COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station

New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 763R

RANDOM CELL CHI-SQUARE DIAGNOSTIC TESTS FOR ECONOMETRIC MODELS: THEORY

Donald W. K. Andrews

September 1985

Revised: June 1986

RANDOM CELL CHI-SQUARE DIAGNOSTIC TESTS FOR ECONOMETRIC MODELS: THEORY

HEADNOTE

This paper extends the Pearson chi-square testing method to non-dynamic 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. In the event of rejection of the null hypothesis, the test provides information concerning the direction of departure from the null. The diagnostics provided by the test are intuitive and particularly easy to interpret.

by

Donald W. K. Andrews

September 1985

Revised: June 1986

1. INTRODUCTION

This paper extends the Pearson chi-square testing method to non-dynamic 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 test yields not only a formal acceptance or rejection of the parametric specification of the model, but also provides information regarding the direction of departure from the null hypothesis, in the event of its rejection. This information is conveyed via easily interpreted diagnostic statistics.

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). The present paper extends it further to most parametric cross-sectional and panel data models used in econometrics. It presents 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 some asymptotic optimality results for the test. An Appendix contains proofs of results given in Sections 3 and 4.

2. DEFINITIONS AND ASSUMPTIONS

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

2.1. Definition of the Chi-Square Test Statistic

The observed sample of size n 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 $\mathcal{Y} \subset R^V$ and $\mathcal{X} \subset R^K$, respectively. P denotes the distribution of $\{(Y_i, X_i) : i = 1, 2, \ldots\}$ under the null hypothesis. The first assumption restricts attention to the case where the model is non-dynamic 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 the following:

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 a σ-finite measure μ, and Θ is a parameter space in R^L.

Note that μ need not be Lebesgue measure, so that Y_i may be discrete, continuous, or mixed. The different alternative hypotheses that may be of interest are discussed below.

Let P_X denote the distribution of X_i under P. Since P_X is not restricted by the null hypothesis, the covariates also may be discrete, continuous, or mixed. Let θ_0 denote the true parameter value, when the null

hypothesis is true.

The proposed chi-square test statistic is constructed by first partitioning the region, $Y \times X$, in which the response variables and covariates lie, into disjoint cells. The test statistic then 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 fixed for all n. (See Section 5 of Andrews (1985b) regarding the choice of J.) Let D be a class of partitions of $Y \times X$, each partition being comprised of J sets from C. That is,

$$(2.1) \qquad D = \left\{ \underline{\gamma} \in C^J : \bigcup_{j=1}^{J} \gamma_j = Y \times X, \ \gamma_j \cap \gamma_k = \phi, \ \forall j \neq k \right\},$$

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

Next, we define a new 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 γ in D. That is,

$$(2.2) \qquad P_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \gamma(Y_i, X_i) \ ,$$

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 <u>conditional empirical measure</u> constructed using the parametric conditional distribution of Y_i given X_i . That is,

$$(2.3) \qquad F_n(\gamma,\theta) = \frac{1}{n}\sum_{i=1}^{n} \int_Y \gamma(y, X_i) f(y|X_i, \theta) d\mu(y) = \frac{1}{n}\sum_{i=1}^{n} F(\gamma, X_i, \theta) \ .$$

DEFINITION: *The <u>conditional empirical process</u>* $\nu_n(\cdot,\theta)$ *indexed by elements* γ *of* D *is defined as*

$$(2.4) \qquad \nu_n(\cdot,\theta) = \sqrt{n}\left[P_n(\cdot) - F_n(\cdot,\theta)\right] \ .$$

Let $\hat{\theta}$ be some estimator of the unknown parameter θ_0 . Then, the random vector $\nu_n(\hat{\Gamma},\hat{\theta})$ 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 asymptotic normal distribution with covariance matrix Σ_0 (defined in Section 3 below).

Let $\hat{W}$ be a consistent estimator of some generalized inverse of Σ_0 . The chi-square test statistic is defined as follows:

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

$$(2.5) \qquad X_n^2(\hat{\Gamma},\hat{\theta}) = \nu_n(\hat{\Gamma},\hat{\theta})'\hat{W}\nu_n(\hat{\Gamma},\hat{\theta}) \ .$$

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 distribution of the test statistic under the null (with degrees of freedom given by the rank of Σ_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 (1985a) for a more detailed discussion.)

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) 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 is may create numerous low probability cells. This strategy has been considered by McFadden (1974) and Horowitz (1985) for discrete choice models using non-random 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 corresponds to strategy (iii) or (iv) where Y is partitioned using non-random 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 non-parametrically by a number of methods: (1) use the sample mean (or median)

plus or minus multiples of the sample standard deviation (or the sample absolute deviation), (2) use the k-means clustering procedure, (3) form cells with equal numbers of observations, or (4) use 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 ellipse, (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" β (i.e., reduce X_i to the scalar $X_i'\hat{\beta}$), or 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 specified linear regression structure, but may have any error distribution that is non-normal. A chi-square test can be constructed with the partitioning based on the least squares residuals, in a manner analogous to that used with chi-square tests of normality in models with an iid response variable (and no non-constant covariates). In fact, it is shown below that

the chi-square test of normality (or any other specified 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 non-normal 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 these models, as in the standard linear regression model, the chi-square statistic provides valuable diagnostics, since the normalized deviations $\nu_{nj}(\hat{\Gamma},\hat{\theta})/\sigma_{nj}$, $j = 1, \ldots, J$, indicate whether the assumed distribution over- or under-predicts the number of outcomes in each cell. Thus, one can detect asymmetry, fat or thin tails, or broad shoulders by inspecting the individual cell deviations.

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 (1978) 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 widely used estimation procedures, e.g., selection models 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 (1985a)).

As a final example, suppose one wishes to test whether some scalar covariate X_{1i} 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_{1i} operates through some nonlinear function $g(X_{1i})$, in which X_{1i} interacts with other covariates linearly or nonlinearly, and in which different values of β are appropriate for different values of X_{1i}. A chi-square test can be formed for this situation by partitioning $\mathbf{X}$ based on the covariate X_{1i} alone, and then partitioning $\mathbf{Y}$ nonparametrically for each $\mathbf{X}$ cell. This test has some power against a wide variety of alternatives, and it provides valuable diagnostics that indicate which region in the range of X_{1i} are, or are not, satisfactorily modelled by the parametric family of distributions.

We mention that, in all of the examples above, the cells can be formed in exactly the same manner whether or not there are restrictions (of any kind) 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 such applications is enhanced greatly.

2.3. Assumptions

To establish asymptotic normality of $\nu_n(\hat{\Gamma}, \hat{\theta})$ and to consistently estimate its asymptotic covariance matrix, we need the conditional density $f(y|x,\theta)$, or equivalently its score function $s(y|x,\theta) = \frac{\partial}{\partial\theta}\log f(y|x,\theta)$, to be smooth in θ near θ_0. We assume:

ASSUMPTION M2: In some neighborhood N_1 of θ_0, the score function $s(y|x,\theta)$ and its partial derivative $\frac{\partial}{\partial\theta}s(y|x,\theta)$ is continuous in θ and dominated by a square integrable function $r(y,x)$ and an integrable function $\bar{r}(y,x)$, respectively. Specifically, $|s(y|x,\theta)_\ell| \le r(y,x)$ and $\left|\frac{\partial}{\partial\theta_k}s(y|x,\theta)_\ell\right| \le \bar{r}(y,x)$, $\forall\theta \in N_1$, $\forall k, \ell = 1, \ldots, L$, where $\int_X \sup_{\theta\in N_1} \int_Y [r^2(y,x) + \bar{r}(y,x)]f(y|x,\theta)d\mu(y)dP_X(x) < \infty$.

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 θ_0. Of this estimator, we assume:

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

$$(2.6)\qquad \sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_0^{-1}\psi(Y_i, X_i, \theta_0) + o_p(1) \quad \text{as} \quad n \to \infty,$$

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 $\int_Y \psi(y, x, \theta_0)f(y|x, \theta_0)d\mu(y) = \underline{0}$,

$\forall x \in X$, and $V_0 = E_P \psi(Y, X, \theta_0)\psi(Y, X, \theta_0)'$ and $D_0 = -E_P \frac{\partial}{\partial \theta}\psi(Y, X, \theta_0)$ are finite and non-singular.

Under assumption E1, $\sqrt{n}(\hat{\theta} - \theta_0)$ has 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 $\frac{\partial}{\partial \theta}\log f(y|x,\theta)$, $D_0 = V_0$, and D_0 is the information matrix evaluated at θ_0 . For M-estimators, $\psi(y,x,\theta)$ is just the defining function of the estimator (e.g., see Andrews (1986b, Theorem 1)). In other cases, e.g., with method of moment estimators, $\psi(y,x,\theta)$ 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 $\nu_n(\hat{\Gamma},\hat{\theta})$, we use a smoothness condition on $\psi(y,x,\theta)$ in θ near θ_0 :

ASSUMPTION E2: In some neighborhood N_2 of θ_0 , $\psi(y,x,\theta)$ and its first and second partial derivatives (with respect to θ) are continuous in θ and are dominated by the functions $r_0(y,x)$, $r_1(y,x)$, and $r_2(y,x)$, respectively. Specifically, $|\psi(y,x,\theta)_\ell| \le r_0(y,x)$,

$\left|\frac{\partial}{\partial \theta_k}\psi(y,x,\theta)_\ell\right| \le r_1(y,x)$, and $\left|\frac{\partial^2}{\partial \theta_k \partial \theta_m}\psi(y,x,\theta)_\ell\right| \le r_2(y,x)$, $\forall \theta \in N_2$, $\forall k, \ell, m = 1, \ldots, L$, where

$$\int_X \sup_{\theta \in N_1} \int_Y [r_0^2(y,x)r(y,x) + r_1(y,x)[r_0(y,x) + r(y,x)] + r_2(y,x)]$$

$$\cdot f(y|x, \theta_0)d\mu(y)dP_X(x) < \infty .$$

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 $\Gamma \in \mathbf{D}$ as $n \to \infty$. To make this assumption meaningful a topology needs to be defined on $\mathbf{D}$. Let F denote the joint distribution of (Y_i, X_i) under P. That is, for all measurable sets C in $\mathbf{Y} \times \mathbf{X}$,

$$F(C) = \int_C \int f(y|x, \theta_0) d\mu(y) dP_X(x) . \tag{2.7}$$

Equip $\mathbf{C}$ with the topology generated by the $L^2(F)$ semi-norm, and give $\mathbf{D}$ the corresponding product topology. With this topology on $\mathbf{C}$, two sets C_1 and C_2 in $\mathbf{Y} \times \mathbf{X}$ are close if $F(C_1 \tilde{\Delta} C_2)$ is small, where $\tilde{\Delta}$ denotes the symmetric difference operator. With this topology on $\mathbf{D}$, convergence in probability of the random elements $\hat{\Gamma}$ to Γ means that for all $\varepsilon > 0$,

$$\begin{aligned} &P\left[F(\hat{\Gamma}_j \tilde{\Delta} \Gamma_j) > \varepsilon\right] \xrightarrow{n\to\infty} 0 , \quad \forall j = 1, \ldots, J , \quad \text{or equivalently,} \\ &F(\hat{\Gamma}_j \tilde{\Delta} \Gamma_j) \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty , \quad \forall j = 1, \ldots, J . \end{aligned} \tag{2.8}$$

ASSUMPTION RC1: $\hat{\Gamma} \xrightarrow{P} \Gamma$ as $n \to \infty$.

This assumption is fulfilled by all of the random cell choices discussed above and in Andrews (1985a). 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 $\mathbf{D}$ converges weakly (as a stochastic process) to a particular tied-down Gaussian process indexed by elements of $\mathbf{D}$. This convergence result does not hold if $\mathbf{D}$ is allowed to contain all measurable partitions of $\mathbf{Y} \times \mathbf{X}$. For the result to hold, we restrict $\mathbf{C}$, and in consequence, $\mathbf{D}$, as follows:

ASSUMPTION RC2: $\mathcal{C}$ is a Vapnik-Cervonenkis (VC) class.

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.

By definition, the class $\mathcal{C}$ 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 $Y \times X$, $\mathcal{C}$ 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 $C \in \mathcal{C}$.

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 partition 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 $\mathcal{C}_1, \ldots, \mathcal{C}_B$, the class of all unions, intersections, differences, and complements of sets in $\mathcal{C}_1, \ldots, \mathcal{C}_B$ 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 in various contexts.

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_i, X_i) constitute a finite dimensional vector space. Hence, cells that are based on residuals that are polynomial functions of the elements of (Y_i, X_i) 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 adopted in this paper.

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 two candidates for the weight matrix used by the test, 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. These steps are outlined as follows: (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

$$(3.1) \qquad q_n(\hat{\Gamma},\hat{\theta}) = \left[\nu_n(\hat{\Gamma},\ \theta_0) - \sqrt{n}\Delta_0'D_0^{-1}\bar{\psi}_n\right]'\Sigma_0^{-}\left[\nu_n(\hat{\Gamma},\ \theta_0) - \sqrt{n}\Delta_0'D_0^{-1}\bar{\psi}_n\right] ,$$

Σ_0 (defined below) is the asymptotic covariance matrix of $\nu_n(\hat{\Gamma},\hat{\theta})$, $\bar{\psi}_n = \frac{1}{n}\sum_{i=1}^{n}\psi(Y_i,\ X_i,\ \theta_0)$, and Δ_0 is defined below. (ii) Next we prove that the stochastic process $(\nu_n(\cdot,\ \theta_0),\ \sqrt{n}\ \bar{\psi}_n,\ \hat{\Gamma})$ indexed by partitions $\gamma \in D$ converges weakly to $(\nu(\cdot),\ \psi,\ \Gamma)$, where $\nu(\cdot)$ is a tied-down Gaussian process indexed by $\gamma \in D$, ψ is a multivariate normal J-vector, and Γ ($\in D$) is the fixed limit of the random partitions $\hat{\Gamma}$. (iii) The desired result is obtained by showing that $q_n(\hat{\Gamma},\hat{\theta})$ is a continuous function $c(\cdot,\cdot,\cdot)$ of $(\nu_n(\cdot,\ \theta_0),\ \sqrt{n}\ \bar{\psi}_n,\ \hat{\Gamma})$. 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 Σ_0.

Since it is shown below that $\nu_n(\hat{\Gamma},\hat{\theta})$ has the same asymptotic distribution as $\nu_n(\Gamma,\ \theta_0) - \sqrt{n}\Delta_0'D_0^{-1}\bar{\psi}_n$, the asymptotic covariance matrix Σ_0 of $\nu_n(\hat{\Gamma},\hat{\theta})$ is given by

$$(3.2) \qquad \Sigma_0 = \Lambda_0 - H_0 - \Delta_0'D_0^{-1}\Pi_0 - (\Delta_0'D_0^{-1}\Pi_0)' + \Delta_0'D_0^{-1}V_0(D_0^{-1})'\Delta_0 ,$$

where $\Lambda_0 = E_P\Gamma(Y,X)\Gamma(Y,X)'$, $H_0 = E_PF(\Gamma,\ X,\ \theta_0)F(\Gamma,\ X,\ \theta_0)'$, $\Delta_0 = E_P\ \frac{\partial}{\partial\theta}\log\ f(Y|X,\ \theta_0)\Gamma(Y,X)'$, $\Pi_0 = E_P\psi(Y,\ X,\ \theta_0)[\Gamma(Y,X) - F(\Gamma,\ X,\ \theta_0)]'$,

D_0 and V_0 are as in Assumption E1, and $F(\Gamma,X,\theta)$ is defined in (2.3).

If $\hat{\theta}$ is the ML estimator (or any other asymptotically efficient estimator), then the covariance matrix Σ_0 simplifies to

$$\Sigma_0 = \Lambda_0 - H_0 - \Delta_0' V_0^{-1} \Delta_0 \ , \tag{3.3}$$

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

To begin the derivation of the asymptotic distribution of $X_n^2(\hat{\Gamma},\hat{\theta})$, 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* $\Delta_n(\gamma,\theta) = \frac{1}{n} \sum_{i=1}^{n} \int_Y \frac{\partial}{\partial \theta} \log f(y|X_i,\theta) \gamma(y,X_i)' f(y|X_i,\theta) d\mu(y)$.

LEMMA 2: *Under assumptions* M1-M2 *and* RC1-RC2,

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

See the Appendix for proofs.

These two lemmas and Assumption E1 immediately give

$$\nu_n(\hat{\Gamma},\hat{\theta}) - [\nu_n(\hat{\Gamma}, \theta_0) - \sqrt{n}\Delta_0' D_0^{-1} \bar{\psi}_n] = o_p(1) \ , \quad \text{as} \quad n \to \infty \ . \tag{3.4}$$

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

$$\hat{W} \xrightarrow{P} \Sigma_0^- \quad \text{as} \quad n \to \infty \ , \tag{3.5}$$

for some generalized inverse $\bar{\Sigma}_0^-$ of Σ_0. In this case, (3.4) gives

$$X_n^2(\hat{\Gamma}, \hat{\theta}) = q_n(\hat{\Gamma}, \hat{\theta}) + o_p(1) \quad \text{as} \quad n \to \infty \,. \tag{3.6}$$

Hence, it suffices to establish the asymptotic distribution of $\nu_n(\hat{\Gamma}, \theta_0) - \sqrt{n}\Delta_0' D_0^{-1} \bar{\psi}_n$. With this in mind, we have

LEMMA 3: *Under assumptions* M1, E1, *and* RC2,

$$\nu_n(\cdot, \theta_0) \xrightarrow{L} \nu(\cdot) \quad \textit{as a process on} \quad D \,, \quad \textit{as} \quad n \to \infty \,,$$

where $\xrightarrow{L}$ *denotes weak convergence (or convergence in law or distribution), and* $\nu(\cdot)$ *is an* R^J*-valued tied-down Gaussian process with bounded uniformly continuous sample paths (almost surely) and covariance structure given by*

$$\begin{aligned} &E_P \nu(\gamma) = 0 \,, \quad \forall \gamma \in D \,, \\ &E_P \nu(\gamma)\nu(\tilde{\gamma})' = E_P \gamma(Y,X)\tilde{\gamma}(Y,X)' - E_P F(\gamma, X, \theta_0) F(\tilde{\gamma}, X, \theta_0) \,, \quad \forall \gamma, \tilde{\gamma} \in D \,, \end{aligned} \tag{3.7}$$

where $F(\cdot, X, \theta)$ *is defined in* (2.3).

COMMENTS: 1. We call the limit process $\nu(\cdot)$ an F-trampoline, where F denotes the joint distribution of (Y_i, X_i). 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_i, X_i) take values in the unit square, the class C just contains sets of

the form $([0,s], [0,t])$, 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 case discussed 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_i, X_i) gives probability mass to any discrete points. For the purposes at hand, however, $L^2(F)$-continuity of the sample paths is the appropriate form of continuity, since $\hat{\Gamma}$ converges in probability to Γ 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^J-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), Gine 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. This follows because we consider a function of $\nu(\cdot)$ below, that is a continuous function 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 $\sqrt{n}\,\bar{\psi}_n$ and the assumption $\hat{\Gamma} \xrightarrow{P} \Gamma$, we find that $\{\sqrt{n}\,\bar{\psi}_n : n = 1, 2, \ldots\}$ and $\{\hat{\Gamma} : n = 1, 2, \ldots\}$ also are uniformly tight. Hence, $\{(\nu_n(\cdot, \theta_0), \sqrt{n}\,\bar{\psi}_n, \hat{\Gamma}) : n = 1, 2, \ldots\}$ viewed as stochastic processes on D are uniformly tight. By the central limit theorem and the assumption $\hat{\Gamma} \xrightarrow{P} \Gamma$, all of the finite dimensional distributions of this process converge weakly to those of $(\nu(\cdot), \psi, \Gamma)$, where $\psi \sim N(\underset{\sim}{0}, V_0)$, $V_0 = E_P\psi(Y_i, X_i, \theta_0)\psi(Y_i, X_i, \theta_0)'$, and

$$(3.8) \qquad E_P\psi\nu(\gamma)' = \Pi(\gamma) = E_P\psi(Y_i, X_i, \theta_0)[\gamma(Y_i, X_i) - F(\gamma, X_i, \theta_0)]' .$$

That is, for all $\gamma \in D$

$$(3.9) \qquad (\nu_n(\gamma, \theta_0), \sqrt{n}\,\bar{\psi}_n, \hat{\Gamma}) \xrightarrow{L} (\nu(\gamma), \psi, \Gamma) \quad \text{as} \quad n \to \infty .$$

These results imply

$$(3.10) \qquad (\nu_n(\cdot, \theta_0), \sqrt{n}\,\bar{\psi}_n, \hat{\Gamma}) \xrightarrow{L} (\nu(\cdot), \psi, \Gamma) \quad \text{as a process on } D \text{ as } n \to \infty ,$$

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

To make use of (3.10), the next result shows that $q_n(\hat{\Gamma}, \hat{\theta})$ is a

continuous function of $(\nu(\cdot, \theta_0), \sqrt{n}\,\bar{\psi}_n, \hat{\Gamma})$ with probability one:

LEMMA 4: *The function* $c(z,w,\gamma)$ *defined by*
$c(z,w,\gamma) = (z(\gamma) - \Delta_0'D_0^{-1}w)'\Sigma_0^-(z(\gamma) - \Delta_0'D_0^{-1}w)$ *for* $z \in g(D)$, $w \in R^J$, $\gamma \in D$, *and arbitrary generalized inverse* Σ_0^- *of* Σ_0, *is continuous (with respect to the product topology on* $g(D) \times R^J \times D$ *) at all points* (z,w,γ) *for which* z *is uniformly continuous.*

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

$$(3.11) \qquad q_n(\hat{\Gamma}, \hat{\theta}) \xrightarrow{L} (\nu(\Gamma) - \Delta_0'D_0^{-1}\psi)'\Sigma_0^-(\nu(\Gamma) - \Delta_0'D_0^{-1}\psi) \quad \text{as} \quad n \to \infty .$$

Since $\mathrm{Var}(\nu(\Gamma) - \Delta_0'D_0^{-1}\psi) = \Sigma_0$, the distribution of the limit above is chi-square with degrees of freedom equal to the rank of Σ_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* Σ_0^- *of* Σ_0. *Then, under assumptions* M1-M2, E1-E2, *and* RC1-RC2,

$$X_n^2(\hat{\Gamma}, \hat{\theta}) \xrightarrow{L} \chi^2_{\Sigma_0} \quad \text{as} \quad n \to \infty ,$$

where $\chi^2_{\Sigma_0}$ *is the chi-square distribution with* $\mathrm{rk}[\Sigma_0]$ *degrees of freedom.*

Let G be the maximal number of groups of cells in Γ such that each covariate value $x \in X$ belongs to cells in one and only one group. For exam-

ple, 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. Or, if the cells partition $Y \times X$ based on residuals, then G equals one. The groups are defined such that the sum of $\Gamma_j(Y_i, X_i) - F(\Gamma_j, X_i, \hat{\theta})$ over all the cells j in any one group is necessarily zero. Thus, if $\underset{\sim}{1}_g$ denotes the J-vector with ones for the elements corresponding to cells in the g^{th} group and zeroes elsewhere, then $\nu_n(\hat{\Gamma},\hat{\theta})'\underset{\sim}{1}_g = 0$ for $g = 1, \ldots, G$. In consequence, $rk[\Sigma_0] \leq J-G$.

In fact, the rank of Σ_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 θ_0.

The principal case where $rk[\Sigma_0] < J-G$ is when $\hat{\theta}$ is a <u>minimum chi-square estimator</u> or some asymptotic equivalent. By definition, a minimum chi-square estimator minimizes the chi-square statistic $\nu_n(\hat{\Gamma},\hat{\theta})'B(\theta)\nu_n(\hat{\Gamma},\theta)$ formed using some $J \times J$ weight matrix $B(\cdot)$. Two examples where estimators are used that are asymptotically equivalent to minimum chi-square estimators are: (1) Pearson's chi-square statistic with the multinomial ML estimator, and (2) 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. In these special cases, the rank of Σ_0 is $J-1-L$ and $J-G-L$, respectively, where G is the number different covariate values. 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 Σ_0 is $J-G$.

3.2. The Weight Matrix

Next we consider the choice of weight matrix $\hat{W}$. Two sample analogues of Σ_0 are introduced, and are shown to be consistent for Σ_0. Under an additional condition, the Moore-Penrose inverses of these estimators are shown to be consistent for the Moore-Penrose inverse of Σ_0. Hence, we have two suitable candidates for $\hat{W}$.

The second result of this section shows that under fairly general conditions, any generalized inverse can replace the Moore-Penrose inverse of the consistent estimator of Σ_0 without affecting the asymptotic distribution of $X_n^2(\hat{\Gamma},\hat{\theta})$. As special cases, this result yields the asymptotic distribution of Pearson's (1900), McFadden's (1974), Moore and Spruill's (1975), Heckman's (1984), and Horowitz's (1985) chi-square test statistics.

The final result of this section provides a simplified computational procedure for calculating $X_n^2(\hat{\Gamma},\hat{\theta})$.

Let

$$\Lambda_n(\gamma,\theta) = \frac{1}{n}\sum_{i=1}^{n} \operatorname{diag}(F(\gamma,X_i,\theta)) \ , \quad H_n(\gamma,\theta) = \frac{1}{n}\sum_{i=1}^{n} F(\gamma,X_i,\theta)F(\gamma,X_i,\theta)' \ ,$$

$$\Pi_n(\gamma,\theta) = \frac{1}{n}\sum_{i=1}^{n} F(\psi\gamma', X_i, \theta) \ , \quad \Delta_{1n}(\gamma,\theta) = \frac{1}{n}\sum_{i=1}^{n} F\left(\left[\frac{\partial}{\partial\theta}\log f\right]\gamma', X_i, \theta\right) \ ,$$

(3.12)

$$D_{1n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} F\left(\left[\frac{\partial}{\partial\theta}\psi\right]', X_i, \theta\right) \ , \quad V_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} F(\psi\psi', X_i, \theta) \ , \quad \text{and}$$

$$\Sigma_{1n}(\gamma,\theta) = \Lambda_n - H_n - \Delta_{1n}'D_{1n}^{-1}\Pi_n - (\Delta_{1n}'D_{1n}^{-1}\Pi_n)' + \Delta_{1n}'D_{1n}^{-1}V_n(D_{1n}^{-1})'\Delta_{1n} \ ,$$

where $F(\cdot, X_i, \theta)$ is defined in (2.3), and for simplicity Λ_n abbreviates $\Lambda_n(\gamma,\theta)$, etc.

Next, let

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

$$b_i = \gamma(Y_i,X_i) - F(\gamma,X_i,\theta) - \Delta_{2n}(\gamma,\theta)'D_{2n}^{-1}(\theta)\psi(Y_i,X_i,\theta) \quad , \quad \text{and} \tag{3.13}$$

$$\Sigma_{2n}(\gamma,\theta) = \frac{1}{n}\sum_{i=1}^{n} b_i b_i' \quad .$$

The two estimators of Σ_0 that we consider are $\hat{\Sigma}_1 = \Sigma_{1n}(\hat{\Gamma},\hat{\theta})$ and $\hat{\Sigma}_2 = \Sigma_{2n}(\hat{\Gamma},\hat{\theta})$. Both of these estimators are sample analogues of Σ_0, but $\hat{\Sigma}_1$ takes conditional expectations using the parametric model wherever possible, whereas $\hat{\Sigma}_2$ is a more pure sample analogue estimator. The relative attributes of the two estimators are unclear. $\hat{\Sigma}_1$ is usually more efficient than $\hat{\Sigma}_2$ under the null. 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 somewhat similar problem, see Efron and Hinkley (1978).

If the estimator $\hat{\theta}$ is the ML estimator (or any other asymptotically efficient estimator), the covariance matrix estimator $\hat{\Sigma}_1$ simplifies considerably. In this case, $\psi(Y_i, X_i, \theta) = \frac{\partial}{\partial\theta}\log f(Y_i|X_i, \theta)$, and so, $\Delta_{1n}(\gamma,\theta) = \Pi_n(\gamma,\theta)$ and the information matrix I_0 equals $D_0 = V_0$. If the Hessian estimator $D_{1n}(\hat{\theta})$ of I_0 is replaced in $\hat{\Sigma}_1$ by the outer product estimator $V_n(\hat{\theta})$ of I_0, then we get

$$\hat{\Sigma}_1 = \Lambda_n(\hat{\Gamma},\hat{\theta}) - H_n(\hat{\Gamma},\hat{\theta}) - \Delta_{1n}(\hat{\Gamma},\hat{\theta})'V_n^{-1}(\hat{\theta})\Delta_{1n}(\hat{\Gamma},\hat{\theta}) \quad . \tag{3.14}$$

To establish consistency of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$, the following lemmas are used:

LEMMA 5: *Under assumptions* M1-M2 *and* E1-E2,

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

LEMMA 6: *Under assumptions* M1, E1, *and* RC1-RC2,

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

The weak law of large numbers and Slutsky's Theorem give

$$(3.15) \qquad \Sigma_{vn}(\Gamma, \theta_0) \xrightarrow{P} \Sigma_0 \quad \text{as} \quad n \to \infty , \quad \text{for} \quad v = 1, 2 .$$

This result, and Lemmas 5 and 6 combine to yield

$$(3.16) \qquad \hat{\Sigma}_v = \Sigma_{vn}(\hat{\Gamma},\hat{\theta}) \xrightarrow{P} \Sigma_0 \quad \text{as} \quad n \to \infty , \quad \text{for} \quad v = 1, 2 .$$

Consistency of $\hat{\Sigma}_v$ for Σ_0 does not imply consistency of $\hat{\Sigma}_v^+$ for Σ_0^+ (where $(\cdot)^+$ denotes the Moore-Penrose inverse), since the Moore-Penrose inverse is not a continuous function. In fact, if $rk[\hat{\Sigma}_v] \neq rk[\Sigma_0]$ with probability bounded away from zero as $n \to \infty$ (where $rk[\cdot]$ denotes the rank of a matrix), then $\|\hat{\Sigma}_v^+\|_2$ is stochastically unbounded, see Andrews (1986a, Theorem 2). If $rk[\hat{\Sigma}_v] = rk[\Sigma_0]$ with probability that converges to one as $n \to \infty$, however, then Andrews (1986a, Theorem 2) gives the desired result $\hat{\Sigma}_v^+ \xrightarrow{P} \Sigma_0^+$ as $n \to \infty$, for $v = 1, 2$. This proves the following Theorem:

THEOREM 2: *Under assumptions* M1-M2, E1-E2, *and* RC1-RC2, *if* $P(rk[\hat{\Sigma}_v] = rk[\Sigma_0]) \xrightarrow{n\to\infty} 1$, *then*

$$\hat{\Sigma}_v^+ \xrightarrow{P} \Sigma_0^+ \quad as \quad n \to \infty , \quad for \quad v = 1, 2 .$$

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

2. Let $\hat{G}$ 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 $rk[\Sigma_0] = J-G$ and $\hat{G} \xrightarrow{P} G$ as $n \to \infty$, then $P(rk[\hat{\Sigma}_v] = rk[\Sigma_0]) \to 1$ as $n \to \infty$ necessarily is satisfied, for $v = 1, 2$. This follows because $\hat{\Sigma}_v \xrightarrow{P} \Sigma_0$ implies $P(rk[\hat{\Sigma}_v] \geq rk[\Sigma_0]) \to 1$ as $n \to \infty$, and $rk[\hat{\Sigma}_v] \leq J-\hat{G}$, for all n, since $\hat{\Sigma}_v$ is orthogonal to $\hat{1}_g$, for $g = 1, \ldots, \hat{G}$, where $\hat{1}_g$ 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 $\hat{W} = \hat{\Sigma}_1^+$ or $\hat{W} = \hat{\Sigma}_2^+$. 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 Σ_0 and $P(\nu_n(\hat{\Gamma},\hat{\theta}) \in M(\hat{\Sigma}), \hat{\Sigma} = \hat{\Sigma}') \xrightarrow{n\to\infty} 1$, where $M(\cdot)$ denotes the column space of a matrix. In this case, the quadratic form $\nu_n(\hat{\Gamma},\hat{\theta})'\hat{\Sigma}^-\nu_n(\hat{\Gamma},\hat{\theta})$ 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, $\nu_n(\hat{\Gamma},\hat{\theta})$ can be written as $\hat{\Sigma}\xi_n(\hat{\Gamma},\hat{\theta})$, for some vector $\xi_n(\hat{\Gamma},\hat{\theta})$, and so,

$$(3.17) \qquad P\left[\nu_n(\hat{\Gamma},\hat{\theta})'\hat{\Sigma}^-\nu_n(\hat{\Gamma},\hat{\theta}) = \xi_n(\hat{\Gamma},\hat{\theta})'\hat{\Sigma}\xi_n(\hat{\Gamma},\hat{\theta})\right] \xrightarrow{n\to\infty} 1 .$$

The right-hand-side of the equality in (3.17) does not depend on the choice of 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* Σ_0 , $P(rk[\hat{\Sigma}] = rk[\Sigma_0], \nu_n(\hat{\Gamma},\hat{\theta}) \in M(\hat{\Sigma}), \hat{\Sigma} = \hat{\Sigma}') \xrightarrow{n\to\infty} 1$, *and* $\hat{W}$ *is taken to be* $\hat{\Sigma}^-$ *for any choice of generalized inverse* $(\cdot)^-$, then

$$X_n^2(\hat{\Gamma},\hat{\theta}) \xrightarrow{L} \chi^2_{\Sigma_0} \quad as \quad n \to \infty ,$$

when the null hypothesis P *is true.*

COMMENTS: 1. If $rk[\hat{\Sigma}] = J-\hat{G}$ and $\hat{\Sigma}$ is orthogonal to $\underset{\sim}{\hat{1}}_g$ for $g = 1, \ldots, \hat{G}$, 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 $\underset{\sim}{\hat{1}}_g$ for $g = 1, \ldots, \hat{G}$. 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 $rk[\Sigma_0] = J-G$, $\hat{G} \xrightarrow{P} G$, and $\hat{\Sigma}$ is taken to be $\hat{\Sigma}_1$ or $\hat{\Sigma}_2$, then the conditions of the Corollary on $\hat{\Sigma}$ are satisfied, since $\hat{\Sigma}$ is consistent by (3.16), $P(rk[\hat{\Sigma}] = rk[\Sigma_0]) \xrightarrow{n\to\infty} 1$ by Comment 2 of Theorem 2, $\hat{\Sigma} = \hat{\Sigma}'$ for all n , and $P(\nu_n(\hat{\Gamma},\hat{\theta}) \in M(\hat{\Sigma})) \xrightarrow{n\to\infty} 1$ 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 $\Lambda_n(\hat{\Gamma},\hat{\theta})^{-1}$, which is a generalized inverse (in the situations they consider) of the consistent estimator $\Sigma_{1n}(\hat{\Gamma},\hat{\theta})$ of Σ_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), $D_{2n}(\hat{\theta})$ can be replaced by $D_{3n}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f(Y_i|X_i, \hat{\theta}) \frac{\partial}{\partial\theta}\log f(Y_i|X_i, \hat{\theta})'$. 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 $\hat{\Gamma}(Y_i, X_i)' - F(\hat{\Gamma}, X_i, \hat{\theta})'$, let B be the $n \times L$ matrix with i^{th} row $\frac{\partial}{\partial\theta}\log f(Y_i|X_i, \hat{\theta})'$, and take $M_B = I_n - B(B'B)^{-1}B'$. Then, $\hat{\Sigma}_2 = \frac{1}{n}(M_B A)'M_B A$. Since $\underset{\sim}{1}'B = \underset{\sim}{0}$, we have $X_n^2(\hat{\Gamma}, \hat{\theta}) = \underset{\sim}{1}'A(A'M_B A)^{+}A'\underset{\sim}{1} = \underset{\sim}{1}'M_B A(A'M_B A)^{+}(M_B A)'\underset{\sim}{1}$. That is, $X_n^2(\hat{\Gamma}, \hat{\theta})$ is the sum of squared residuals from the projection of $\underset{\sim}{1}$ on the space spanned by $M_B A$.

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

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

$$(3.18) \qquad X_n^2(\hat{\Gamma},\hat{\theta}) = \underset{\sim}{1}'H(H'H)^{+}H'\underset{\sim}{1} \ .$$

That is, $X_n^2(\hat{\Gamma},\hat{\theta})$ is the explained sum of squares from the regression of $\underset{\sim}{1}$ 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 $X_n^2(\hat{\Gamma},\hat{\theta})$ is obtained from standard regression packages by deleting the redundant columns of H and performing the indicated regression.)

4. LOCAL POWER AND ASYMPTOTIC OPTIMALITY

In this section we investigate the large sample power of random cell chi-square tests by considering their local power properties. We also discuss their consistency and asymptotic optimality properties, and compare the tests to others in the literature.

4.1. Local Power

The local alternatives considered here apply to all alternative distributions of Y_i given X_i. Suppose one is interested in an approximation to the power of the chi-square test for a particular sample size n_0, and against an alternative conditional distribution of Y_i given X_i defined by the density $q(y|x)$ with respect to some σ-finite measure $\tilde{\mu}$. (Without loss of generality, assume $\tilde{\mu}$ dominates μ.) Since $q(\cdot|\cdot)$ is an alternative density, $q(\cdot|\cdot) \notin \{f(\cdot|\cdot,\theta) : \theta \in \Theta\}$. The marginal distribution of X under the alternative of interest is arbitrary, just as under the null, so we adopt the same notation for it, viz., $P_X(\cdot)$. 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_i given X_i, and marginal distribution P_X of X_i.

Define a function of scaled deviations $d(y, x, \theta_0)$ of the conditional density under the null hypothesis, from the conditional density under the alternative of interest:

$$(4.1) \qquad d(y, x, \theta_0) = \sqrt{n_0}(q(y|x) - f(y|x, \theta_0)) \ .$$

Next, define the following sequence of local alternative conditional densities:

$$(4.2) \qquad q_n(y|x) = f(y|x, \theta_0) + d(y, x, \theta_0)/\sqrt{n} \ , \quad \text{for} \quad n = 1, 2, \ldots \ .$$

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_i given X_i and marginal distribution P_X of X_i.

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 of interest 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 we use to establish the asymptotic local power results are analogous to those used in Section 3 to determine asymptotic results under 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* σ*-finite measure* $\tilde{\mu}$ *) of* Y_i *given* X_i *and marginal distribution* P_X *of* X_i .

ASSUMPTION M2': *The parametric conditional densities* $f(y|x,\theta)$ *satisfy* M2, *and* $E_Q[r^2(Y,X) + \bar{r}(Y,X)] < \infty$.

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* $E_Q\left[r_0(Y,X)[r_0(Y,X) + r_1(Y,X)] + r_2(Y,X)\right] < \infty$.

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 $\{Q_n : n = 1, 2, \ldots\}$.

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 often is 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*

$$X_n^2(\hat{\Gamma},\hat{\theta}) \xrightarrow{L} \chi^2_{\Sigma_0}(\delta) \quad \text{as} \quad n \to \infty ,$$

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 $\chi^2_{\Sigma_0}(\delta)$ *denotes the non-central chi-square distribution with* $rk[\Sigma_0]$ *degrees of freedom and non-centrality parameter* δ .

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

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

(d) *Given any estimator* $\hat{\Sigma}$ *that satisfies* $\hat{\Sigma} \xrightarrow{Q_n} \Sigma_0$ *as* $n \to \infty$ *and* $Q_n(rk[\hat{\Sigma}] = rk[\Sigma_0], \nu_n(\hat{\Gamma},\hat{\theta}) \in M(\hat{\Sigma}), \hat{\Sigma} = \hat{\Sigma}') \xrightarrow{n\to\infty} 1$, *if* $\hat{W} = \hat{\Sigma}^-$, *for any generalized inverse* $(\cdot)^-$, *then*

$$X_n^2(\hat{\Gamma},\hat{\theta}) \xrightarrow{L} \chi^2_{\Sigma_0}(\delta) \quad \text{as} \quad n \to \infty ,$$

under the sequence of local alternatives $\{Q_n : n = 1, 2, \ldots\}$.

COMMENTS: 1. The null and local power approximations for the Pearson chi-square test, with and without adjusted weight matrix and random cells, have been found via Monte Carlo experimentation to be remarkably accurate in models

without covariates. These results appear in a long series of papers in the statistical literature (see Andrews (1985b) for a brief survey). Although these results are suggestive for models with covariates, the extent to which they carry over is an open question that will require a similarly large array of Monte Carlo results to answer.

2. It is not possible to give general results stating that the use of either the minimum chi-square or ML estimator, to form the chi-square statistic, dominates the other in terms of local power. For 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 and Asymptotic Optimality

We now state the consistency properties of chi-square tests. Let $Q = Q(y|x)$ denote a conditional distribution of Y given X that is not in the parametric family $\{f(y|x,\theta) : \theta \in \Theta\}$. Let Γ_1 denote the limit partition of $\hat{\Gamma}$, and θ_1 the limit vector of $\hat{\theta}$, under Q and P_X as $n \to \infty$. If

$$(4.3) \qquad \int_X \int_Y \Gamma_{1j}(y,x) f(y|x, \theta_1) d\mu(y) dP_X(x) \neq \int_X \int_Y \Gamma_{1j}(y,x) dQ(y|x) dP_X(x) ,$$

for some $j = 1, \ldots, J$, then the chi-square test is consistent against Q. (See Andrews (1985b) 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.

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)).[2]

For composite null hypotheses, general 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 sub-class of alternatives, such as a W, LR, or LM test, or one can choose a procedure that promises to have good power against a wider variety of alternatives, though not necessarily optimal power over any parametric sub-class, such as a chi-square test. For models without covariates this option has been available for many years, and both types of procedures have been used extensively for such purposes. For models with covariates, however, general test procedures for the latter option have not been available until quite recently (e.g., see Heckman (1984), Horowitz (1985), Newey (1985), Tauchen (1985), and White (1982)). A goal of this paper is to make such tests (and their accompanying diagnostics) 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.

We mention that the consistency properties of the classical W, LR, and LM tests vis-a-vis those of chi-square tests illustrate that neither dominates

the other in general, for nonparametric classes of alternatives. The reason is simply that each is consistent for certain alternatives for which the other is inconsistent. Thus, depending on the loss ascribed to type II errors for different alternatives, either chi-square or classical W, LR, or LM tests may be preferred in a given situation where a nonparametric class of alternatives is of interest.

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 $\mathbf{Y} \times \mathbf{X}$, 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 increasingly high 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. And 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 holds for tests of univariate and multivariate normality (or any other specified distribution) of the errors in linear single equation, seemingly unrelated, and multivariate regression models with intercept terms, provided the cells $\hat{\Gamma}$ are determined by the residuals alone, and θ is estimated by least squares, maximum likelihood, pseudo-maximum likelihood, Zellner's feasible Aitken estimator, a classical M-estimator, or any estimator whose function $\psi(Y,X,\theta)$ is of the form $\begin{bmatrix} \psi_1(Y - X\theta_1, \theta_2)\begin{pmatrix} 1 \\ X \end{pmatrix} \\ \psi_2(Y - X\theta_1, \theta_2) \end{bmatrix}$, 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 is shown in the Appendix that the chi-square test formed using the weight matrix $\hat{\Sigma}_1^+$ 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 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 Γ_j contains more than one value of the covariates. If we are interested in the predictive accuracy of the model for the cells Γ , a more inclusive null hypothesis than $H_0 : f(y|x) = f_0(y|x)$ is appropriate. Specifically, we consider $H_0' : \underset{\sim}{f} = \underset{\sim}{f}_0$, where, given any conditional distribution $h(y|x)$, $\underset{\sim}{h}$ denotes the $(J-G)$-dimensional vector of

conditional probabilities under $h(y|x)$ of the cells in Γ (with G redundant cells, due to conditional probabilities summing to one, omitted). In this context, the chi-square test based on $X_n^2(\hat{\Gamma})$ is asymptotically equivalent to certain Wald tests under the null and local alternatives, and possesses various nonparametric asymptotic optimality properties.

To elaborate, given any conditional distribution $h(y|x)$, consider the parametric family:

$$(4.4) \qquad H_h = \{h(y|x, \underset{\sim}{p}) : \underset{\sim}{p} \in P\},$$

where $h(y|x, \underset{\sim}{p}) = \prod_{j=1}^{J} \left[h(y|x) \frac{p_j}{h_j} \right]^{\Gamma_j(y,x)}$, $\underset{\sim}{p} = (p_1, \ldots, p_{J-G})'$, h_j is the j^{th} element of $\underset{\sim}{h}$,

$$P = \left\{ \underset{\sim}{p} \in R^{J-G} : 0 < p_j < 1, \ \forall j; \ 0 < \sum_{i=J_g}^{J_{g+1}-1} p_j < 1, \text{ for } g = 1, \ldots, G \right\}, \text{ and}$$

J_g denotes the index of the first cell involving the g^{th} value of the covariates. (It is assumed that the G cells of Γ 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 $H_0 : \underset{\sim}{f} = \underset{\sim}{f}_0$ is satisfied for this parametric family only if $\underset{\sim}{p} = \underset{\sim}{f}_0$. The Wald statistic for testing $\underset{\sim}{p} = \underset{\sim}{f}_0$ is a quadratic form in the vector $\hat{\underset{\sim}{p}} - \underset{\sim}{f}_0$, where $\hat{\underset{\sim}{p}}$ is the ML estimator of $\underset{\sim}{p}$. This test statistic is asymptotically equivalent to the chi-square test statistic $X_n^2(\hat{\Gamma})$ under the null and local alternatives, see the Appendix.

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 H_h, 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 uncountably infinite number of parametric families H_h. Since every alternative conditional density is included in some parametric family H_h, the chi-square test exhibits optimality properties that apply to the entire nonparametric class of alternatives to H_0'.

Lastly, we briefly discuss the relative attributes of White's (1982) information matrix (IM) test, Newey's (1985) conditional moments (CM) tests, and chi-square tests. The IM test has been advocated for use as a general specification test, so it is a natural alternative to chi-square tests. As shown in Chesher (1983), the IM test is a Lagrange multiplier test against a class of alternatives that maintain the structure of the specified parametric model, but allow the parameter vector to be random rather than fixed. For this class of alternatives, the IM test possesses Wald's asymptotic optimality properties. Thus, the choice between the IM test and a chi-square test, depends on whether one is more concerned with random coefficient alternatives or alternatives that yield predictive inaccuracy, and whether one is interested in the diagnostics produced by the chi-square test.

Newey (1985, p. 1059) suggests a CM test for use as a general specification test.[3] This test is based on a vector of products of exogenous variables and elements of the score function of the parametric model. This test often has power against a wide variety of alternatives. In certain directions, how-

ever, the test's power is low. For example, it has no power for testing normality in linear and nonlinear regression models, nor for testing multivariate normality in linear simultaneous equations models, and linear and nonlinear multivariate and seemingly unrelated regressions models. Such results lead one to question its power against non-normality in models based on normally distributed latent errors, such as censored, truncated, and switching regression models, and probit models.[4] In fact, when the true regression function in these models consists of just a constant, CM tests have no power.

In sum, Newey's general specification test has greater power, the greater is the correlation between the exogenous variables and the score function, as caused by misspecification. In contrast, chi-square tests have power that is greater, the greater is the predictive inaccuracy caused by misspecification. The choice between the two depends upon which alternatives are of more interest, and which diagnostics are more helpful.

Cowles Foundation, Department of Economics, Yale University

APPENDIX

We begin the Appendix by describing the measure theoretic framework used for the results of this paper. Let $\tilde{F}$ denote the Borel σ-algebra on $Y \times X \subset R^{V+K}$. The underlying probability space of the whole random experiment, whose outcome yields $\{(Y_i, X_i) : i = 1, 2, \ldots\}$, is denoted $(\tilde{\Omega}, B, P)$, where $\tilde{\Omega} = [Y \times X]^{\infty}$ (that is, $\tilde{\Omega}$ is the product space of an infinite number of copies of $Y \times X$), B is the σ-algebra on $\tilde{\Omega}$ generated by the infinite sequence of σ-algebras $\tilde{F}$, and P is the probability distribution given in assumption M1 above.

For given n, the random element $\hat{\Gamma}$ (which depends on n in general) is taken to be a map from $\tilde{\Omega}$ to D that is measurable with respect to the σ-algebra in $\tilde{\Omega}$ that is generated by the first n observations, and the Borel σ-algebra in D.

Let $g(D)$ be the set of all R^J-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 σ-field generated by the supremum norm. In the case of the standard empirical process, this has 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 σ-field than the Borel σ-field. We choose the σ-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 σ-field. Finally, as Dudley (1978) and Pollard (1984) show,

the basic asymptotic results we desire hold with this choice of σ-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 $S_n(\theta) = \sup_{\gamma \in D} |F_n(\gamma,\theta) - F_n(\gamma,\theta_0) - \Delta_n(\gamma,\theta_0)'(\theta-\theta_0)|$. We prove the result for an arbitrary element j of the vector $\sqrt{n}\, S_n(\theta)$. To simplify notation the subscript j is suppressed below. Suppose we can show:

$$(A.1) \qquad P\left[\overline{\lim_{n\to\infty}} \sup_{\theta \in N(\delta_n)} \sqrt{n}\, S_n(\theta) = 0\right] = 1 ,$$

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

$$(A.2) \qquad 1 = \underline{\lim_{n\to\infty}} P(A_{1n} \cap A_{2n}) \le \underline{\lim_{n\to\infty}} P(\sqrt{n}\, S_n(\hat{\theta}) \le \varepsilon) ,$$

and the Lemma is proved.

To show (A.1), we first note that for all θ in some neighborhood N_1

of θ_0 ,

$$\text{(A.3)} \qquad \left|\frac{\partial^2}{\partial\theta_k\partial\theta_m} f(y|x,\theta)\right| - \left|\frac{\partial}{\partial\theta_k} s(y|x,\theta)_m + s(y|x,\theta)_k s(y|x,\theta)_m\right| \cdot f(y|x,\theta)$$

$$\leq [\bar{r}(y,x) + r^2(y,x)] \cdot f(y|x,\theta) \ ,$$

where $\int_X \sup_{\theta\in N_1} \int_Y [\bar{r}(y,x) + r^2(y,x)] f(y|x,\theta) d\mu(y) dP_X(x) < \infty$, by M2.

This gives,

$$\sup_{\theta\in N(\delta_n)} \sqrt{n}\, S_n(\theta) = \sup_{\theta\in N(\delta_n)} \sup_{C\in\mathcal{C}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_Y 1_C(y, X_i) \cdot \left[f(y|X_i, \theta) - f(y|X_i, \theta_0) \right.\right.$$

$$\left.\left. - \frac{\partial}{\partial\theta'} f(y|X_i, \theta_0)(\theta - \theta_0) \right] d\mu(y) \right|$$

$$\leq \sup_{\theta\in N(\delta_n)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_Y \left| \frac{1}{2}(\theta - \theta_0)' \frac{\partial^2}{\partial\theta\partial\theta'} f(|X_i, \bar{\theta})(\theta - \theta_0) \right| d\mu(y) \ ,$$

$$\text{(A.4)} \qquad \text{for } \bar{\theta} \in N(\delta_n) \ ,$$

$$\leq \sup_{\theta\in N(\delta_n)} \frac{1}{2\sqrt{n}} \sum_{i=1}^{n} \int_Y |\theta-\theta_0|' \underset{\sim}{11}' |\theta-\theta_0| \cdot \sup_{k,m\leq L} \left| \frac{\partial^2}{\partial\theta_k\partial\theta_m} f(y|X_i, \bar{\theta}) \right| d\mu(y) \ ,$$

$$\leq \frac{L^2}{2n^{1/2+2a}} \sum_{i=1}^{n} \sup_{\theta\in N_1} \int_Y [\bar{r}(y,x) + r^2(y,x)] f(y|X_i, \theta) d\mu(y)$$

$$\xrightarrow[n\to\infty]{a.s.} 0 \ ,$$

where the third inequality holds for n sufficiently large and uses the definition of $N(\delta_n)$ 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 $\Delta_n(\Gamma, \theta_0) = \Delta_0 + o_p(1)$, and the proof of Lemma 6 below shows that $\Delta_n(\hat{\Gamma}, \theta_0) - \Delta_n(\Gamma, \theta_0) = o_p(1)$. Hence, $\Delta_n(\hat{\Gamma}, \theta_0) = \Delta_0 + o_p(1)$. *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 uniformly continuous functions, as desired. The empirical process and the limit P-bridge process can be replaced by the conditional empirical process $\nu_n(\cdot, \theta_0)$ and the conditional F-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 $\eta > 0$ and $\varepsilon > 0$ there exists a $\delta > 0$ such that

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

where $R_\delta = \{r : r = 1_{C_1} - 1_{C_2};\ C_1, C_2 \in C;$ and $F(C_1 \tilde{\Delta} C_2) < \delta^2\}$.

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 rv's $\{X_i : i = 1, 2, \ldots\}$, $\{Y_i : i = 1, 2, \ldots\}$, and $\{Y_i' : i = 1, 2, \ldots\}$ on an enlarged probability space, say (Ω, B, P) , such that $\{X_i\}$ are iid rv's with distribution P_X , and $\{Y_i\}$ and $\{Y_i'\}$ are sequences of independent rv's with conditional distributions of Y_i and Y_i' given X_i that are independent and have the same conditional density $f(y|X_i, \theta_0)$ (with respect to the measure μ) . Form two identically distributed conditional empirical processes $\nu_n(\cdot)$ and $\nu_n'(\cdot)$ indexed by sets in C , defined by

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

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

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 $F(C, X_i, \theta_0) = \int 1_C(y, X_i) f(y|X_i, \theta_0) d\mu(y)$ is the conditional probability

of C given X_i. (To simplify notation, and wlog, we suppress the dependence of $\nu_n(\cdot,\theta)$ on θ, 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

$$S_n = \left\{\omega \in \Omega : P(|\nu_n'(r)| \leq \eta/2 | (Y_i, X_i)) \geq 1/2, \ \forall r \in R_\delta\right\},$$

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

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

Define the random element τ ($= \tau(\nu_n)$) on the set $(\sup_r |\nu_n(r)| > \eta)$ such that τ takes values in R_δ and $|\nu_n(\tau)| > \eta$, where $\sup_r$ denotes $\sup_{r\in R_\delta}$. Then,

$$\begin{aligned} 1_{S_n} \cdot 1\left[\sup_r |\nu_n(r)| > \eta\right] &\leq 1\left[\sup_r |\nu_n(r)| > \eta\right] \cdot 2P\left[|\nu_n'(\tau)| \leq \eta/2 \,\middle|\, (Y_i, X_i)\right] \\ &= 2P\left[\sup_r |\nu_n(r)| > \eta, \ |\nu_n'(\tau)| \leq \eta/2 \,\middle|\, (Y_i, X_i)\right] \\ &\leq 2P\left[|\nu_n(\tau)| > \eta, \ |\nu_n'(\tau)| \leq \eta/2 \,\middle|\, (Y_i, X_i)\right] \\ &\leq 2P\left[\sup_r |\nu_n(r) - \nu_n'(r)| \geq \eta/2 \,\middle|\, (Y_i, X_i)\right], \end{aligned} \tag{A.7}$$

where the equality above holds because τ is a random function of $\nu_n(\cdot)$, and hence, (Y_i, X_i), only. Taking expectations yields

$$P\left[S_n \cap \sup_r |\nu_n(r)| > \eta\right] \leq 2P\left[\sup_r |\nu_n(r) - \nu_n'(r)| \geq \eta/2\right].$$

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:

$$(A.8) \qquad P\left[\sup_r |\nu_n(r)| > \eta\right] \leq 2P\left[\sup_r \sqrt{n}|P_n(r) - P_n'(r)| \geq \eta/2\right] + 1 - P(S_n) .$$

(The measurability difficulties overlooked in the argument above that establishes (A.8) need to be handled using the permissibility 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 δ small: The proof of tightness for the standard empirical process uses maximal inequalities of the form: Given $\eta > 0$ and $\varepsilon > 0$, $\exists \delta > 0$ such that

$$\limsup_{n\to\infty} P\left[\sup_{r\in R_\delta} \sqrt{n}|P_n(r) - P_n'(r)| \geq \eta\right] < \varepsilon ,$$

where $P_n(\cdot)$ and $P_n'(\cdot)$ are independent copies of the empirical measure (e.g., see Pollard (1984, Equicontinuity 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 $\{(Y_i', X_i) : i = 1, \ldots, n\}$ are no longer identically distributed. Fortunately, the tightness result for the standard empirical process (or symmetrized process) can be extended to independent non-identically distributed (inid) rv's without great difficulty, provided the marginal distributions do not fluctuate too widely (see Alexander (1984) for explicit treatment of the inid case). In particular, in the case

of indexing by a VC class, one only needs to have control of the variances of $r = 1_{C_1} - 1_{C_2}$, for all $r \in R_\delta$, for different observations i (e.g., see Theorem 2.8 of Alexander (1984)). That is, we need to show: Given $\varepsilon > 0$, $\exists \delta > 0$ such that

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

This follows, with probability one, because

$$(A.9) \qquad G_n \leq \sup_r \frac{1}{n} \sum_{i=1}^{n} F(r^2, X_i, \theta_0) \xrightarrow[n\to\infty]{a.s.} \sup_r F(r^2) \leq \delta^2 \ ,$$

using the uniform SLLN and the definition of R_δ , since $F(r^2) = F(C_1 \tilde{\Delta} C_2)$.

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

$$\limsup_{n\to\infty} P\left(\sup_{r \in R_\delta} \sqrt{n} |P_n(r) - P_n'(r)| > \eta \,\middle|\, \{X_i\} \right) < \varepsilon \ .$$

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

$$(A.10) \qquad \limsup_{n\to\infty} P\left(\sup_{r \in R_\delta} \sqrt{n} |P_n(r) - P_n'(r)| > \eta \right) < \varepsilon \ .$$

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

$$(A.11)\qquad P\left(\left|\nu_n'(r)\right| > \eta/2 \,\middle|\, (Y_i,\ X_i)\right) \le \frac{4}{n}\sum_{i=1}^{n} F(r^2,\ X_i,\ \theta_0)/\eta^2 \ , \quad \forall r \in R_\delta \ .$$

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

$$\sup_r \left|\frac{1}{n}\sum_{i=1}^{n} F(r^2,\ X_i,\ \theta_0) - F(r^2)\right| \le \int_{Y^\infty} \sup_r \left|\frac{1}{n}\sum_{i=1}^{n} r^2(Y_i,\ X_i) - F(r^2)\right| dP((Y_\ell))$$

$$\xrightarrow{n\to\infty} 0 \quad a.s.$$

Since $\sup_{r\in R_\delta} F(r^2) \le \delta^2$, the above results combine to give:

$$(A.12)\qquad P(S_n) \ge P\left[\frac{4}{n}\sum_{i=1}^{n} F(r^2,\ X_i,\ \theta_0) < \eta^2/2,\ \forall r \in R_\delta\right] \xrightarrow{n\to\infty} 1 \ ,$$

for δ sufficiently small. *Q.E.D.*

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 $\zeta > 0$ such that $|\gamma_1 - \gamma| < \zeta$ implies $|z(\gamma_1) - z(\gamma)| < \varepsilon/2$. We can choose a neighborhood $\tilde{N}$ of $(z(\cdot),\ \gamma)$ such that for all $(z_1(\cdot),\ \gamma_1)$ in $\tilde{N}$, we have $|\gamma_1 - \gamma| < \zeta$ and $\|z_1(\cdot) - z(\cdot)\| < \varepsilon/2$, where $\|\cdot\|$ denotes the sup norm. Then, $|z_1(\gamma_1) - z(\gamma)| \le |z_1(\gamma_1) - z(\gamma_1)| + |z(\gamma_1) - z(\gamma)| = \varepsilon$, and h is continuous at $(z(\cdot),\ \gamma)$. *Q.E.D.*

PROOF OF LEMMA 5: The desired result for $\Sigma_{1n}(\hat{\Gamma},\cdot)$ follows if we can show $\Lambda_n(\hat{\Gamma},\hat{\theta}) - \Lambda_n(\hat{\Gamma},\ \theta_0) = o_p(1)$, and the analogous results for $H_n(\hat{\Gamma},\cdot)$, $\Delta_{1n}(\hat{\Gamma},\cdot)$, $\Pi_n(\hat{\Gamma},\cdot)$, $D_{1n}(\cdot)$, and $V_n(\cdot)$. It suffices to show:

(A.13) $$\xi_n(\hat{\theta}) \equiv \frac{1}{n}\sum_{i=1}^{n} g_n[F(h_n(\hat{\theta}), X_i, \hat{\theta}) - F(h_n(\theta_0), X_i, \theta_0)] = o_p(1) ,$$

where g_n and $h_n(\theta)$ are rv's that satisfy (i) $|g_n| \leq 1$, (ii) $h_n(\theta) = h(y, X_i, \theta)$ or $h_n(\theta) = h(y, X_i, \theta)\hat{\Gamma}(y, X_i)_j$, where $\hat{\Gamma}(y\ X_i)_j$ denotes the indicator function of $(y, X_i) \in \hat{\Gamma}_j$, (iii) for all $\delta > 0$ sufficiently small, $c_1(h,\delta) \equiv \sup_{\theta\in N(\delta)} |h(y,x,\theta) - h(y, x, \theta_0)| \leq v(y,x)\cdot\delta$, for some $v(y,x)$ such that $K_1 \equiv \int_X \sup_{\theta\in N(\delta)} \int_Y v(y,x)f(y|x,\theta)d\mu(y)dP_X(x) < \infty$, where $N(\delta) \equiv \{\theta : \|\theta - \theta_0\| < \delta\}$, and

(iv) $K_2 \equiv \int_X \sup_{\theta\in N(\delta)} \int_Y |h(y, x, \theta_0)|\cdot r(y,x)\cdot f(y|x,\theta)d\mu(y)dP_X(x) < \infty$.

The sufficiency of this condition follows because (1) the j^{th} diagonal element of $\Lambda_n(\hat{\Gamma},\hat{\theta}) - \Lambda_n(\hat{\Gamma}, \theta_0)$ equals $\xi_n(\hat{\theta})$ with $g_n = 1$, $h_n(\theta) = \hat{\Gamma}_j$, $c_1(h,\delta) = 0$, and $K_2 < \infty$ by M2; (2) the $(j,k)^{th}$ element of $H_n(\hat{\Gamma},\hat{\theta}) - H_n(\hat{\Gamma}, \theta_0)$ can be written as $a_{1n} + a_{2n}$, where $a_{1n} \equiv \frac{1}{n}\sum_{i=1}^{n} F(\hat{\Gamma}, X_i, \hat{\theta})_j[F(\hat{\Gamma}, X_i, \hat{\theta})_k - F(\hat{\Gamma}, X_i, \theta_0)_k]$ equals $\xi_n(\hat{\theta})$ with $g_n = F(\hat{\Gamma}, X_i, \hat{\theta})_j \leq 1$, $h_n(\theta) = \hat{\Gamma}_k$, $c_1(h,\theta) = 0$, and $K_2 < \infty$ by M2, and $a_{2n} \equiv \frac{1}{n}\sum_{i=1}^{n} F(\hat{\Gamma}, X_i, \theta_0)_k[F(\hat{\Gamma}, X_i, \hat{\theta})_j - F(\hat{\Gamma}, X_i, \theta_0)_j]$ equals $\xi_n(\hat{\theta})$ with $g_n = F(\hat{\Gamma}, X_i, \theta_0)_k \leq 1$, $h_n(\theta) = \hat{\Gamma}_j$, $c_1(h,\theta) = 0$, and $K_2 < \infty$ by M2; (3) the $(\ell,j)^{th}$ element of $\Delta_{1n}(\hat{\Gamma},\hat{\theta}) - \Delta_{1n}(\hat{\Gamma}, \theta_0)$ equals $\xi_n(\hat{\theta})$ with $g_n = 1$, $h_n(\theta) = s(y|X_i, \theta)_\ell\hat{\Gamma}(y, X_i)_j$, $c_1(h,\theta) = \sup_{\theta\in N(\delta)} |s(y|x,\theta)_\ell - s(y|x,\theta_0)_\ell| \leq \sup_{\theta\in N(\delta)} \bar{r}(y,x)\underset{\sim}{1}'|\theta-\theta_0| \leq \bar{r}(y,x)\cdot\sqrt{L}\cdot\delta$, using the mean value theorem and M2 for the first inequality, $K_1 < \infty$ by M2, and $K_2 < \infty$ by M2; (4) the $(\ell,j)^{th}$ element of $\Pi_n(\hat{\Gamma},\hat{\theta}) - \Pi_n(\hat{\Gamma}, \theta_0)$ equals $\xi_n(\hat{\theta})$ with $g_n = 1$, $h_n(\theta) = \psi(y, X_i, \theta)_\ell\hat{\Gamma}(y, X_i)_j$,

$c_1(h,\theta) = \sup_{\theta\in N(\delta)} |\psi(y,x,\theta)_\ell - \psi(y,x,\theta_0)_\ell| \le r_1(y,x)\cdot\sqrt{L}\cdot\delta$, using the mean value theorem and E2 as above, $K_1 < \infty$ by E2, and $K_2 < \infty$ by E2; (5) the $(\ell,m)^{th}$ element of $D_{1n}(\hat{\theta}) - D_{1n}(\theta_0)$ equals $\xi_n(\hat{\theta})$ with $g_n = 1$, $h_n(\theta) = \frac{\partial}{\partial\theta_m}\psi(y,X_i,\theta)_\ell$, $c_1(h,\theta) = \sup_{\theta\in N(\delta)} \left|\frac{\partial}{\partial\theta_m}\psi(y,X_i,\theta)_\ell - \frac{\partial}{\partial\theta_m}\psi(y,X_i,\theta_0)_\ell\right|$ $\le r_2(y,x)\cdot\sqrt{L}\cdot\delta$, using the mean value theorem and E2 as above, $K_1 < \infty$ by E2, and $K_2 < \infty$ by E2; and (6) the $(\ell,m)^{th}$ element of $V_n(\hat{\theta}) - V_n(\theta_0)$ equals $\xi_n(\hat{\theta})$ with $g_n = 1$ and $h_n(\theta) = \psi(y, X_i, \theta)_\ell\psi(y, X_i, \theta)_m$, $c_1(h,\theta) = \sup_{\theta\in N(\delta)} |\hat{\psi}_\ell\hat{\psi}_m - \psi_\ell\psi_m| \le \sup_{\theta\in N(\delta)} |\hat{\psi}_\ell|\cdot|\hat{\psi}_m - \psi_m| + \sup_{\theta\in N(\delta)} |\psi_m|\cdot|\hat{\psi}_\ell - \psi_\ell|$ $\le 2\cdot r_0(y,x)\cdot r_1(y,x)\cdot\sqrt{L}\cdot\delta$, where $\hat{\psi}_\ell$ and ψ_ℓ denote $\psi(y,x,\hat{\theta})_\ell$ and $\psi(y,x,\theta_0)_\ell$, respectively, 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:

$$
\begin{aligned}
\sup_{\theta\in N(\delta)} |\xi_n(\theta)| &\le \sup_{\theta\in N(\delta)} \frac{1}{n}\sum_{i=1}^{n} \int_Y |h(y, X_i, \theta) - h(y, X_i, \theta_0)|\cdot f(y|X_i, \theta)\,d\mu(y) \\
&\quad + \sup_{\theta\in N(\delta)} \frac{1}{n}\sum_{i=1}^{n} \int_Y |h(y, X_i, \theta_0)|\cdot|f(y|X_i, \theta) - f(y|X_i, \theta_0)|\,d\mu(y) \\
&\le \frac{1}{n}\sum_{i=1}^{n} \sup_{\theta\in N(\delta)} \int_Y v(y, X_i) f(y|X_i, \theta)\,d\mu(y)\cdot\delta \qquad \text{(A.14)} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \sup_{\theta\in N(\delta)} \int_Y |h(y, X_i, \theta_0)|\cdot r(y, X_i)\cdot f(y|X_i, \theta)\cdot\sqrt{L}\cdot\delta\,d\mu(y) \\
&\xrightarrow{a.s.} \delta(K_1 + \sqrt{L}\,K_2) = K_3\cdot\delta ,
\end{aligned}
$$

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 $v(y,x)$ and $h(y, x, \theta_0)$.

Let $B_{1n} = \{ \sup_{\theta \in N(\delta)} |\xi_n(\theta)| \leq (K_3+1)\cdot\delta \}$ and $B_{2n} = \{\hat{\theta} \in N(\delta)\}$. By (A.14) and the consistency of $\hat{\theta}$ (assumption E1), we have $\lim_{n\to\infty} P(B_{1n}) = 1$ and $\lim_{n\to\infty} P(B_{2n}) = 1$. Hence,

$$(A.15) \qquad 1 = \lim_{n\to\infty} P(B_{1n} \cap B_{2n}) \leq \lim_{n\to\infty} P(|\xi_n(\hat{\theta})| \leq (K_3+1)\cdot\delta) ,$$

and (A.13) is established.

The proof with $\Sigma_{2n}(\hat{\Gamma},\cdot)$ in place of $\Sigma_{1n}(\hat{\Gamma},\cdot)$ is similar. *Q.E.D.*

PROOF OF LEMMA 6: The desired result follows for $\Sigma_{1n}(\cdot, \theta_0)$ if we can show $\Lambda_n(\hat{\Gamma}, \theta_0) - \Lambda_n(\Gamma, \theta_0) = o_p(1)$, and the analogous results for $H_n(\cdot, \theta_0)$, $\Pi_n(\cdot, \theta_0)$, and $\Delta_{1n}(\cdot, \theta_0)$. It suffices to show

$$(A.16) \qquad B_{jn} = \frac{1}{n}\sum_{i=1}^{n} g_n[F(h\hat{\Gamma}_j, X_i, \theta_0) - F(h\Gamma_j, X_i, \theta_0)] = o_p(1) , \forall j = 1, \ldots, J ,$$

for some rv's g_n and $h(Y_i, X_i)$, where $|g_n| \leq 1$, and $E_p h^2(Y,X) < \infty$.

The sufficiency of this condition follows because (1) the j^{th} diagonal element of $\Lambda_n(\hat{\Gamma}, \theta_0) - \Lambda_n(\Gamma, \theta_0)$ equals B_{jn} with $g_n = h(Y_i, X_i) = 1$; (2) the $(j,k)^{th}$ element of $H_n(\hat{\Gamma}, \theta_0) - H_n(\Gamma, \theta_0)$ can be written as $\frac{1}{n}\sum_{i=1}^{n} F(\hat{\Gamma},i)_j[F(\hat{\Gamma},i)_k - F(\Gamma,i)_k] + \frac{1}{n}\sum_{i=1}^{n} F(\Gamma,i)_k[F(\hat{\Gamma},i)_j - F(\Gamma,i)_j]$, where $F(\hat{\Gamma},i)_j$ denotes the j^{th} element of $F(\hat{\Gamma}, X_i, \theta_0)$, the first summand equals B_{kn} with $g_n = F(\hat{\Gamma},i)_j$ and $h(Y_i, X_i) = 1$, and the second summand equals B_{jn} with $g_n = F(\Gamma,i)_k$ and $h(Y_i, X_i) = 1$; (3) the $(\ell,j)^{th}$

element of $\Pi_n(\hat{\Gamma}, \theta_0) - \Pi_n(\Gamma, \theta_0)$ equals B_{jn} with $g_n = 1$, $h(Y_i, X_i) = \psi(Y_i, X_i, \theta_0)_\ell$, and $E_P\psi^2(Y, X, \theta_0)_\ell < \infty$ by E2; and (4) the $(\ell,j)^{th}$ element of $\Delta_{1n}(\hat{\Gamma}, \theta_0) - \Delta_{1n}(\Gamma, \theta_0)$ equals B_{jn} with $g_n = 1$, $h(Y_i, X_i) = s(Y_i|X_i, \theta_0)_\ell$, and $E_P s^2(Y|X, \theta_0) < \infty$ by M2.

To show (A.20), we have

$$(A.17) \quad B_{jn} \le \frac{1}{n}\sum_{i=1}^{n} \int_Y |h(y,X_i)| \cdot |\hat{\Gamma}(y,X_i)_j - \Gamma(y,X_i)_j| f(y|X_i,\theta_0) d\mu(y)$$

$$\le \left[\frac{1}{n}\sum_{i=1}^{n} \int_Y h^2(y,X_i) f(y|X_i,\theta_0) d\mu(y)\right]^{1/2} \cdot \left[\frac{1}{n}\sum_{i=1}^{n} \int_Y (\hat{\Gamma}_j - \Gamma_j)^2 f(y|X_i,\theta_0) d\mu(y)\right]^{1/2}$$

$$= \left[E_P h^2(Y,X) + o_p(1)\right]^{1/2} \cdot \left[\frac{1}{n}\sum_{i=1}^{n} F(\hat{\Gamma}_j \tilde{\Delta} \Gamma_j, X_i, \theta_0)\right]^{1/2},$$

using the Cauchy-Schwartz inequality and the WLLN, where $\tilde{\Delta}$ denotes the symmetric difference operator.

To show the second multiplicand of (A.17) is $o_p(1)$, we use the result of Lemma 3 above, with C replaced by $\tilde{C} = \{D : D = C_1 \tilde{\Delta} C_2$, for some $C_1, C_2 \in C\}$. Note that $\tilde{C}$ is a VC class, since C is. Lemma 3 gives

$$(A.18) \quad (\nu_n(\cdot, \theta_0), \hat{\Gamma}_j \tilde{\Delta} \Gamma_j) \xrightarrow{L} (\nu(\cdot), \phi)$$

as a process indexed by $\gamma \in \tilde{D}$ as $n \to \infty$,

where $\tilde{D}$ is the analogue of D with C replaced by $\tilde{C}$, and ϕ is the null set. With P-probability one, $\nu_n(\hat{\Gamma}\tilde{\Delta}\Gamma, \theta_0)_j$ is a continuous function of $(\nu_n(\cdot, \theta_0), \hat{\Gamma}_j \tilde{\Delta} \Gamma_j)$, by Lemma 4. Hence, the continuous mapping theorem gives

$$(A.19) \quad \nu_n(\hat{\Gamma}\tilde{\Delta}\Gamma, \theta_0)_j \xrightarrow{L} \nu(\phi)_j = 0 \text{ as } n \to \infty, \ \forall j = 1, \ldots, J.$$

That is,

$$(A.20) \qquad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1_{\hat{\Gamma}_j \tilde{\Delta}\Gamma_j}(Y_i, X_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F(\hat{\Gamma}_j \tilde{\Delta}\Gamma_j, X_i, \theta_0) = o_p(1) \quad \text{as} \quad n \to \infty .$$

Also, by the analogous result for the standard empirical process $\eta_n(\cdot, \theta_0)$, we have

$$(A.21) \qquad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1_{\hat{\Gamma}_j \tilde{\Delta}\Gamma_j}(Y_i, X_i) - \sqrt{n}\, F(\hat{\Gamma}_j \Delta \tilde{\Gamma}_j) = o_p(1) \quad \text{as} \quad n \to \infty .$$

Since $F(\hat{\Gamma}_j \tilde{\Delta}\Gamma_j) = o_p(1)$ by RC1, equations (A.20) and (A.21) yield

$$(A.22) \qquad \frac{1}{n} \sum_{i=1}^{n} F(\hat{\Gamma}_j \tilde{\Delta}\Gamma_j, X_i, \theta_0) = o_p(1) \quad \text{as} \quad n \to \infty .$$

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

The proof with $\Sigma_{2n}(\cdot, \theta_0)$ in place of $\Sigma_{1n}(\cdot, \theta_n)$ is similar.

Q.E.D.

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 $\{Q_n\}$. Its proof goes through unchanged (except the CLT is replaced by a triangular array CLT), because (i) $n^a(\hat{\theta} - \theta_0) = o_p(1)$ under $\{Q_n\}$, since $a < 1/2$ and $\sqrt{n}(\hat{\theta}-\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_0^{-1}[\psi(Y_i,X_i,\theta_0) - E_{Q_n}\psi(Y,X,\theta_0)] + \sqrt{n_0} D_0^{-1} E_Q \psi(Y,X,\theta_0)$, and (ii) the neighborhood $N(\delta_n)$ is of radius n^{-a} , 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, E1, and the assumption $\hat{W} \xrightarrow{Q_n} \Sigma_0^-$, equations (3.4) and (3.6) hold under $\{Q_n\}$. Further, the result of Lemma 3 holds under $\{Q_n\}$ with the limit process $\nu(\cdot)$ replaced by $\tilde{\nu}(\cdot)$, where $\tilde{\nu}(\cdot)$ is exactly the same as $\nu(\cdot)$ except that $E\tilde{\nu}(\gamma) = \sqrt{n_0}(E_Q\gamma(Y,X) - E_P\gamma(Y,X))$. The proof of tightness of $\{\nu_n(\cdot, \theta_0) : n = 1, 2, \ldots\}$ under $\{Q_n : n = 1, 2, \ldots\}$ proceeds exactly as under P above. The finite dimensional distributions of $\nu_n(\cdot, \theta_0)$ converge to those of $\tilde{\nu}(\cdot)$ by application of the triangular array CLT.

By the same arguments as given in (3.8) through (3.10), $(\nu_n(\cdot, \theta_0), \sqrt{n}\,\bar{\psi}_n, \hat{\Gamma}) \xrightarrow{L} (\tilde{\nu}(\cdot), \bar{\psi}, \Gamma)$ as $n \to \infty$ under $\{Q_n\}$, where $\tilde{\nu}(\cdot)$ is as above, $\bar{\psi} \sim N(\sqrt{n_0}E_Q\psi(Y, X, \theta_0), V_0)$, and $E\bar{\psi}\tilde{\nu}(\gamma)' = \Pi(\gamma)$. Hence, by the continuous mapping theorem, the asymptotic covariance matrix of $\nu_n(\hat{\Gamma}, \hat{\theta})$ under $\{Q_n\}$ equals that of $\bar{\nu}_n(\Gamma, \theta_0) - \sqrt{n}\Delta_0'D_0^{-1}\bar{\psi}_n$, viz., Σ_0. In addition, Lemma 4 and the continuous mapping theorem imply that the approximating quadratic $q_n(\hat{\Gamma}, \hat{\theta})$ of (3.1) satisfies

$$(A.23)\quad q_n(\hat{\Gamma}, \hat{\theta}) \xrightarrow{L} (\tilde{\nu}(\Gamma) - \Delta_0'D_0^{-1}\bar{\psi})'\Sigma_0^+(\tilde{\nu}(\Gamma) - \Delta'D_0^{-1}\bar{\psi}) \quad \text{as} \quad n \to \infty \quad \text{under} \quad \{Q_n\}.$$

The right-hand-side of (A.23) is a quadratic form in normal variates, and by Rao and Mitra (1971, Theorem 9.2.3) it has $\chi^2_{\Sigma_0}(\delta)$ 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 $\Sigma_{1n}(\hat{\Gamma}, \cdot)$ requires no changes, because $\hat{\theta}$ is consistent for θ_0 under $\{Q_n\}$, and the rv $\xi_n(\hat{\theta})$

does not depend on $\{Y_i : i = 1, \ldots, n\}$ (except through $\hat{\theta}$ and $\hat{\Gamma}$). The proof of Lemma 5 for $\Sigma_{2n}(\hat{\Gamma}, \cdot)$ holds with the SLLN replaced by the triangular-array SLLN, and the latter applies provided $E_Q r_2(Y,X) < \infty$, $E_Q \bar{r}(Y,X) < \infty$, $E_Q r_1(Y,X) < \infty$, and $E_Q r_0(Y,X) r_1(Y,X) < \infty$. Assumptions E2′ and M2′ include these conditions.

Under assumption RC1′, the proof of Lemma 6 for $\Sigma_{1n}(\hat{\Gamma}, \cdot)$ holds unchanged because B_{jn} does not depend on $\{Y_i : i = 1, \ldots, n\}$ (except through $\hat{\theta}$ and $\hat{\Gamma}$), and $\hat{\theta}$ is consistent for θ_0 under $\{Q_n\}$. The proof of Lemma 6 for $\Sigma_{2n}(\hat{\Gamma}, \cdot)$ holds using the triangular-array SLLN provided $E_Q r_0^2(Y,X) < \infty$ and $E_Q r^2(Y,X) < \infty$. Assumptions E2′ and M2′ include the latter conditions.

As in equation (3.15), the triangular-array WLLN gives $\Sigma_{vn}(\Gamma, \theta_0) \xrightarrow{Q_n} \Sigma_0$ for $v = 1, 2$, provided $E_Q r_0(Y,X) < \infty$, $E_Q r(Y,X) < \infty$, $E_Q \bar{r}(Y,X) < \infty$, and $E_Q r_0^2(Y,X) < \infty$, as is guaranteed by E2′ and M2′. Combining these results with those of Lemmas 5 and 6 gives the results of part (b).

Part (c) holds under the given assumptions by Theorem 1 of Andrews (1986a).

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 $\mu \in M(\Sigma_0)$ is satisfied under the assumption $Q_n(\nu_n(\hat{\Gamma}, \hat{\theta}) \in M(\Sigma_0)) \to 1$ as $n \to \infty$. *Q.E.D.*

We now establish the claim of Section 4.2 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 4.2 specifies the models in question.) This result follows because the difference between the asymptotic distribution of $X_n^2(\hat{\Gamma},\hat{\theta})$ in these two cases depends only on the difference in the asymptotic covariance matrix terms $\Delta_0'D_0^{-1}\Pi_0$ and $\Delta_0'D_0^{-1}V_0(D_0^{-1})'\Delta_0$ of $\nu_n(\hat{\Gamma},\hat{\theta})$. Without loss of generality, the model with covariates and intercept terms can be re-parametrized such that each element of X_i has mean zero. Then, the elements of the score function that correspond to the parameters on X_i factor into the product of terms that depend on the errors U_i and a linear function of X_i. In addition, the limit cells Γ depend only on U_i. Hence, by the fact that $E_P X_i = \underset{\sim}{0}$, we find that the rows of Δ_0 that correspond to parameters on X_i consist of zeroes. Furthermore, for the estimators considered, the matrix D_0 has a block diagonal structure between the parameters on X_i and the remaining parameters. And the matrices Π_0 and V_0, with the rows and columns removed that correspond to parameters on X_i, are the same as Π_0 and V_0 in the no covariate case. These results combine to establish the equivalence of $\Delta_0'D_0^{-1}\Pi_0$ in the covariate and no covariate cases; likewise with $\Delta_0'D_0^{-1}V_0(D_0^{-1})'\Delta_0$. *Q.E.D.*

Next we show that the chi-square statistic using $\hat{\Sigma}_1$ is the LM statistic (denoted LM_n) against the class of all multinomial conditional distributions, when the response variable and covariates take on a finite number of values, and the cells completely cross-classify these values (as claimed in Section 4.2).

Let G and R be the number of different values of the covariates and response variables, respectively. Thus, $J = G \cdot R$. Let Γ_{gr} denote the cell where $X = g$ and $Y = r$, for $g = 1, \ldots, G$, $r = 1, \ldots, R$. Let N_{gr}

be the number of observations (Y_i, X_i) that fall in Γ_{gr}. Let M_g be the number of observations with $X_i = g$. Let f_{gr} denote the conditional probability of Γ_{gr} given $X_i = g$, as determined by the specified parametric model with $\theta = \hat{\theta}$, i.e., f_{gr} equals $F(\Gamma_{gr}, X_i, \hat{\theta})$ when $X_i = g$.

We establish the desired result by showing that

$$(A.24) \qquad X_n^2(\hat{\Gamma}, \hat{\theta}) = \sum_{g=1}^{G} \sum_{r=1}^{R} \frac{(N_{gr} - M_g f_{gr})^2}{M_g f_{gr}} = LM_n ,$$

To show the first equality of (A.24), we note that some algebra yields

$$(A.25) \qquad \hat{\Sigma}_1 = \Lambda^{1/2}(I_J - P_T - P_{\Lambda^{-1/2}\Delta'})\Lambda^{1/2} ,$$

where the subscript " ln " and the suffix " $(\hat{\Gamma}, \hat{\theta})$ " have been dropped from Λ and Δ for simplicity, P_T is the projection matrix onto the column space of T and likewise for $P_{\Lambda^{-1/2}\Delta'}$, and T is a $J \times G$ matrix with g^{th} column $(\underline{0}', \sqrt{f_{g1}}, \ldots, \sqrt{f_{gR}}, \underline{0}')'$. Since $T'\Lambda^{-1/2}\Delta' = \underline{0}$, $I_J - P_T - P_{\Lambda^{-1/2}\Delta'}$ is a projection matrix, and

$$(A.26) \qquad \hat{\Sigma}_1^{+} = \Lambda^{-1/2}(I_J - P_T - P_{\Lambda^{-1/2}\Delta'})\Lambda^{-1/2} .$$

Algebraic manipulation, and use of the likelihood equations, now yields the first equality of (A.24).

To show the second equality of (A.24), let p_{gr} denote the conditional probability that $(Y_i, X_i) \in \Gamma_{gr}$, unrestricted by the parametric model. Define $\underline{p} = (\underline{p}_1', \underline{p}_2', \ldots, \underline{p}_G')'$, where $\underline{p}_g = (p_{g1}, \ldots, p_{gR-1})'$. Let $p(y|x, \underline{p})$ denote the unrestricted density of Y given X (with respect to

some measure $\tilde{\mu}$), and let $I(\underset{\sim}{p})$ denote the $G(R-1) \times G(R-1)$ dimensional information matrix corresponding to the likelihood function $\prod_{i=1}^{n} p(Y_i|X_i, \underset{\sim}{p})$. Define

$$(A.27) \qquad LM_n(\underset{\sim}{p}) = \frac{\partial}{\partial \underset{\sim}{p}} \log \prod_{i=1}^{n} p(Y_i|X_i, \underset{\sim}{p})' I(\underset{\sim}{p})^{-1} \frac{\partial}{\partial \underset{\sim}{p}} \log \prod_{i=1}^{n} p(Y_i|X_i, \underset{\sim}{p}) .$$

The LM test statistic LM_n is given by $LM_n(\underset{\sim}{f})$.

Now, the elements of the score function simplify as follows:

$$(A.28) \qquad s_{gr} = \frac{\partial}{\partial p_{gr}} \log \prod_{i=1}^{n} p(Y_i|X_i, \underset{\sim}{p}) = (N_{gr} - M_g p_{gr})/p_{gr} - (N_{gR} - M_g p_{gR})/p_{gR} .$$

Further, the information matrix $I(\underset{\sim}{p})$ is block diagonal with G blocks of dimension $(R-1) \times (R-1)$ on the diagonal, denoted $I_g(\underset{\sim}{p})$, for $g = 1, \ldots, G$. Hence, $LM_n(\underset{\sim}{p}) = \sum_{g=1}^{G} \underset{\sim}{s}_g' I_g(\underset{\sim}{p})^{-1} \underset{\sim}{s}_g$, where $\underset{\sim}{s}_g = (s_{g1}, \ldots, s_{gR-1})'$. Some algebra yields $I_g(\underset{\sim}{p}_g)^{-1} = \frac{1}{M_g}[\mathrm{diag}(\underset{\sim}{p}_g) - \underset{\sim}{p}_g \underset{\sim}{p}_g']$, and

$$(A.29) \qquad \underset{\sim}{s}_g' I_g(\underset{\sim}{p})^{-1} \underset{\sim}{s}_g = \sum_{r=1}^{R} (N_{gr} - M_g p_{gr})^2 / p_{gr} .$$

Evaluating $\underset{\sim}{s}_g' I_g(\underset{\sim}{p})^{-1} \underset{\sim}{s}_g$ at $\underset{\sim}{p} = \underset{\sim}{f}$, yields the desired result. *Q.E.D.*

Finally, we show that the chi-square statistic with weight matrix $\hat{\Sigma}_2^+$ (defined using $D_{3n}(\hat{\theta})$ in place of $D_{2n}(\hat{\theta})$) is asymptotically equivalent to the Wald statistic for testing $\underset{\sim}{p} = \underset{\sim}{f}_0$, for the situation described in Section 4.2. The Wald statistic is a quadratic form based on $\sqrt{n}(\hat{\underset{\sim}{p}} - \underset{\sim}{f}_0)$, where $\hat{\underset{\sim}{p}}$ is the ML estimator of $\underset{\sim}{p}$. It is easy to show that $\hat{\underset{\sim}{p}}$ has j^{th} element $\hat{p}_j = \frac{1}{M_j} \sum_{i=1}^{n} \Gamma(Y_i, X_i)_j$, for $j = 1, \ldots, J-G$, where M_j is the

number of observations with X_i-value equal to the single covariate value contained in Γ_j . The weight matrix of the Wald statistic is consistent for the inverse of the non-singular asymptotic covariance matrix of $\sqrt{n}(\hat{\underset{\sim}{p}} - \underset{\sim}{f}_0)$.

On the other hand, the chi-square statistic $X_n^2(\hat{\Gamma})$ is asymptotically equivalent to $X_n^2(\Gamma)$ 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 $X_n^2(\Gamma) = \underset{\sim}{1}'\bar{A}(\bar{A}'P_B\bar{A})^{+}\bar{A}'\underset{\sim}{1}$ and $(1/\sqrt{n})\bar{A}'\underset{\sim}{1} = \sqrt{n}\hat{S}(\hat{\underset{\sim}{p}} - \underset{\sim}{f}_0)$ in this context, where $\bar{A}$ is the same as $\tilde{A}$ of Section 3.2, except that it is defined using Γ and G rather than $\hat{\Gamma}$ and $\hat{G}$, and $\hat{S}$ denotes the $(J-G) \times (J-G)$ diagonal matrix with diagonal elements M_j/n for $j = 1, \ldots, J-G$. Hence, $X_n^2(\Gamma)$ also is a quadratic form in $\sqrt{n}(\hat{\underset{\sim}{p}} - \underset{\sim}{f}_0)$. Its weight matrix has the same probability limit as that of the Wald statistic, because the asymptotic covariance matrix of $\sqrt{n}(\hat{\underset{\sim}{p}} - \underset{\sim}{f}_0)$ 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 $X_n^2(\hat{\Gamma})$ in this context.

Q.E.D.

FOOTNOTES

1. I would like to thank David Pollard for his helpful suggestions, and for pointing out an error in the proofs of an earlier draft. I also thank several referees and the participants of the econometrics workshops at Berkeley, Stanford, Harvard/M.I.T., Chicago, and Princeton, for their comments and suggestions. The financial support of the National Science Foundation, through Grant No. SES-8419789, is acknowledged gratefully. This paper was presented at the 1985 annual meetings of the Institute of Mathematical Statistics in Las Vegas, and the 1985 Canadian Econometrics Study Group in London, Ontario.

2. Even within this sub-class of alternatives, W, LR, and LM tests are not unambiguously asymptotically optimal. This is evident from the fact that the Wald tests based on two nested parametric families (both of which contain the parametric null distributions), both possess certain optimality properties even though they are not asymptotically equivalent.

3. For the most part, Newey's (1985) CM tests are constructed with finite dimensional parametric alternatives in mind. For such tests, the same sort of comparison with chi-square tests applies as that discussed above between W, LR, and LM tests, and chi-square tests. The comparison differs, however, to the extent that CM tests have (proven) asymptotic optimality properties only in the case where the best CM test coincides with an LM test. In other cases, the best CM test is best only in the restricted class of CM tests, whereas W, LR, and LM tests possess optimality properties with respect to the class of all sequences of tests.

4. The example provided by Newey (1985, p. 1062), for testing against non-normality in probit models, is more appropriately termed a test against functional form of the regression function (since the "nonnormal" alternatives he considers actually are probit models with nonlinear regression functions). Thus, his example is not one in which the CM test coincides with an LM test against non-normality.

REFERENCES

Alexander, K. S. (1984): "Probability Inequalities for Empirical Processes and a Law of the Iterated Logarithm," Annals of Probability, 12, 1041-1067.

Andrews, D. W. K. (1985a): "Random Cell Chi-Square Diagnostic Tests for Econometric Models: I. Introduction and Applications," Cowles Foundation Discussion Paper No. 762, Yale University, New Haven, CT.

_____ (1985b): "Random Cell Chi-Square Diagnostic Tests for Econometric Models: II. Theory," Cowles Foundation Discussion Paper No. 763, Yale University, New Haven, CT.

_____ (1986a): "Asymptotic Results for Generalized Wald Tests," Cowles Foundation Discussion Paper No. 761R, Yale University, New Haven, CT.

_____ (1986b): "Stability Comparisons of Estimators," Econometrica, 54, forthcoming.

Billingsley, P. (1968): Convergence of Probability Measures. New York: Wiley.

Chesher, A. (1983): "The Information Matrix Test: Simplified Calculation via a Score Test Interpretation," Economic Letters, 13, 45-48.

Chibisov, D. M. (1970): "Certain Chi-Square Type Tests for Continuous Distributions," Theory of Probability and Its Applications, 16, 1-22.

Dudley, R. M. (1978): "Central Limit Theorems for Empirical Measures," Annals of Probability, 6, 899-929.

_____ (1984): "A Course on Empirical Processes," in École d'Été de Probabilités de Saint-Flour XII-1982, Springer Lecture Notes in Mathematics No. 1097, ed. by P. L. Hennequin. New York: Springer-Verlag.

Durbin, J. (1973): "Weak Convergence of the Sample Distribution Function When Parameters Are Estimated," Annals of Statistics, 1, 279-290.

Efron, B. and D. V. Hinkley (1978): "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information," Biometrika, 65, 475-487.

Gaenssler, P. (1984): Empirical Processes. Institute of Mathematical Statistics Lecture Notes-Monograph Series, Vol. 3. Hayward, CA: I.M.S.

Giné, E. and J. Zinn (1984): "On the Central Limit Theorem for Empirical Processes," Annals of Probability, 12, 929-989.

Hájek, J. and Z. Sidák (1967): Theory of Rank Tests. New York: Academic Press.

Hartigan, J. A. (1975): Clustering Algorithms. New York: Wiley.

Heckman, J. J. (1984): "The χ^2 Goodness of Fit Statistic for Models with Parameters Estimated from Microdata," Econometrica, 52, 1543-1547.

Horowitz, J. L. (1985): "Testing Probabilistic Discrete Choice Models of Travel Demand by Comparing Predicted and Observed Aggregate Choice Shares," Transportation Research-B, 19B, 17-38.

Kendall, M. G. and A. Stuart (1973): The Advanced Theory of Statistics, Volume II, 3rd edition. New York: Hafner.

Klein, L. R. (1974): A Textbook of Econometrics, 2nd edition. Englewood Cliffs, NJ: Prentice-Hall.

LeCam, L. (1960): "Locally Asymptotically Normal Families of Distributions," University of California Publications in Statistics, 3, 37-98.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behaviour," in Frontiers of Econometrics, ed. by P. Zarembka. New York: Academic Press.

Moore, D. S. (1971): "A Chi-Square Statistic with Random Cell Boundaries," Annals of Mathematical Statistics, 42, 147-156.

Moore, D. S. and M. C. Spruill (1975): "Unified Large-sample Theory of General Chi-squared Statistics for Tests of Fit," Annals of Statistics, 3, 599-616.

Moore, D. S. and J. B. Stubblebine (1981): "Chi-square Tests for Multivariate Normality with Application to Common Stock Prices," Communications in Statistics-Theory and Methods, A10(8), 713-738.

Nakamura, A. and M. Nakamura (1983): "Part-time and Full-time Work Behaviour of Married Women: A Model with a Doubly Truncated Dependent Variable," Canadian Journal of Economics, 16, 229-257.

_____ (1985): "Dynamic Models of the Labor Force Behaviour of Married Women Which Can Be Estimated Using Limited Amounts of Past Information," Journal of Econometrics, 27, 273-298.

Newey, W. K. (1985): "Maximum Likelihood Specification Testing and Conditional Moment Tests," Econometrica, 53, 1047-1070.

Nikulin, M. S. (1973): "Chi-square Test for Continuous Distributions with Shift and Scale Parameters," Theory of Probability and its Applications, 18, 559-568.

Pearson, K. (1900): "On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling," The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, 50, 157-175.

Pollard, D. (1979): "General Chi-Square Goodness-of-Fit Tests with Data-dependent Cells," Z. Wahrscheinlichkeitstheorie verw. Gebiete, 50, 317-331.

_____ (1984): Convergence of Stochastic Processes. New York: Springer-Verlag.

Rao, C. R. and S. K. Mitra (1971): Generalized Inverse of Matrices and Its Applications. New York: Wiley.

Rao, K. C. and D. S. Robson (1974): "A Chi-square Statistic for Goodness-of-Fit Tests within the Exponential Family," Communications in Statistics, 3, 1139-1153.

Romesburg, H. C. (1984): Cluster Analysis for Researchers. New York: van Nostrand Reinholdt.

Spath, H. (1985): Cluster Dissection and Analysis. New York: Wiley.

Tauchen, G. E. (1985): "Diagnostic Testing and Evaluation of Maximum Likelihood Models," Journal of Econometrics, 30, 415-433.

Vapnik, V. N. and A. Ya. Cervonenkis (1971): "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," Theory of Probability and Its Applications, 16, 264-280.

_____ (1981): "Necessary and Sufficient Conditions for the Uniform Convergence of Means to Their Expectations," Theory of Probability and Its Applications, 26, 532-553.

Veall, M. R. (1986): "On Estimating the Effects of Peak Demand Pricing," Journal of Applied Econometrics, 1, 81-93.

Wald, A. (1943): "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large," Transactions of the American Mathematical Society, 54, 426-482.

Watson, G S. (1957): "The χ^2 Goodness-of-Fit Test for Normal Distributions," Biometrika, 44, 336-348.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," Econometrica, 50, 1-25.