The
chi-square test statistic is defined as follows:

**Definition:** The random cell chi-square test statistic is given by

\[(2.5) \quad X_n^2(\hat{\epsilon}, \hat{\theta}) = \rho_n(\hat{\epsilon}, \hat{\theta})'\hat{W}p_n(\hat{\epsilon}, \hat{\theta}) .\]

The test based on this statistic rejects the null hypothesis if the statistic is
sufficiently large, where large is determined by the asymptotic chi-square distri-
bution of the test statistic under the null (with degrees of freedom given by the
rank of $\Sigma_0$).

### 2.2. Cell Choices

We now outline several different choices of cells for testing general goodness-of-fit, as well as for testing goodness-of-fit of particular aspects of a parametric
model. (See Andrews (1988) for a more detailed discussion.)

By definition, a nonparametric partitioning method is one that does not rely on
the parametric model specified in the null hypothesis to form the cells. For
general goodness-of-fit tests, four basic nonparametric partitioning strategies are
possible: (i) group all variables together and nonparametrically partition $Y \times X$,
(ii) nonparametrically partition $Y$ and $X$ separately and form cross-product cells,
(iii) first partition $X$, then separately partition $Y$ for each $X$ cell, and (iv) do
likewise with $Y$ partitioned first. Method (i) is the least structured approach.
Method (ii) allows one to see which regions in $Y$ or $X$ are modelled inadequately
by the parametric model. It has the disadvantage, however, that it may create
numerous low probability cells. This strategy has been considered by McFadden
(1974) and Horowitz (1985) for discrete choice models using nonrandom cells.
The third and fourth strategies allow one to see which regions in $X$ and $Y$ are
modelled inadequately, respectively. Heckman's (1984) partitioning scheme corre-
sponds to strategy (iii) or (iv) where $Y$ is partitioned using nonrandom cells and
each partition of $X$ consists of a single cell. Strategy (iv) is natural in discrete
choice models.

The range space of a single real-valued variable can be partitioned nonparamet-
rically by a number of methods: (1) using the sample mean (or median) plus
or minus multiples of the sample standard deviation (or the sample absolute
deviation), (2) using the $k$-means clustering procedure, (3) forming cells with
equal numbers of observations, or (4) using some other clustering procedure
(e.g., see Hartigan (1975), Romesburg (1984), and Spath (1985)).

For the case of partitioning the space of vector-valued variables, method (1)
generalizes by considering concentric ellipses centered at the sample mean with
shape determined by the sample covariance matrix, perhaps also partitioned
along the axes of the ellipses, (2) generalizes without change, (3) does not
generalize unambiguously, and (4) applies here as well. In addition, one can use a procedure that reduces a vector-valued variable to a real-valued variable and then apply one of the methods above for partitioning the space of a real-valued variable. Such a reduction can be obtained by using the first principal component, by using an estimator of a “regression parameter” $\beta$ (i.e., reduce $X_i$ to the scalar $X_i'\beta$), or by using any of a number of other methods. The random partitioning methods that fall under the assumptions given in Section 3.3 below are quite flexible, so one can choose the partitions that are of greatest interest with minimal encumbrances.

Next we discuss several examples of tests in which part of the specified parametric model is maintained and part is tested against a nonparametric class of alternatives. First, suppose one wishes to test the assumption of normality in a linear regression model. Here the alternatives of interest maintain the linear regression structure, but may have any nonnormal error distribution. A chi-square test can be constructed with the partition based on the least squares residuals in a manner analogous to that used with chi-square tests of normality in models with iid response variables (and no nonconstant covariates). In fact, it is shown below that the chi-square test of normality (or any other distribution) has the same asymptotic properties in these two models. For the latter model with no covariates, the usefulness of the chi-square statistic for testing against the nonparametric class of all nonnormal distributions has been demonstrated in the statistics literature.

The extension of the above test of normality to censored or truncated regression models is straightforward. One just needs to adjust the cells to account for the fact that the residuals are not observed when the response variable is too small and/or too large.

In linear seemingly unrelated and multivariate regression and simultaneous equations models, one can test for multivariate normality of errors by partitioning the space of residuals into concentric ellipses (possibly partitioned along their axes). This procedure is analogous to that used by Moore and Stubblebine (1981) for testing multivariate normality of iid random vectors. As with the linear regression model, the chi-square tests have the same asymptotic properties whether or not covariates are present in the first two models listed above. Again, these tests can be extended to models with censoring or truncation.

In a number of models, the assumption of bivariate normality is crucial for the consistency of common estimation procedures, e.g., in selection and switching regression models. For these models as well, bivariate normality can be tested by forming cells based on residuals. One needs to form the cells, however, so that only the partial residual information that is provided by the sample is used in the partitioning scheme (see Andrews (1988)).

As a final example, suppose one wishes to test whether some scalar covariate $X_i$, affects the conditional distribution of $Y_i$ given $X_i$ in a more complex fashion than simply through a “regression function” $X_i'\beta$. The class of alternatives of interest in this case clearly is nonparametric, since it includes models in which $X_i$ operates through some nonlinear function $g(X_i)$, in which $X_i$ interacts with
other covariates linearly or nonlinearly, and in which different values of $\beta$ are appropriate for different values of $X_i$. A chi-square test can be formed for this situation by partitioning $X$ based on the covariate $X_i$, alone and then partitioning $Y$ nonparametrically for each $X$ cell. This test has some power against a wide variety of alternatives.

In all of the examples above the cells can be formed in exactly the same way whether or not there are restrictions on the parameters under the null hypothesis. For example, with cells based on residuals, one merely estimates the model with the restrictions imposed and forms cells in the same way.

For several applications of chi-square tests in the econometrics literature, see Klein (1974), Moore and Stubblebine (1981), Nakamura and Nakamura (1983, 1985), Tauchen (1985), and Veall (1986). These applications either rely on fixed cells or apply to models without covariates. Using the results of the present paper, the potential range of applications is enhanced greatly.

2.3. Assumptions

To establish asymptotic normality of $r_n(\hat{\theta}, \hat{\theta})$ and to consistently estimate its asymptotic covariance matrix, we need the conditional density $f(y|x, \theta)$, or equivalently its score function $s(y|x, \theta) = \partial \log f(y|x, \theta)/\partial \theta$, to be smooth in $\theta$ near $\theta_0$. We assume:

ASSUMPTION M2: In some neighborhood $N_1$ of $\theta_0$, the score function $s(y|x, \theta)$ and its partial derivative $\partial s(y|x, \theta)/\partial \theta$ are continuous in $\theta$ and dominated by a square integrable function $r(y, x)$ and an integrable function $\tilde{r}(y, x)$, respectively. Specifically, $|s(y|x, \theta)| \leq r(y, x)$ and $|\partial s(y|x, \theta)/\partial \theta| \leq \tilde{r}(y, x)$. $\forall \theta \in N_1$, $\forall k, l = 1, \ldots, L$, where

$$\int_X \sup_{\theta \in N_1} \int_Y \left[ r^2(y, x) + \tilde{r}(y, x) \right] f(y|x, \theta) \, d\mu(y) \, dP_X(x) < \infty.$$ 

Most parametric models that are used in practice and that satisfy the regularity conditions for ML estimation also satisfy M2.

The chi-square test statistic defined in (2.5) relies on an estimator $\hat{\theta}$ of $\theta_0$. Of this estimator, we assume:

ASSUMPTION E1: The estimator $\hat{\theta}$ satisfies

$$\sqrt{n} \left( \hat{\theta} - \theta_0 \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n D_0^{-1} \psi(Y, X_i, \theta_0) + o_p(1) \quad \text{as} \quad n \to \infty$$

under the null hypothesis $P$, where $\psi(y, x, \theta)$ is a measurable function from $Y \times X \times \Theta$ to $R^l$ that satisfies $f y \psi(y, x, \theta_0) f(y|x, \theta_0) d\mu(y) = 0$, $\forall x \in X$, and $V_0 \equiv E_P \psi(Y, X, \theta_0) \psi(Y, X, \theta_0)'$ and $D_0 \equiv -E_P (\partial \psi(Y, X, \theta_0)/\partial \theta)'$ are finite and nonsingular.

Under Assumption E1, $\sqrt{n} \left( \hat{\theta} - \theta_0 \right)$ has an asymptotic normal distribution with covariance matrix $D_0^{-1} V_0 D_0^{-1}'$. Assumption E1 is not very restrictive. It is
fulfilled by most estimators currently used in practice. For the ML estimator (or any asymptotically efficient estimator), \( \psi(y, x, \theta) \) is just the score function \[ \frac{\partial \log f(y|x, \theta)}{\partial \theta} \], \( D_0 = V_0 \), and \( D_0 \) is the information matrix evaluated at \( \theta_0 \). For M-estimators, \( \psi(y, x, \theta) \) is just the defining function of the estimator (e.g., see Andrews (1986, Theorem 1)). In other cases, \( \psi(y, x, \theta) \) can be deduced straightforwardly from the definition of the estimator (usually, using the first order conditions of the optimization problem).

To consistently estimate the asymptotic covariance matrix of \( \psi(y, x, \theta) \), we use a smoothness condition on \( \psi(y, x, \theta) \) in \( \theta \) near \( \theta_0 \):

**Assumption E2**: In some neighborhood \( N_2 \) of \( \theta_0 \), \( \psi(y, x, \theta) \) and its first and second partial derivatives (with respect to \( \theta \)) are continuous in \( \theta \) and are dominated by the functions \( r_0(y, x) \), \( r_1(y, x) \), and \( r_2(y, x) \), respectively. Specifically,
\[
|\psi(y, x, \theta)| \leq r_0(y, x), \quad |\partial \psi(y, x, \theta)/\partial \theta_k| \leq r_1(y, x),
\]
and
\[
|\partial \psi(y, x, \theta)/\partial \theta_k \partial \theta_l| \leq r_2(y, x), \quad \forall \theta \in N_2, \forall k, l, m = 1, \ldots, L.
\]

where
\[
\int \sup_{N_2} \left[ r_0^2(y, x) r(y, x) + r_1(y, x) [r_0(y, x) + r(y, x)] + r_2(y, x) \right]
\]
\[
\cdot f(y|x, \theta_0) \, d\mu(y) \, dP_X(x) < \infty.
\]

Next we consider the random cells \( \hat{\Gamma} \) (which depend on \( n \) in general). Below we assume that \( \hat{\Gamma} \) converges in probability to some fixed partition of cells \( \Gamma \in D \) as \( n \to \infty \). To make this assumption meaningful a topology needs to be defined on \( D \). Let \( F \) denote the joint distribution of \( (Y, X) \) under \( P \). That is, for all measurable sets \( C \) in \( Y \times X \),
\[
(2.7) \quad F(C) = \int \int f(y|x, \theta_0) \, d\mu(y) \, dP_X(x).
\]

Equip \( C \) with the topology generated by the \( L^2(F) \) semi-norm and give \( D \) the corresponding product topology. With this topology on \( C \), two sets \( C_1 \) and \( C_2 \) in \( Y \times X \) are close if \( F(C_1 \Delta C_2) \) is small, where \( \Delta \) denotes the symmetric difference operator. With this topology on \( D \), convergence in probability of the random elements \( \hat{\Gamma} \) to \( \Gamma \) means that for all \( \epsilon > 0 \),
\[
(2.8) \quad P\left( F(\hat{\Gamma}, \Delta \Gamma_j) > \epsilon \right) \to 0 \text{ as } n \to \infty, \quad \forall j = 1, \ldots, J. \quad \text{or equivalently,}
\]
\[
F(\hat{\Gamma}, \Delta \Gamma_j) \xrightarrow{P} 0 \quad \text{as } n \to \infty, \quad \forall j = 1, \ldots, J.
\]

**Assumption RC1**: \( \hat{\Gamma} \xrightarrow{P} \Gamma \text{ as } n \to \infty \).
This assumption is fulfilled by all of the random cell choices discussed above and in Andrews (1988). It is sufficiently flexible to allow many other choices of random cells as well.

The asymptotic distribution of the chi-square test statistic defined in (2.5) is obtained by proving that the conditional empirical process indexed by elements of $D$ converges weakly (as a stochastic process) to a particular tied-down Gaussian process indexed by elements of $D$. This convergence result does not hold if $D$ is allowed to contain all measurable partitions of $Y \times X$. For the result to hold we restrict $C$ as follows:

**Assumption RC2:** $C$ is a Vapnik-Cervonenkis (VC) class.

By definition, the class $C$ of subsets of $Y \times X$ is a VC class if there exists a polynomial $\rho(\cdot)$ such that, for every set of $v$ points in $Y \times X$, $C$ picks out at most $\rho(v)$ distinct subsets. That is, if $S \subset Y \times X$ contains $v$ points, then there are at most $\rho(v)$ distinct sets of the form $S \cap C$ for $C \in C$. The VC class condition is particularly convenient, because it does not depend on the underlying distribution $P$ of the data. The condition can be relaxed, if need be, by replacing it with a condition that depends on $P$.

Examples of VC classes of sets and methods for generating scores of such classes are given in Pollard (1979, 1984), Dudley (1978, 1984), and Vapnik and Cervonenkis (1971, 1981). We mention that the class of polyhedrons in $Y \times X$ with at most $r$ sides is a VC class for each fixed $r$. Thus, the partitioning of $Y \times X$ via (i) rectangles or rectangular cylinder sets (with respect to any coordinate system), (ii) $k$-means clustering, (iii) principal component analysis, or (iv) any other algorithm that yields cells with a finite number of straight edges, satisfies RC2.

The class of hyperellipsoids in $Y \times X$ also is a VC class of sets. Furthermore, given any finite number of VC classes $C_1, \ldots, C_g$, the class of all unions, intersections, differences, and complements of sets in $C_1, \ldots, C_g$ forms a VC class (e.g., see Pollard (1984)). This result can be used to create greatly expanded VC classes from existing ones. For example, by intersecting hyperellipsoids and finite-sided polyhedrons, we obtain VC classes of cells of the form suggested above for testing multivariate normality.

The class of graphs of functions that form a finite dimensional vector space comprises a VC class. Polynomial functions (of given degree) of the elements of $(Y, X)$ constitute a finite dimensional vector space. Hence, cells that are based on residuals that are polynomial functions of the elements of $(Y, X)$ form a VC class. This establishes Assumption RC2 for numerous examples including those described above that involve linear simultaneous equations, multiple regression, and SUR models. The same argument applies with restrictions of any kind on the parameters.

We note that an alternative method of defining and analyzing random cells is to require the cells to depend on the data through a finite dimensional parameter; see Moore and Spruill (1975) and Tauchen (1985). This approach is less general
than the one adopted here, since it does not cover some nonparametric partitioning schemes, but it has the advantage of more clearly illustrating the reason why the use of random cells does not affect the asymptotic distribution of the test statistic.

The Appendix describes the measure theoretic framework used here.

3 THE CHI-SQUARE TEST UNDER THE NULL HYPOTHESIS

The first part of this section derives the asymptotic distribution of the chi-square test statistic under the null hypothesis. The second part introduces three candidates for the weight matrix used by the test statistic and shows that they satisfy the requisite consistency property. The second part also provides a simplified computational procedure for calculating the test statistic.

3.1. Asymptotic Distribution under the Null

The asymptotic null distribution of $X_n^2(\hat{\Gamma}, \hat{\theta})$ is derived in several steps. (i) First we show that the asymptotic distributions of $X_n^2(\hat{\Gamma}, \hat{\theta})$ and the approximating quadratic $q_n(\hat{\Gamma}, \hat{\theta})$ are equivalent, where

$$q_n(\hat{\Gamma}, \hat{\theta}) = \left( v_n(\hat{\Gamma}, \theta_0) - \sqrt{n} \Delta_0^r D_0^{-1} \bar{\psi}_n \right) \Sigma_0^{-1} \left( v_n(\hat{\Gamma}, \theta_0) - \sqrt{n} \Delta_0^r D_0^{-1} \bar{\psi}_n \right).$$

$\Sigma_0$ (defined below) is the asymptotic covariance matrix of $v_n(\hat{\Gamma}, \hat{\theta})$, $\bar{\psi}_n = (1/n) \sum_{i=1}^{n} \psi(Y_i, X_i, \theta_0)$, and $\Delta_0$ is defined below. (ii) Next we prove that the stochastic process $(v_n(\cdot, \theta_0), \sqrt{n} \bar{\psi}_n, \hat{\Gamma})$ indexed by partitions $\gamma \in D$ converges weakly to $(\nu(\cdot), \psi, \Gamma)$, where $\nu(\cdot)$ is a tied-down Gaussian process indexed by $\gamma \in D$. $\psi$ is a multivariate normal $J$-vector, and $\Gamma$ is the nonrandom limit of $\hat{\Gamma}$. (iii) The desired result is obtained by showing that the function $c(\cdot, \cdot, \cdot)$ that maps $(v_n(\cdot, \theta_0), \sqrt{n} \bar{\psi}_n, \hat{\Gamma})$ into $q_n(\hat{\Gamma}, \hat{\theta})$ is continuous with $\nu(\cdot)$-probability one. The continuous mapping theorem and result (ii) then imply that $q_n(\hat{\Gamma}, \hat{\theta})$ (and hence, $X_n^2(\hat{\Gamma}, \hat{\theta})$) converges weakly to $c(\cdot, \cdot, \cdot)$ evaluated at $(\nu(\cdot), \psi, \Gamma)$. The latter has the desired chi-square distribution with degrees of freedom given by the rank of $\Sigma_0$.

Since it is shown below that $v_n(\hat{\Gamma}, \hat{\theta})$ has the same asymptotic distribution as $v_n(\Gamma, \theta_0) - \sqrt{n} \Delta_0^r D_0^{-1} \bar{\psi}_n$, the asymptotic covariance matrix $\Sigma_0$ of $v_n(\hat{\Gamma}, \hat{\theta})$ is

$$\Sigma_0 = \Lambda_0 - H_0 - \Delta_0 D_0^{-1} \Pi_0 - (\Delta_0 D_0^{-1} \Pi_0)' + \Delta_0 D_0^{-1} V_0 (D_0^{-1})' \Delta_0,$$

where

$$\Lambda_0 = \mathbb{E}_\rho \Gamma(Y, X) \Gamma(Y, X)',$$

$$H_0 = \mathbb{E}_\rho F(\Gamma, X, \theta_0) F(\Gamma, X, \theta_0)',$$

$$\Delta_0 = \mathbb{E}_\rho \frac{\partial}{\partial \theta} \log f(Y|X, \theta_0) \Gamma(Y, X)' ,$$

$$\Pi_0 = \mathbb{E}_\rho \psi(Y, X, \theta_0) [\Gamma(Y, X) - F(\Gamma, X, \theta_0)]',$$

$D_0$ and $V_0$ are as in Assumption E1, and $F(\Gamma, X, \theta)$ is as defined in (2.3).
If $\hat{\theta}$ is the ML estimator (or any other asymptotically efficient estimator), then the covariance matrix $\Sigma_0$ simplifies to

$$\Sigma_0 = \Lambda_0 - H_0 - \Delta_0 V_0^{-1} \Delta_0,$$

because $\Pi_0 = \Delta_0$ and $D_0 = V_0$, the information matrix.

To begin the derivation of the asymptotic distribution of $\chi^2_n(\hat{F}, \hat{\theta})$, we state two lemmas:

**Lemma 1:** Under Assumptions M1–M2 and E1,

$$\sup_{\gamma \in D} \sqrt{n} \left| F_n(\gamma, \hat{\theta}) - F_n(\gamma, \theta_0) - \Delta_n(\gamma, \theta_0)(\hat{\theta} - \theta_0) \right| = o_p(1)$$

as $n \to \infty$, where

$$\Delta_n(\gamma, \theta) = \frac{1}{n^2} \sum_{i=1}^{n} \int_{\gamma} \frac{\partial}{\partial \theta} \log f(y|X_i, \theta) \gamma(y, X_i)' f(y|X_i, \theta) \, d\mu(y).$$

**Lemma 2:** Under Assumptions M1–M2 and RC1–RC2,

$$\Delta_n(\hat{F}, \theta_0) = \Delta_0 + o_p(1) \quad \text{as } n \to \infty.$$

See the Appendix for proofs.

These two lemmas and Assumption E1 immediately give

$$v_n(\hat{F}, \hat{\theta}) - \left[ v_n(\hat{F}, \theta_0) - \sqrt{n} \Delta_n(\hat{F}, \theta_0) \right] = o_p(1) \quad \text{as } n \to \infty.$$

Now, suppose $\hat{W}$ is an estimated weight matrix that satisfies

$$\hat{W} \overset{P}{\rightarrow} \Sigma_0^{-} \quad \text{as } n \to \infty$$

for some generalized inverse $\Sigma_0^{-}$ of $\Sigma_0$. In this case, (3.4) gives

$$\chi^2_n(\hat{F}, \hat{\theta}) = q_n(\hat{F}, \hat{\theta}) + o_p(1) \quad \text{as } n \to \infty.$$

Hence, it suffices to establish the asymptotic distribution of $v_n(\hat{F}, \theta_0) - \sqrt{n} \Delta_n(\hat{F}, \theta_0) \hat{\psi}_n$. With this in mind, we have the following lemma.

**Lemma 3:** Under Assumptions M1, E1, and RC2,

$$v_n(\cdot, \theta_0) \overset{d}{\rightarrow} v(\cdot) \quad \text{as a process on } D \quad \text{as } n \to \infty,$$

where $\overset{d}{\rightarrow}$ denotes weak convergence (or convergence in distribution) and $v(\cdot)$ is an $R^d$-valued tied-down Gaussian process with bounded uniformly continuous sample paths (almost surely) and covariance structure given by

$$E_P v(\gamma) = 0, \quad \forall \gamma \in D,$$

$$E_P v(\gamma) v(\tilde{\gamma})' = E_P F(\gamma, X, \theta_0) F(\gamma, X, \theta_0)'$$

$$\quad \forall \gamma, \tilde{\gamma} \in D,$$

where $F(\cdot, X, \theta)$ is defined in (2.3).
CHI-SQUARE DIAGNOSTIC TESTS

COMMENTS: 1. We call the limit process \( \nu(\cdot) \) an \( F \)-trampoline, where \( F \) denotes the joint distribution of \((Y, X_\cdot)\). This terminology extends that of Pollard (1984), who calls the limit process of the standard empirical process an \( F \)-bridge. Pollard's terminology is chosen to be more or less consistent with the widespread use of the term Brownian bridge, which is the limit process of the standard empirical process when \( F \) is a uniform \( (0,1) \) distribution.

The appropriateness of the term trampoline is evident in the case where \((Y, X_\cdot)\) takes values in the unit square. the class \( C \) just contains sets of the form \([0, s] \times [0, t]\) for \( 0 \leq s, t \leq 1 \) and \( D \) is taken to equal \( C \). With this choice of \( C \), the limit process can be indexed by points \((s, t)\) in the unit square rather than by sets or partitions. In this case, the limit process of the conditional empirical process is identically zero on three sides of the unit square and has continuous surface with probability one. (If \( F \) is absolutely continuous with respect to Lebesgue measure, continuity of the surface is the usual Euclidean continuity.) Hence, realizations of the conditional empirical process resemble the bouncing of a trampoline—with one side broken. In contrast, the limit of the standard empirical process has only two adjacent sides and the opposite vertex identically zero, and hence, more closely resembles a bridge than a trampoline.

2. The sample paths of \( \nu(\cdot) \) are uniformly continuous with respect to the topology of \( D \) generated by the \( L^2(F) \) semi-norm on \( C \). In the example above, where \( \nu(\cdot) \) can be indexed by points in Euclidean space, \( L^2(F) \)-continuity does not correspond to Euclidean continuity if the distribution \( F \) of \((Y, X_\cdot)\) gives probability mass to any points. For the purposes at hand, however, \( L^2(F) \)-continuity of the sample paths is the appropriate form of continuity, since \( \hat{T} \) converges in probability to \( T \) in terms of the \( L^2(F) \) semi-norm.

3. By definition, weak convergence of \( \nu_n(\cdot, \theta_0) \) to \( \nu(\cdot) \) requires convergence of the expectations of all bounded continuous functions of \( \nu_n(\cdot, \theta_0) \) to those of \( \nu(\cdot) \) as \( n \to \infty \). In the present context, continuity of such functions is defined with respect to the supremum norm on \( g(D) \), where \( g(D) \) is the set of all \( R^2 \)-valued functions defined on \( D \).

4. The proof of Lemma 3 uses recent results of Pollard (1984) that establish the weak convergence of the standard empirical process indexed by sets or functions. Also see the related work by Dudley (1978, 1984), Gaenssler (1984), Giné and Zinn (1984), and Alexander (1984). Note that the establishment in Lemma 3 of almost surely uniformly continuous sample paths of \( \nu(\cdot) \) is important, because it allows the continuous mapping theorem to be applied below. We consider a function of \( \nu(\cdot) \) below that is continuous only at realizations of \( \nu(\cdot) \) that have uniformly continuous sample paths.

Lemma 3 implies that \( \{\nu_n(\cdot, \theta_0): n = 1, 2, \ldots\} \) are uniformly tight. By the central limit theorem applied to \( \sqrt{n} \psi_n \) and the assumption \( \hat{T} \overset{d}{=} T \), we find that \( \{\sqrt{n} \psi_n, n = 1, 2, \ldots\} \) and \( \{\hat{T}, n = 1, 2, \ldots\} \) also are uniformly tight. Hence, \( \{(\nu_n(\cdot, \theta_0), \sqrt{n} \psi_n, \hat{T}): n = 1, 2, \ldots\} \) viewed as stochastic processes on \( D \) are uniformly tight. By the central limit theorem and the assumption \( \hat{T} \overset{d}{=} T \), all of the finite dimensional distributions of this process converge weakly to those of
(\psi(\cdot), \psi, \Gamma), \text{ where } \psi \sim N(0, V_0),

(3.8) \quad V_0 = E_p \psi(Y, X, \theta_0) \psi(Y, X, \theta_0)', \quad \text{and}

E_p \psi \gamma = \Pi(\gamma) = E_p \psi(Y, X, \theta_0)[\psi(Y, X, \theta_0) - F(Y, X, \theta_0)]',

That is, for all \( \gamma \in D, \)

(3.9) \quad (v_n(\gamma, \theta_0), \sqrt{n} \psi_n, \hat{\gamma}) \xrightarrow{d} (v(\gamma), \psi, \Gamma) \quad \text{as } n \to \infty.

These results imply

(3.10) \quad (v_n(\cdot, \theta_0), \sqrt{n} \psi_n, \hat{\gamma}) \xrightarrow{d} (v(\cdot), \psi, \Gamma) \quad \text{as a process on } D \text{ as } n \to \infty,

where \( v(\cdot) \) is the process defined in Lemma 3.

To make use of (3.10), the next result shows that the function that maps \((v(\cdot, \theta_0), \sqrt{n} \psi_n, \hat{\gamma})\) into \(q_n(\hat{\gamma}, \hat{\theta})\) is continuous with \(v(\cdot)\)-probability one:

**Lemma 4:** The function \(c(z, w, \gamma)\) defined by

\[
c(z, w, \gamma) = (z(\gamma) - \Delta_0 D_0^{-1} w) \Sigma_0^{-1} (z(\gamma) - \Delta_0 D_0^{-1} w)
\]

for \(z \in g(D), w \in R^l, \gamma \in D,\) and arbitrary generalized inverse \(\Sigma_0^{-1}\) of \(\Sigma_0\) is continuous (with respect to the product topology on \(g(D) \times R^l \times D\)) at all points \((z, w, \gamma)\) for which \(z\) is uniformly continuous.

Now, Lemma 4, the weak convergence of \((v_n(\cdot, \theta_0), \sqrt{n} \psi_n, \hat{\gamma})\) as a process on \(D,\) and the fact that the set of uniformly continuous sample paths of \(v(\cdot)\) is separable and occurs with probability one allow us to apply the continuous mapping theorem of Pollard (1984, Theorem IV.12) to yield

(3.11) \quad q_n(\hat{\gamma}, \hat{\theta}) \xrightarrow{d} \left( v(\Gamma) - \Delta_0 D_0^{-1} \psi \right) \Sigma_0^{-1} \left( v(\Gamma) - \Delta_0 D_0^{-1} \psi \right) \quad \text{as } n \to \infty.

Since \(\text{Var}(v(\Gamma) - \Delta_0 D_0^{-1} \psi) = \Sigma_0,\) the distribution of the limit above is chi-square with degrees of freedom equal to the rank of \(\Sigma_0\) by Theorem 9.2.2 of Rao and Mitra (1971). Hence, we have proved the following theorem:

**Theorem 1:** Suppose the null hypothesis \(P\) is true and the estimated weight matrix \(\hat{W}\) converges in probability to some generalized inverse \(\Sigma_0\) of \(\Sigma_0.\) Then, under Assumptions M1-M2, E1-E2, and RC1-RC2,

\[
X_n^2(\hat{\gamma}, \hat{\theta}) \xrightarrow{d} \chi_2^{\Sigma_0}_0 \quad \text{as } n \to \infty,
\]

where \(\chi_2^{\Sigma_0}_0\) is the chi-square distribution with rk[\(\Sigma_0\)] degrees of freedom.

Let \(G\) be the maximal number of groups of cells in \(\Gamma\) such that each covariate value \(x \in X\) belongs to cells in one and only one group. For example, if the cells in \(Y \times X\) are formed via a cross-classification of cells that partition \(Y\) and \(X,\) then \(G\) equals the number of cells in the partition of \(X.\) Alternatively, if the cells partition \(Y \times X\) based on residuals, then \(G\) equals one. The groups are defined such that the sum of \(\Gamma_j(Y, X) - F(\Gamma_j, X, \theta)\) over all the cells \(j\) in any one group is necessarily zero. Thus, if \(1_x\) denotes the \(J\)-vector with ones for the
elements corresponding to cells in the gth group and zeroes elsewhere, then \( \nu_a(T, \hat{\theta})^1 \), \( g = 1, \ldots, G \). In consequence, \( \text{rk}[\Sigma_0] \leq J - G \).

In fact, the rank of \( \Sigma_0 \) generally is \( J - G \). In some special cases, however, its rank is less than \( J - G \). The key factor is the method of estimation of \( \theta_0 \). The principal case where \( \text{rk}[\Sigma_0] < J - G \) is when \( \hat{\theta} \) is a minimum chi-square estimator or some asymptotic equivalent. By definition, a minimum chi-square estimator minimizes the chi-square statistic \( \nu_a(\hat{T}, \theta)^B(\theta) \nu_a(\hat{T}, \theta) \) formed using some \( J \times J \) weight matrix \( B(\cdot) \). The first order conditions for this minimization put \( L \) constraints on the partial derivatives of \( \nu_a(\hat{T}, \theta) \). In consequence, \( X_a^2(\hat{T}, \theta) \) behaves in large samples like a quadratic form in \( J \) normal variates that are subject to \( L \) linear constraints due to estimation and \( G \) linear constraints due to the "sum to zero" property within each of the \( G \) groups. Hence, when a minimum chi-square estimator or some asymptotic equivalent is used, the rank of \( \Sigma_0 \) is \( J - G - L \). (See Pollard (1979) for more details in the case where no covariates are present.) Two examples where estimators are used that are asymptotically equivalent to minimum chi-square estimators are: (i) Pearson’s chi-square statistic with the multinomial ML estimator and (ii) McFadden’s (1974) chi-square statistic for multinomial logit models with covariates that take on at most a finite number of different values with estimation by maximum likelihood. Except in such special cases, however, minimum chi-square estimators and their asymptotic equivalents are unnatural and inefficient, and hence, are unlikely to be used. Thus, in most cases, the rank of \( \Sigma_0 \) is \( J - G \).

3.2. The Weight Matrix

Next we consider three choices for the weight matrix \( \mathbf{W} \). Let
\[
\Sigma_{1n}(\gamma, \theta) = \Lambda_n - H_n - \Delta_n D_n^{-1} \Pi_n - (\Delta_n D_n^{-1} \Pi_n)^t
\]
\[+ D_n^{-1} V_n (D_n^{-1}) \Delta_n.\]

\[
\Lambda_n(\gamma, \theta) = \frac{1}{n} \sum_{i=1}^{n} \text{diag}(F(\gamma, X_i, \theta)).
\]

\[
H_n(\gamma, \theta) = \frac{1}{n} \sum_{i=1}^{n} F(\gamma, X_i, \theta) F(\gamma, X_i, \theta)^t.
\]

(3.12) \[ \Pi_n(\gamma, \theta) = \frac{1}{n} \sum_{i=1}^{n} F(\psi', X_i, \theta). \]

\[
\Delta_n(\gamma, \theta) = \frac{1}{n} \sum_{i=1}^{n} F\left(\frac{\partial}{\partial \theta} \log f \right)' \gamma', X_i, \theta \right).
\]

\[
D_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} F\left(\frac{\partial}{\partial \psi} \theta \right)' \gamma, X_i, \theta \right) \text{ and}
\]

\[
V_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} F(\psi', X_i, \theta),
\]

where \( F(\cdot, X_i, \theta) \) is defined in (2.3) and \( \Lambda_n \) abbreviates \( \Lambda_n(\gamma, \theta) \), etc.
Define $\Sigma_{2n}(\gamma, \theta)$ exactly as $\Sigma_{1n}(\gamma, \theta)$ is defined, but with $\Delta_{1n}(\gamma, \theta)$ and $\Pi_{1n}(\gamma, \theta)$ replaced by

$$
\Delta_{2n}(\gamma, \theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(Y_i|X_i, \theta) \gamma(Y_i, X_i)' \quad \text{and}
$$

$$
\Pi_{2n}(\gamma, \theta) = \frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, X_i, \theta) \gamma(Y_i, X_i)', \quad \text{respectively.}
$$

Next, let

$$
b_i = \gamma(Y_i, X_i) - F(\gamma, X_i, \theta) - \Delta_{2n}(\gamma, \theta)'D_{2n}^{-1}(\theta)\psi(Y_i, X_i, \theta).
$$

(3.13)

$$
\Sigma_{3n}(\gamma, \theta) = \frac{1}{n} \sum_{i=1}^{n} b_i b_i', \quad \text{and} \quad D_{2n}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \psi(Y_i, X_i, \theta).
$$

The three estimators of $\Sigma_0$ that we consider are $\hat{\Sigma}_1 = \Sigma_{1n}(\hat{\Gamma}, \hat{\theta})$, $\hat{\Sigma}_2 = \Sigma_{2n}(\hat{\Gamma}, \hat{\theta})$, and $\hat{\Sigma}_3 = \Sigma_{3n}(\hat{\Sigma}, \hat{\theta})$. Each is a sample analogue of $\Sigma_0$, but $\hat{\Sigma}_1$ takes conditional expectations using the parametric model wherever possible, whereas $\hat{\Sigma}_3$ is a more pure sample analogue estimator. $\hat{\Sigma}_2$ is a variant of $\hat{\Sigma}_1$ that is easier to calculate in certain contexts. The relative attributes of the estimators are unclear. $\hat{\Sigma}_3$ has the advantage of always being positive semi-definite. $\hat{\Sigma}_1$ is the most efficient estimator under the null (since it is the ML estimator of $\Sigma_{1n}(\Gamma, \theta_0)$). This does not imply, however, that its use will yield greater power or less discrepancy between the nominal asymptotic size and the true size of the test. For a discussion of a similar problem, see Efron and Hinkley (1978). The limited Monte Carlo results of Section 5 indicate that $\hat{\Sigma}_2$ is preferred to $\hat{\Sigma}_3$ in terms of size, but vice versa in terms of power, in certain contexts.

If the estimator $\hat{\theta}$ is an asymptotically efficient estimator, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ simplify considerably. In this case,

$$
\psi(Y_i, X_i, \theta) = \frac{\partial}{\partial \theta} \log f(Y_i|X_i, \theta),
$$

$$
\Delta_{1n}(\gamma, \theta) = \Pi_{1n}(\gamma, \theta), \quad \Delta_{2n}(\gamma, \theta) = \Pi_{2n}(\gamma, \theta),
$$

and the information matrix $I_0$ equals $D_0 = V_0$. If the Hessian estimator $D_{1n}(\hat{\theta})$ of $I_0$ is replaced by the outer product estimator $V_n(\hat{\theta})$, then we get

(3.14)

$$
\hat{\Sigma}_1 = \Lambda_n(\hat{\Gamma}, \hat{\theta}) - H_n(\hat{\Gamma}, \hat{\theta}) - \Delta_{1n}(\hat{\Gamma}, \hat{\theta})'V_n^{-1}(\hat{\theta})\Delta_{1n}(\hat{\Gamma}, \hat{\theta})
$$

and $\hat{\Sigma}_2$ is defined analogously with $\Delta_{1n}(\hat{\Gamma}, \hat{\theta})$ replaced by $\Delta_{2n}(\hat{\Gamma}, \hat{\theta})$.

To establish consistency of $\hat{\Sigma}_v$ for $v = 1, 2, 3$, two lemmas are used:

**Lemma 5:** Under Assumptions M1–M2 and E1–E2,

$$
|\Sigma_{vn}(\hat{\Gamma}, \hat{\theta}) - \Sigma_{vn}(\hat{\Gamma}, \theta_0)| = o_p(1) \quad \text{as } n \to \infty \text{ for } v = 1, 2, 3.
$$

**Lemma 6:** Under Assumptions M1, E1, and RC1–RC2,

$$
|\Sigma_{vn}(\hat{\Gamma}, \theta_0) - \Sigma_{vn}(\Gamma, \theta_0)| = o_p(1) \quad \text{as } n \to \infty \text{ for } v = 1, 2, 3.
$$
The weak law of large numbers and Slutsky's Theorem give
(3.15) \[ \Sigma_{v_n}(\hat{\Gamma}, \theta_0) \xrightarrow{p} \Sigma_0 \quad \text{as } n \to \infty \text{ for } v = 1, 2, 3. \]
This result and Lemmas 5 and 6 combine to yield
(3.16) \[ \hat{\Sigma}_v = \Sigma_{v_n}(\hat{\Gamma}, \theta) \xrightarrow{p} \Sigma_0 \quad \text{as } n \to \infty \text{ for } v = 1, 2, 3. \]
Consistency of \( \hat{\Sigma}_v \) for \( \Sigma_0 \) does not imply consistency of \( \hat{\Sigma}_v^- \) for \( \Sigma_0^- \) (where \( (\cdot)^- \) denotes the Moore-Penrose inverse), since the Moore-Penrose inverse is not a continuous function. In fact, if \( \text{rank}[\hat{\Sigma}_v] \neq \text{rank}[\Sigma_0] \) with probability bounded away from zero as \( n \to \infty \) (where \( \text{rank}[(\cdot)] \) denotes the rank of a matrix), then \( \|\hat{\Sigma}_v^-\|_2 \) is stochastically unbounded; see Andrews (1987, Theorem 2). If \( \text{rank}[\hat{\Sigma}_v] = \text{rank}[\Sigma_0] \) with probability that converges to one as \( n \to \infty \), however, then Andrews (1987, Theorem 2) gives the desired result \( \hat{\Sigma}_v^- \xrightarrow{p} \Sigma_0^- \) as \( n \to \infty \) for \( v = 1, 2, 3 \). This proves the following Theorem:

**THEOREM 2**: Under Assumptions M1–M2, E1–E2, and RC1–RC2. If \( P(\text{rank}[\hat{\Sigma}_v] = \text{rank}[\Sigma_0]) \to 1 \) as \( n \to \infty \), then
(3.17) \[ \hat{\Sigma}_v^- \xrightarrow{p} \Sigma_0^- \quad \text{as } n \to \infty \text{ for } v = 1, 2, 3. \]

**COMMENTS**: 1. Theorem 2 provides three candidates for the weight matrix \( \hat{W} \) used in the definition of \( X_n^2(\hat{\Gamma}, \hat{\theta}) \).

2. Let \( G \) be the maximum number of groups of cells in \( \hat{\Gamma} \) such that each covariate value \( x \) in \( X \) belongs to cells in one and only one group. If \( \text{rank}[\Sigma_0] = J - G \) and \( \hat{\Sigma}_v \geq G \) as \( n \to \infty \), then \( P(\text{rank}[\hat{\Sigma}_v] = \text{rank}[\Sigma_0]) \to 1 \) as \( n \to \infty \) necessarily is satisfied for \( v = 1, 2, 3 \). This follows because \( \hat{\Sigma}_v \geq G \), \( \Sigma_0 \) implies \( P(\text{rank}[\hat{\Sigma}_v] \geq \text{rank}[\Sigma_0]) \to 1 \) as \( n \to \infty \) and \( \text{rank}[\hat{\Sigma}_v] \leq J - G \) for all \( n \) since \( \hat{\Sigma}_v \) is orthogonal to \( 1_f \) for \( g = 1, \ldots, G \), where \( 1_f \) is the \( J \)-vector with ones for the elements corresponding to the \( g \)th group of cells in \( \hat{\Gamma} \) and zeroes elsewhere.

Theorems 1 and 2 above combine to show that \( X_n^2(\hat{\Gamma}, \hat{\theta}) \) has asymptotic chi-square distribution when \( \hat{W} = \hat{\Sigma}_v^- \) for \( v = 1, 2, 3 \). We now show that the choice of Moore-Penrose generalized inverse often is not necessary for this result. Suppose \( \hat{\Sigma} \) is an estimator of \( \Sigma_0 \) and \( P(v_n(\hat{\Gamma}, \hat{\theta}) \in M(\hat{\Sigma})) \xrightarrow{p} \Sigma_0^- \) 1, where \( M(\cdot) \) denotes the column space of a matrix. In this case, the quadratic form \( v_n(\hat{\Gamma}, \hat{\theta})' \hat{\Sigma}_v^- v_n(\hat{\Gamma}, \hat{\theta}) \) is numerically identical for all choices of generalized inverse \( (\cdot)^- \) with probability that goes to one as \( n \to \infty \). This follows because, with probability that goes to one as \( n \to \infty \), \( v_n(\hat{\Gamma}, \hat{\theta}) \) can be written as \( \hat{\Sigma}\xi_n(\hat{\Gamma}, \hat{\theta}) \) for some vector \( \xi_n(\hat{\Gamma}, \hat{\theta}) \), and so,

(3.17) \[ P(v_n(\hat{\Gamma}, \hat{\theta})' \hat{\Sigma}_v^- v_n(\hat{\Gamma}, \hat{\theta}) = \xi_n(\hat{\Gamma}, \hat{\theta})' \hat{\Sigma}_v^- \xi_n(\hat{\Gamma}, \hat{\theta})) \xrightarrow{p} 1. \]

The right-hand-side of the equality in (3.17) does not depend on the generalized inverse. Hence, we have the following Corollary to Theorems 1 and 2:

**COROLLARY**: Under Assumptions M1–M2, E1–E2, and RC1–RC2, if \( \hat{\Sigma} \) is a consistent estimator of \( \Sigma_0 \), \( P(\text{rank}[\hat{\Sigma}] = \text{rank}[\Sigma_0], v_n(\hat{\Gamma}, \hat{\theta}) \in M(\hat{\Sigma}), \hat{\Sigma} = \hat{\Sigma}_v^- \) \( \xrightarrow{p} 1 \).
and \( \hat{W} \) is taken to be \( \hat{\Sigma}^{-} \) for any generalized inverse \((\cdot)^{\ast}\), then
\[
X_n^{2}(\hat{\Gamma}, \hat{\theta}) \xrightarrow{d} \chi^2_{g_0} \quad \text{as} \quad n \to \infty
\]
when the null hypothesis \( P \) is true.

**Comments:** 1. If \( \text{rk}[\hat{\Sigma}] = J - G \) and \( \hat{\Sigma} \) is orthogonal to \( \hat{1}_g \) for \( g = 1, \ldots, G \) for all \( n \), then \( \nu_n(\hat{\Gamma}, \hat{\theta}) \in M(\hat{\Sigma}) \) for all \( n \), because \( \nu_n(\hat{\Gamma}, \hat{\theta}) \) is necessarily orthogonal to \( \hat{1}_g \) for \( g = 1, \ldots, G \). In this case, the test statistic is identical for all choices of \( g \)-inverse, not only with probability that goes to one as \( n \to \infty \), but for all \( n \) and all sample realizations. Hence, \( X_n^{2}(\hat{\Gamma}, \hat{\theta}) \) can be calculated using whichever generalized inverse is easiest to compute.

2. If \( \text{rk}[\Sigma_0] = J - G \), \( \hat{\Sigma} \xrightarrow{d} \Sigma \), and \( \hat{\Sigma} \) is taken to be \( \hat{\Sigma}_v \) for \( v = 1, 2, \) or \( 3 \), then the conditions of the Corollary on \( \hat{\Sigma} \) are satisfied, since \( \hat{\Sigma} \) is consistent by (3.16), \( P(\text{rk}[\hat{\Sigma}] = \text{rk}[\Sigma_0]) \xrightarrow{d} 1 \) by Comment 2 of Theorem 2, \( \hat{\Sigma} = \hat{\Sigma}_v \) for all \( n \), and \( P(\nu_n(\hat{\Gamma}, \hat{\theta}) \in M(\hat{\Sigma})) \xrightarrow{d} 1 \) by the argument of Comment 1 above.

3. The Corollary establishes the asymptotic distributions of Pearson's (1900) and McFadden's (1974) test statistics as special cases. These test statistics just correspond to \( X_n^{2}(\hat{\Gamma}, \hat{\theta}) \) with the weight matrix \( \hat{W} \) taken to be \( \Delta_n(\hat{\Gamma}, \hat{\theta})^{-1} \), which is a generalized inverse (in the situations they consider) of the consistent estimator \( \Sigma_n(\hat{\Gamma}, \hat{\theta}) \) of \( \Sigma_0 \).

4. The Corollary also establishes the asymptotic distributions of Moore and Spruill's (1975), Heckman's (1984), and Horowitz's (1985) test statistics as special cases.

When \( X_n^{2}(\hat{\Gamma}, \hat{\theta}) \) is based on \( \hat{\Sigma}_2 \) and the ML estimator (or any other asymptotically efficient estimator), \( D_n(\hat{\theta}) \) can be replaced by
\[
D_n(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(Y_i|X_i, \hat{\theta}) \frac{\partial}{\partial \theta} \log f(Y_i|X_i, \hat{\theta})'
\]
In this case, \( X_n^{2}(\hat{\Gamma}, \hat{\theta}) \) is particularly straightforward to compute. To see this, let \( A \) be the \( n \times J \) matrix with \( i \)-th row \( \hat{\Gamma}'(Y_i, X_i, \hat{\theta}) - F(\hat{\Gamma}, X_i, \hat{\theta})' \), let \( B \) be the \( n \times L \) matrix with \( i \)-th row \( \partial \log f(Y_i|X_i, \hat{\theta}) / \partial \theta \), and take \( M_B = I_n - B(B'B)^{-1}B' \). Then, \( \hat{\Sigma}_2 = (M_B A)' M_B A/n \). Since \( 1'B = 0 \), we have \( X_n^{2}(\hat{\Gamma}, \hat{\theta}) = 1'A(A'M_B A)^{\ast}A'1 = 1'M_B A(A'M_B A)^{\ast}(M_B A)'1 \). That is, \( X_n^{2}(\hat{\Gamma}, \hat{\theta}) \) is the sum of squared residuals from the projection of \( 1 \) on the space spanned by \( M_B A \).

Let \( \hat{A} \) denote \( A \) with any one column removed from each of the \( G \) groups of columns of \( A \) that correspond to the \( G \) groups of \( \hat{\Sigma} \) (defined in Comment 2 following Theorem 2). Since \( M_B A \hat{1}_g = 0 \) for \( g = 1, \ldots, G \), the space spanned by the columns of \( M_B A \) equals that spanned by those of \( M_B A \) and \( X_n^{2}(\hat{\Gamma}, \hat{\theta}) = 1'A(A'M_B A)^{\ast}A'1 \). Thus, \( X_n^{2}(\hat{\Gamma}, \hat{\theta}) \) is invariant to the dropping of any one cell from each of the \( G \) groups of columns of \( \hat{\Sigma} \).

Further, let \( ESS(\cdot) \) denote the explained sum of squares from the projection of \( 1 \) onto the space spanned by the columns of the matrix \( \cdot \). We have: \( M([A:B]) = M([A:B]) = M([M_B A:B]) \), where \( M_B A \) and \( B \) have orthogonal columns. Hence,
\[ ESS([\hat{\theta}]) = ESS(M_B A) + ESS(B) = X^2_n(\hat{\theta}, \hat{\theta}) + 0, \] since \( B'1 = 0 \). And so, letting \( H = \{ A : B \} \), we have
\[ X^2_n(\hat{\theta}, \hat{\theta}) = 1_H (H'H)^{-1} H'1. \]

That is, \( X^2_n(\hat{\theta}, \hat{\theta}) \) is the explained sum of squares from the regression of 1 on \( H \) or is \( n \) times the \( R^2 \) from this regression. (If \( H \) has less than full column rank, then the appropriate value of \( X^2_n(\hat{\theta}, \hat{\theta}) \) is obtained by deleting redundant columns of \( H \) and performing the indicated regression.)

### 4 LOCAL POWER AND ASYMPTOTIC OPTIMALITY

In this section we investigate the local power, consistency, and asymptotic optimality properties of chi-square tests.

#### 4.1. Local Power

The local alternatives considered here apply to any alternative distribution of \( Y \), given \( X \). Suppose one is interested in an approximation to the power of the chi-square test for sample size \( n_0 \) and against an alternative conditional distribution of \( Y \), given \( X \) defined by the density \( q(y|x) \) with respect to some \( \sigma \)-finite measure \( \mu \). (Without loss of generality, assume \( \bar{\mu} \) dominates \( \mu \).) Since \( q(\cdot \mid \cdot) \) is an alternative density, \( q(\cdot \mid \cdot) \in \{ f(\cdot \mid \cdot, \theta) : \theta \in \Theta \} \). The marginal distribution of \( X \) under the alternative of interest is arbitrary, just as under the null, so we adopt the same notion for it, viz., \( P_X \). Let \( Q \) denote the distribution of \( \{(Y_i, X_i) : i = 1, 2, \ldots \} \) when \( (Y_i, X_i) \) are iid with conditional density \( q(y|x) \) of \( Y \), given \( X \), and marginal distribution \( P_X \) of \( X \). Let
\[ d(y, x, \theta_0) = \sqrt{n_0} (q(y|x) - f(y|x, \theta_0)). \]

Define the following sequence of local alternative conditional densities:
\[ q_n(y|x) = f(y|x, \theta_0) + d(y, x, \theta_0) / \sqrt{n} \quad \text{for} \quad n = 1, 2, \ldots. \]

Note that \( q_n(y|x) \) is a proper density for all \( n \) greater than or equal to \( n_0 \). Let \( Q_n \) denote the distribution of \( \{(Y_i, X_i) : i = 1, 2, \ldots \} \) when \( (Y_i, X_i) \) are iid with conditional density \( q_n(y|x) \) of \( Y \), given \( X \), and marginal distribution \( P_X \) of \( X \).

The sequence of local alternative distributions we consider is \( \{Q_n : n = 1, 2, \ldots \} \). These distributions approach \( P \) as \( n \to \infty \) and the \( n_0 \)th term of the sequence is \( Q \), the alternative of interest. Although any alternative \( Q \) can be considered, the asymptotic power approximations are local in nature, and hence, their accuracy is best when \( Q \) is “close” to \( P \) and \( n_0 \) is large.

The assumptions used for the local power results are analogous to those used in Section 3 for the null:

**Assumption M1':** \( \{(Y_i, X_i) : i = 1, 2, \ldots \} \) are distributed under the sequence of local alternatives \( \{Q_n : n = 1, 2, \ldots \} \) as iid rv's with conditional density \( q_n(y|x) \) (with respect to the \( \sigma \)-finite measure \( \bar{\mu} \)) of \( Y \), given \( X \), and marginal distribution \( P_X \) of \( X \).
ASSUMPTION M2': The parametric conditional densities \( f(y|x, \theta) \) satisfy M2 and \( E_Q[r^2(Y, X) + r(Y, X)] < \infty \).

ASSUMPTION E1': The estimator \( \hat{\theta} \) satisfies E1 with equation (2.6) holding under the sequence of local alternatives \( \{Q_n; n = 1, 2, \ldots \} \).

ASSUMPTION E2': The defining function \( \psi(y, x, \theta) \) of \( \hat{\theta} \) satisfies E2 and
\[
E_Q[r_0(Y, X)[r_0(Y, X) + r_1(Y, X) + r_2(Y, X)] < \infty.
\]

ASSUMPTION RC1': The random partitions \( \hat{\Gamma} \) satisfy \( \hat{\Gamma} \xrightarrow{Q_n} \Gamma \) as \( n \to \infty \).

Note that "\( \xrightarrow{Q_n} \)" denotes convergence in probability under \( \{Q_n\} \).

Under the moment conditions of E2', Assumption E1' holds for most estimators that satisfy E1. For example, it holds for ML estimators. As an alternative to E1, one could adopt an assumption such as A1 of Durbin (1973, p. 281). Close inspection of E1 and A1, however, shows that they are analogous, so we adopt the one that is more suited to the present development. Assumption RC1' necessarily holds if \( \{Q_n\} \) are contiguous to \( P \), as is usually the case (see LeCam (1960) or Hajek and Sidák (1967)).

The local power results are given in the following Theorem:

**Theorem 3:** Suppose Assumptions M1'-M2', E1'-E2', and RC1'-RC2 hold.

(a) If \( \hat{W} \xrightarrow{Q_n} \Sigma_0^+ \) as \( n \to \infty \), then
\[
X_n^2(\hat{\Gamma}, \hat{\theta}) \xrightarrow{d} \chi^2_{2\delta}(\delta) \quad \text{as } n \to \infty
\]
under the sequence of local alternatives \( \{Q_n; n = 1, 2, \ldots \} \), where
\[
\delta = n_0 \left( E_Q \Gamma(Y, X) - E_P \Gamma(Y, X) - \Delta_0 D_0^{-1} E_Q \psi(Y, X, \theta_0) \right)'
\cdot \Sigma_0^+ \left( E_Q \Gamma(Y, X) - E_P \Gamma(Y, X) - \Delta_0 D_0^{-1} E_Q \psi(Y, X, \theta_0) \right)
\]
and \( \chi^2_{2\delta}(\delta) \) denotes the noncentral chi-square distribution with \( \text{rk}[\Sigma_0] \) degrees of freedom and noncentrality parameter \( \delta \).

(b) \( \Sigma_{en}(\hat{\Gamma}, \hat{\theta}) \xrightarrow{Q_n} \Sigma_0 \) as \( n \to \infty \) for \( v = 1, 2, 3 \).

(c) If \( Q_n(\text{rk}[\Sigma_{en}(\hat{\Gamma}, \hat{\theta})] = \text{rk}[\Sigma_0]) \to 1 \) as \( n \to \infty \), then \( \Sigma_{en}(\hat{\Gamma}, \hat{\theta})^+ \xrightarrow{Q_n} \Sigma_0^+ \) as \( n \to \infty \) for \( v = 1, 2, 3 \).

(d) Given any estimator \( \hat{\Sigma} \) that satisfies \( \hat{\Sigma} \xrightarrow{Q_n} \Sigma_0 \) as \( n \to \infty \) and \( Q_n(\text{rk}[\hat{\Sigma}] = \text{rk}[\Sigma_0], \nu_{\hat{\Sigma}}(\hat{\Gamma}, \hat{\theta}) \in M(\hat{\Sigma}), \hat{\Sigma} \in \hat{\Sigma}^+ \) \to 1 as \( n \to \infty \), if \( \hat{W} = \hat{\Sigma}^{-} \) for some generalized inverse \( (\cdot)^- \), then
\[
X_n^2(\hat{\Gamma}, \hat{\theta}) \xrightarrow{d} \chi^2_{2\delta}(\delta) \quad \text{as } n \to \infty
\]
under the sequence of local alternatives \( \{Q_n; n = 1, 2, \ldots \} \).
COMMENT: It is not possible to give general results stating that the use of the minimum chi-square estimator or the ML estimator in forming the chi-square statistic dominates the other in terms of local power. In special cases (where no covariates are present), it has been shown that their local power functions criss-cross. See Chibisov (1970) and Moore and Spruill (1975).

4.2. Consistency

We now state the consistency properties of chi-square tests. Let \( Q = Q(y|x) \) be a conditional distribution of \( Y \) given \( X \) that is not in the parametric family \( \{ f(y|x, \theta): \theta \in \Theta \} \). Let \( \Gamma_i \) be the limit partition of \( \hat{\Gamma} \) and \( \theta_i \) the limit vector of \( \hat{\theta} \) under \( Q \) and \( P_X \) as \( n \to \infty \). If

\[
\int_X \int_{\Gamma_{ij}} f(y, x)f(y|x, \theta_1) \mu(y) \, dP_X(x)
\]

for some \( j = 1, \ldots, J \), then the chi-square test is consistent against \( Q \). (See Andrews (1985) for regularity conditions under which this result holds.) That is, the chi-square test is consistent against any alternative that renders the predictions of the parametric model for the cells chosen to be inaccurate in large samples.

4.3. Asymptotic Optimality

Next we discuss the power properties of chi-square tests relative to those of Wald (W), likelihood ratio (LR), and Lagrange multiplier (LM) tests. Chi-square tests are designed to test against nonparametric families of alternative distributions. For example, all conditional distributions not in \( \{ f(y|x, \theta): \theta \in \Theta \} \) are of interest when testing for general goodness-of-fit. In contrast, W, LR, and LM tests are designed to have high power within some finite dimensional parametric model that includes the null hypothesis. As is well known, these tests possess certain asymptotic optimality properties with respect to power over these finite dimensional classes of alternatives (see Wald (1943)).

Test procedures that possess asymptotic optimality properties for power against nonparametric classes of alternatives do not exist (except in special cases). Thus, one has the option of choosing a test that has some optimality properties against a restricted subclass of alternatives, such as a W, LR, or LM test, or of choosing a test that promises to have good power against a wider variety of alternatives, though not necessarily optimal power over any parametric subclass, such as

---

2 Even within this subclass of alternatives, W, LR, and LM tests are not unambiguously asymptotically optimal. This is evident from the fact that the tests based on two nested parametric families (both of which contain the same parametric null distributions) both possess certain optimality properties even though they are not asymptotically equivalent.
chi-square test. The goal of this paper is to make tests of the latter sort available for testing a broad class of parametric distributions against a flexible array of nonparametric alternatives, with emphasis on detecting predictive inaccuracies of specified models.

It remains to explain the asserted promise of good power for chi-square tests against a wide variety of alternatives. First, if one's loss function is related to predictive accuracy over certain regions in $Y \times X$, then a chi-square test can be constructed that is consistent against all alternatives that cause the model to yield inaccurate predictions for such regions. Furthermore, the test has higher power, the greater is the expected inaccuracy.

Second, in certain cases the chi-square test has the same asymptotic properties in models with covariates as in analogous models without covariates. The usefulness of chi-square tests for testing against wide varieties of alternatives in the latter models has been demonstrated in the statistical literature. In particular, this equality holds for tests of univariate and multivariate normality (or any other specified distribution) of the errors in single equation, seemingly unrelated, and multivariate regression linear models with intercept terms, provided the cells $\hat{X}$ are determined by the residuals alone and $\theta$ is estimated by least squares, maximum likelihood, pseudo-maximum likelihood, Zellner’s feasible Aitken estimator, a classical M-estimator, or any estimator whose influence function $\psi(y, x, \theta)$ is of the form

$$
\begin{bmatrix}
\psi_1(y - x'\theta_1, \theta_2)(1)
\psi_2(y - x'\theta_1, \theta_2)
\end{bmatrix},
$$

where $\theta = (\theta_1', \theta_2')'$. (See the Appendix for a proof.)

Third, in certain cases the chi-square test can be shown to possess optimality properties against wide varieties of alternatives. In fact, in one such case it is optimal against nonparametric families of alternatives. These cases are suggestive of its power in more general contexts, even though analogous optimality results may not be obtainable.

The first case consists of discrete response models in which the covariates take on a finite number of values and the cells completely cross-classify the response and covariate values. For example, McFadden’s (1974) chi-square statistic is designed for such a model, where the particular form of the model is multinomial logit. The totality of alternative distributions in this context consists of multinomial conditional distributions of $Y$ given $X$—a class that has finite, but potentially large, dimension. It can be shown that the chi-square test formed using the weight matrix $\hat{\Sigma}_x$ is precisely the LM test of the parametric null hypothesis against the class of all alternatives, i.e., all multinomial conditional distributions. Thus, the chi-square test possesses the standard asymptotic optimality properties of having asymptotically best weighted average power and best constant power over certain ellipses and of being asymptotically most stringent with respect to the class of all alternative distributions (see Wald (1943)).
The second case consists of models that involve a single distribution, i.e.,
\( f(y|x) = f_0(y|x) \), where the covariates take on a finite number \( G \) of values and
no one cell \( \Gamma_i \) contains more than one value of the covariates. If we are interested
in the predictive accuracy of the model for the cells \( \Gamma_i \), a more inclusive null
hypothesis than \( H_0: f(y|x) = f_0(y|x) \) is appropriate. Specifically, consider
\( H_0^* : f = f_0 \) where, given any conditional distribution \( h(y|x) \), \( h \) denotes the
\((J - G)\)-dimensional vector of conditional probabilities under \( h(y|x) \) of the cells
in \( \Gamma \) (with \( G \) redundant cells, due to conditional probabilities summing to one,
omitted). In this context, the chi-square test based on \( X^2_n(F) \) is asymptotically
equivalent to certain Wald tests under the null and local alternatives and
possesses various nonparametric asymptotic optimality properties. See the end of
the Appendix for details.

5 MONTE CARLO RESULTS

Here we describe a Monte Carlo experiment in which the chi-square test of
Section 2.1 is compared with several information matrix and Hausman specification
tests. The parametric model is a censored regression model:

\[
Y_i = \begin{cases} 
  c + X_i'\beta + U_i & \text{if } c + X_i'\beta + U_i > 0 \\
  0 & \text{otherwise}
\end{cases} \quad (i = 1, \ldots, n).
\]

where \( Y_i, U_i, c \in \mathbb{R}^1, X_i, \beta \in \mathbb{R}^4, \{ (Y_i, X_i): i = 1, \ldots, n \} \) is an independent
sequence, and the conditional distribution of \( U_i \) given \( X_i \) is \( N(0, \sigma^2) \).

Results under the null hypothesis are obtained in three cases that yield
censoring probabilities of 1/2, 2/3, and 1/3. In each case, the data are generated
with \( \beta = (1, 1, 1, 1)' \), \( X_i \) and \( U_i \) independent, \( U_i \sim N(0, 1) \), and \( X_i \sim N(\xi, \Sigma_X) \).
where \( \xi = (-.5, .25, 0, .25)' \) and \( \Sigma_X \) is a diagonal band (Toeplitz) matrix with first
row \((1, .5, .2, .1)\). To obtain the desired censoring probabilities the intercept \( c \) is
set equal to 0.0, -.4307, and .4307, respectively. The sample size \( n = 250 \). Some
additional results for \( n = 100 \) also are given.\(^3\) The number of repetitions is 5,000.

Four different alternative distributions are used. Since the test statistics under
study are general specification tests, the alternatives are chosen to yield quite
different directions of departure from the null hypothesis. The first three
alternatives are exactly the same as the null but with error distributions that are thick
tailed \( (t_2) \), thin tailed (uniform \([ - \sqrt{3}, \sqrt{3}]\)), and asymmetric (lognormal stan-
dardized to have mean zero and variance one).

The fourth alternative corresponds to a model for censored data suggested by
Cragg (1971, eqns. (7) and (9), p. 831). In this model, the value of \( Y_i \) when \( Y_i \) is
uncensored is determined separately from the censor/uncensor decision. In
particular, whether \( Y_i \) is censored is determined by the same model as under the

\(^3\) Results for larger sample sizes, to reflect the large samples available in some cross-sectional
economic applications, would be desirable but it is too expensive and time consuming to obtain the
requisite large number of repetitions. Note that the increased expense with larger samples is due more
to increased costs of estimation than to increased costs of calculating the test statistics under study.
null. Then, the value of $Y_i$ for uncensored observations is determined by the same model as under the null but with independently drawn $(X_i, U_i)$ and with the distribution of $U_i$ being a $N(0, 1)$ distribution conditioned on $U_i + X_i'\beta = Y_i > 0$ (i.e., a truncated normal distribution). The regressors used to estimate $(c, \beta, \sigma^2)$ in this model are the second set of regressors that determine the value of $Y_i$ when $Y_i$ is uncensored.

In order to assess power separately from discrepancies between nominal and true size, all power calculations are made using exact critical values obtained from the Monte Carlo results under the null (using the same censoring probabilities). The same sample sizes are considered as under the null. One thousand repetitions are used for each of the alternatives.\footnote{More repetitions are used under the null than under the alternatives, because the null results are used to obtain exact critical values for the alternatives.}

Three information matrix tests suggested by White (1982) are considered. The first, denoted IM20, is based on twenty of the twenty-one nonredundant elements of the information matrix with the main diagonal element corresponding to the intercept parameter omitted to eliminate a singularity. The second, IM5, is based on the five main diagonal elements excluding the intercept parameter and the third, IM4, on the four main diagonal elements excluding the intercept and the variance parameters. The exact form of the test is that given by Lancaster (1984).

Two Hausman specification tests suggested by Ruud (1984) are considered. Both tests are based on the vector of differences $(\hat{\eta}, \hat{\beta})/\hat{\sigma} - (\hat{\eta}, \hat{\beta})$, where $(\hat{\eta}, \hat{\beta}/\hat{\sigma}^2)$ are the ML estimators from the censored regression model and $(\hat{\eta}, \hat{\beta})$ are the ML estimators from the probit model obtained by letting $Y_i = 1$ if $X_i'\hat{\beta} + U_i > 0$. (In the latter model $\sigma^2$ is normalized to equal one.) The first test, denoted HR1, has weight matrix given by the inverse of the difference between the inverse of the true conditional information matrices of the probit and censored regression models (i.e., expectations are taken with respect to $Y_i$ but not $X_i$) both evaluated at $(\hat{\eta}, \hat{\beta}/\hat{\sigma})$. The second test, denoted HR2, uses the same weight matrix except the two information matrices are evaluated at $(\hat{\eta}, \hat{\beta}/\hat{\sigma})$ and $(\hat{\eta}, \hat{\beta})/\hat{\sigma}$, respectively.

Two chi-square tests are considered. The first, denoted $X^2 - 1$, uses the weight matrix based on $\hat{\Sigma}_2$ and the second, denoted $X^2 - 2$, uses the weight matrix based on the more pure sample analogue estimator $\hat{\Sigma}_3$.\footnote{$\hat{\Sigma}_3$ is used rather than $\hat{\Sigma}_1$ is used for the first test because it is more convenient computationally in this model.} Both tests use the same random cells. Seven uncensored cells and three censored cells are used. The uncensored cells are based on the scaled residuals $(Y_i - X_i'\beta)/\hat{\sigma}$. Since large negative errors are censored, we form uncensored cells that are not symmetric about the estimated regression line. In a graph with $Y_i/\hat{\sigma}$ and $X_i'\beta/\hat{\sigma}$ on the vertical and horizontal axes, respectively, the uncensored cells consist of regions where $Y_i > 0$ that are parallel to the $Y_i/\hat{\sigma} = X_i'\beta/\hat{\sigma}$ line. From left to right, seven regions are formed whose (uncensored) probabilities of occurrence given $X_i$ are .05, .06, .13, .13, .13, .13, and .37 under a standard normal error distribution. The
### TABLE I

**Exact Rejection Probabilities under the Null Hypothesis for Tests with Asymptotic Significance Levels of 1%, 5%, and 10%**

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>$P(\text{cen}) = 1/2^a$</th>
<th>$n = 250$</th>
<th>$P(\text{cen}) = 2/3$</th>
<th>$n = 250$</th>
<th>$P(\text{cen}) = 1/3$</th>
<th>$n = 250$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
<td>5%</td>
<td>10%</td>
<td></td>
<td>1%</td>
<td>5%</td>
</tr>
<tr>
<td>IM20</td>
<td>67.3</td>
<td>82.2</td>
<td>87.9</td>
<td>79.4</td>
<td>89.8</td>
<td>93.6</td>
</tr>
<tr>
<td>IM5</td>
<td>16.9</td>
<td>30.6</td>
<td>39.6</td>
<td>19.6</td>
<td>33.3</td>
<td>43.5</td>
</tr>
<tr>
<td>IM4</td>
<td>5.3</td>
<td>15.5</td>
<td>24.0</td>
<td>6.2</td>
<td>16.8</td>
<td>26.6</td>
</tr>
<tr>
<td>HR1</td>
<td>1.4</td>
<td>4.3</td>
<td>8.3</td>
<td>1.3</td>
<td>4.9</td>
<td>8.2</td>
</tr>
<tr>
<td>HR2</td>
<td>3.8</td>
<td>7.1</td>
<td>10.6</td>
<td>4.0</td>
<td>6.5</td>
<td>8.9</td>
</tr>
<tr>
<td>$X^2 - 1$</td>
<td>0.9</td>
<td>3.2</td>
<td>5.9</td>
<td>1.6</td>
<td>3.8</td>
<td>6.6</td>
</tr>
<tr>
<td>$X^2 - 2$</td>
<td>3.2</td>
<td>10.8</td>
<td>17.6</td>
<td>5.0</td>
<td>13.7</td>
<td>21.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>$P(\text{cen}) = 1/2$</th>
<th>$n = 100$</th>
<th>$P(\text{cen}) = 1/2$</th>
<th>$n = 100$</th>
<th>$P(\text{cen}) = 1/3$</th>
<th>$n = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
<td>5%</td>
<td>10%</td>
<td></td>
<td>1%</td>
<td>5%</td>
</tr>
<tr>
<td>IM20</td>
<td>81.8</td>
<td>94.2</td>
<td>97.1</td>
<td></td>
<td>81.8</td>
<td>94.2</td>
</tr>
<tr>
<td>IM5</td>
<td>31.2</td>
<td>47.3</td>
<td>58.0</td>
<td></td>
<td>31.2</td>
<td>47.3</td>
</tr>
<tr>
<td>IM4</td>
<td>8.2</td>
<td>22.5</td>
<td>34.9</td>
<td></td>
<td>8.2</td>
<td>22.5</td>
</tr>
<tr>
<td>HR1</td>
<td>3.4</td>
<td>7.6</td>
<td>12.2</td>
<td></td>
<td>3.4</td>
<td>7.6</td>
</tr>
<tr>
<td>HR2</td>
<td>3.9</td>
<td>6.6</td>
<td>9.3</td>
<td></td>
<td>3.9</td>
<td>6.6</td>
</tr>
<tr>
<td>$X^2 - 1$</td>
<td>3.2</td>
<td>6.6</td>
<td>9.8</td>
<td></td>
<td>3.2</td>
<td>6.6</td>
</tr>
<tr>
<td>$X^2 - 2$</td>
<td>11.2</td>
<td>21.3</td>
<td>30.4</td>
<td></td>
<td>11.2</td>
<td>21.3</td>
</tr>
</tbody>
</table>

* $P(\text{cen})$ denotes the probability of censoring in the model under consideration.

Censored cells are divided into three groups based on the magnitude of $X^2/\hat{\beta}$. Since the conditional probability of censoring is near zero for large values of $X^2/\hat{\beta}$, we divide the censored observations into three unequal cells. The cell corresponding to the largest $X^2/\hat{\beta}$ values contains half of the censored observations. The second censored cell contains the quarter of the censored observations with the smallest $X^2/\hat{\beta}$ values and the third censored cell contains the rest.

The Monte Carlo results are given in Tables I and II. Table I shows the discrepancies between the exact and nominal sizes of the test statistics under study. Table II provides power comparisons between the statistics. The results can be summarized as follows: The IM tests exhibit huge differences between exact and nominal rejection probabilities under the null. This is especially true of the first two IM tests. For these two tests the asymptotic critical values are of no use for the sample sizes considered. The HR1 and $X^2 - 1$ statistics exhibit the smallest differences between exact and nominal null rejection probabilities with HR1 being superior. The $X^2 - 1$ statistic has the desirable feature that its direction of error tends to be that of rejecting too seldom rather than too often when the asymptotic critical value is used. On the other hand, $X^2 - 2$ and HR2 consistently reject too often when the asymptotic critical value is used.

With regard to power, we discuss first the results for the sample size two hundred and fifty. The $X^2 - 2$ test is clearly the most powerful test considered. It is nearly uniformly most powerful over the disparate alternatives considered. In addition, its power is very good in an absolute sense in all cases. Following $X^2 - 2$, the $X^2 - 1$, IM20, and IM5 tests are the next best tests in terms of all around power. Each has relatively good power against all alternatives. The
### TABLE II

**Exact Rejection Probabilities under Alternatives for Tests with Exact Null Rejection Probabilities of 1%, 5%, and 10%**

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>Uniform</th>
<th>Logistic</th>
<th>Cragg Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha = 1/2$</td>
<td>$\alpha = 1/10$</td>
<td>$\alpha = 1/2$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>$\alpha = 1/2$</td>
<td>$\alpha = 1/10$</td>
<td>$\alpha = 1/2$</td>
</tr>
<tr>
<td>$\pi$</td>
<td>$\alpha = 1/2$</td>
<td>$\alpha = 1/10$</td>
<td>$\alpha = 1/2$</td>
</tr>
<tr>
<td>IM20</td>
<td>37.8</td>
<td>55.0</td>
<td>65.5</td>
</tr>
<tr>
<td>IM3</td>
<td>41.0</td>
<td>65.9</td>
<td>83.2</td>
</tr>
<tr>
<td>IM4</td>
<td>48.1</td>
<td>62.0</td>
<td>69.2</td>
</tr>
<tr>
<td>HR1</td>
<td>75.7</td>
<td>89.4</td>
<td>92.5</td>
</tr>
<tr>
<td>HR2</td>
<td>0.6</td>
<td>2.3</td>
<td>2.5</td>
</tr>
<tr>
<td>$X^2 - 1$</td>
<td>81.7</td>
<td>92.2</td>
<td>96.0</td>
</tr>
<tr>
<td>$X^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_2$</td>
<td>$\alpha = 2/3$</td>
<td>$\alpha = 1/10$</td>
<td>$\alpha = 2/3$</td>
</tr>
<tr>
<td>$\pi$</td>
<td>$\alpha = 2/3$</td>
<td>$\alpha = 1/10$</td>
<td>$\alpha = 2/3$</td>
</tr>
<tr>
<td>IM20</td>
<td>39.5</td>
<td>65.8</td>
<td>78.9</td>
</tr>
<tr>
<td>IM3</td>
<td>32.4</td>
<td>69.8</td>
<td>85.5</td>
</tr>
<tr>
<td>IM4</td>
<td>64.8</td>
<td>87.8</td>
<td>94.2</td>
</tr>
<tr>
<td>HR1</td>
<td>97.4</td>
<td>99.9</td>
<td>100.0</td>
</tr>
<tr>
<td>HR2</td>
<td>33.1</td>
<td>48.5</td>
<td>48.7</td>
</tr>
<tr>
<td>$X^2 - 1$</td>
<td>32.0</td>
<td>61.3</td>
<td>76.0</td>
</tr>
<tr>
<td>$X^2$</td>
<td>60.4</td>
<td>81.3</td>
<td>87.2</td>
</tr>
<tr>
<td>$f_2$</td>
<td>$\alpha = 5$</td>
<td>$\alpha = 1/10$</td>
<td>$\alpha = 5$</td>
</tr>
<tr>
<td>$\pi$</td>
<td>$\alpha = 5$</td>
<td>$\alpha = 1/10$</td>
<td>$\alpha = 5$</td>
</tr>
<tr>
<td>IM20</td>
<td>14.0</td>
<td>38.7</td>
<td>60.6</td>
</tr>
<tr>
<td>IM3</td>
<td>7.2</td>
<td>79.2</td>
<td>92.5</td>
</tr>
<tr>
<td>IM4</td>
<td>73.4</td>
<td>87.3</td>
<td>91.5</td>
</tr>
<tr>
<td>HR1</td>
<td>95.1</td>
<td>98.4</td>
<td>99.3</td>
</tr>
<tr>
<td>HR2</td>
<td>17.8</td>
<td>48.6</td>
<td>56.4</td>
</tr>
<tr>
<td>$X^2 - 1$</td>
<td>97.0</td>
<td>99.3</td>
<td>99.4</td>
</tr>
<tr>
<td>$X^2$</td>
<td>96.0</td>
<td>99.7</td>
<td>99.8</td>
</tr>
</tbody>
</table>

* $\pi$ denotes the probability of censoring under the null distribution that corresponds to the alternative distribution under consideration.
$X^2 - 1$ test does better against the $t$, and lognormal alternatives, but worse against the uniform alternative in comparison with the IM20 and IM5 tests. Each of the other tests exhibits very low power against one or more alternatives for each of the three levels of censoring. The IM4 and HR1 tests have trivial power against the uniform alternative for all levels of censoring. The HR2 test has trivial power against the $t_2$ and uniform alternatives for all levels of censoring. On the other hand, it is worth noting that the HR1 test has quite high power against each of the three alternatives other than the uniform.

For the sample size one hundred, all of the tests are much less powerful than with the larger sample size. The $X^2 - 2$ test is still the best all around test in terms of power, but it is not uniformly most powerful. Each of the other tests exhibits poor power against one or more alternatives for all three levels of censoring. (We mention that the results of Table II were not chosen from a larger set of results with the aim of showing chi-square tests in a good light.)

In conclusion, the Monte Carlo results of this section indicate that chi-square tests exhibit good power properties in censored regression models both absolutely and relative to other specification tests provided the sample size is not too small.

_Cowles Foundation, Department of Economics, Yale University, New Haven, CT, U.S.A._

*Manuscript received September, 1985, final revision received April, 1988.*

**APPENDIX**

We begin the Appendix by describing the measure theoretic framework used for the results of this paper. Let $\mathcal{F}$ denote the Borel $\sigma$-algebra on $Y \times X \subset \mathbb{R}^{k+1}$. The underlying probability space of the whole random experiment, whose outcome yields $\{(Y_i, X_i); i = 1, 2, \ldots\}$, is denoted $(\Omega, \mathcal{G}, \mathcal{P})$, where

$\Omega = \{Y \times X\}^\infty$ (that is, $\Omega$ is the product space of an infinite number of copies of $Y \times X$). $\mathcal{Q}$ is the $\sigma$-algebra on $\Omega$ generated by the infinite sequence of $\sigma$-algebras $\mathcal{F}_i$. and $\mathcal{P}$ is the probability distribution given in Assumption M1.

The random element $F$ (which depends on $n$ in general) is taken to be a map from $\Omega$ to $D$ that is measurable with respect to the $\sigma$-algebra in $\Omega$ that is generated by the first $n$ observations and the Borel $\sigma$-algebra in $D$.

Let $g(D)$ be the set of all $R^l$-valued functions defined on $D$. We consider the supremum norm on $g(D)$. The conditional empirical process is not necessarily measurable with respect to the Borel $\sigma$-field generated by the supremum norm. In the case of the standard empirical process, this had led authors, e.g., Billingsley (1968), to replace the supremum norm by some other norm or metric, such as the Skorokhod metric. Instead, we follow Dudley (1978) and Pollard (1984) and adopt the more natural supremum norm, but consider a smaller $\sigma$-field than the Borel $\sigma$-field. We choose the $\sigma$-field on $g(D)$ generated by the coordinate projection maps of the conditional empirical process. The conditional empirical process is necessarily measurable with respect to this $\sigma$-field. Finally, as Dudley (1978) and Pollard (1984) show, the basic asymptotic results we desire hold with this choice of $\sigma$-field.

Finally, to ensure that various functions of the conditional empirical process are measurable (in particular, those functions used in the proofs of results given below), it is necessary to place additional technical conditions on $C$. These conditions are given in Appendix C of Pollard (1984) and are not reproduced here because of their strictly technical nature. For precision of the results stated in Sections 3 and 4, assume that the definition of VC classes incorporates these measurability conditions.

**PROOF OF LEMMA 1:** Let $S_n(\theta) = \sup_{\theta \in \Theta} \{F_n(\gamma, \theta) - F_n(\gamma, \theta_0) - \Delta_n(\gamma, \theta_0)\}$. We prove the result for an arbitrary element $j$ of the vector $\sqrt{n}S_n(\theta)$. To simplify notation the subscript $j$ is
suppose we can show

\[(A.1) \quad P\left( \lim_{n \to \infty} \sup_{\theta \in N(\delta_{\theta_0})} \sqrt{n} S_n(\theta) = 0 \right) = 1.\]

where \(N(\delta_{\theta_0}) = \{ \theta : \| \theta - \theta_0 \| < \delta_{\theta_0} \}\) and \(\delta_{\theta} = n^{-a}\) for some \(a \in (1/4, 1/2)\). Let \(A_{1n} = \{ \sup_{\theta \in N(\delta_{\theta_0})} \sqrt{n} S_n(\theta) < \epsilon \}\) and \(A_{2n} = \{ \delta \in N(\delta_{\theta_0}) \}\). By (A.1), \(\lim_{n \to \infty} P(A_{1n}) = 1\). Assumptions E1 and M1 and the central limit theorem imply that \(n^{a}(\theta - \theta_0) = o_p(1)\) as \(n \to \infty\). That is, \(\lim_{n \to \infty} P(A_{2n}) = 1\). Hence, \(\forall \epsilon > 0\)

\[(A.2) \quad \lim_{n \to \infty} P(A_{1n}) < \lim_{n \to \infty} P(\sqrt{n} S_n(\theta) < \epsilon)\]

and the Lemma is proved.

To show (A.1), note that for all \(\theta\) in some neighborhood \(N_{\theta_0}\) of \(\theta_0\),

\[(A.3) \quad \frac{\partial^2}{\partial \theta_k \partial \theta_m} f(y|x, \theta) = \frac{\partial}{\partial \theta_k} s(y|x, \theta) + s(y|x, \theta) \sum_{i \neq k} \frac{\partial}{\partial \theta_i} s(y|x, \theta) \cdot f(y|x, \theta) \quad \leq \left[ f(y|x) + r^2(y|x) \right] f(y|x, \theta).\]

where

\[
\int \sup_{\theta \in N_{\theta_0}} \left[ f(y|x) + r^2(y|x) \right] f(y|x) \, d\mu(y) \, dP_X(x)
\]

by M2

This gives,

\[
\sup_{\theta \in N(\delta_{\theta_0})} \sqrt{n} S_n(\theta)
\]

\[
= \sup_{\theta \in N(\delta_{\theta_0})} \sup_{x \in C} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int f(y|x_i) \cdot \left[ f(y|x_i, \theta) - f(y|x_i, \theta_0) - \frac{\partial}{\partial \theta} f(y|x_i, \theta_0)(\theta - \theta_0) \right] \, d\mu(y) \right]
\]

\[
\leq \sup_{\theta \in N(\delta_{\theta_0})} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int \left| \frac{1}{2} (\theta - \theta_0) \frac{\partial^2}{\partial \theta \partial \theta} f(y|x_i, \theta)(\theta - \theta_0) \right| \, d\mu(y)
\]

\[
\leq \sup_{\theta \in N(\delta_{\theta_0})} \frac{1}{2\sqrt{n}} \sum_{i=1}^{n} \int \left| (\theta - \theta_0) \frac{\partial^2}{\partial \theta \partial \theta} f(y|x_i, \theta)(\theta - \theta_0) \right| \, d\mu(y)
\]

\[
\leq \frac{L^2}{2\sqrt{n}} \sum_{i=1}^{n} \sup_{\theta \in N_{\theta_0}} \left[ \int f(y|x_i, \theta) + r^2(y|x_i) \right] \, d\mu(y)
\]

\[
\lim_{n \to \infty} = 0.
\]

where \(\bar{\delta} \in N(\delta_{\theta_0})\), the third inequality holds for \(n\) sufficiently large and uses the definition of \(N(\delta_{\theta_0})\) and equation (A.3), and the almost sure convergence follows using Assumptions M1–M2 by the strong law of large numbers and by the choice of "\(a\)" such that \(1/2 + 2a > 1\). Q.E.D.
PROOF OF LEMMA 2: The weak law of large numbers implies \( \Delta_n(\bar{S}, \theta_0) = \Delta_0 + o_p(1) \) and the proof of Lemma 6 below shows that \( \Delta_n(\bar{S}, \theta_0) - \Delta_n(\bar{S}, \theta_0) = o_p(1) \). Hence, \( \Delta_n(\bar{S}, \theta_0) = \Delta_0 + o_p(1) \). Q.E.D.

PROOF OF LEMMA 3: All of the finite dimensional distributions of \( \nu_n(\cdot, \theta_0) \) converge weakly to those of \( \nu(\cdot) \) by the multivariate central limit theorem. Thus, weak convergence of \( \nu_n(\cdot, \theta_0) \) to \( \nu(\cdot) \) as a process follows if we can establish uniform tightness of \( \{\nu_n(\cdot, \theta_0); n = 1, 2, \ldots\} \) (see Pollard (1984, Compactness Theorem 29, p. 82)). Furthermore, the sample paths of the limit process \( \nu(\cdot) \) are bounded and uniformly continuous with probability one, provided the compact sets constructed in the proof of tightness contain only bounded uniformly continuous elements (see Pollard (1984, Ch. IV, Sec 5, and Ch. VII, Sec. 5)).

Pollard's (1984, p. 157) Theorem 21 establishes uniform tightness of the (standard) empirical process, defined as a process indexed by functions, provided a condition holds, viz., his equation (22), that bounds the increments of the process. The compact sets used in his proof of uniform tightness contain only bounded continuous functions, as desired. The empirical process and the limit \( \bar{S} \)-bridge process can be replaced by the conditional empirical process \( \nu_n(\cdot, \theta_0) \) and the conditional \( \bar{S} \)-bridge \( \nu(\cdot) \), respectively, in Pollard's condition (22) and in his proof of the Theorem 21 and the proof goes through unchanged. Using the conditional empirical process, Pollard's condition (22) is: For each \( \epsilon > 0 \) and \( \eta > 0 \), there exists a \( \delta > 0 \) such that

\[
(\text{A.5}) \quad \limsup_{n \to \infty} \mathbb{P} \left( \sup_{r \in R_\delta} \left| \nu_n(r, \theta_0) \right| > \eta \right) < \epsilon,
\]

where \( R_\delta = \{ r: r = 1 - C_i - C_j \cup C_i \cup C_j \in \mathbb{C}; \text{ and } F(C_i, \hat{C}_j) < \delta^2 \} \).

It remains to establish (A.5), as well as the other conditions of Pollard's Theorem 21. These other conditions require \( C \) to be a totally bounded "permissible" subset of \( L^2(F) \), where permissible sets satisfy certain conditions needed to handle measurability difficulties (see Pollard (1984, Appendix C)). Since \( C \) is a VC class, total boundedness follows immediately from Pollard's (1984, p. 34) Lemma 36.

To show (A.5), construct \( \nu_n(\cdot, \theta_0) \) on \( \Omega \), \( \mathbb{F} \), \( \nu \), and \( \mu \) on an enlarged probability space, say \((\Omega', \mathbb{F}', \nu', \mu')\), such that \( \nu \) is a continuous distribution with \( \nu \) and \( \mu \) on \( \mathbb{F}' \) and \( \Omega' \) are sequences of independent \( r \)'s with conditional distributions of \( \nu_n(\cdot, \theta_0) \) and \( \nu_n'(\cdot) \) indexed by sets in \( C \) defined by

\[
\nu_n(\cdot) = \sqrt{n} \left( P_n(\cdot) - \frac{1}{n} \sum_{i=1}^{n} F(\cdot, X, \theta_0) \right) \quad \text{and}
\]

\[
\nu_n'(\cdot) = \sqrt{n} \left( P_n'(\cdot) - \frac{1}{n} \sum_{i=1}^{n} F(\cdot, X, \theta_0) \right),
\]

where \( P_n(\cdot) \) and \( P_n'(\cdot) \) are the empirical measures based on \( \{(Y_i, X_i); i = 1, \ldots, n\} \) and \( \{(Y_i', X_i); i = 1, \ldots, n\} \), respectively, and \( F(C, X, \theta_0) = \int_{C} f(y|X, \theta_0) d\mu(y) \) is the conditional probability of \( C \) given \( X \). (To simplify notation and wlog, we suppress the dependence of \( \nu_n(\cdot, \theta_0) \) on \( \theta_0 \), since we are considering only the case \( \theta = \theta_0 \), and we index the empirical process by sets rather than partitions.)

We now establish a symmetrization result by extending, rather straightforwardly, Pollard's (1984, p. 14) Symmetrization Lemma 8. Let

\[
\mathcal{S}_n = \left\{ \omega \in \Omega: \mathbb{P}(|\nu_n'(\cdot)| < \eta/2 \{ (Y_i, X_i); i = 1/2, r \in R_\delta \} \right\},
\]

where \( \{(Y_i, X_i); i = 1, 2, \ldots, n\} \). We have

\[
(\text{A.6}) \quad 1_{\mathcal{S}_n} \cdot \left( \sup_{r \in R_\delta} |\nu_n(r)| > \eta \right) \leq 1_{\mathcal{S}_n} \cdot \left( \sup_{r \in R_\delta} |\nu_n(r)| > \eta \right) \cdot 2\mathbb{P}(|\nu_n'(\cdot)| < \eta/2 \{ (Y_i, X_i) \})
\]

Define a random element \( \pi = \tau(\nu_n) \) on the set \( \{ \sup_{r \in R_\delta} |\nu_n(r)| > \eta \} \) such that \( \pi \) takes values in \( R_\delta \) and
\(|r_n(r)| > \eta\), where \(\sup\), denotes \(\sup_{r \in R_n}\). Then,

\[
1_{r_n} \left\{ \sup_{r} |r_n(r)| > \eta \right\} \leq \left\{ \sup_{r} |r_n(r)| > \eta \right\} + 2P(\left| r_n^* - r_n \right| > \eta) \leq \left\{ \sup_{r} |r_n^* - r_n| \right\} \leq \eta/2, \quad \{Y, X\}
\]

(A.7)

where the equality above holds because \(r\) is a random function of \(r_n(\cdot)\), and hence, \(\{Y, X\}\), only taking expectations yields

\[
P\left( S_n \cap \sup_{r} |r_n(r)| > \eta \right) \leq 2P(\sup_{r} |r_n(r) - r_n^*(r)| > \eta/2).
\]

For any set \(B \in \mathcal{B}, \ P(S_n \cap B) \geq P(B) - P(S_n^c)\). Hence, we have the following symmetrization result:

\[
P\left( \sup_{r} |r_n(r)| > \eta \right) \leq 2P(\sup_{r} \sqrt{n}|P_n(r) - P_n^*(r)| > \eta/2) + 1 - P(S_n)
\]

(The measurability difficulties overlooked in the argument above that establishes (A.8) need to be handled using the permmissibility assumption on the VC class of sets, as done for Pollard's Symmetrization Lemma in his Appendix C. Note that for countable VC classes no such measurability problems arise.)

Next we show that the limit supremum as \(n \to \infty\) of the first term of the right-hand-side of (A.8) can be made arbitrarily small by taking \(\delta\) small. The proof of tightness for the standard empirical process uses maximal inequalities of the form: Given \(\eta > 0\) and \(\varepsilon > 0\), there is \(\delta > 0\) such that

\[
\lim\sup_{n \to \infty} P\left( \sup_{r \in R_n} \sqrt{n}|P_n(r) - P_n^*(r)| > \eta \right) < \varepsilon.
\]

where \(P_n(\cdot)\) and \(P_n^*(\cdot)\) are independent copies of the empirical measure (e.g., see Pollard (1984, Equicoinefinitude Lemma, p. 150)). In our case, \(P_n(\cdot)\) and \(P_n^*(\cdot)\) are not independent, because they are based on the same \(\{X_i\}\) rv's. Conditional on \(\{X_i\}\), however, they are independent, though the underlying rv's \((Y_i, X_i), i = 1, \ldots, n\) and \((Y_{i'}, X_{i'}) = (1, \ldots, n)\) are no longer identically distributed. Fortunately, the tightness result for the standard empirical process (or symmetrized process) can be extended to independent nonidentically distributed (ind) rv's without great difficulty provided the marginal distributions do not fluctuate too widely (see Alexander (1984) for explicit treatment of the ind case). In particular, in the case of indexing by a VC class, one only needs to have control of the variances of \(r = 1, \ldots, n\) for different observations \(i\) (e.g., see Theorem 2.8 of Alexander (1984)). That is, we need to show: Given \(\varepsilon > 0\), there is \(\delta > 0\) such that

\[
G_n = \sup_{r \in R_n} \frac{1}{n} \sum_{i=1}^{n} \text{Var}[r(Y_i, X_i)] \leq \varepsilon \quad \text{for all large } n
\]

This follows, with probability one, because

\[
G_n \leq \sup_{r \in R_n} \frac{1}{n} \sum_{i=1}^{n} F(r, X_i, \theta) \overset{\text{as } n \to \infty}{\longrightarrow} \sup_{r \in R_n} F(r^2) \leq \delta^2
\]

using the uniform SLLN and the definition of \(R_n\), since \(F(r^2) = F(C_1, \delta C_2)\).

Thus, for \(\{X_i\}\) in a set with probability one and given \(\eta > 0\) and \(\varepsilon > 0\), there exists \(\delta > 0\) that does not depend on \(\{X_i\}\) such that

\[
\lim\sup_{n \to \infty} P\left( \sup_{r \in R_n} \sqrt{n}|P_n(r) - P_n^*(r)| > \eta \right) < \varepsilon
\]

By the bounded convergence theorem, we can integrate out \(\{X_i\}\) to get

\[
\lim\sup_{n \to \infty} P\left( \sup_{r \in R_n} \sqrt{n}|P_n(r) - P_n^*(r)| > \eta \right) < \varepsilon
\]
Below we show $\lim_{n \to \infty} P(S_n) = 1$. Combining this result with (A.8) and (A.10) gives (A.5) and the proof is complete.

To show $\lim_{n \to \infty} P(S_n) = 1$, use Chebyshev's inequality to get

$$ P\left(|\chi^2_n(\gamma)| > \eta/2; \{Y_i, X_i\}\right) \leq \frac{\sum_{i=1}^{n} F(r^2, X_i, \theta_i)}{\eta^2} \quad \forall r \in R \tag{A.11} $$

Also, by the uniform SLLN and the bounded convergence theorem,

$$ \sup_r \left| \frac{1}{n} \sum_{i=1}^{n} F(r^2, X_i, \theta_i) - F(r^2) \right| \leq \int_Y \sup_r \left| \frac{1}{n} \sum_{i=1}^{n} r^2(Y_i, X_i) - F(r^2) \right| dP(\{Y_i\}) \to 0 \quad a.s $$

Since $\sup_r F(r^2) \leq \delta^2$, the above results combine to give

$$ P(S_n) \geq P\left(\frac{\sum_{i=1}^{n} F(r^2, X_i, \theta_i) < \eta^2/2; \forall r \in R}{} \right) \to 1 \quad \text{for } \delta \text{ sufficiently small.} \tag{A.12} $$

**Proof of Lemma 4:** It suffices to show that $h(z(\cdot), \gamma) = z(\gamma)$ is continuous at all $(z(\cdot), \gamma)$ such that $z(\cdot)$ is uniformly continuous. Given $\varepsilon > 0$, uniform continuity of $z(\cdot)$ guarantees the existence of a constant $\delta > 0$ such that $|\gamma - \gamma'| < \delta$ implies $|f(\gamma) - z(\gamma')| < \varepsilon/2$. We can choose a neighborhood $\mathcal{N}$ of $(z(\cdot), \gamma)$ such that for all $(z(\cdot), \gamma)$ in $\mathcal{N}$, we have $|\gamma - \gamma'| < \delta$ and $|f(\cdot, z(\cdot))| < \varepsilon/2$, where $\| \cdot \|$ denotes the norm. Then, $|z(\gamma) - z(\gamma')| < \varepsilon/2$ and $h$ is continuous at $(z(\cdot), \gamma)$. $Q.E.D$

**Proof of Lemma 5:** The desired result for $\Sigma_n(\hat{F}, \hat{\theta})$ follows if we can show $\Lambda_n(\hat{F}, \hat{\theta}) \to \Lambda_n(\hat{F}, \hat{\theta})$ and the analogous results for $H_n(\hat{F}, \hat{\theta})$, $\Delta_n(\hat{F}, \hat{\theta})$, $\Pi_n(\hat{F}, \hat{\theta})$, and $V_n(\hat{F}, \hat{\theta})$. It suffices to show:

$$ \xi_n(\delta) = \frac{1}{n} \sum_{i=1}^{n} F(h_n(\delta), X_i, \delta) - F(h_n(\delta), X_i, \delta) = o_n(1) \tag{A.13} $$

where $g_n$ and $h_n(\theta)$ are rv's that satisfy (i) $|g_n| < 1$, (ii) $h_n(\delta) = h(\cdot, X_i, \hat{\theta})$ or $h_n(\delta) = h(\cdot, X_i, \hat{\theta})$, $\hat{F}(\cdot, X_i)$, denotes the indicator function of $(\cdot, X_i) \in \hat{F}$, and $V_n(\hat{F}, \hat{\theta})$ is sufficiently small.

$$ c_1(\delta) = \sup_{v \in \mathcal{N}(\delta)} |h(y, x, \theta) - h(y, x, \theta_0)| \leq v(y, x, \delta) $$

for some $v(y, x)$ such that

$$ K_1 = \int_{X \in \mathcal{N}(\delta)} \sup_{v \in \mathcal{N}(\delta)} \int_Y v(y, x) f(y|x, \theta) d\mu(y) dP_x(x) < \infty $$

where $\mathcal{N}(\delta) = (\theta - \hat{\theta}_0, \|\theta - \hat{\theta}_0\| \leq \delta)$, and

$$ K_2 = \int_{X \in \mathcal{N}(\delta)} \sup_{v \in \mathcal{N}(\delta)} \int_Y \left| h(y, x, \theta_0) - r(y, x) f(y|x, \theta) d\mu(v) dP_x(x) < \infty \right| $$

The sufficiency of this condition follows because (1) the greatest diagonal element of $\Lambda_n(\hat{F}, \hat{\theta}) - \Lambda_n(\hat{F}, \hat{\theta})$ equals $\xi_n(\delta)$ with $g_n = 1, h_n(\delta) = \hat{F}, c_1(\delta) = 0$, and $K_2 < \infty$ by M2, and (2) the (j, k)th
element of \( H_n(\hat{\theta}; \theta_0) - H_n(\hat{\theta}; \theta_0) \) can be written as \( a_{1n} + a_{2n} \), where

\[
a_{1n} = \frac{1}{n} \sum_{i=1}^{n} F(\hat{\theta}; X, \theta) - \sum_{i=1}^{n} F(\hat{\theta}; X, \theta_0) \approx \xi_n(\theta)
\]
equals \( \xi_n(\theta) \) with

\[
g_n = F(\hat{\theta}; X, \theta) - \sum_{i=1}^{n} F(\hat{\theta}; X, \theta_0) \approx \xi_n(\theta)
\]
equals \( \xi_n(\theta) \) with

\[
a_{2n} = \frac{1}{n} \sum_{i=1}^{n} F(\hat{\theta}; X, \theta_0) - \sum_{i=1}^{n} F(\hat{\theta}; X, \theta_0) \approx \xi_n(\theta)
\]
equals \( \xi_n(\theta) \) with

\[
g_n = F(\hat{\theta}; X, \theta_0) - \sum_{i=1}^{n} F(\hat{\theta}; X, \theta_0) \approx \xi_n(\theta)
\]
equals \( \xi_n(\theta) \) with

(3) the \((i, j)\)th element of \( \Delta_n(\hat{\theta}; \theta_0) - \Delta_n(\hat{\theta}; \theta_0) \) equals \( \xi_n(\theta) \) with \( g_n = 1 \), \( h_n(\theta) = s(\theta) \cdot \tilde{f}(\theta) \cdot \tilde{f}(\theta) \),

\[
c_i(\theta, \theta) = \sup_{\theta \in \Omega(\theta)} |s(y) - s(y)| \leq \sup_{\theta \in \Omega(\theta)} \tilde{f}(y) \cdot \tilde{f}(y) \cdot v_{\theta} \cdot \delta
\]
using the mean value theorem and E2 for the first inequality, \( K_1 < \infty \) by E2, and \( K_2 < \infty \) by E2;

(4) the \((i, j)\)th element of \( \Pi_n(\hat{\theta}; \theta_0) - \Pi_n(\hat{\theta}; \theta_0) \) equals \( \xi_n(\theta) \) with \( g_n = 1 \), \( h_n(\theta) = \psi(\theta) \cdot \tilde{f}(\theta) \cdot \tilde{f}(\theta) \),

\[
c_i(\theta, \theta) = \sup_{\theta \in \Omega(\theta)} |\psi(y) - \psi(y)| \leq \sup_{\theta \in \Omega(\theta)} \psi(y) \cdot \tilde{f}(y) \cdot v_{\theta} \cdot \delta
\]
using the mean value theorem and E2 as above, \( K_1 < \infty \) by E2, and \( K_2 < \infty \) by E2;

(5) the \((i, m)\)th element of \( \Delta_n(\hat{\theta}; \theta_0) - \Delta_n(\hat{\theta}; \theta_0) \) equals \( \xi_n(\theta) \) with \( g_n = 1 \), \( h_n(\theta) = \frac{\partial}{\partial \theta_m} \psi(\theta) \cdot \tilde{f}(\theta) \cdot \tilde{f}(\theta) \),

\[
c_i(\theta, \theta) = \sup_{\theta \in \Omega(\theta)} \left| \frac{\partial}{\partial \theta_m} \psi(y) - \frac{\partial}{\partial \theta_m} \psi(y) \right| \leq \sup_{\theta \in \Omega(\theta)} \psi(y) \cdot \tilde{f}(y) \cdot v_{\theta} \cdot \delta
\]
using the mean value theorem and E2 as above, \( K_1 < \infty \) by E2, and \( K_2 < \infty \) by E2;

(6) the \((i, m)\)th element of \( \Delta_n(\hat{\theta}; \theta_0) - \Delta_n(\hat{\theta}; \theta_0) \) equals \( \xi_n(\theta) \) with \( g_n = 1 \), \( h_n(\theta) = \psi(\theta) \cdot \tilde{f}(\theta) \cdot \tilde{f}(\theta) \),

\[
c_i(\theta, \theta) = \sup_{\theta \in \Omega(\theta)} \left| \frac{\partial}{\partial \theta_m} \psi(y) - \frac{\partial}{\partial \theta_m} \psi(y) \right| \leq \sup_{\theta \in \Omega(\theta)} \psi(y) \cdot \tilde{f}(y) \cdot v_{\theta} \cdot \delta
\]
using the mean value theorem and E2 as above, \( K_1 < \infty \) by E2, and \( K_2 < \infty \) by E2.

Now we show that \((A.13)\) holds. Straightforward manipulations using the assumed properties of \( g_n \) and \( h_n(\theta) \) give:

\[
\sup_{\theta \in \Omega(\theta)} |\xi_n(\theta)| \leq \sup_{\theta \in \Omega(\theta)} \frac{1}{n} \sum_{i=1}^{n} \int |h(y, X, \theta) - h(y, X, \theta_0)| \cdot f(y | X, \theta_0) \, dm(y) \]

\[+ \sup_{\theta \in \Omega(\theta)} \frac{1}{n} \sum_{i=1}^{n} \int |h(y, X, \theta_0) - h(y, X, \theta_0)| \cdot f(y | X, \theta_0) \, dm(y) \]

\[\leq \frac{1}{n} \sum_{i=1}^{n} \int v(y, X) \cdot f(y | X, \theta_0) \, dm(y) \cdot \delta \]

\[+ \frac{1}{n} \sum_{i=1}^{n} \int v(y, X) \cdot r(y, X) \cdot f(y | X, \theta_0) \cdot \sqrt{L} \cdot \delta \, dm(y) \]

\[
\lim_{n \to \infty} \delta(K_1 + \sqrt{L} K_2) = K_3, \delta.
\]
where the second inequality holds for the first summand by the assumed properties of \( h(y, x, \theta) \) and for the second summand using the mean value theorem and assumption M2 and the almost sure convergence follows by the strong law of large numbers (SLLN) and the assumed properties of \( v(y, x) \).

Let \( B_{\infty} = \{ \sup_{\theta \in \Theta} |\xi_n(\theta)| \leq (K_1 + 1) \cdot \delta \} \) and \( B_{\infty} = \{ \in \cap N(\delta) \} \). By (A.14) and the consistency of \( \hat{\theta} \) (Assumption E1), we have \( \lim_{n \to \infty} P(B_{\infty}) = 1 \) and \( \lim_{n \to \infty} P(B_{\infty}) = 1 \). Hence,

\[
\lim_{n \to \infty} P(B_{\infty} \cap B_{\infty}) \leq \lim_{n \to \infty} P(|\xi_n(\hat{\theta})| \leq (K_1 + 1) \cdot \delta) \]

and (A.13) is established.

The proof with \( \Sigma_{\infty}(\hat{\Gamma}, \cdot) \) or \( \Sigma_{\infty}(\hat{\Gamma}, \cdot) \) in place of \( \Sigma_{\infty}(\hat{\Gamma}, \cdot) \) is similar.

Q.E.D.

**Proof of Lemma 6:** The desired result follows for \( \Sigma_{\infty}(\cdot, \cdot, \theta_0) \) if we can show \( \Lambda_{\infty}(\hat{\Gamma}, \theta_0) - \Lambda_{\infty}(\Gamma, \theta_0) = \phi(1) \) and the analogous results for \( H_{\infty}(\cdot, \theta_0), \Pi_{\infty}(\cdot, \theta_0), \) and \( \Delta_{\infty}(\cdot, \theta_0) \). It suffices to show

\[
B_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \left[ F(h_{\Gamma}, X, \theta_0) - F(h_{\Gamma}, X, \theta_0) \right] = \phi(1), \quad \forall j = 1, \ldots, J.
\]

for some rv's \( g_i \) and \( h(Y, X), \) where \( |g_i| \leq 1 \) and \( E_{\mu} h^2(Y, X) < \infty \).

The sufficiency of this condition follows because (1) the \( j \)th diagonal element of \( \Lambda_{\infty}(\hat{\Gamma}, \theta_0) - \Lambda_{\infty}(\Gamma, \theta_0) \) equals \( B_n \), with \( g_i = h(Y, X) \); (2) the \( (j, k) \)th element of \( H_{\infty}(\hat{\Gamma}, \theta_0) - H_{\infty}(\Gamma, \theta_0) \) can be written as

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} F(h_{\Gamma}, i, i) - F(h_{\Gamma}, i, i) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F(h_{\Gamma}, i, i) - F(h_{\Gamma}, i, i).
\]

where \( F(h_{\Gamma}, i, i) \) denotes the \( i \)th element of \( F(h_{\Gamma}, X, \theta_0) \), the first summand equals \( B_n \) with \( g_i = F(h_{\Gamma}, i, i) \), and \( h(Y, X) = 1 \); and the second summand equals \( B_n \) with \( g_i = F(h_{\Gamma}, i, i) \), and \( h(Y, X) = 1 \); (3) the \( (j, k) \)th element of \( \Pi_{\infty}(\hat{\Gamma}, \theta_0) - \Pi_{\infty}(\Gamma, \theta_0) \) equals \( B_n \), with \( g_i = 1 \), \( h(Y, X) = \psi(Y, X, \theta_0) \), and \( E_{\mu} \psi^2(Y, X, \theta_0) < \infty \) by E2; and (4) the \( (j, k) \)th element of \( \Delta_{\infty}(\hat{\Gamma}, \theta_0) - \Delta_{\infty}(\Gamma, \theta_0) \) equals \( B_n \), with \( g_i = 1 \), \( h(Y, X) = s(Y|X, \theta_0) \), and \( E_{\mu} \psi^2(Y, X, \theta_0) < \infty \) by M2.

To show (A.20), we have

\[
B_n \leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int h(y, X) \{ F(h, Y, X), Y, X, \theta_0 \} \, d\mu(y)
\]

\[
\leq \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int h^2(Y, X) f(y|X, \theta_0) \, d\mu(y) \right)^{1/2}
\]

\[
\leq \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int (h_{\Gamma} - \gamma)^2 f(y|X, \theta_0) \, d\mu(y) \right)^{1/2}
\]

\[
= \left( E_{\mu} h^2(Y, X) + \phi(1) \right)^{1/2} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F(h_{\Gamma}, \hat{\Delta}_{\Gamma}, Y, \theta_0) \right)^{1/2}
\]

using the Cauchy-Schwarz inequality and the WLLN, where \( \hat{\Delta} \) denotes the symmetric difference operator.

To show the second multiplicative of (A.17) is \( \phi(1) \), we use the result of Lemma 3 above with \( C \) replaced by \( \tilde{C} \equiv \{ G: \ G = C_1 \tilde{C} C_2 \} \), for some \( C_1, C_2 \subseteq C \). Note that \( \tilde{C} \) is a VC class, since \( C \) is Lemma 3 gives

\[
\tilde{\nu}_n(\cdot, \theta_0, \hat{\Gamma}, \hat{\Delta}_{\Gamma}) \overset{d}{\longrightarrow} (\nu(\cdot), \phi) \quad \text{as a process indexed by } \gamma \in \tilde{D} \text{ as } n \to \infty,
\]

where \( \tilde{D} \) is the analogue of \( D \) with \( C \) replaced by \( \tilde{C} \) and \( \phi \) is the null set. With \( \nu(\cdot) \)-probability one, the function that maps \( (\tilde{\nu}_n(\cdot, \theta_0, \hat{\Gamma}, \hat{\Delta}_{\Gamma}) \) into \( \tilde{\nu}_n(\hat{\Delta}_{\Gamma}, \theta_0) \), is continuous by Lemma 4. Hence, the continuous mapping theorem gives

\[
\tilde{\nu}_n(\hat{\Delta}_{\Gamma}, \theta_0) \overset{d}{\longrightarrow} (\nu(\phi), 0) \quad \text{as } n \to \infty, \forall j = 1, \ldots, J
\]
That is,
\[(A\ 20) \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\tilde{Y}, \tilde{Y}}(Y_i, X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F(\tilde{\theta}, \tilde{\Gamma}, X_i, \theta_i) = o_p(1) \quad \text{as } n \to \infty \]

Also, by the analogous result for the standard empirical process \(\bar{\pi}(\cdot, \theta_i)\), we have
\[(A\ 21) \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\tilde{Y}, \tilde{Y}}(Y_i, X_i) - \sqrt{n} F(\tilde{\theta}, \tilde{\Gamma}) = o_p(1) \quad \text{as } n \to \infty \]

Since \(F(\tilde{\theta}, \tilde{\Gamma}) = o_p(1)\) by RC1, equations (A.20) and (A.21) yield
\[(A\ 22) \quad \frac{1}{n} \sum_{i=1}^{n} F(\tilde{\theta}, \tilde{\Gamma}, X_i, \theta_i) = o_p(1) \quad \text{as } n \to \infty . \]

Equations (A.17) and (A.22) combine to give (A.16), as desired.

The proof with \(\Sigma_n(\cdot, \theta_i)\) or \(\Sigma_{1,n}(\cdot, \theta_i)\) in place of \(\Sigma_n(\cdot, \theta_i)\) is similar.

**PROOF OF THEOREM 3.** The proof of part (a) follows the proof of Theorem 1 given in Section 3 and the Appendix above. The result of Lemma 1 holds under the local alternatives \((Q_n)\), and the proof goes through unchanged (except that the CLT is replaced by a triangular array CLT; because \(n(\tilde{\theta} - \theta_i)^{-1} = o_p(1)\) under \((Q_n)\), since \(n < 1/2\) and
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_{\Gamma, \theta_i} [Y_i, X_i, \theta_i] - E_{Q_n} [Y_i, X_i, \theta_i] = o_p(1) \quad \text{as } n \to \infty \]

under \((Q_n)\), and in the neighborhood \(N(\tilde{\theta})\) of radius \(n^{-1/2}\), which is greater than \(n^{-1/2}\).

The result of Lemma 2 also holds under \((Q_n)\). The proof requires no change, given the result discussed below that Lemma 5 holds under \((Q_n)\), except that the WLLN needs to be replaced by a triangular array WLLN.

Given Lemmas 1 and 2, EL, and the assumption \(\bar{\pi} \to \Sigma_0\) in equations (34) and (36) hold under \((Q_n)\). Further, the result of Lemma 3 holds under \((Q_n)\), with the limit process \(\bar{\pi}(\cdot)\) replaced by \(\tilde{\pi}(\cdot)\), where \(\tilde{\pi}(\cdot)\) is exactly the same as \(\bar{\pi}(\cdot)\) except that \(E_{Q_n} \tilde{\pi}(Y_i) - E_{\tilde{\pi}}(Y_i)\). The proof of tightness of \((\bar{\pi}_n(\cdot, \theta_i)\) under \((Q_n)\), with \(n = 1, 2, \ldots\) under \((Q_n)\), proceeds exactly as under \(P\) above. The finite dimensional distributions of \(\bar{\pi}_n(\cdot, \theta_i)\) converge to those of \(\tilde{\pi}(\cdot)\) by application of the triangular array CLT.

By the same arguments as given in (38) through (40), \(\bar{\pi}_n(\cdot, \theta_i) \to \Sigma_0(\cdot, \theta_i)\) as \(n \to \infty\) under \((Q_n)\), where \(\tilde{\pi}(\cdot)\) is as above, \(\tilde{\pi} \sim N(\Sigma_0(\cdot, \theta_i), V_0)\), and \(E_{Q_n} \tilde{\pi}(Y_i) = \Pi(Y_i)\). Moreover, by the continuous mapping theorem, the asymptotic covariance matrix of \(\tilde{\pi}_n(\cdot, \theta_i)\) under \((Q_n)\) equals that of \(\tilde{\pi}(\cdot, \theta_i)\) as \(n \to \infty\), viz. \(\Sigma_0\). In addition, Lemma 4 and the continuous mapping theorem imply that the approximating quadratic \(q_n(\tilde{\Gamma}, \tilde{\theta})\) of (3.1) satisfies
\[(A\ 23) \quad q_n(\tilde{\Gamma}, \tilde{\theta}) \to \tilde{\pi}(\Gamma) - \Sigma_0 \tilde{\theta}\quad \text{as } n \to \infty \quad \text{under } (Q_n) \]

The right-hand side of (A.23) is a quadratic form in normal variables and by Rao and Mitra (1971, Theorem 9.2.3) it has a \(\chi^2(n)\) distribution. Using (3.6), this establishes part (a).

To establish part (b) of the Theorem, we show that the results of Lemmas 5 and 6 hold under \((Q_n)\). The proof of Lemma 5 for \(\tilde{\pi}_n(\cdot, \theta_i)\) requires no changes because \(\tilde{\theta}\) is consistent for \(\theta_i\) under \((Q_n)\), and the \(\psi_\theta(\cdot)\) does not depend on \(Y_i, i = 1, \ldots, n\) (except through \(\tilde{\theta}\) and \(\tilde{\Gamma}\)). The proof of Lemma 5 for \(\tilde{\pi}_n(\cdot, \theta_i)\) holds under a triangular array WLLN provided \(E_{Q_n} \tilde{\pi}_n(Y, X) < \infty\), \(E_{Q_n} \tilde{\pi}_n(Y, X) \tilde{\pi}_n(Y, X) < \infty\), and \(E_{Q_n} \tilde{\pi}_n(Y, X) \in \tilde{\pi}_n(Y, X) < \infty\). Assumptions E2 and M'2 include these conditions.

Under Assumption RC1, the proof of Lemma 6 for \(\tilde{\pi}_n(\cdot, \theta_i)\) remains valid since \(\tilde{\theta}\) does not depend on \(Y_i, i = 1, \ldots, n\) (except through \(\tilde{\theta}\) and \(\tilde{\Gamma}\)) and \(\tilde{\theta}\) is consistent for \(\theta_i\) under \((Q_n)\). The proof of Lemma 6 for \(\tilde{\pi}_n(\cdot, \theta_i)\) holds under a triangular array WLLN provided \(E_{Q_n} \tilde{\pi}_n(Y, X) < \infty\), \(E_{Q_n} \tilde{\pi}_n(Y, X) \tilde{\pi}_n(Y, X) < \infty\), and \(E_{Q_n} \tilde{\pi}_n(Y, X) \in \tilde{\pi}_n(Y, X) < \infty\). Assumptions E2 and M'2 include the latter conditions. As in (3.15), a triangular array WLLN gives \(\tilde{\pi}_n(\Gamma) \to \tilde{\pi}_n\) for \(r = 1, 2, 3\) provided \(E_{Q_n} \tilde{\pi}_n(Y, X) < \infty\), \(E_{Q_n} \tilde{\pi}_n(Y, X) \tilde{\pi}_n(Y, X) < \infty\), and \(E_{Q_n} \tilde{\pi}_n(Y, X) \in \tilde{\pi}_n(Y, X) < \infty\), and as is guaranteed by E2 and M'2. Combining these results with those of Lemmas 5 and 6 gives the results of part (b).
Part (c) holds under the assumptions by Theorem 1 of Andrews (1987).

Part (d) holds under the given assumptions by the same argument as used in Section 3 to establish the Corollary and by Theorem 9.2.3 of Rao and Mitra (1971), noting that their condition \( \mu \in \mathcal{M}(\Sigma_0) \) is satisfied under the assumption \( Q_n(\nu, (\hat{\theta}, \hat{\theta}) \in \mathcal{M}(\Sigma_0)) \to 1 \) as \( n \to \infty \).

We now establish the claim of Section 4.3 that chi-square tests based on residuals have the same asymptotic distribution under the null and local alternatives in certain models with covariates, as in the analogous models without covariates (Section 4.3 specifies the models in question.) This result follows because the difference between the asymptotic distribution of \( \chi^2_n(\tilde{\Gamma}, \tilde{\theta}) \) in these two cases depends only on the difference in the asymptotic covariance matrix terms \( \Delta_0 \Delta_0^{-1} \Pi_0 \) and \( \Delta_0 \Delta_0^{-1} \Pi_0 \delta_0 (\Delta_0^{-1}) \Delta_0 \) of \( \chi^2_n(\tilde{\Gamma}, \tilde{\theta}) \). Without loss of generality, the model with covariates and intercept terms can be reparameterized such that each element of \( X \), mean zero. Then, the elements of the score function that correspond to the parameters on \( X \), factor into the product of terms that depend on the errors \( \epsilon \) and a linear function of \( X \). In addition, the limit cells \( \Gamma \) depend only on \( U \). Hence, by the fact that \( E_{\tilde{\theta}} X_i = \theta \), we find that the rows of \( \Delta_0 \) that correspond to parameters of \( X \) consist of zeroes. Furthermore, for the estimators considered, the matrix \( \Delta_0 \) has a block diagonal structure between the parameters of \( X \) and the remaining parameters. Also, the matrices \( \Pi_0 \) and \( V_0 \), with the rows and columns removed that correspond to parameters of \( X \), are the same as \( \Pi_0 \) and \( V_0 \) in the no covariate case. These results combine to establish the equivalence of \( \Delta_0 \Delta_0^{-1} \Pi_0 \) in the covariate and no covariate cases; likewise with \( \Delta_0 \Delta_0^{-1} V_0(\Delta_0^{-1}) \Delta_0 \).

Finally, we elaborate on the asymptotic optimality properties discussed in the final paragraph of Section 4.3. Given any conditional distribution \( h(y|x) \), consider the parametric family

\[
\mathcal{H}_n = \{ h(y|x, \mathcal{P}) : \mathcal{P} \in \mathcal{P} \},
\]

where

\[
h(y|x, \mathcal{P}) = \prod_{j=1}^J \left( h(y|x) \frac{p_j}{h_j} \right)^{r_j}, \quad \mathcal{P} = (p_1, \ldots, p_J),
\]

\( h_j \) is the \( j \)th element of \( h \),

\[
\mathcal{P} = \left\{ \mathcal{P} \in \mathbb{R}^J : 0 < p_j < 1, \forall j ; 0 < \sum_{j=1}^{J} p_j < 1, \text{ for } g = 1, \ldots, G \right\},
\]

and \( J \) denotes the index of the first cell involving the \( g \)th value of the covariates (it is assumed that the \( G \) cells of \( \Gamma \) with redundant conditional probabilities are numbered \( J = G + 1, \ldots, J \) and the remaining cells are numbered such that cells with the same covariate values are numbered consecutively.)

The null hypothesis \( H_0: f = f_0 \) is satisfied for this parametric family only if \( \mathcal{P} = f_0 \). The Wald statistic for testing \( \mathcal{P} = f \) is a quadratic form in the vector \( \hat{\theta} - \hat{f}_0 \), where \( \hat{\theta} \) is the ML estimator of \( \theta \). This test statistic is asymptotically equivalent to the chi-square test statistic \( \chi^2_n(\hat{\theta}) \) under the null and local alternatives; see below.

Since the chi-square test has the proper asymptotic size for all distributions in the null hypothesis \( H_0 \) (where the distribution of the covariates is arbitrary), it possesses Wald's asymptotic optimality properties for testing \( H_0 \) against the non-null conditional distributions in \( \mathcal{H}_n \), coupled with any marginal distribution \( P_X \) of the covariates that gives positive probability to each covariate value. These optimality properties hold for all conditional distributions \( h(y|x) \). Hence, the chi-square test enjoys optimality properties for testing against the non-null distributions in each of an uncountable infinite number of parametric families \( \mathcal{H}_n \). Since every alternative conditional density is included in some parametric family \( \mathcal{H}_n \), the chi-square test exhibits optimality properties that apply to the entire nonparametric class of alternatives to \( H_0 \).

Now we show that the chi-square statistic with weight matrix \( \Sigma^* \) is asymptotically equivalent to the Wald statistic for testing \( \mathcal{P} = f \). The Wald statistic is a quadratic form based on \( \sqrt{n} (\hat{\theta} - \hat{f}_0) \) where \( \hat{\theta} \) is the ML estimator of \( \theta \). It is easy to show that \( \hat{\theta} \) has \( J \)th element \( \hat{\theta}_j = (1/M_j \Sigma^*_{-j})^t (\bar{y}_j, X_j) \), for \( j = 1, \ldots, J \), where \( M_j \) is the number of observations with \( X \)-value equal to the \( j \)th covariate value contained in \( \Gamma_j \). The weight matrix of the Wald statistic is consistent for the inverse of the nonsingular asymptotic covariance matrix of \( \sqrt{n} (\hat{\theta} - \hat{f}_0) \).
On the other hand, the chi-square statistic $X^2_i(G)$ is asymptotically equivalent to $X^2_j(I)$ under the null and local alternatives by the arguments of Sections 3.1 and 4.1. Furthermore, the results of Section 3.2 show that $X^2_i(G) = (1/n)^2 A_1 (A_1 + M) A_1^T$ and $(1/n)^2 A_1 = \sqrt{n} S (\hat{\beta} - \beta_0)$ in this context, where $A$ is the same as $A$ of Section 3.2 except that it is defined using $I$ and $G$ rather than $I$ and $G$ and $S$ denotes the $(J-G) \times (J-G)$ diagonal matrix with diagonal elements $M_j/n$ for $j = 1, \ldots, J-G$. Hence, $X^2_i(G)$ also is a quadratic form in $\sqrt{n} (\hat{\beta} - \beta_0)$. Its weight matrix has the same probability limit as that of the Wald statistic, because the asymptotic covariance matrix of $\sqrt{n} (\hat{\beta} - \beta_0)$ is nonsingular and both statistics have the same asymptotic distribution (under the null and local alternatives). This establishes the asymptotic equivalence of the Wald statistic and $X^2_i(G)$ in this context.

REFERENCES


