ISSN 0316-5078

INSTITUTE FOR ECONOMIC RESEARCH

QUEEN'S UNIVERSITY

Kingston, Ontario, Canada K7L 3N6

TESTING FOR CONSISTENCY USING ARTIFICIAL REGRESSIONS[1]

by

James G. MacKinnon
Russell Davidson
Queen's University

DISCUSSION PAPER NO. 687

Testing for Consistency Using Artificial Regressions[1]

Russell Davidson

James G. MacKinnon

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

[1] This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to seminar participants at GREQE (Marseilles), Free University of Brussels, the University of Bristol and Nuffield College, Oxford, for comments on an earlier version.

July, 1987

Abstract

We consider several issues related to what Hausman [1978] called "specification tests", namely tests designed to verify the consistency of parameter estimates. We first review a number of results about these tests in linear regression models, and present some new material on their distribution when the model being tested is false, and on a simple way to improve their power in certain cases. We then show how in a general nonlinear setting they may be computed as "score" tests by means of slightly modified versions of any artificial linear regression that can be used to calculate Lagrange Multiplier tests, and explore some of the implications of this result. In particular, we show how to create a variant of the information matrix test that tests for parameter consistency. We examine both the conventional information matrix test and our new version in the context of binary choice models, and provide a simple way to compute both tests based on artificial regressions. Some Monte Carlo evidence is also presented; it suggests that the most common form of the information matrix test can be extremely badly behaved in samples of even quite large size.

Key Words and Phrases: Durbin-Hausman tests, information matrix tests, binary choice models, outer-product-of-the-gradient regression.

1. Introduction

There are at least two distinct questions we may ask when we test an econometric model. The first is simply whether certain restrictions hold, i.e. whether the model is specified "correctly". This question is what standard t and F tests attempt to answer in the case of regression models, and what the three classical tests, Wald, LM and LR, attempt to answer in models estimated by maximum likelihood. The second is whether the parameters of the model have been estimated consistently. Hausman [1978], in a very influential paper, introduced a family of tests designed to answer this second question and called them "specification tests". The basic idea of Hausman's tests, namely that one may base a test on a "vector of contrasts" between two sets of estimates, one of which will be consistent under weaker conditions than the other, dates back to a relatively neglected paper by Durbin [1954]. We shall therefore refer to all tests of this general type as Durbin-Hausman or DH tests.

There has been a good deal of work on DH tests in recent years; see the survey paper by Ruud [1984]. In this paper we consider several issues related to tests of this type. In section 2, we review a number of results on DH tests in linear regression models. The primary function of this section is to present results for the simplest possible case; these should then serve as an aid to intuition. We also present some new material on the distribution of DH test statistics when the model being tested is false, and on a simple way to improve the power of the tests in certain cases.

In section 3 we provide a simple and intuitive exposition of results, originally due to Ruud [1982, 1984] and Newey [1985], on the calculation of DH tests in nonlinear models as "score" tests by means of artificial linear regressions. We go beyond previous work by showing that *any* artificial

regression which can be used to compute LM tests can be modified so as to compute DH tests. An immediate implication of our argument is Holly's [1982] result on the equivalence of DH and classical tests in certain cases. They will be equivalent whenever the number of restrictions tested by the classical test is no greater than the number of parameters the consistency of which is being tested by the DH test, if those parameters would be estimated inconsistently when the restrictions were incorrect. We also show that there are circumstances in which the DH and classical tests will be equivalent (in finite samples) even when the parameters in question would be estimated consistently when the restrictions are incorrect. Thus rejection of the null by a DH test does not always indicate parameter inconsistency.

In section 4, we build on results of Davidson and MacKinnon [1987] to show how to compute a DH version of *any* score-type test based on an artificial regression, even one not designed against any explicit alternative. We show how this procedure may be applied to tests such as the information matrix test (White [1982], Chesher [1984]), and Newey's [1985] conditional moment tests.

In section 5, we discuss the power of DH tests as compared with classical tests, in the case where the two are not identical. Finally, in section 6, we discuss the information matrix test and its DH version in the context of binary choice models. We provide a simple way to compute both tests based on artificial regressions. Some Monte Carlo evidence is also presented; among other things, it suggests that the most common form of the information matrix test can be so badly behaved in samples of even quite large size as to be totally useless in practice.

2. The Case of Linear Regression Models

Suppose the model to be tested is

$$y = X\beta + u , \qquad u \sim IID(0,\sigma^2) , \tag{1}$$

where there are n observations and k regressors. When conducting asymptotic analysis, we shall assume that $\text{plim}(X^Tu/n) = 0$ and that $\text{plim}(X^TX/n)$ is a positive definite matrix. When conducting finite-sample analysis, we shall further assume that X is fixed in repeated samples and that the u_t's are normally distributed.

The basic idea of the DH test is to compare the OLS estimator

$$\tilde{\beta} = (X^TX)^{-1}X^Ty \tag{2}$$

with some other linear estimator

$$\hat{\beta} = (X^TAX)^{-1}X^TAy , \tag{3}$$

where A is a symmetric $n \times n$ matrix assumed for simplicity to have rank no less than k (otherwise, not all elements of $\hat{\beta}$ could be estimated and we would only be able to compare the estimable part of $\hat{\beta}$ with the corresponding subvector of $\tilde{\beta}$, as in Davidson, Godfrey and MacKinnon [1985]). If (1) actually generated the data, these two estimates will have the same probability limit; they will have the same expectation if X is fixed in repeated samples or independent of u. To see the former result, observe that

$$\text{plim}(\hat{\beta}) = \left[\text{plim}\left((X^TAX/n)^{-1}\right)\right]\left[\text{plim}(X^TAX/n)\beta + \text{plim}(X^TAu/n)\right] ,$$

which equals β provided that $\text{plim}(X^TAu/n) = 0$.

The test is based on the vector of contrasts

$$\begin{aligned}\hat{\beta} - \tilde{\beta} &= (X^TAX)^{-1}X^TAy - (X^TX)^{-1}X^Ty \\ &= (X^TAX)^{-1}\left(X^TAy - (X^TAX)(X^TX)^{-1}X^Ty\right) \\ &= (X^TAX)^{-1}\left(X^TA(I - X(X^TX)^{-1}X^T)y\right) \\ &= (X^TAX)^{-1}X^TAM_Xy ,\end{aligned} \tag{4}$$

where $M_X \equiv I - X(X^TX)^{-1}X^T$ is the orthogonal projection onto the orthogonal complement of the span of the columns of the matrix X. The complementary

orthogonal projection will be denoted P_X, and throughout the paper the notations P and M subscripted by a matrix expression will denote orthogonal projections onto and off the span of the columns of that expression.

The first factor in (4), $(X^TAX)^{-1}$, is simply a $k \times k$ matrix with full rank. Its presence will obviously have no effect on any test statistic that we might compute. Hence what we really want to do is test whether

$$\text{plim}(X^TAM_Xy/n) = 0 . \tag{5}$$

The vector X^TAM_Xy has k elements, but even if AX has full rank, not all those elements may be random variables, because M_X may annihilate some columns of AX. Suppose that k^* is the number of linearly independent columns of AX which are not annihilated by M_X. Then if we let the corresponding k^* columns of X be denoted by X^*, testing (5) is equivalent to testing

$$\text{plim}(X^{*T}AM_Xy/n) = 0 . \tag{6}$$

Now consider the artificial regression

$$y = X\beta + AX^*\delta + \text{errors} . \tag{7}$$

It is easily shown that the OLS estimate of δ is

$$\ddot{\delta} = (X^{*T}AM_XAX^*)^{-1}X^{*T}AM_Xy ,$$

and it is evident that $\text{plim}(\ddot{\delta}) = 0$ iff (6) holds.

The ordinary F statistic for $\delta = 0$ in (7) is

$$\frac{y^TP_{M_XAX^*}y/k^*}{y^TM_{[X\ M_XAX^*]}y/(n-k-k^*)} . \tag{8}$$

If (1) actually generated the data, this statistic will certainly be valid asymptotically, since the denominator will then consistently estimate σ^2. It will be exactly distributed as $F(k^*,n-k-k^*)$ in finite samples if the u_t's in (1) are normally distributed.

There are many possible choices for A. In the case originally studied by Durbin [1954], $\hat{\beta}$ is an IV estimator formed by first projecting X onto the space spanned by a matrix of instruments W, so that $A = P_W$. The test is then often *interpreted* as a test for the exogeneity of those components of X not in the space spanned by W; see Wu [1973], Hausman [1978], Nakamura and Nakamura [1981] and Fisher and Smith [1985]. This interpretation is misleading, since what is being tested is not the exogeneity or endogeneity of some components of X, but rather the effect of *possible* endogeneity on the estimates of β.

Alternatively, $\hat{\beta}$ may be the OLS estimator for β in the model

$$y = X\beta + Z\gamma + u , \tag{9}$$

where Z is an $n \times \ell$ matrix of regressors not in the span of X, so that $A = M_Z$. This form of the test is thus asking whether the estimates of β when Z is excluded from the model are consistent. This is a simple example of the case examined, in a much more general context, by Holly [1982]. As is now well-known, the DH test is equivalent to an ordinary F test for $\gamma = 0$, provided that $k \geq \ell$ and a certain matrix has full rank. This is easily seen from regression (7), which in this case is

$$y = X\beta + M_Z X\delta + \text{errors} \tag{10}$$

$$= X(\beta+\delta) - P_Z X\delta + \text{errors} . \tag{11}$$

It is evident from (11) that whenever the matrix Z^TX has rank ℓ, regression (10) will have exactly the same explanatory power as regression (9), since X and $P_Z X = Z(Z^TZ)^{-1}Z^TX$ will span the same space as X and Z. The F test for $\delta = 0$ in (10) will thus be identical to the F test for $\gamma = 0$ in (9), which is Holly's result specialized to the linear regression case. A necessary but not sufficient condition for Z^TX to have rank ℓ is that $k \geq \ell$.

There is an interesting relationship between the "exogeneity" and omitted-variables variants of the DH test. In the former, $A = P_W$ and $P_W X^*$ consists of all columns of $P_W X$ that do not lie in the space spanned by X, so that the test regression is

$$y = X\beta + P_W X^* \delta + \text{errors} . \tag{12}$$

In the latter, $M_Z X^* = M_Z X$, so long as the matrix $[X \; Z]$ has full rank. Now suppose that we expand Z so that it equals W; i.e. it includes at least as many variables as X, including some variables that are in the span of X. Evidently X^* will then consist of those columns of X which are not in the span of W, so that the test regression is

$$y = X\beta + M_W X^* \delta + \text{errors} . \tag{13}$$

But it is evident that (12) and (13) will have exactly the same explanatory power, since the matrices $[X \; P_W X]$ and $[X \; M_W X]$ span the same space. This means that the test which is *interpreted* as a test for exogeneity and the test which is *interpreted* as a test for the consistency of parameter estimates when certain variables have been omitted, are in fact exactly the same test.

Although (12) and (13) yield the same test statistics, they yield different estimates of β. As an illustration, consider the case where $X^* = X$. In this case

$$M_{M_W X} X = P_W X \tag{14}$$

and

$$M_{P_W X} X = M_W X , \tag{15}$$

so that the estimate of β from (13) is $(X^T P_W X)^{-1} X^T P_W y$, which is the IV estimate, and the estimate of β from (12) is $(X^T M_W X)^{-1} X^T M_W y$, which is the OLS estimate from the unrestricted model (9). In the more usual case where $X^* \neq X$, one or both of the equalities (14) and (15) will not hold, depending

on whether $X^* \neq X$ because some columns of X are in the span of W or because X contains fewer columns than W.

The matrix A could also be almost any sort of $n \times n$ covariance matrix, so that (3) would then be a GLS estimator. It is a familiar result that OLS and GLS estimators have the same probability limit if the regression model is specified correctly, but not in general otherwise. Thus this form of the DH test is *not* testing for a non-scalar covariance matrix, but rather for misspecification of the regression model. One can use a similar procedure when the null hypothesis involves estimation by GLS; see Boothe and MacKinnon [1986].

Yet another example is the differencing specification test, where A is an ingeniously chosen matrix such that $\hat{\beta}$ is a vector of estimates based on first-differenced data (see Plosser, Schwert and White [1982] and Davidson, Godfrey and MacKinnon [1985]). In this case there are a few minor complications caused by the fact that X^TAX does not have full rank. For still more examples, and discussion, see Breusch and Godfrey [1986].

One of the unique and potentially valuable features of DH tests is that they may be used when the null hypothesis is *not* that the data were generated by (1), but simply that the OLS estimates $\tilde{\beta}$ from (1) are consistent. However, if in fact neither (1) nor (7) represents the actual data generating process, or DGP, the denominator of (8) will provide an overestimate of the amount of noise in the actual DGP, so that (8) will have actual size less than its nominal size, with consequent loss of power when the null is false. Specifically, if the data are generated by the process

$$y = X\beta_0 + a_0 + u , \quad u \sim N(0, \sigma_0^2 I) ,$$

where a_0 may be thought of as a linear combination of omitted variables, the F-statistic for $\delta = 0$ in (7) will be distributed as *doubly* non-central

$F(k^*, n-k-k^*)$ with numerator and denominator non-centrality parameters

$$a_0^T P_{M_X AX^*} a_0/\sigma_0^2 \tag{16}$$

and

$$(a_0^T M_{[X\ M_X AX^*]} a_0)/\sigma_0^2 \tag{17}$$

respectively. These non-centrality parameters are evidently $1/\sigma_0^2$ times the explained sum of squares and the residual sum of squares from the artificial regression

$$M_X a_0 = M_X AX^* \eta + \text{errors} . \tag{18}$$

This explained sum of squares is of course the *reduction* in the sum of squared residuals in the regression

$$a_0 = X\alpha + AX^*\eta + \text{errors} \tag{19}$$

which is due to AX^* . When regression (18) fits perfectly, this means that X and AX^* in (19) jointly explain all the variation in a_0 . The numerator NCP (16) then simplifies to

$$a_0^T M_X a_0/\sigma_0^2 \tag{20}$$

and the denominator NCP (17) is equal to zero. The test will then have as much power as any test with k^* degrees of freedom could have. However, when (18) fits less than perfectly, the numerator NCP (16) is smaller than (20) and the denominator NCP (17) is greater than zero, both of which cause the test to have less power. For more on test statistics which are distributed as doubly non-central F , see Thursby and Schmidt [1977].

In certain cases it may be possible to improve the estimate of σ^2 , thus reducing the denominator NCP and hence increasing power. Consider again the case where $A = M_Z$. Whenever $\rho(AX) \equiv \rho(M_Z X) < \rho(Z)$, so that the DH test differs from the classical test for $\gamma = 0$ in (9), regression (10) must fit less well than regression (9), because the latter has $k+\ell$ regressors while the former only has $k+k^*$ (= 2k in regular cases). Instead of using the

ordinary F statistic, then, one might use the test statistic

$$\frac{y^T P_{M_X M_Z X} y / k^*}{y^T M_{[X\ Z]} y / (n-k-\ell)} . \tag{21}$$

The numerator of (21) is thus the same as the numerator of the ordinary F statistic for $\delta = 0$ in (10), while the denominator is the estimate of σ^2 from (9). It is obvious that this statistic will be asymptotically valid whenever (9) generated the data. It is also easy to see that it will actually have the $F(k^*, n-k-\ell)$ distribution in finite samples whenever (9) generated the data and the null hypothesis that $E(\hat{\beta}) = E(\tilde{\beta})$ is true, assuming of course that we are dealing with fixed regressors and normally distributed u_t's . This is because the quadratic forms in the numerator and denominator of (21) are independent, which follows from the fact that $M_X M_Z X$ is in the null space of $[X\ Z]$.

By itself, reducing the number of degrees of freedom in the denominator of an F test has the effect of reducing power (see Das Gupta and Perlman [1974]). Thus if the data were generated by (10), the modified F test (21) would have slightly less power than the ordinary F test for $\delta = 0$ in (10) (unless $n-k-\ell$ is very small, in which case the loss in power may not be slight). However, in some cases where (1) is false, (9) may fit much better than (10), thus yielding a much lower estimate of σ^2 . In such cases, the modified F test (21) will be much more powerful than the ordinary one.

These tests are easily modified so as to test the consistency of a subvector of the parameters rather than the whole parameter vector. A simple expression for the k-vector of contrasts was given by (4). We can select any k_1 of these by premultiplying (4) by a $k_1 \times k$ matrix B consisting of zeroes and ones; for example, B would consist of a $k_1 \times k_1$ identity matrix placed beside a $k_1 \times (k-k_1)$ matrix of zeroes if we wanted to

select the contrasts corresponding to the first k_1 elements of β. The vector of contrasts we are interested in is thus

$$B(X^TAX)^{-1}X^TAM_Xy \ . \tag{22}$$

Evidently a test that (22) is asymptotically zero may be based in the usual way on the artificial regression

$$y = X\beta + AX(X^TAX)^{-1}B^T\mu + \text{errors} \ .$$

The test will have k_1 degrees of freedom, unless some columns of $AX(X^TAX)^{-1}B^T$ have to be dropped because the matrix $[X \ \ AX(X^TAX)^{-1}B^T]$ does not have full rank. Whether this test would have more or less power for a given DGP than a joint test of all k parameters will depend on the extent to which the columns of $AX(X^TAX)^{-1}$ are collinear, just as the individual t-tests for a set of collinear variables may or may not be more powerful than a single F test.

3. General Nonlinear Models

Since the work of Hausman [1978], it has been well known that DH tests may be used in the context of very general classes of models involving maximum likelihood estimation. There are three principal theoretical results in this literature. The first, due to Hausman, is that the (asymptotic) covariance matrix of a vector of contrasts is equal to the difference between the (asymptotic) covariance matrices of the two vectors of parameter estimates, provided that one of the latter is (asymptotically) efficient under the null hypothesis. This is essentially a corollary of the Cramér-Rao bound.

The second principal result, due to Holly [1982], is that when the two parameter vectors being contrasted correspond to restricted and unrestricted ML estimates (the vectors consisting only of those parameters which are estimated under the restrictions), the DH test will under certain

circumstances be equivalent to the three classical test statistics, Wald, LM and LR. Whether this equivalence holds or not will depend on the number of parameters in the restricted and unrestricted models, and on the rank of certain matrices; as we show below, the results are completely analogous to those on whether the DH test based on (10) is equivalent to the ordinary F test based on (9).

The third principal result, due to Ruud [1982, 1984] and Newey [1985], is that tests asymptotically equivalent to DH tests can be computed as score tests. This implies that various artificial regressions can be used to compute these tests. The only artificial regression which has been explicitly suggested for this purpose is the so-called outer-product-of-the-gradient or OPG regression, in which a vector of ones is regressed on the matrix of contributions from single observations to the gradient of the loglikelihood function. This regression is widely used for calculating LM tests (see Godfrey and Wickens [1981]), and has more recently been suggested by Newey [1985] as an easy way to calculate his "conditional moment" tests, including some which are DH tests. Unfortunately, the OPG regression is known to have poor finite-sample properties (see Davidson and MacKinnon [1983, 1985] and Bera and McKenzie [1986]). As we shall now show, *any* artificial regression that can be used to compute LM tests can also be used to compute DH tests. In view of the undesirable properties of the OPG regression (a dramatic example of how bad these can be is presented in Section 6), this result may be important for applied work.

There are many classes of models for which artificial linear regressions other than the OPG regression are available. These include univariate and multivariate nonlinear regression models (Engle [1982, 1984]), probit and logit models (Davidson and MacKinnon [1984b]) and a rather general class of

nonlinear models, with nonlinear transformations on the dependent variable(s), for which "double-length" artificial regressions with 2n "observations" are appropriate (Davidson and MacKinnon [1984a]). To the extent that evidence is available, these all appear to have better finite-sample properties than the OPG regression.

We shall deal with the following general case. There is a sample of size n which gives rise to a loglikelihood function

$$\mathcal{L}(\theta_1, \theta_2) = \sum_{t=1}^{n} \ell_t(\theta_1, \theta_2) \ , \tag{23}$$

where θ_1 is a k-vector and θ_2 an ℓ-vector of parameters, the latter equal to zero if the model is correctly specified. Maximum likelihood estimates of the vector $\theta = [\theta_1^T \ \ \theta_2^T]^T$ under the restriction $\theta_2 = 0$ will be denoted $\tilde{\theta}$, while unrestricted estimates will be denoted $\hat{\theta}$. The scores with respect to θ_1 and θ_2 are denoted by $\mathbf{g}_1(\theta)$ and $\mathbf{g}_2(\theta)$; thus

$$\mathbf{g}_i(\theta) = \sum_{t=1}^{n} \frac{\partial \ell_t(\theta_1, \theta_2)}{\partial \theta_i} \ , \quad i = 1,2 \ .$$

A "^" or a "~" over any quantity indicates that it is evaluated at $\hat{\theta}$ or $\tilde{\theta}$ respectively.

The model represented by (23) is assumed to satisfy all the usual conditions for maximum likelihood estimation and inference to be asymptotically valid (see, for example, Amemiya [1985, Chapter 4]). In particular, we assume that $\tilde{\theta}$ is interior to a compact parameter space, and that the information matrix $\mathcal{I} \equiv \lim(E(\mathbf{g}\mathbf{g}^T/n))$ is a finite, non-singular matrix. The submatrix of $\mathcal{I}$ corresponding to θ_i will be denoted $\mathcal{I}_{ii}$; the corresponding submatrix of $\mathcal{I}^{-1}$ will be denoted $(\mathcal{I}^{-1})_{ii}$.

Taking Taylor series approximations to the first-order conditions for $\tilde{\theta}_1$ and $(\hat{\theta}_1, \hat{\theta}_2)$ around the true parameter vector θ^0, and applying a suitable law of large numbers, we find that

$$n^{\frac{1}{2}}(\tilde{\theta}_1 - \theta_1^0) \cong \mathcal{I}_{11}^{-1}[I_k \ \ 0]n^{-\frac{1}{2}}\mathbf{g}(\theta^0)$$

and

$$n^{\frac{1}{2}}(\hat{\theta}_1-\theta_1^0) \cong [I_k \quad 0]\mathcal{I}^{-1}n^{-\frac{1}{2}}g(\theta^0) ,$$

where I_k is a $k \times k$ identity matrix and 0 is a $k \times \ell$ matrix of zeroes. It follows that

$$\begin{aligned} n^{\frac{1}{2}}(\tilde{\theta}_1-\hat{\theta}_1) &\cong \left[\mathcal{I}_{11}^{-1}[I_k \quad 0] - [I_k \quad 0]\mathcal{I}^{-1}\right]n^{-\frac{1}{2}}g(\theta^0) \\ &= \left[\mathcal{I}_{11}^{-1} - (\mathcal{I}^{-1})_{11}\right]n^{-\frac{1}{2}}g_1(\theta^0) - \left[(\mathcal{I}^{-1})_{12}\right]n^{-\frac{1}{2}}g_2(\theta^0) . \end{aligned} \tag{24}$$

From (24) it is easy to show that the asympototic covariance matrix of $n^{\frac{1}{2}}(\tilde{\theta}_1-\hat{\theta}_1)$ is

$$\begin{aligned} &\left[(\mathcal{I}_{11})^{-1}[I_k \quad 0] - [I_k \quad 0]\mathcal{I}^{-1}\right]\mathcal{I}\left[(\mathcal{I}_{11})^{-1}[I_k \quad 0] - [I_k \quad 0]\mathcal{I}^{-1}\right]^T \\ &= [\mathcal{I}^{-1}]_{11} - (\mathcal{I}_{11})^{-1} . \end{aligned} \tag{25}$$

The first term in (25) is the asymptotic covariance matrix of $n^{\frac{1}{2}}(\hat{\theta}_1-\theta_1^0)$ and the second is the asymptotic covariance matrix of $n^{\frac{1}{2}}(\tilde{\theta}_1-\theta_1^0)$, so that (25) is a special case of Hausman's principal result.

Standard results on partitioned matrices tell us that

$$(\mathcal{I}^{-1})_{11} = [\mathcal{I}_{11} - \mathcal{I}_{12}(\mathcal{I}_{22}^{-1})\mathcal{I}_{21}]^{-1}$$

and

$$(\mathcal{I}^{-1})_{12} = - [\mathcal{I}_{11} - \mathcal{I}_{12}(\mathcal{I}_{22}^{-1})\mathcal{I}_{21}]^{-1}\mathcal{I}_{12}\mathcal{I}_{22}^{-1} .$$

Substituting these into (24) yields the following expression for $n^{\frac{1}{2}}(\tilde{\theta}_1-\hat{\theta}_1)$:

$$\mathcal{I}_{11}^{-1}n^{-\frac{1}{2}}g_1 + [\mathcal{I}_{11} - \mathcal{I}_{12}(\mathcal{I}_{22}^{-1})\mathcal{I}_{21}]^{-1}\left[\mathcal{I}_{12}\mathcal{I}_{22}^{-1}n^{-\frac{1}{2}}g_2 - n^{-\frac{1}{2}}g_1\right] . \tag{26}$$

This expression allows us to derive easily computed test statistics based on the general notion of an artificial regression.

In the usual case of testing restrictions in the context of maximum likelihood estimation, an artificial regression involves two things: a regressand, say $r(\theta)$, and a matrix of regressors, say $R(\theta)$, partitioned as $[R_1 \quad R_2]$, which have the properties that

(i) $R^T(\theta)r(\theta)$ is the gradient of the loglikelihood function at θ,

(ii) $R^T(\ddot{\theta})R(\ddot{\theta})/n$ consistently estimates the information matrix whenever $\ddot{\theta}$ consistently estimates θ.

Replacing the gradients and information sub-matrices in (26) by their finite-sample analogues, evaluated at $\tilde{\theta}$, and ignoring factors of n, yields the expression

$$(\tilde{R}_1^T\tilde{R}_1)^{-1}\tilde{R}_1^T\tilde{r} - (\tilde{R}_1^T\tilde{M}_2\tilde{R}_1)^{-1}\tilde{R}_1^T\tilde{M}_2\tilde{r} = -(\tilde{R}_1^T\tilde{M}_2\tilde{R}_1)^{-1}\tilde{R}_1^T\tilde{M}_2\tilde{M}_1\tilde{r} \quad , \tag{27}$$

where $\tilde{M}_i$ denotes $M_{\tilde{R}_i}$. Notice that the left-hand side of (27) resembles the expression for a restricted OLS estimator minus an unrestricted one: think of $\tilde{r}$ as the regressand, $\tilde{R}_1$ as the matrix of regressors for the null hypothesis and $\tilde{M}_2$ as the matrix which projects off the space spanned by the additional regressors whose coefficients are zero under the null.

Now consider the artificial regression

$$\tilde{r} = \tilde{R}_1 b_1 + \tilde{M}_2\tilde{R}_1^* b_2 + \text{errors} \quad , \tag{28}$$

where the $n \times k^*$ matrix $\tilde{R}_1^*$ consists of as many columns of $\tilde{R}_1$ as possible subject to the condition that the matrix $[\tilde{R}_1 \;\; \tilde{M}_2\tilde{R}_1^*]$ have full rank. The explained sum of squares from this regression is

$$\tilde{r}^T P_{[\tilde{R}_1 \; \tilde{M}_2\tilde{R}_1^*]}\tilde{r} = \tilde{r}^T P_{\tilde{M}_1\tilde{M}_2\tilde{R}_1^*}\tilde{r} \quad ,$$

since $\tilde{R}_1^T\tilde{r} = 0$ by the first-order conditions. Under suitable regularity conditions it is easily shown that this statistic is asymptotically distributed as $\chi^2(k^*)$ under the null hypothesis that $\theta_2 = 0$. This result also extends to any situation where the data are generated by a sequence of local DGP's with $\theta_2\mathcal{I}_{21} = 0$ which tends to $\theta_2 = 0$, *provided* that $\mathcal{I}_{21}$ has full rank; we discuss this important proviso below.

Notice that (28) may be rewritten as

$$\tilde{r} = \tilde{R}_1(b_1+b_2) - \tilde{R}_2(\tilde{R}_2^T\tilde{R}_2)^{-1}\tilde{R}_2^T\tilde{R}_1 b_2 + \text{errors} \quad .$$

Thus, as with the linear case, it makes no difference whether we use (28) or

$$\tilde{r} = \tilde{R}_1 c_1 + \tilde{P}_2\tilde{R}_1^* c_2 + \text{errors} \tag{29}$$

for the purpose of computing a test.

The classical LM test can of course be computed as the explained sum of squares from the artificial regression

$$\tilde{r} = \tilde{R}_1 b_1 + \tilde{R}_2 b_2 + \text{errors} . \tag{30}$$

The equivalence result of Holly [1982] is now obvious. Suppose that $\ell < k$, so that there are fewer restrictions than parameters under the null hypothesis, and that $\tilde{R}_2^T \tilde{R}_1$ has rank ℓ. Then it must be the case that $\tilde{R}_1$ and $\tilde{P}_2 \tilde{R}_1 = \tilde{R}_2 (\tilde{R}_2^T \tilde{R}_2)^{-1} \tilde{R}_2^T \tilde{R}_1$ span the same space as $\tilde{R}_1$ and $\tilde{R}_2$, so that (30) and (29) will have exactly the same explanatory power. The LM and DH tests will then be numerically identical. Provided that $\mathcal{I}_{21} = \text{plim}(\tilde{R}_2^T \tilde{R}_1 / n)$ has full rank ℓ, the asymptotic equivalence of all forms of classical and DH tests, which is Holly's result, then follows immediately from the numerical equality of these two tests.

When $\mathcal{I}_{21}$ does not have full rank, some elements (or linear combinations of elements) of θ_1 will be estimated consistently by $\tilde{\theta}_1$ even when the restrictions are false, and regardless of the actual values of θ_2. In this situation, the results of the DH test may easily be misinterpreted. The appropriate thing to do when $\mathcal{I}_{21}$ does not have full rank is to drop as many columns of $\tilde{P}_2 \tilde{R}_1$ as necessary, and reduce the degrees of freedom for the test accordingly. In practice, however, $\tilde{R}_2^T \tilde{R}_1$ may well have full rank even though $\mathcal{I}_{21}$ does not, so that the investigator may not realize there is a problem. As a result, he may well reject the null hypothesis of consistency even when $\tilde{\theta}_1$ is in fact consistent. The key to understanding this is to recognize that, even though the null hypothesis of the DH version of a classical test is $\theta_2 \mathcal{I}_{21} = 0$ rather than $\theta_2 = 0$, the test is still testing a hypothesis about θ_2 and *not* a hypothesis about $\mathcal{I}_{21}$. When the test is done by an artificial regression, the latter is simply estimated by $\tilde{R}_2^T \tilde{R}_1 / n$,

and if $\mathcal{I}_{21}$ does not have full rank, the estimate will almost never reveal that fact.

To see this clearly, consider the following very simple case. Suppose that the restricted model is

$$y = X\beta + u \tag{31}$$

and the unrestricted one is

$$y = X\beta + \gamma z + u , \tag{32}$$

with the $n \times k$ random matrix X and the $n \times 1$ random vectors z and u being distributed in such a way that $\text{plim}(X^T z/n) = 0$ and $\text{plim}(X^T u/n) = 0$. Under these circumstances it is clear that OLS estimation of (31) will yield consistent estimates of β. Now consider the DH test, which may be based on the regression

$$y = X\beta + z(z^T z)^{-1} z^T X^* \delta + u , \tag{33}$$

where X^* may be any column of X. Unless $z^T X^*$ happens to be exactly equal to zero, in which case the test cannot be computed, a t-test for $\delta = 0$ in (33) will be numerically identical to a t-test for $\gamma = 0$ in (32). Thus if $\gamma \neq 0$ and the sample is large enough, the DH test will reject the null hypothesis with probability one, even though $\tilde{\beta}$ is in fact consistent.

The reason for this apparently puzzling result is that in a finite sample we have computed a DH test which it would have been impossible to compute asymptotically. Unfortunately, it is often possible to do this. In such circumstances, the finite-sample test results will not mean what they ordinarily mean. This is true of all forms of the DH test, and not simply the score form. In cases where the information matrix is block-diagonal between the parameters which are estimated under the null and the parameters which are restricted, the former will always be estimated consistently even when the restrictions are false. This implies that the covariance matrix of the

vector of contrasts, expression (25), must be a zero matrix. But the finite-sample analogue of (25) will almost never be a zero matrix, and it is usually computed in such a way as to ensure that it is positive semi-definite. As a result, it will be just as possible to compute, and misinterpret, the DH statistic in its original form as in its score form.

4. DH Tests in Other Directions

In Davidson and MacKinnon [1987], we showed that the Holly result is perfectly general when the null hypothesis is estimated by maximum likelihood. The reason for this is that when one set of estimates is asymptotically efficient if the model is correctly specified, the other set is always asymptotically equivalent (locally) to ML estimates with *some* set of restrictions removed; Holly's result then shows that, when the number of restrictions removed is no greater than the number of parameters estimated under the null, and the information matrix satisfies certain conditions, a DH test is equivalent to a classical test of those restrictions.

As a corollary of this result, we can start with any score-type test and derive a DH variant of it, similar to the test based on regression (28). Consider an artificial regression analogous to (30) , but with $\tilde{R}_2$ replaced by an $n \times m$ matrix $\tilde{Z} \equiv Z(\tilde{\theta})$:

$$\tilde{r} = \tilde{R}_1 c_1 + \tilde{Z} c_2 + \text{errors} . \tag{34}$$

The matrix $\tilde{Z}$ must satisfy certain conditions, which essentially give it the same properties as $\tilde{R}_2$; these are discussed below. Provided it does so, and assuming that the matrix $[\tilde{R}_1 \;\; \tilde{Z}]$ has full rank, the explained sum of squares from this regression will be asymptotically distributed as $\chi^2(m)$ when the DGP is (23) with $\theta_2 = 0$.

The variety of tests covered by (34) is very great. In addition to LM tests based on all known artificial regressions, tests of this form include

Newey's [1985] conditional moment tests, all the score-type DH tests discussed in sections 2 and 3 above, White's [1982] information matrix test in the OPG form suggested by Lancaster [1984], and Ramsey's [1969] RESET test.

We now briefly indicate how to prove the above proposition. The proof is similar to standard proofs for LM tests based on artificial regressions, and the details are therefore omitted. As noted above, it is necessary that $\tilde{Z}$ satisfy certain conditions, so that it essentially has the same properties as $\tilde{R}_2$. First, we require that $\operatorname{plim}(\tilde{r}^T\tilde{Z}/n) = 0$ under the null hypothesis; if this condition were not satisfied, we obviously could not expect c_2 in (34) to be zero. Second, we require that

$$\operatorname{plim}(\tilde{Z}^T\tilde{r}\tilde{r}^T\tilde{Z}/n) = \operatorname{plim}(\tilde{Z}^T\tilde{Z}/n) \tag{35}$$

and

$$\operatorname{plim}(\tilde{Z}^T\tilde{r}\tilde{r}^T\tilde{R}_1/n) = \operatorname{plim}(\tilde{Z}^T\tilde{R}_1/n) \ , \tag{36}$$

which are similar to the condition that

$$\operatorname{plim}(\tilde{R}_1\tilde{r}\tilde{r}^T\tilde{R}_1/n) = \operatorname{plim}(\tilde{R}_1^T\tilde{R}_1/n) \ ; \tag{37}$$

(37) does not have to be assumed because it is a consequence of property (ii) and the consistency of $\tilde{\theta}$. Finally, we require that a central limit theorem be applicable to the vector

$$n^{-\frac{1}{2}}\tilde{Z}^T\tilde{M}_1\tilde{r} \ , \tag{38}$$

and that laws of large numbers be applicable to the quantities whose probability limits appear on the right-hand sides of (35), (36) and (37).

Consider the vector (38). Asymptotically, it has mean zero under the null hypothesis, and its asymptotic covariance matrix is

$$\operatorname{plim}(\tilde{Z}^T\tilde{M}_1\tilde{r}\tilde{r}^T\tilde{M}_1\tilde{Z}/n) \ ,$$

which is equal to

$$\operatorname{plim}\left\{\frac{1}{n}\left(\tilde{Z}^T\tilde{r}\tilde{r}^T\tilde{Z} - \tilde{Z}^T\tilde{r}\tilde{r}^T\tilde{R}_1(\tilde{R}_1^T\tilde{R}_1)^{-1}\tilde{R}_1^T\tilde{Z} - \tilde{Z}^T\tilde{R}_1(\tilde{R}_1^T\tilde{R}_1)^{-1}\tilde{R}_1^T\tilde{r}\tilde{r}^T\tilde{Z}\right.\right.$$

$$+ \tilde{Z}^T\tilde{R}_1(\tilde{R}_1^T\tilde{R}_1)^{-1}\tilde{R}_1^T\tilde{r}\tilde{r}^T\tilde{R}_1(\tilde{R}_1^T\tilde{R}_1)^{-1}\tilde{R}_1^T\tilde{Z})\Big)\Big] . \tag{39}$$

Rewriting (39) so that each term is a product of probability limits which are O(1), using (35), (36) and (37), and simplifying, we find that

$$\text{plim}(\tilde{Z}^T\tilde{M}_1\tilde{r}\tilde{r}^T\tilde{M}_1\tilde{Z}/n) = \text{plim}(\tilde{Z}^T\tilde{M}_1\tilde{Z}/n) .$$

This plus the asymptotic normality of (38) implies that the statistic

$$(n^{-\frac{1}{2}}\tilde{r}^T\tilde{M}_1\tilde{Z})(\text{plim}(\tilde{Z}^T\tilde{M}_1\tilde{Z}/n))^{-1}(n^{-\frac{1}{2}}\tilde{Z}^T\tilde{M}_1\tilde{r}) \tag{40}$$

is asymptotically distributed as $\chi^2(m)$. But since our assumptions imply that a law of large numbers can be applied to $\tilde{Z}^T\tilde{M}_1\tilde{Z}/n$, the explained sum of squares from regression (34), which is

$$\tilde{r}^T\tilde{M}_1\tilde{Z}(\tilde{Z}^T\tilde{M}_1\tilde{Z})^{-1}\tilde{Z}^T\tilde{M}_1\tilde{r} ,$$

will asymptotically be the same random variable as (40).

It is obvious how to construct a DH version of this test, and it is now obvious that such a test will be asymptotically valid. We obtain the DH version by simply replacing $\tilde{Z}$ in (34) with $\tilde{M}_Z\tilde{R}_1$ or $\tilde{P}_Z\tilde{R}_1$. It is evident that if $\tilde{Z}$ satisfies the conditions imposed on it above, then so will $\tilde{P}_Z\tilde{R}_1$, because it is simply the projection of $\tilde{R}_1$ onto the space spanned by $\tilde{Z}$. As usual, the number of degrees of freedom of the test will in regular cases be m if $m \leq k$, in which case the DH and ordinary score test statistics will be numerically identical. When $m > k$, however, the DH test will have fewer degrees of freedom than the ordinary score test (i.e., at most k).

The DH versions of score tests may be particularly useful when m is large. Consider White's [1982] information matrix (IM) test. As Lancaster [1984] has shown, this can easily be computed via the OPG regression, which is a special case of regression (34). In this case, $\tilde{r}$ is an n-vector of ones, $\tilde{R}_1$ is the matrix $\tilde{G}_1$, the ti^{th} element of which is $\partial\ell_t(\theta)/\partial\theta_i$, evaluated at $\tilde{\theta}$, and $\tilde{Z}$ is a matrix of which a typical element is

$$\frac{\partial^2 \ell_t(\theta)}{\partial\theta_i \partial\theta_j} + \left[\frac{\partial \ell_t(\theta)}{\partial\theta_i}\right]\left[\frac{\partial \ell_t(\theta)}{\partial\theta_j}\right], \quad i = 1,\ldots,m, \quad j = 1,\ldots,i, \qquad (41)$$

evaluated at $\tilde{\theta}$. The number of columns in $\tilde{Z}$ is k^2+k, although in practice some columns often have to be dropped if $[\tilde{G}_1 \;\; \tilde{Z}]$ has less-than-full rank.

Except when k is very small, the IM test is likely to involve a very large number of degrees of freedom. Various ways to reduce this have been suggested; one could, for example, simply restrict attention to the diagonal elements of the information matrix, setting $j = i$ in (41). But this seems arbitrary. Moreover, as Chesher [1984] has shown, the implicit alternative of the IM test is a form of random parameter variation which will not necessarily be of much economic interest. People frequently employ the test not to check for this type of parameter variation, but because it is thought to have power against a wide range of types of model misspecification. Model misspecification is often of little concern if it does not affect parameter estimates. An attractive way to reduce the number of degrees of freedom of the IM test, then, is to use a DH version of it. This can easily be accomplished by replacing $\tilde{Z}$ in the artificial regression by $\tilde{P}_Z\tilde{G}_1$.

In many circumstances, we believe, the DH version of the IM test will be more useful than the original. Instead of asking whether there is evidence that the gradient outer product and Hessian estimates of the information matrix differ, the test asks whether there is evidence that they differ for a reason which affects the parameter estimates. One would expect the DH version of the test to have more power in many cases, since it will have at most k degrees of freedom, instead of $\frac{1}{2}(k^2+k)$ for the usual IM test; see Section 5. Note, however, that it will still be impossible to compute the test when $n < \frac{1}{2}(k^2+k)$, since $\tilde{P}_Z\tilde{G}_1$ would then equal $\tilde{G}_1$. Even in its DH version, then, the IM test remains a procedure to be used only when the sample size is

reasonably large.

Of course, it only makes sense to do a DH version of the IM test when the latter is testing in directions which affect parameter consistency. This is not so in the case of linear regression models, where it is easy to see that the IM test is implicitly testing for certain forms of heteroskedasticity, skewness and kurtosis (see Hall [1987]). For a linear regression model with normal errors, the contribution to the loglikelihood function from the t^{th} observation is

$$\ell_t = \tfrac{1}{2}\log(2\pi) - \log(\sigma) - \left[(y_t - X_t\beta)^2\right]/\left[2\sigma^2\right] , \tag{42}$$

where β is a p-vector so that $k = p+1$. The contributions to the gradient for β_i and σ respectively are

$$G_{ti} = (y_t - X_t\beta)X_{ti}/\sigma^2 , \tag{43}$$

$$G_{t,k} = -1/\sigma + (y_t - X_t\beta)^2/\sigma^3 , \tag{44}$$

and the second derivatives of (42) are

$$\frac{\partial^2 \ell_t}{\partial\sigma\partial\sigma} = 1/\sigma^2 - 3(y_t - X_t\beta)^2/\sigma^4 ,$$

$$\frac{\partial^2 \ell_t}{\partial\sigma\partial\beta_i} = -2(y_t - X_t\beta)X_{ti}/\sigma^3 ,$$

and

$$\frac{\partial^2 \ell_t}{\partial\beta_i\partial\beta_j} = -X_{ti}X_{tj}/\sigma^2 .$$

The OPG regression consists of p regressors $\tilde{G}_{ti}$, which correspond to the β_i's , and one regressor $\tilde{G}_{t,k}$ which corresponds to σ , plus the test regressors $\tilde{Z}$. The first two of these are expressions (43) and (44) respectively, evaluated at OLS estimates $\tilde{\beta}$ and $\tilde{\sigma}$ (the latter using n rather n-p in the denominator). The test regressor corresponding to any pair of parameters is the sum of the second derivative of ℓ_t with respect to those parameters and the product of the corresponding first derivatives, again evaluated at $\tilde{\beta}$ and $\tilde{\sigma}$. We simplify all these expressions by using

the fact that, since the test statistic is an explained sum of squares, multiplying any regressor by a constant will have no effect on it, and by defining e_t as $\tilde{u}_t/\tilde{\sigma}$.

The regressors for the OPG version of the IM test are thus seen to be:

$$\text{for } \beta_i: \quad e_t X_{ti} \tag{45}$$

$$\text{for } \sigma: \quad e_t^2 - 1 \tag{46}$$

$$\text{for } \beta_i, \beta_j: \quad (e_t^2 - 1) X_{ti} X_{tj} \tag{47}$$

$$\text{for } \beta_i, \sigma: \quad (e_t^3 - 3e_t) X_{ti} \tag{48}$$

$$\text{for } \sigma, \sigma: \quad e_t^4 - 5e_t^2 + 2 \quad . \tag{49}$$

When the original regression contains a constant term, (47) will be perfectly collinear with (46) when i and j both refer to the constant term, so that one of them will have to be dropped and the degrees of freedom for the test reduced by one to $\frac{1}{2}(p^2+3p)$.

It is evident that the (β_i, β_j) regressors are testing in directions which correspond to heteroskedasticity of the type that White's [1980] test is designed to detect (namely heteroskedasticity that affects the consistency of the OLS covariance matrix estimator) and that the (β_i, σ) regressors are testing in directions that correspond to skewness interacting with the X_{ti}'s . If we subtract (46) from (49), the result is $e_t^4 - 6e_t^2 + 3$, from which we see that the linearly independent part of the (σ, σ) regressor is testing in the kurtosis direction. The IM test is thus seen to be testing for heteroskedasticity, skewness and kurtosis, none of which prevent $\tilde{\beta}$ from being consistent. Hence it would make no sense to compute a DH variant of the IM test in this case, and indeed it would be impossible to do so asymptotically. If one did do such a test in practice, one would run into precisely the problem discussed in the previous section: the test might well reject if the model suffered from heteroskedasticity, skewness or kurtosis,

but the rejection would not say anything about the consistency of $\tilde{\beta}$.

5. The Power of DH and Classical Tests

When the DH version of a classical test differs from the original, the former may or may not be more powerful than the latter. Although this fact and the reasons for it are reasonably well-known, it seems worthwhile to include a brief discussion which, we hope, makes the issues clear. We shall deal with the general case of section 3, and will rely heavily on results in Davidson and MacKinnon [1987].

Suppose the data are generated by a sequence of local DGP's which tends to the point $\theta^0 \equiv (\theta_1^0, 0)$. The direction in which the null is incorrect can always be represented by a vector

$$M_1(R_2 w_2 + R_3 w_3) ,$$

where $M_1 \equiv M_1(\theta^0)$, $R_2 \equiv R_2(\theta^0)$ and R_3 is a matrix with the same properties as R_1 and R_2 , which represents directions other than those contained in the alternative hypothesis. The vectors w_2 and w_3 indicate the weights to be given to the various directions; one can think of w_2 as being proportional to θ_2 . Following Davidson and MacKinnon [1987], it is possible to show that under such a sequence any of the classical test statistics for the hypothesis $\theta_2 = 0$ will be asymptotically distributed as noncentral $\chi^2(\ell)$ with noncentrality parameter (or NCP)

$$\operatorname{plim}\left(\frac{1}{n}(w_2^T R_2^T + w_3^T R_3^T) M_1 R_2\right) \left(\operatorname{plim}(R_2^T M_1 R_2 / n)\right)^{-1} \operatorname{plim}\left(\frac{1}{n} R_2^T M_1 (R_2 w_2 + R_3 w_3)\right) . \qquad (50)$$

This NCP is the probability limit of $1/n$ times the explained sum of squares from the artificial regression

$$M_1(R_2 w_2 + R_3 w_3) = M_1 R_2 b + \text{errors} . \qquad (51)$$

When the DGP belongs to the alternative hypothesis, so that $w_3 = 0$, this regression fits perfectly and (50) simplifies to

$$\text{plim}\left(\frac{1}{n}w_2^T R_2 M_1 R_2 w_2\right) ,$$

which is equivalent to expressions for noncentrality parameters found in standard references such as Engle [1984].

Similarly, the noncentrality parameter for the DH variant of the classical test against $\theta_2 = 0$ will be the probability limit of $1/n$ times the explained sum of squares from the artificial regression

$$M_1(R_2 w_2 + R_3 w_3) = M_1 P_2 R_1 b^* + \text{errors} . \tag{52}$$

If we make the definition

$$C \equiv (R_2^T R_2)^{-1} R_2^T R_1 ,$$

regression (52) can be rewritten as

$$M_1(R_2 w_2 + R_3 w_3) = M_1 R_2 C b^* + \text{errors} . \tag{53}$$

From (51) and (53) it is clear that the DH and classical tests will have the same NCP in two circumstances. The first of these is when $\ell = k$ and the matrix C has full rank, which is the familiar case where the classical and DH tests are equivalent. The second is when

$$R_2 w_2 = R_2 C w^* , \tag{54}$$

where w^* is a k-vector. In both these cases regressions (51) and (53) will have the same explained sum of squares.

When the DH test is not equivalent to the classical tests and condition (54) does not hold, it must have a smaller NCP than the classical tests. This will be true whether or not $w_3 = 0$, since $R_2 C$ can never have more explanatory power than R_2. Whether the DH test will have more or less power than the classical test then depends on whether its reduced number of degrees of freedom more than offsets its smaller NCP.

6. Binary Choice Models: An Example

In this section we consider a simple example where a DH variant of the

IM test does make sense. Failures of distributional assumptions, of the sort which do not affect the consistency of least squares estimates, do render ML estimates of binary choice models inconsistent. It is therefore both important to test for these and interesting to see if they are affecting the parameter estimates.

We shall be concerned with the simplest type of binary choice model, in which the dependent variable y_t may be either zero or one and

$$\Pr(y_t = 1) = F(X_t\beta) \ , \tag{55}$$

where $F(x)$ is a thrice continuously differentiable function which maps from the real line to the 0-1 interval, is weakly increasing in x, and has the properties

$$F(x) \geq 0 \ ; \ F(-\infty) = 0 \ ; \ F(\infty) = 1 \ ; \ F(-x) = 1 - F(x) \ . \tag{56}$$

Two examples are the probit model, where $F(x)$ is the cumulative standard normal distribution function, and the logit model, where $F(x)$ is the logistic function. The contribution to the loglikelihood of the t^{th} observation is

$$\ell_t(\beta) = y_t \log[F(X_t\beta)] + (1-y_t)\log[F(-X_t\beta)] \ .$$

The contributions to the gradient for $y_t = 1$ and $y_t = 0$ are respectively

$$f(X_t\beta)X_{ti}/F(X_t\beta)$$

and

$$-f(-X_t\beta)X_{ti}/F(-X_t\beta) \ ,$$

where $f(x)$ is the first derivative of $F(x)$. Thus the corresponding elements of the matrix $\mathbf{G}^T\mathbf{G}$ are

$$[f(X_t\beta)/F(X_t\beta)]^2 X_{ti}X_{tj} \tag{57}$$

and

$$[f(-X_t\beta)/F(-X_t\beta)]^2 X_{ti}X_{tj} \ . \tag{58}$$

The second derivatives of $\ell_t(\beta)$ for $y_t = 1$ and $y_t = 0$ are respectively

$$\left[f'(X_t\beta)F(X_t\beta) - f^2(X_t\beta)\right]X_{ti}X_{tj}/\left[F(X_t\beta)^2\right] \tag{59}$$

and

$$\left[-f'(X_t\beta)F(-X_t\beta) - f^2(-X_t\beta)\right]X_{ti}X_{tj}/\left[F(-X_t\beta)^2\right] , \tag{60}$$

where $f'(x)$ denotes the derivative of $f(x)$ and we have used the symmetry property of (56) which implies that $f'(x) = -f'(x)$. The sum of (57) and (59) is

$$f'(X_t\beta)X_{ti}X_{tj}/F(X_t\beta) \tag{61}$$

and the sum of (58) and (60) is

$$-f'(X_t\beta)X_{ti}X_{tj}/F(-X_t\beta) . \tag{62}$$

The expectation of the random variable whose two possible realizations are (61) and (62) should be zero if the model is correctly specified, and this is what the IM test would be testing. This expectation is

$$F(X_t\beta)\left[f'(X_t\beta)X_{ti}X_{tj}/F(X_t\beta)\right] + F(-X_t\beta)\left[-f'(X_t\beta)X_{ti}X_{tj}/F(-X_t\beta)\right]$$

$$= f'(X_t\beta)X_{ti}X_{tj} - f'(X_t\beta)X_{ti}X_{tj} = 0 .$$

The IM test may be based on the OPG regression, as usual, or it may be based on the artificial regression proposed by Engle [1984] and Davidson and MacKinnon [1984b] specifically for binary choice models, which we shall refer to as the PL (for probit/logit) regression. Computing the IM test by means of an artificial regression other than the OPG regression may be attractive because of the poor finite-sample properties of the latter (see below!). Unless one counts White's [1980] heteroskedasticity test for regression models as an IM test, this does not seem to have been suggested previously.

The regressand for the PL artificial regression is

$$\tilde{r}_t = y_t\left[\frac{F(-X_t\tilde{\beta})}{F(X_t\tilde{\beta})}\right]^{\frac{1}{2}} + (y_t-1)\left[\frac{F(X_t\tilde{\beta})}{F(-X_t\tilde{\beta})}\right]^{\frac{1}{2}} , \tag{63}$$

and the regressors corresponding to the β_i's are

$$\tilde{R}_{ti} = \left[F(X_t\tilde{\beta})F(-X_t\tilde{\beta})\right]^{-\frac{1}{2}} f(X_t\tilde{\beta})X_{ti} . \tag{64}$$

We want to construct the test regressors so that the ij^{th} test regressor times (63) yields (61) when $y_t = 1$ and (62) when $y_t = 0$. It is thus easily seen that the ij^{th} test regressor must be

$$\tilde{Z}_{t,ij} = \left[F(X_t\tilde{\beta})F(-X_t\tilde{\beta})\right]^{-\frac{1}{2}} f'(X_t\tilde{\beta})\, X_{ti}X_{tj} . \tag{65}$$

In the probit case, this artificial regression has a very interesting interpretation. Since $f(X_t\tilde{\beta})$ is the standard normal density,

$$f'(X_t\beta) = -(2\pi)^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}(X_t\beta)^2\right]X_t\beta = -X_t\beta\, f(X_t\beta),$$

so that (65) becomes

$$-\left[F(X_t\tilde{\beta})F(-X_t\tilde{\beta})\right]^{-\frac{1}{2}} f(X_t\tilde{\beta})\, X_t\tilde{\beta}\, X_{ti}X_{tj} . \tag{66}$$

This is identical to the test regressor one would get if one did an LM test of the model (55) against the alternative

$$\Pr(y_t = 1) = F\left[(X_t\beta)/\exp\left[\sum_{i=1}^{k}\sum_{j=1}^{i} X_{ti}X_{tj}\gamma_{ij}\right]\right] , \tag{67}$$

which can be derived from the latent variable model

$$\begin{aligned} &y_t^* = X_t\beta + u_t , \quad u_t \sim N\left[0, \exp\left[2\sum_{i=1}^{k}\sum_{j=1}^{i} X_{ti}X_{tj}\gamma_{ij}\right]\right] , \\ &y_t = 1 \text{ if } y_t^* > 0 ; \quad y_t = 0 \text{ otherwise} . \end{aligned} \tag{68}$$

The model (68) is thus a special case of a model which incorporates a natural form of heteroskedasticity. The general model was considered by Davidson and MacKinnon [1984b], who derived the appropriate LM test. This model is special because the variance of u_t depends exclusively on the cross-products of the X_{ti}'s. It is clear that the implicit alternative of the IM test is precisely this heteroskedastic model. Moreover, just as for ordinary regression models it is only heteroskedasticity related to the cross-products of the regressors which affects the consistency of the covariance matrix estimates, so for probit models it is only heteroskedasticity of this type which (locally) prevents the information matrix equality from holding and which thus renders ML probit estimates inconsistent. This is purely a local result, of course;

if a DGP involving any form of heteroskedasticity were some fixed distance from the probit model, one could not expect ML estimates based on homoskedasticity to be consistent.

Notice that if one of the X_{ti}'s , say X_{tj} , is a constant term, the test regressor (66) which corresponds to X_{tj}^2 is

$$-\left[F(X_t\tilde{\beta})F(-X_t\tilde{\beta})\right]^{-\frac{1}{2}} f(X_t\tilde{\beta})\, X_t\tilde{\beta} ,$$

which is a linear combination of the regressors (64) that correspond to the $\tilde{\beta}_i$'s . This test regressor must therefore be dropped, and the degrees of freedom of the test reduced to $\frac{1}{2}k(k+1) - 1$.

Newey [1985] recognized that the IM test implicitly tests against heteroskedasticity in the case of probit models, and suggested that this test may be particularly attractive for such models. We now report the results of a small Monte Carlo experiment designed to shed light on this conjecture. There are two main results. First, we find that the OPG form of the IM test for probit models rejects the null far too often in samples of moderate or even rather large size, while the PL form of the IM test proposed above performs much better. Second, we find that in some realistic cases the DH version of the IM test may have significantly more power than the ordinary version.

In all our experiments the matrix X consisted of a constant term and one or more other regressors, which were normally distributed and equi-correlated with correlation one half. Only one set of realizations of these variables was generated, and only for 100 observations. For larger sample sizes this set of observations was replicated as many times as necessary. This scheme reduced the costs of the simulation, made it easy to calculate NCP's (which for a given test depend only on X and on the parameters of the DGP), and ensured that any changes as n was increased

were not due to changes in the pattern of the exogenous variables.

We first investigated the performance under the null of the ordinary IM test and its DH version, calculated by both the OPG and PL regressions, for samples of size 100, 200, 400, 800 and 1600. We let the number of parameters under the null hypothesis, k, vary from 2 to 4, so that the number of degrees of freedom for the ordinary IM test was 2, 5 or 9, and for the DH version 2, 3 or 4. The DH and ordinary IM test are thus identical when $k = 2$.

Results for samples of size 100, 400 and 1600 are shown in Table 1. The most striking result is the extreme tendency to over-reject of the OPG tests, which worsens rapidly as k increases, and diminishes only slowly as the sample size increases. For $k = 4$, the OPG IM test rejects over 98% of the time at the nominal 5% level when $n = 100$, and over 50% of the time even when $n = 1600$! It is clear that the sample would have to be enormous for this test's true size to be anywhere close to its nominal one. The DH version of the OPG test is slightly better behaved than the original, but the improvement is marginal. Previous results on the finite-sample performance of the OPG test have generally not been favorable to it, but the present results are far worse than those reported previously. Since most applications are likely to involve many more than four parameters, it seems doubtful that the OPG form of the IM test for probit models can ever yield even approximately reliable results in samples of the size that are typically used by econometricians.

The tests based on the PL regression are far better behaved than the OPG tests, but are still a long way from their asymptotic distribution even in samples of 1600. They have roughly the right mean, but their standard deviations are too high because very large values occur much more often than

they should by chance. As a result, they tend to under-reject at the 10% level and over-reject at the 1% level, while being fairly close to their nominal size at 5% . Curiously, the problem of too many outliers appears initially to get worse as n increases; for $k = 4$ (the worst case), the standard deviation for both the ordinary and DH versions is largest for $n = 400$, as is the rejection frequency at the nominal 1% level.

Since the OPG test rejects so often as to be completely useless, there is apparently no choice but to use the PL version; however, these results suggest that even it should be regarded with considerable suspicion, especially if there are more than a very few parameters and the sample size is not very large indeed.

Our second set of experiments was designed to investigate power when the data were generated by (67). Calculation of NCP's, using the artificial regression (51), showed that for a wide range of γ_{ij}'s chosen so that all cross-products contributed very roughly the same amount to the variance, the NCP for the DH version was only slightly smaller than the NCP for the ordinary IM test. In more extreme cases, such as when only one γ_{ij} was non-zero, the NCP for the DH version could be less than half as large. In the former case, the DH version should be more powerful asymptotically, since a slight reduction in the NCP is more than offset by what can be a substantial reduction in degrees of freedom, but in the latter the ordinary IM test would be more powerful.

The object of the Monte Carlo experiments was to see how accurately the asymptotic analysis of Section 5 predicted finite-sample power. We considered a single "plausible" pattern for the γ_{ij}'s, and then scaled the latter to the sample size so that the tests would have power somewhere around 50% at the nominal 5% level. The resulting NCP's, which are

of course invariant to the sample size, were 5.15 for $k = 2$, 6.18 and 5.97 (DH version) for $k = 3$, and 9.16 and 8.57 (DH version) for $k = 4$.

Results for the PL tests only are shown in Table 2; results for the OPG tests are not shown because, as one would expect from the results in Table 1, they always rejected far more often than asymptotic theory predicted. The table also shows, in rows labelled "Asymp", the values that would be expected if the test statistics actually had their asymptotic non-central chi-squared distributions.

The behavior of the PL tests when the null is false is broadly consistent with their behavior when it is true. In particular, they reject much too often at the 1% level, and have means which are often far too large, because there are many more extremely large values than asymptotic theory predicts. However, they do not consistently under-reject at the 10% level, and the pattern as n increases is not always monotonic. For the case considered here, asymptotic analysis predicts that the DH version will have a modest power advantage. This is usually the case in the experimental results as well, although the ordinary IM test is sometimes more powerful when n is small, especially at the 1% level.

Based on these results, we find it difficult to endorse Newey's [1985] recommendation of the IM test for probit models. The conventional OPG form of the test should clearly not be used. Among the tests we studied, the DH version computed via the PL regression generally performs the best, both under the null and under the alternatives we studied, but even it generates far too many realizations in the right-hand tail. It might well be more productive to test for particular, relatively simple forms of hetero-skedasticity which do not involve many degrees of freedom, especially those

which seem plausible for the model at hand, rather than to calculate any form of the IM test.

7. Conclusion

This paper has dealt with several aspects of Durbin-Hausman tests of parameter consistency. Its main contribution has been to show that tests of parameter consistency may be based on any artificial regression that can be used to compute score-type tests, and that any test based on such a regression can be converted into a test of parameter consistency. In particular, we have shown that this is true for the information matrix test, and we have also shown that, for the case of binary choice models, the IM test may be computed by means of more than one artificial regression. The latter is a valuable result, because our Monte Carlo work suggests that the usual OPG form of this test has appallingly bad finite-sample properties when applied to probit models, even when the sample size is quite large.

References

Amemiya, T. (1985), *Advanced Econometrics*. Cambridge, Mass.: Harvard University Press.

Bera, A.K., and McKenzie, C.R. (1986), "Alternative forms and properties of the score test," *Journal of Applied Statistics*, 13, 13-25.

Boothe, P., and MacKinnon, J.G. (1986), "A specification test for models estimated by GLS," *Review of Economics and Statistics*, 68, 711-714.

Breusch, T.S., and Godfrey, L.G. (1986), "Data transformation tests," *Economic Journal*, 96, 47-58.

Chesher, A. (1984), "Testing for neglected heterogeneity," *Econometrica*, 52, 865-872.

Das Gupta, S. and Perlman, M.D. (1974), "Power of the noncentral F test: effect of additional variates on Hotelling's T^2 test," *Journal of the American Statistical Association*, 69, 174-180.

Davidson, R., Godfrey, L.G., and MacKinnon, J.G. (1985), "A simplified version of the differencing test," *International Economic Review*, 26, 639-647.

Davidson, R. and MacKinnon, J.G. (1983), "Small sample properties of alternative forms of the Lagrange Multiplier test," *Economics Letters*, 12, 269-275.

Davidson, R., and MacKinnon, J.G. (1984a), "Model specification tests based on artificial linear regressions," *International Economic Review*, 25, 485-502.

Davidson, R., and MacKinnon, J.G. (1984b), "Convenient specification tests for logit and probit models," *Journal of Econometrics*, 25, 241-262.

Davidson, R., and MacKinnon, J.G. (1985), "Testing linear and loglinear regressions against Box-Cox alternatives," *Canadian Journal of Economics*, 25, 499-517.

Davidson, R. and MacKinnon, J.G. (1987), "Implicit alternatives and the local power of test statistics," *Econometrica*, 55, forthcoming.

Durbin, J. (1954), "Errors in variables," *Review of the International Statistical Institute*, 22, 23-32.

Engle, R.F. (1982), "A general approach to Lagrange Multiplier model diagnostics," *Journal of Econometrics*, 20, 83-104.

Engle, R.F. (1984), "Wald, likelihood ratio and Lagrange multiplier tests in econometrics," in Z. Griliches and M. Intriligator, ed., *Handbook of Econometrics*. Amsterdam, North Holland.

Fisher, G.R. and Smith, R.J. (1985), "Least squares theory and the Hausman specification test," Queen's University, Institute for Economic Research, Discussion Paper No. 641.

Godfrey, L.G. and Wickens, M.R. (1981), "Testing linear and log-linear regressions for functional form," *Review of Economic Studies*, 48, 487-496.

Hall, A. (1987), "The information matrix test for the linear model," *Review of Economic Studies*, 54, 257-263.

Hausman, J.A. (1978), "Specification tests in econometrics," *Econometrica*, 46, 1251-1272.

Holly, A. (1982), "A remark on Hausman's specification test," *Econometrica*, 50, 749-759.

Lancaster, T. (1984), "The covariance matrix of the information matrix test," *Econometrica*, 52, 1051-1053.

Nakamura, A., and Nakamura, M. (1981), "On the relationships among several specification error tests presented by Durbin, Wu and Hausman," *Econometrica*, 49, 1583-1588.

Newey, W.K. (1985), "Maximum likelihood specification testing and conditional moment tests," *Econometrica*, 53, 1047-1070.

Plosser, C.I., Schwert, G.W., and White, H. (1982), "Differencing as a test of specification," *International Economic Review*, 23, 535-552.

Ramsey, J.B. (1969), "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society*, Series B, 31, 350-371.

Ruud, P.A. (1982), "Score tests of consistency," mimeo, University of California at Berkeley.

Ruud, P.A. (1984), "Tests of specification in econometrics," *Econometric Reviews*, 3, 221-106.

Thursby, J.G., and Schmidt, P. (1977), "Some properties of tests for specification error in a linear regression model," *Journal of the American Statistical Association*, 72, 635-641.

White, H. (1980), "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817-838.

White, H. (1982), "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1-25.

Wu, D. (1973), "Alternative tests of independence between stochastic regressors and disturbances," *Econometrica*, 41, 733-750.

Table 1 Performance of Alternative Tests Under the Null

k	Obs.	Test	Mean	Std. Dev.	Rejection Frequencies at Nominal Levels 10%	5%	1%
2	100	OPG	6.94**	7.36**	46.5**	40.1**	28.2**
		PL	1.81*	2.67**	7.8*	5.2	2.3**
	400	OPG	4.11**	5.45**	28.0**	21.4**	12.4**
		PL	1.91	2.04	8.2*	4.1	1.2
	1600	OPG	2.74**	3.60**	17.3**	11.0**	4.9**
		PL	1.96	1.95	9.3	4.3	1.0
3	100	OPG	21.71**	11.27**	84.5**	79.5**	67.6**
		PL	4.14**	5.87**	6.6**	4.9	2.7**
		OPG-DH	15.79**	11.07**	76.7**	70.1**	57.5**
		PL-DH	2.44**	4.69**	6.0**	4.4	2.3**
	400	OPG	13.40**	10.82**	55.7**	47.5**	32.2**
		PL	4.79	4.87**	9.8	6.4*	3.1**
		OPG-DH	9.31**	9.09**	51.2**	44.0**	30.3**
		PL-DH	2.81*	3.51**	8.8	5.7	2.4**
	1600	OPG	8.53**	6.94**	33.6**	25.3**	14.2**
		PL	4.91	4.04**	9.7	6.2*	2.9**
		OPG-DH	5.55**	5.81**	29.3**	22.9**	12.7**
		PL-DH	2.95	3.21**	9.6	5.8	2.4**
4	100	OPG	35.37**	8.70**	99.4**	98.3**	94.2**
		PL	6.64**	5.72**	6.2**	4.5	2.6**
		OPG-DH	22.09**	10.01**	92.0**	88.7**	79.4**
		PL-DH	2.60**	3.19**	4.8**	3.0**	1.6*
	400	OPG	37.48**	21.15**	88.9**	84.4**	75.2**
		PL	8.22**	7.92**	9.1	6.6*	3.7**
		OPG-DH	24.72**	19.82**	82.0**	75.8**	65.4**
		PL-DH	3.57**	5.17**	8.0*	5.6	2.8**
	1600	OPG	21.94**	15.81**	62.5**	53.1**	38.5**
		PL	8.78	5.51**	9.9	6.0*	2.6**
		OPG-DH	12.92**	13.35**	55.7**	47.3**	33.3**
		PL-DH	3.76*	3.63**	8.9	5.8	2.2**

Notes: All results are based on 2000 replications.

* and ** indicate that a quantity differs from what it should be asymptotically at the .05 and .001 levels respectively.

Degrees of freedom for the ordinary IM tests are 2 for $k = 2$, 5 for $k = 3$ and 9 for $k = 4$.

The standard deviations of χ^2 random variables with 2, 3, 4, 5 and 9 degrees of freedom are respectively 2, 2.45, 2.83, 3.16 and 4.24.

Table 2 Power of Alternative Tests

k	Obs.	Test	Mean	Rejection Frequencies at Nominal Levels 10%	5%	1%
2	Asymp.	PL	7.15	64.0	51.6	28.3
	100	PL	7.40	47.4*	40.3**	27.1*
	200	PL	7.49*	53.6**	44.0**	27.7
	400	PL	7.77**	59.5**	48.9*	30.0
	800	PL	7.90**	62.1	51.0	30.1
	1600	PL	7.76**	65.4	53.2	31.4*
3	Asymp.	PL	11.18	57.4	44.5	22.6
		PL-DH	8.97	64.0	51.5	28.5
	100	PL	35.19**	55.1*	52.1**	46.8**
		PL-DH	31.43**	57.2**	51.5	44.9**
	200	PL	37.29**	63.2**	58.4**	51.6**
		PL-DH	33.02**	65.5	58.8**	49.0**
	400	PL	30.76	61.3**	56.4**	48.9**
		PL-DH	26.62**	64.9	57.5**	47.2**
	800	PL	23.22**	65.2**	58.5**	47.8**
		PL-DH	19.27**	67.2*	60.4**	46.3**
	1600	PL	17.84**	64.6*	56.5**	42.0**
		PL-DH	14.47**	67.1*	58.1**	43.1**
4	Asymp.	PL	18.16	64.5	51.9	28.6
		PL-DH	12.57	75.0	63.9	40.1
	100	PL	114.08**	50.8**	48.9*	45.9**
		PL-DH	114.18**	51.7**	50.1**	47.3**
	200	PL	159.94**	67.0*	64.3**	59.6**
		PL-DH	141.66**	69.8**	66.2*	61.0**
	400	PL	120.85**	68.7**	64.1**	57.8**
		PL-DH	107.29**	73.0*	68.8**	61.1**
	800	PL	69.29**	66.0	61.2**	51.4**
		PL-DH	60.12**	72.6*	67.2*	56.9**
	1600	PL	33.78**	60.0**	53.4	40.5**
		PL-DH	27.15**	68.5**	60.6*	48.6**

Notes: All results are based on 2000 replications.

* and ** indicate that a quantity differs from what it should be asymptotically at the .05 and .001 levels respectively.

Degrees of freedom for the ordinary IM tests are 2 for $k = 2$, 5 for $k = 3$ and 9 for $k = 4$.