els is a particular instance of Holly's more
general result. (I have ignored certain
special cases in the statement of Holly's
result.)

The DWH test based on (70) will not
be the same as the classical F test based
on (69) when \( k < r \). The former regres-
sion will have \( 2k \) regressors while the
latter will have \( k + r \), so that the DWH
test will have \( k \) degrees of freedom and
the classical test will have \( r \). When \( r \)
is larger than \( k \), the DWH test may there-
fore be more or less powerful than the
classical F tests; see Section IV.1. As
Hausman (1978) stressed, this type of
DWH test tends to be particularly useful
when the unrestricted model involves
many more parameters than the re-
stricted model.

In addition to the two cases already
considered, the matrix \( \Lambda \) could be almost
any sort of \( n \times n \) positive definite sym-
metric matrix. Breusch and Godfrey
(1986) discuss a variety of tests based on
different choices of \( \Lambda \) and call them "data
transformation tests"; see also Godfrey
(1996, Section 4.10). One special case is
the differencing specification test, in
which \( \Lambda \) is a matrix such that \( \beta \) is a vector
of estimates based on first-differenced
data. (Charles Plosser, G. William
Schwert, and White 1982; Davidson,
Godfrey, and MacKinnon 1986; Paul
Booth and MacKinnon 1986).

There has been a great deal of research
on DWH tests during the past decade,
and numerous variants of these tests exist
in addition to the ones I have discussed.
It is possible to base a test on the differ-
bance between two vectors of estimates
even when neither of them is efficient
(White 1982a; Ruud 1984). It is also pos-
sible to use artificial regressions to com-
pute DWH tests for nonlinear regres-
sion models and more general types of nonlin-
ear models (Ruud 1984; Newey 1988; Da-
v davidson and MacKinnon 1989). DWH
tests are particularly useful in two situa-
tions. They provide a convenient way to
test for inconsistency due to possibly en-
dogenous regressors, and they can be
preferable to more conventional specifi-
cation tests when the latter would involve
an excessively large number of degrees
of freedom.

10. Multivariate Regression Models

All of the discussion so far has, for
simplicity, dealt with the GNR for uni-
variate regression models. However, ex-
tensions to the case of multivariate non-
linear regression models, such as
demand systems, are fairly straightforward.
A multivariate nonlinear regres-
sion model can be written as

\[
y_t = f_t(\beta) + u_t, \quad E(u_t|u_{t-1}) = \Sigma,
E(u_t | u_{t-1}) = 0 \quad \forall t \neq s. \tag{71}
\]

Here \( y_t, f_t(\beta) \), and \( u_t \) are \( 1 \times m \) vectors
of observations on the dependent vari-
able, regression functions and error
terms respectively. If the \( m \times m \) con-
temporaneous covariance matrix \( \Sigma \) of the
error terms is known, or can be estimated,
up to a factor of proportionality, a system
like (71) is typically estimated by general-
ized least squares (GLS) or feasible GLS.
Otherwise, maximum likelihood (ML)
based on the assumption of normal errors
is generally used. Let \( \hat{F}_t \) denote the
\( k \times m \) matrix of derivatives of \( f_t(\beta) \)
with respect to the elements of \( \beta \), evaluated
at the estimates \( \hat{\beta} \) obtained using the
covariance matrix \( \Sigma \); these may be
CLS, feasible GLS or ML estimates. Let \( \hat{f}_t \)
de note \( f_t(\hat{\beta}) \), and \( \hat{\Sigma}_t \) denote an \( r \times m \) matrix
of derivatives of a more general model
\( f_t(\beta, \gamma) \) with respect to the elements of
\( \gamma \), evaluated at \( \hat{\beta}, \hat{\gamma} \). Then the GNR that
corresponds to (10) is

\[
(y_t - \hat{f}_t)^\top \psi = b^\top F_t \hat{\beta} + c^\top \hat{\Sigma}_t \psi + \text{residuals}, \tag{72}
\]

where \( b \) is a \( k \)-vector, \( c \) is an \( r \)-vector
and \( \psi \) is an \( m \times m \) matrix (usually trian-
regular) with the property that $\hat{\Theta}^2 = \Sigma^{-1}$.

The elements of equation (72) are row vectors. Transposing these and stacking them into a single regression with $mm$ observations yields the desired GNR. It has essentially all the same properties as the univariate GNR (10). In particular, tests for $\gamma = 0$ may be based on the ordinary $F$ statistic for $c = 0$ in (72), or on the quantity $mmR^2$. For more details, see Engle (1984).

III. Tests Based on Other Artificial Regressions

Tests based on the Gauss-Newton regression are designed for testing the specification of regression models. However, artificial regressions can be used to test the specification of many other econometric models as well. These artificial regressions are simply convenient ways to implement tests that are usually one of two types: Lagrange Multiplier (LM) tests and Conditional Moment (CM) tests. These two families of tests will be discussed very briefly in the next two sections, and the use of artificial regressions to compute them will be discussed in the remaining three sections of this paper.

1. Lagrange Multiplier Tests

Lagrange Multiplier tests are commonly used to test the specification of models that are estimated by the method of maximum likelihood, in which estimates are obtained by maximizing the value of a loglikelihood function that depends on the data and the parameters of the model. The key to understanding LM tests is to recognize that, in the context of specification testing, they almost never involve Lagrange multipliers at all! They are actually what many statisticians call score tests, that is, tests based on the gradient (or score) of the loglikelihood function evaluated under the null hypothesis. What these tests do is to check whether the part of the score vector that corresponds to restricted parameters, when evaluated at the restricted estimates, is sufficiently close to zero. On the other hand, LM tests, in the strict sense, check whether the values of the Lagrange multipliers associated with maximum likelihood estimation subject to equality restrictions are sufficiently close to zero. But it turns out that LM tests are often numerically identical to score tests of the same restrictions, even though the latter are calculated in a very different way. Perhaps because economists are so familiar with Lagrange multipliers, econometricians have chosen to call all these tests LM tests, even when they are actually computed as score tests. This is curious, the more so because score tests (Radakrishna Rao 1948) were developed prior to LM tests (Samuel Silvey 1955).

The key components of an LM test (in its score form) are the score vector or gradient, for the unrestricted model and the information matrix, which is defined as the expectation of the outer product of the gradient. Let $g(\theta, \gamma)$ denote the $(k+r) \times 1$ score vector, where $\theta$ is a $k$-vector of parameters estimated under the null hypothesis, and $\gamma$ is an $r$-vector of parameters that are zero under the null. Similarly, let $I(\theta, \gamma)$ denote the $(k+r) \times (k+r)$ information matrix. The restricted estimates of $\theta$ and $\gamma$ are $\hat{\theta}$ and $\hat{\gamma}$ respectively. Then an LM statistic can be written as

$$g^T I^{-1} g = g(\hat{\theta}, 0)(I(\hat{\theta}, 0))^{-1} g(\hat{\theta}, 0).$$

This is a quadratic form in the random vector $g$, which is the score vector evaluated at the estimates under the null. It will be asymptotically distributed as $\chi^2(r)$ under the null hypothesis that $\gamma = 0$.

The $k$ elements of $g$ that correspond to $\theta$ will be equal to zero, by the first
order conditions for $\theta$. The remaining $r$ elements will generally be nonzero, but should be close to zero if the restrictions do in fact hold. That is what the LM statistic (73) is testing. This may be clearer if we rewrite (73) as follows:

$$L_q^T(I_{q q} - L_{q q}L_{q q}^{-1}L_{q q}^{-1})^{-1}L_q^T.$$  

(74)

Here the subscripts $q$ and $g$ denote subvectors or submatrices corresponding to the parameter vectors $q$ and $g$ respectively. Expression (74) uses the fact that $g_0 = 0$, together with some standard results on partitioned matrices. The middle matrix in (74) is simply an estimate of the covariance matrix of the score subvector $g_q$. Let us call it $V(g_q)$. Then (74) can be rewritten as

$$L_q^TV(g_q)^{-1}L_q.$$  

(75)

When the LM statistic is written in this form, it becomes clear that it is simply testing whether the score subvector $g_q$ is significantly different from zero.

If there were only one restriction, so that $r = 1$, (75) would collapse to

$$L_q^TV(g_q)^{-1}L_q,$$

which looks like the square of an (asymptotic) $t$ statistic. In this simple case, the test would reject the null hypothesis if $g_q$ were large enough relative to its estimated variance. The intuition is straightforward. The score $g_q$ measures how rapidly the loglikelihood function changes as $g$ is allowed to move away from zero. If this change is large, relative to the standard error of $g_q$, then the model can be made to fit significantly better by relaxing the restriction that $g = 0$. Hence the restriction should be rejected.

In some cases, it is quite convenient to calculate the LM statistic as a quadratic form like (73) or (74). In others, it is more convenient to use an artificial regression. Consider the general formulation

$$a_i(q, \gamma) = \sum_{i=1}^k R_i^T(q, \gamma)\gamma_i + \sum_{j=1}^r R_j(q, \gamma)\eta_j + \text{residual},$$  

(76)

where $a_i(q, \gamma)$ denotes the regressand of the artificial regression, $R_i^T(q, \gamma)$ denotes regressors that correspond to parameters which are estimated unrestricted under the null hypothesis (i.e., $\theta$), and $R_j(q, \gamma)$ denotes regressors that correspond to parameters which are set equal to zero under the null (i.e., $\gamma$). The regressand and regressors of (76) must have the property that

$$g_i(q, \gamma) = \sum_{i=1}^k a_i(q, \gamma)R_i(q, \gamma)$$

for all $i$. Thus the inner product of the regressand with each of the regressors must yield an element of the score vector. It must also be the case that for all $i$ and $j$ a valid estimate of the $j$th component of the information matrix $I(q, \gamma)$ is provided by the quantity

$$\sum_{i=1}^k R_i(q, \gamma)R_i(q, \gamma).$$

Provided that regression (76) satisfies these properties, along with certain regularity conditions (Davidson and MacKinnon 1990), the explained sum of squares, when all the components of the regression are evaluated at restricted estimates (i.e., $\theta, \gamma$), will be an LM statistic. There are many artificial regressions that satisfy the necessary properties; we shall encounter some of them in Sections III.3, III.4, and III.5.

This treatment of LM tests has been very brief. More complete discussions of these tests, and of their relationships to other classical tests (i.e., Likelihood Ratio and Wald tests), may be found in Breusch and Pagan (1980), Engle (1984), Holly (1987), and Godfrey (1988).
2. Conditional Moment Tests

A relatively new approach to testing model specification is the "moment specification testing" approach that has developed from the work of Newey and Tauchen (1989). Tests based on this approach are called conditional moment (or CM) tests. The basic idea is that, if a model is correctly specified, many random quantities should have conditional expectations of zero. The empirical counterparts of these quantities will not be exactly zero, but they should be close to zero. One can test whether they are sufficiently close to zero by using procedures similar to those used to calculate LM tests. An artificial regression that can be used to compute a very wide variety of CM and LM tests will be discussed in the next section.

Conditional moment tests can be used to test almost any sort of moment condition on a stochastic model. Suppose that economic or statistical theory says that the expectation of some function of the data and model parameters, say \( E(m(y, \theta)) \), is equal to zero. Here \( y \) is an observation on the dependent variable, and \( \theta \) is a vector of parameters. The empirical counterpart of \( E(m(y, \theta)) \) is

\[
n^{-1} \sum_{i=1}^{n} m(y_{i}, \hat{\theta}). \quad (77)
\]

A one-degree-of-freedom conditional moment test would be computed by dividing (77) by an estimate of its standard deviation.

As a simple example, suppose that a random variable \( u_i \) is supposed to be distributed as \( N(0, \sigma^2) \). The expectations of \( u_i^2 \) and \( u_i^2 - 3\sigma^4 \) should both be zero, because the normal distribution is symmetric, and the fourth moment of a normally distributed random variable is equal to three times the square of its variance. If \( \hat{u}_i \) is an estimate of \( u_i \) and \( \hat{\theta} \) is an estimate of its standard deviation, conditional moment tests for skewness and excess kurtosis could be based on the quantities

\[
\sum_{i=1}^{n} \hat{u}_i^3 \quad \text{and} \quad \sum_{i=1}^{n} (\hat{u}_i^4 - 3\sigma^4)
\]

respectively. These are not test statistics, since we need to divide them by estimates of their standard deviations before we can say whether they are significantly different from zero. They are simply the empirical counterparts of the moment conditions that we wish to test. Actual tests for skewness and kurtosis in regression models are slightly more complicated (White and Glenn MacDonald 1980; Carlos Jarque and Bera 1980; Bera and Jarque 1981, 1982; Nicholas Kiefer and Mark Salmon 1983).

This example is so simple that it may be misleading. Conditional moment tests are not restricted to testing hypotheses about the higher moments of error terms. Consider the moment conditions:

\[
E(u_i) = 0 \quad \text{and} \quad E(u_i^2) = 0, \quad \text{where} \quad u_i \quad \text{is an error term and} \quad \gamma_i \quad \text{is an observation on some variable that is supposed to be uncorrelated with both the error terms and their squares. In the context of regression models, we have already seen how to use the GNR (or the HRTR) to test whether moment conditions like these hold. The empirical counterparts of the moments that are supposed to be zero in this case are simply}
\]

\[
\sum_{i=1}^{n} u_i \gamma_i \quad \text{and} \quad \sum_{i=1}^{n} u_i^2 \gamma_i^2
\]

The first of these would be the numerator of the \( t \) statistic for \( c = 0 \) in the GNR (14). The second would be the numerator of a test statistic for heteroskedasticity associated with the variable \( \gamma_i \) calculated using a GNR-like test. Thus it seems plausible to assert that any test based on
the GNR or the BRTR can be regarded as a form of CM test.
Some authors, notably White (1987, 1989), have argued forcefully that essentially all specification tests should be regarded as conditional moment tests. For example, an LM test can be thought of as a special sort of CM test in which the moment conditions are that
\[ E(g_i(0,0)) = 0 \quad \text{for } i = k + 1, \ldots, k + r. \]
The "moment conditions" in this case are that the elements of the score vector which correspond to parameters that were set to zero under the null hypothesis should have expectation zero.
An important class of tests that can be regarded as a special case of CM tests is the class of Instrumental Variables, or IM, tests suggested by White (1982a, 1987). These tests rely on the fact that the information matrix for a model estimated by maximum likelihood can be estimated consistently in several different ways, which will in general differ asymptotically if the model is not specified correctly. The moment conditions that are tested are the differences between the elements of two different information matrix estimates. In the case of linear regression models, an IM test for all the parameters of the model (i.e., \( \sigma^2 \) as well as \( \beta \)) is actually testing for correlation between the squared residuals and the squares and cross-products of the regressors, plus correlation between the cubed residuals and the regressors, plus excess kurtosis (Alastair Hall 1987). Further results on IM tests of regression models may be found in Beran and Sangkyu Lee (1990). More generally, Checher (1984) has shown that the IM test can be interpreted as a test for neglected heterogeneity, that is, random parameter variation across the sample.
Although LM tests can be useful, they do have one serious practical disadvantage. If a model has \( k \) parameters, an IM test will in general have \( k(k + 1)/2 \) degrees of freedom (although it may have fewer than this in special cases). This number can be very large, so that IM tests can be seriously lacking in power when \( k \) is not quite small; see Section IV.2.

3. Tests Based on the OPG Regression

The most widely applicable type of artificial regression is the Outer Product of the-Gradient, or OPG, regression, equation (71), below. It can be used in almost every case where maximum likelihood estimation is employed, and variants of it can be used in other cases as well. Suppose that the loglikelihood function for the alternative model can be written as
\[
\sum_{i=1}^{n} \ell_i(y_i, \theta, \gamma).
\] (78)
Here \( y_i \) is the \( i \)th observation on the dependent variable (which may be a vector or a scalar), \( \theta \) is a \( k \)-vector of parameters that characterize the model to be tested, and \( \gamma \) is an \( r \)-vector of parameters that appear in the alternative but not in the null.

The key to the OPG regression is that the loglikelihood function (78) is written as the sum of \( n \) contributions \( \ell_i \). It is always possible, although not always easy, to write loglikelihood functions in this way. For a dynamic model, for example, it is necessary to express the joint density as a product of marginal and conditional densities in order to write the loglikelihood function as a sum of \( n \) contributions. Let \( C_i^{(\theta,\gamma)} \) and \( C_i^{(\theta,\gamma)} \) denote the derivatives of \( \ell_i(y_i, \theta, \gamma) \) with respect to the \( i \)th element of \( \theta \) and the \( j \)th element of \( \gamma \) respectively. Then if \( g(\theta, \gamma) \) denotes the gradient of (78), \( C(\theta, \gamma) \) is the \( n \times (k + r) \) matrix of derivatives of \( \ell_i \) with respect to all the elements of \( \theta \) and \( \gamma \), and \( i \) is an \( n \)-vector of ones, we can write
Two results make it possible to compute LM tests via the OPG regression. One is the equality (79), and the other is the fact that the matrix $n^{-1} \mathbf{C}^{-1} \mathbf{C}^\prime$, where $\mathbf{C} = \mathbf{C}(\beta, \gamma)$, provides a valid way to estimate the information matrix. The OPG regression gets its name from the fact that it uses this matrix, which is often called the OPG estimator of the information matrix. Using these two results, it is easy to see that a statistic equivalent to the LM statistic (73) is

$$\mathbf{t}' \mathbf{C}^{-1} \mathbf{C}' \mathbf{t}.$$  

(80)

This is just the explained sum of squares from the OPG regression, in which a vector of ones is regressed on the matrix $\mathbf{C}$:

$$\mathbf{t} = \mathbf{C} \mathbf{b}^* + \text{residuals}. $$  

(81)

Here $\mathbf{b}^*$ is a $(k + \gamma)$-vector of coefficients. Because the total sum of squares from (81) is $n$, the test statistic (80) is equal to $n$ minus the sum of squared residuals, and this is usually the easiest way to calculate it. Godfrey and Michael Wickens (1981) was probably the first published paper in econometrics to calculate an LM test in this way.

It is natural to partition $\mathbf{C}$ as $[\mathbf{C}_0 \mathbf{C}_1]$. The OPG regression (81) can then be rewritten as

$$\mathbf{t} = \mathbf{C}_0 \mathbf{b} + \mathbf{C}_1 \mathbf{c} + \text{residuals}, $$  

(82)

where $\mathbf{b}$ and $\mathbf{c}$ are the first and second parts of $\mathbf{b}^*$. The resemblance to the Gauss-Newton regression (10) is now striking. As with the CNR, the first $k$ regressors correspond to the parameters estimated under the null and depend only on the model being tested. They are orthogonal to the regressand because the first-order conditions for $\theta$ require that $\mathbf{t}' \mathbf{C}_0$ must be zero. The remaining $\gamma$ regressors correspond to the parameters of the alternative that are set to zero under the null. The explained sum of squares from (82) is equal to

$$\mathbf{t}' \mathbf{C}_1^{-1} \mathbf{C}' \mathbf{t} - \mathbf{G}_0' \mathbf{G}_0 (\mathbf{G}_0' \mathbf{C}_0)^{-1} \mathbf{G}_0' \mathbf{C}_1^{-1}. $$  

(83)

The resemblance between this expression and expression (74) for the LM statistic is striking and by no means coincidental. A quantity like $\mathbf{G}_0' \mathbf{C}_1^{-1} \mathbf{C}'$ simply estimating $\mathbf{F}_0$, the $(k, \gamma)$ submatrix of the information matrix.

One must be careful when constructing the regressors of (82). That $\mathbf{C}_0$ has been constructed correctly can be verified by regressing $\mathbf{t}$ on it alone and checking that all coefficients are zero; this is also a good way to verify that $\theta$ satisfies the first-order conditions for a maximum. It is less easy to check that $\mathbf{C}_1$ has been constructed correctly, and errors in doing so can easily produce totally invalid test statistics. For example, accidentally including the equivalent of a constant vector in $\mathbf{C}_1$ would make the regression fit perfectly, and would apparently yield a test statistic equal to the sample size! As with the CNR, it is generally necessary to include the columns of $\mathbf{G}_0$ as regressors in the OPG regression, even though they are orthogonal to the regressand by construction. If they were incorrectly excluded, the calculated test statistic would be too small, except in special cases where the information matrix is block-diagonal between the blocks corresponding to $\theta$ and $\gamma$. This can be seen from (83), since omitting $\mathbf{G}_0$ from the regression would cause the middle matrix there to be replaced by $(\mathbf{G}_0' \mathbf{C}_0)^{-1}$.

The OPG regression can be used to calculate CM tests as well as LM tests. One simply replaces the matrix $\mathbf{C}_0$ with any matrix of random variables, say $\mathbf{Z}$, chosen so that the sum over all $\mathbf{t}$ of the elements of each column of $\mathbf{Z}$ would be the empirical counterpart of some mo-
ment condition to be tested. As before, the simplest test statistic is \( n \) minus the sum of squared residuals, and it would be asymptotically distributed as \( \chi^2 \) with as many degrees of freedom as \( Z \) has columns. This result (in a more rigorous and more general form) was proved by Newey (1985).

The main advantage of conditional moment tests based on the OPC regression is that they work for a very wide variety of models, and for any type of moment restriction. As Pagan and Frank Vella (1980) point out in a valuable recent survey, they are particularly useful for models involving limited-dependent variables, such as tobit models and models relying on censoring or sample selectivity, because specification tests for models of this type such as those proposed by Forrest Nelson (1981), Bera, Jarque, and Lung-Fei Lee (1984), Lung-Fei Lee and G. S. Maddala (1986), Peter Robinson, Bera, and Jarque (1982), Christian Gourieroux et al. (1987), Chesher and Marget-Irish (1987) and Joel Horowitz and George Neumann (1980) can be recast in this framework and calculated via the OPC regression. Simon Peters and Smith (1991) and Smith (1990) explicitly use the OPC regression to calculate a variety of tests, some of them DWH tests, for misspecification of limited dependent variable models.

Some authors, following Tauchen (1985), prefer to run the OPC regression backwards; a notable example is Pagan and Vella (1989). Suppose that \( \bar{Z} \) has only one column, say \( z \). Then the regression suggested by Tauchen is

\[
\bar{z} = \alpha + \beta \bar{x} + \text{residuals},
\]

which is (82) with \( \bar{x} \), replaced by \( \bar{z} \) and its role interchanged with that of \( \bar{x} \). The test statistic is the \( t \) statistic for \( \alpha = 0 \).

The advantage of this formulation is that the OLS estimate of \( \alpha \) will provide a consistent estimate of the mean of the elements of \( \bar{z} \), since \( \bar{z} \) is orthogonal to the columns of \( \bar{G}_\beta \). If the mean of \( \bar{z} \) is of interest, the Tauchen approach may be attractive. However, when \( \bar{Z} \) has more than one column, this approach requires estimation of as many equations as there are columns of \( \bar{Z} \), followed by calculation of a quadratic form. This requires a good deal more effort than simply running an OPG regression. It can be shown that we will get essentially the same results whether we run the OPC regression (82) or the Tauchen regression (84). In fact, the \( t \) statistic for \( \alpha = 0 \) in (84) is numerically equal to the \( t \) statistic obtained by reversing the roles of \( \bar{z} \) and \( \bar{x} \).

There is unfortunately one serious problem with specification tests based on the OPG regression. In finite samples they very frequently (but not always) reject the null hypothesis too often when it is true. This tendency has been documented by numerous authors, including Davidson and MacKinnon (1984b, 1985b), Bera and Colin James (1982) and Godfrey, McEwan, and McKenzie (1985). The problem seems to be particular acute for IM tests computed using the OPG regression, a procedure that was suggested by Chesher (1983) and Tony Lancaster (1984). Monte Carlo results demonstrate the dreadful finite-sample performance of the OPG version of the IM test may be found in Larry Taylor (1987), John Kenoyer and Neu-Real (1988), Davidson and MacKinnon (1992), Christopher Orme (1990a), Hall (1980), and Chesher and Richard Spady (1991). In some of these papers, there are cases where OPG IM tests reject correct null hypotheses virtually all the time at the nominal 5 percent level.

Because of its generally poor performance in finite samples, the OPG regression should be used with caution. Unless its properties are known to be acceptable for the type of model and sample size in question, large values of test statistics
based upon it cannot safely be interpreted as evidence of misspecification. The procedures to be discussed in Section IV.1 can, however, be very useful in deciding whether an OPG test statistic is large enough to be significant.

4. An Artificial Regression for Binary Response Models

For binary response models such as the logit and probit models, there exists a very simple artificial regression that can be derived as an extension of the Gaussian-Newton regression and works much better than the OPG regression. It was suggested by Engle (1984) and Davidson and MacKinnon (1984b), and is discussed by Godfrey (1988, Section 6.2). However, many standard references on logit and probit models, such as Maddala (1983) and Amemiya (1985), were written before it became widely known.

Suppose that the dependent variable \( y_i \) is a binary variable which is either zero or one. The object of a binary response model is to predict the probability that \( y_i = 1 \) conditional on some information set, say \( \Omega_i \). A useful class of binary response models can be written as

\[
E(y_i|\Omega_i) = \Pr(y_i = 1) = F(x_i\beta). \tag{85}
\]

Here \( x_i \) is a row vector of observations on independent variables that belong to \( \Omega_i \), \( \beta \) is the vector of parameters to be estimated, and \( F(x) \) is a continuous function with the following properties:

\[
F(-\infty) = 0, \quad F(\infty) = 1, \quad \frac{\partial F(x)}{\partial x} > 0.
\]

For the probit model, \( F(x) \) is the standard normal distribution function and \( f(x) \) is the standard normal density. For the logit model, \( F(x) \) is the logistic function

\[
f(x) = \frac{\exp(x)}{1 + \exp(x)}.
\]

which has first derivative

\[
f(x) = \frac{\exp(x)}{(1 + \exp(x))^2}.
\]

For these two models \( F(x) \) satisfies the symmetry condition that \( F(-x) = 1 - F(x) \), which implies that \( \Pr(y_i = 0) = 1 - F(x_i\beta) = F(-x_i\beta) \). It is not at all essential that binary response models satisfy this symmetry condition, but in what follows I shall assume for simplicity that they do.

The loglikelihood function for this class of binary response models is

\[
\Lambda(\beta) = \sum_{i=1}^{n} \left( 1 - y_i \right) \log(F(x_i\beta)) + y_i \log(1 - F(x_i\beta)). \tag{86}
\]

A typical element of the information matrix corresponding to (86) is

\[
I_\phi(\beta) = \sum_{i=1}^{n} X_i X_i' \left( \frac{f(x_i\beta)^2}{F(x_i\beta)F(-x_i\beta)} \right).
\]

This looks like the information matrix for a linear regression model with error terms that follow a particular pattern of heteroskedasticity, a fact that can be used to derive the artificial regression for binary response models.

Suppose we rewrite the model (85) as a nonlinear regression model:

\[
y_i = F(x_i\beta) + \nu_i. \tag{87}
\]

The error term \( \nu_i \) here is evidently non-normal and heteroskedastic. Since \( \nu_i \) is like a Bernoulli trial with probability \( p \) given by \( F(x_i\beta) \), and the variance of a Bernoulli trial is \( p(1 - p) \), it is not surprising that the variance of \( \nu_i \) can be shown to be

\[
V(x_i\beta) = F(x_i\beta)F(-x_i\beta). \tag{88}
\]

The ordinary GNR for (87) would be

\[
y_i - F(x_i\beta) = f(x_i\beta)x_i\beta + \text{residual},
\]

but it is inappropriate because of the heteroskedasticity of the \( \nu_i \)'s. Multiplying both sides by the square root of the in-
verse of (88) yields the artificial regression:

\[ (V(x, \beta)^{-1/2}(y_t - F(x_0, \beta)) + \text{residual} \]

This regression has all the usual properties of artificial regressions. In particular, when it is evaluated at restricted estimates, the explained sum of squares is an LM test statistic for testing the restrictions. Other test statistics, such as the ordinary F statistic and \( nR^2 \), are also valid, but seem to have slightly poorer finite-sample properties (Davidson and MacKinnon 1984a). Thus to test for possibly omitted regressors \( z_0 \), one can compute the explained sum of squares from the artificial regression

\[ \hat{\beta}^{-1/2}(y_t - F(x_0, \beta)) = \hat{\beta}^{-1/2}f(x_0, \beta)(x_0, \beta) + z_0 \epsilon + \text{residual}, \quad (90) \]

where \( \hat{\beta} \) denotes the ML estimates of \( \beta \) for the original model (88) and \( \hat{\beta}^{-1/2} \) denotes \( F(x_0, \beta)(x_0, \beta)^{-1} \). This is surely the simplest way to test for omitted variables in probit and logit models. After one has estimated the original model and obtained the vectors with typical elements \( f(x_0, \beta) \) and \( F(x_0, \beta) \), estimating (88) by OLS is trivial.

One can use artificial regressions very similar to (90) to test (88) against a wide variety of alternative models. The regressand and the regressors corresponding to \( \beta \) remain unaltered; only the regressors corresponding to the parameter(s) that are zero under the null would be different. Davidson and MacKinnon (1984b) suggest one such alternative model, in which the binary response model is derived from an underlying latent variable model with error terms that exhibit heteroskedasticity. It is even possible to test for serial correlation, although that is generally not an interesting alternative when one is using cross-section data, by using a slightly modified version of the approach suggested by Courtière, Alain Monfort, and Alain Trugnan (1985). One could also test the functional form of \( F(\cdot) \), either by testing against a nonnested alternative with a different functional form, or by testing against a more general functional form such as the ones discussed by Therese Stukel (1988) and Smith (1989) for the logit case.

5. Other Artificial Regressions

The OPG regression should not be used if better alternatives, such as the GMR and the artificial regression for binary response models discussed in the previous section, are available. These examples of specialized artificial regressions are less widely applicable than the OPG regression but better behaved in finite samples. Several other specialized artificial regressions exist now, and more will undoubtedly be discovered in the future.

One such artificial regression is the double-length regression, or DLR (Davidson and MacKinnon 1984a, 1985). The class of models to which it applies may be written as

\[ \mathbf{y}_t = \mathbf{X}_t \mathbf{\beta} + \mathbf{e}_t, \quad t = 1, \ldots, n. \]

\[ \mathbf{e}_t \sim \text{NID}(0, \mathbf{I}). \quad (90) \]

where each \( \mathbf{e}_t \) is a smooth function that depends on the random variable \( y_t \), and \( \mathbf{X}_t \) depends on \( \mathbf{y}_t \) and \( \mathbf{\beta} \) (implicitly) on exogenous and/or predetermined variables. This may seem at first sight to be a rather restrictive class of models, but it is actually quite general. For example, a linear regression model with normal errors can be written in the form of (90) by making the definitions

\[ f(\mathbf{\theta}, y_t) = (y_t - \mathbf{x}_t \mathbf{\beta}) / \sigma, \quad \mathbf{\theta} = [\mathbf{\beta}' \sigma]. \]

The class of models defined by (90) also includes many that are not regression models, such as those in which the dependent variable is subject to a nonlinear
IV. Two Issues in Misspecification Testing

In this part I shall discuss two issues that arise whenever one uses model specification tests. Section IV.1 deals with their finite-sample properties, and Section IV.2 discusses what determines their power.

1. Finite-sample Properties

With few exceptions, the distributions of the test statistics discussed in this paper are known only asymptotically. This creates a serious problem for investigators who are trying to interpret the results of one or more specification tests when the sample size is not very large. The best way to interpret a test statistic is generally to calculate its p-value. The p-value associated with a test statistic is the probability, under the null hypothesis, of obtaining by chance a value of the statistic as or more extreme than the one actually obtained. If the p-value associated with a test statistic is sufficiently small, there is evidence that something is wrong with the model. However, when the finite-sample distribution of a test statistic is unknown, p-values are generally calculated on the basis of the asymptotic distribution. If the asymptotic and finite-sample distributions are substantially different, this can lead to serious errors of inference. Equally serious errors can occur if one uses the, perhaps more conventional, approach of deciding in advance on a significance level for the test; that is, the probability of a Type I error, and then rejecting the null if the value of the test statistic exceeds the asymptotic critical value associated with that significance level.

Although there is no simple solution to this problem, one can minimize its consequences. First of all, it is important to know something about the properties of the type of test being used for the size of sample in question. Published studies that have employed Monte Carlo and/or analytic methods to study the finite-sample distributions of various types of test statistics can be very useful. References to several studies of this type have been given above, and Godfrey (1988) summarizes what is known about the properties of many tests and provides additional references. For most specification tests, the asymptotic p-value tends to be smaller than the true p-value, so that the null hypothesis is likely to be rejected too frequently in finite samples. This is especially true for OFC tests. However, in other cases, such as tests for heteroskedasticity, the asymptotic p-value tends to be somewhat larger than the true one.

There are often several different but asymptotically equivalent ways to calculate specification tests against the same alternative. The resulting test statistics will usually have different finite-sample properties, and it generally makes sense to use the one for which the finite-sample distribution is as close as possible to the asymptotic distribution. Of course, one might wish to violate this general rule if a test with better properties under the null hypothesis were significantly less powerful or much harder to compute than a test with worse properties under the null.

In most cases, an investigator will have in mind a significance level α such as .01 or .05 (as noted in Part I, the conventional .05 level is probably too high when several test statistics are being calculated). A p-value below α would lead the investigator to conclude that the model is incompatible with the data. It is important to have an accurate estimate of the p-value associated with a test statistic only when that p-value may be near
transformation (see below). Models with non-normal errors can be put into the form of (90) by transforming the distribution of the errors to the normal distribution. Even many multivariate models can be handled. The DLR cannot handle models with limited dependent variables, however.

The key feature of the DLR, and the one that gives it its name, is that the artificial regression has twice as many "observations" as the data set. For models of the class (90), the loglikelihood function and its gradient involve two separate summations: one that corresponds to an ordinary sum-of-squares term and one that corresponds to a Jacobian term. The latter arises because the dependent variable is subject to a nonlinear transformation. The way that the DLR takes both of these summations into account is to incorporate two "observations" into the artificial regression for each genuine observation.

The DLR can be used to perform a wide variety of specification tests. It is particularly useful when the alternative model involves a nonlinear transformation of the dependent variable, so that the GNR is not applicable. For example, Davidson and MacKinnon (1985b) show how to use a DLR to test the linear regression model

\[ y_t = \sum_i \beta_i x_{it} + \sum_j \gamma_j z_{it} + u_t \]  

(91)

or the loglinear regression model

\[ \log(y_t) = \sum_i \beta_i \log(x_{it}) + \sum_j \gamma_j z_{it} + u_t \]  

(92)

against one in which the regressors and some of the regressors are subjected to the Box-Cox transformation \( z(\lambda) = (x^\lambda - 1)/\lambda; \)

\[ y(\lambda) = \sum_i \beta_i x_{it}(\lambda) + \sum_j \gamma_j z_{it} + u_t \]  

(93)

Davidson and MacKinnon (1985b) and Godfrey, McAlister, and McKenzie (1988) provide Monte Carlo evidence that, in finite samples, tests of (91) and (92) against (93) that are based on the DLR generally perform better under the null hypothesis than tests based on the OPG regression, such as the one suggested by Godfrey and Wickens (1961).

The DLR can be used to calculate many other types of specification test, including nonnested tests and information matrix tests (Davidson and MacKinnon 1984a, 1988, 1992. MacKinnon and Magee 1990). One possible disadvantage of the procedure, however, is its dependence on strong distributional assumptions. For example, if the data were actually generated by the linear model (92) with heteroskedastic and/or nonnormal errors, a DLR test of (92) against the Box-Cox alternative (93) might well result in the null hypothesis being rejected. The same is true of OPG tests for this example, of course. When certain distributional assumptions are an integral part of the model under test, this feature of the DLR and OPG tests is desirable. However, when these assumptions are merely made for convenience, as is often the case, it is not.

Several other artificial regressions have been proposed. In particular, Colin Cameron and Trivedi (1990a) use an artificial regression to test the hypothesis of mean-variance equality in Poisson regression models. Cameron and Trivedi (1990b) propose a class of artificial regressions for computing conditional moment tests, and Orme (1990b) develops some rather novel artificial regressions that have 2n observations in some cases and 3n in others, in order to compute information matrix tests and other forms of GM test. No doubt many useful artificial regressions have yet to be discovered and applied.
Thus if, for example, the asymptotic $P$ value of a test statistic that is known to over-reject in finite samples is greater than $\alpha$, so that the test does not reject, the fact that the actual $P$ value is unknown and may well be larger still is of little concern to the investigator. Similarly, if the asymptotic $P$ value of a test statistic is .00002, the investigator may well feel that finding the exact $P$ value is of little interest, unless this particular test is known to be capable of over-rejecting very severely for samples of the size in question.

On the other hand, if the asymptotic $P$ value is near $\alpha$, or if it is substantially less than $\alpha$ but the test is known to be capable of severe over-rejection, it is important to find the exact $P$ value. The conscientious investigator therefore has little choice but to resort to a limited simulation experiment. Such an experiment need not be as elaborate as most published Monte Carlo experiments, because its purpose is simply to estimate the $P$ value(s) associated with the realized value(s) of one or more test statistics. A few hundred replications may be all that it takes to establish whether the $P$ value(s) in question are greater or less than $\alpha$.

On each replication of the experiment one generates a set of data, estimates the model using that data set, and computes one or more test statistics. For each test statistic, one then sees if the value computed on that replication exceeds the actual value computed from the real data set. The fraction of the replications for which this occurs, say $\hat{p}$, then provides an estimate of the $P$ value of interest.

One can estimate a confidence interval for this estimate very easily using the normal approximation to the binomial distribution. A 100(1 - $\beta$) percent confidence interval is

$$\hat{p} - z_{\beta}(\hat{p}(1 - \hat{p})/N)^{1/2} < p < \hat{p} + z_{\beta}(\hat{p}(1 - \hat{p})/N)^{1/2}, \quad (94)$$

where $p$ and $\hat{p}$ are the true and estimated $P$ values, $N$ is the number of replications, and $z_{\beta}$ is the value of a standard normal random variable such that the upper tail probability is $\beta$. Using (94) one can easily determine how large $N$ has to be to obtain a sufficiently accurate estimate of $p$. If $\hat{p}$ is far from $\alpha$, $N$ may not have to be very large at all.

Aside from computational cost, which is becoming less and less of a constraint as computer technology improves, the only real difficulty with this procedure is deciding how to generate the data for the simulation experiment(s). There are two distinct problems: choosing parameter values and generating error terms. Suppose that the model being tested is characterized by a parameter vector $\theta$. The natural thing to do is to use the estimated parameter vector $\hat{\theta}$ when generating the data. This will often work fine, but may yield misleading results if the distribution of the test statistic of interest is not invariant to $\theta$, i.e., if changing $\theta$ affects the distribution. When the distribution of the test statistic does depend on $\theta$, it may be wise to repeat the experiment several times using different values of $\theta$ within the confidence region based on $\hat{\theta}$ and its estimated covariance matrix.

If the estimated $P$ values are always similar, the experimental results are almost certainly reliable. However, if they differ substantially, and are sometimes both larger and smaller than $\alpha$, it may be impossible to estimate the $P$ value with sufficient accuracy by simulation.

When the model imposes strong distributional assumptions on the error terms,
MacKinnon: Model Specification Tests and Artificial Regressions

It is natural to draw the error terms for the simulation from the specified distribution. When the model does not do so, and most regression models do not, it is conventional to use the normal distribution. However, if the error terms are believed to be independent but possibly nonnormal or heterogeneous, it may in such cases be desirable to "bootstrap" them. Bootstrapping, in the restrictive sense in which I am using it here, simply means resampling with replacement either from the empirical distribution of the residuals, perhaps after those residuals have been adjusted to take account of their tendency to underestimate error terms, or from the joint distribution of the data. There is an enormous literature on bootstrapping. Bradley Efron and Cail Gong (1983) provide a good introduction to the subject, Efron and Robert Tibshirani (1986) and David Hinkley (1986) provide more recent surveys, and Michael Veall (1987) provides an interesting econometric application.

The type of simulation experiment discussed above does not involve any sophisticated techniques. However, such techniques can sometimes be very useful, and one should consult some of the large literature on Monte Carlo methods before undertaking any sort of simulation experiment. Hendry (1984) provides a readable introduction to the subject, but contains some recent developments. Two useful books are Brian Ripley (1987) and Peter Lewis and John Grat (1988).

2. The Power of Specification Tests

Specification tests are useless if they do not have sufficient power to reject false null hypotheses. A proper treatment of test power is necessarily quite technical, and is therefore beyond the scope of this paper; Davidson and MacKinnon (1987) and Sakkonen (1995) provide advanced treatments, and Davidson and MacKinnon (1995a) an elementary one. In this section I shall deal with a very simple example that continues the discussion of Part I. Almost everything that is true for this example is, in essence, true asymptotically for any test that can be computed by means of an artificial regression.

Suppose the model to be tested is the linear regression model

\[ y = X\beta + u, \quad u \sim N(0, \sigma^2 I) \]  

(95)

where \( X \) is an \( n \times k \) matrix that is treated as fixed, and \( \sigma^2 \) is assumed to be known. The assumption that \( \sigma^2 \) is known simplifies the finite-sample analysis of power enormously. If the analysis were asymptotic, no such assumption would need to be made, and we could obtain essentially the same results without making it. Jerry Thursby and Schmidt (1977) deal with a case where the sample size is finite and \( \sigma^2 \) is unknown.

The model (95) is to be tested against the alternative

\[ y = X\beta_0 + Z\gamma + u, \quad u \sim N(0, \sigma^2 I) \]  

(96)

where \( Z \) is an \( n \times r \) matrix, although the data are actually generated by the DGP

\[ y = X\beta_0 + \delta w + u, \quad u \sim N(0, \sigma^2 I) \]  

(97)

where \( w \) is an \( n \)-vector. Intuitively, we would expect tests of (95) against (96) to have more power the hotter the latter is able to mimic the DGP (97). The following discussion makes this intuition a little more formal.

To test the hypothesis that \( \gamma = 0 \), one can use the test statistic

\[ y'M_y(Z'M_yZ)^{-1}Z'M_yy/\sigma^2 \]  

(98)

where \( M_y = I - X(X'X)^{-1}X' \). This test statistic is \( r \) times the numerator of the usual \( F \) statistic, divided by \( \sigma^2 \). It would be distributed as \( \chi^2(r) \) if the DGP were a special case of (96). But the DGP is actually (97). Substituting (97) into (98), and using the fact that \( M_yX = 0 \), the test statistic becomes
(\delta_{yw} + w'M_2Z'Z'M_2Z)^{-1} \\
Z'M_2(\delta_{yw} + w/\sigma_2^2). \tag{99}

It can be shown that this quantity is distributed as noncentral \(\chi^2\) with \(r\) degrees of freedom and noncentrality parameter \( \text{nCP} \), or NCP.

\[
\delta_{yw}'M_2Z'Z'M_2Z^{-1}Z'M_2w/\sigma_2^2. \tag{100}
\]

The power of the test statistic (98) will depend on two things: the NCP, expression (100), and the number of degrees of freedom, \(r\). The larger the NCP, the larger, on average, will be the test statistic. In fact, its mean will be equal to \(r\) plus the NCP. The larger the value of the test statistic, the smaller will be the associated \(p\) value and the more likely that we will reject the null hypothesis.

Thus for a given number of degrees of freedom, power is maximized by making the NCP as large as possible. For a given NCP, on the other hand, power will decline as the number of degrees of freedom increases, because the test statistic will be compared to larger critical values.

The NCP (100) can be computed as the explained sum of squares from the extremely artificial regression

\[
(\delta_{yw}/\sigma_2)M_{ww} = M_2Zb + \text{residuals}. \tag{101}
\]

This explained sum of squares can be rewritten as

\[
R^2_{yw}/(\sigma_2)w'M_2w, \tag{102}
\]

where \(R^2_{yw}\) denotes the uncentered \(R^2\) from regression (101). The second and third factors here depend only on the model that is being tested and the DGP. The third factor is the sum of squared residuals from a regression of \(w\) on \(X\). Essentially, these two factors measure how far the DGP is from the closest linear prediction on the null hypothesis, relative to the variance of the error terms. If the product of these two factors is small, because the sample size is small, the error terms are noisy, or the null hypotheses (95) does not differ much from the DGP (97), the test will have much power.

It is only through the first factor, \(R^2_{yw}\), that \(Z\) affects the NCP (102) at all. Ideally, we would like to pick \(Z\) so that \(M_2Z\) explains all the variation in \(M_{ww}\), making \(R^2_{yw}\) unity and the NCP as large as possible. This will happen whenever the DGP is a special case of the alternative we test against. At the other extreme, when \(M_2Z\) is unable to explain any of the variation in \(M_{ww}\), the NCP will be zero and the test statistic (98) will have power equal to its significance level, that is, no useful power at all. In between these two extremes, every test for which \(M_2Z\) explains some of the variation in \(M_{ww}\) will have some usable power, and all such tests will reject the null hypothesis with probability one when \((\delta_{yw}/\sigma_2)w'M_2w\) is sufficiently large. Thus it is quite possible that many different tests will lead us to reject the hypothesis that the original model (85) is correctly specified, even when none of the alternatives is correctly specified either.

Suppose that our objective in choosing \(Z\) is to maximize the power of the test, given a specified significance level. There is necessarily a tradeoff involved in doing this. By adding extra columns to \(Z\) we increase the NCP for the test, or at worst leave it unchanged, because regression (101) will generally fit better. But doing so also increases the number of degrees of freedom. The former effect increases power, while the latter decreases it. Notice that \(w'M_2w\), since it is essentially a sum of squared residuals, will tend to be proportional to the sample size. Thus from (102) we see that a given increase in \(R^2_{yw}\) will have a greater effect on the NCP the larger is the sample size. This suggests that, when the sample size is large, it may be desirable to include a lot of columns in \(Z\), in the hope of making \(R^2_{yw}\) large. When the sample size is small, on the other hand, this strategy
will be less attractive, because even if we succeed in making $R^2$ large, the NCP may not be large enough to give adequate power to a test with many degrees of freedom.

Although the example is a very special one, the basic insights it yields are true very generally. In place of (96), think of the GNR (10) or the general artificial regression (76). Then everything that has been said about the choice of $Z$ in (96) applies equally well, as far as asymptotic analysis is concerned, to the choice of $Z$ in (10). The example suggests three principal conclusions. First, no test will have much power if the sample size is too small and/or the model being tested is not sufficiently misspecified. Second, when the sample size is small, the most powerful tests will tend to be ones with few degrees of freedom. Finally, when the sample size is large and/or the model is seriously misspecified, tests with many degrees of freedom may be attractive, because for such tests the analogue of $R^2$ is likely to be relatively large. In such cases it may be attractive to test against two or more alternatives at the same time by including in the regression the test regressors suggested by all of them.

V. Final Remarks

In this article I have taken the point of view that investigators should test every econometric model they estimate for any type of misspecification that might reasonably be expected to be present. If they find evidence of misspecification, they should, if at all possible, revise their models to eliminate it. This is of course not the only view of specification testing, as evidenced by the stimulating articles of Pagan (1987) and Hendry, Edward Leamer, and Dale Poirier (1990). For reasons of space, however, I have not attempted to discuss objections to this point of view, or alternatives to it, or indeed any fundamental issues of econometric methodology at all.

The literature on model specification testing is much larger than a single review article can do justice to. To keep the size of this one manageable, I have chosen to emphasize tests that are based on artificial regressions, although in many cases these same tests may also be thought of as Lagrange Multiplier tests, C(a) tests, or conditional moment tests. The artificial regressions approach has a number of merits. In many cases it provides the easiest way to compute specification tests, and in most cases it provides a reasonably easy way to do so. Moreover, it is fruitful to think about most aspects of model specification in terms of a linear regression model with possibly omitted variables. By recasting every specification test as a test of whether some coefficients in a regression are zero, the artificial regressions approach fits perfectly with this view of specification testing, even though it is in no way limited to testing the specification of regression functions.

The artificial regressions approach is particularly convenient when, as is usually the case, one wants to test a model against several different alternatives, because a large number of tests against different alternatives can generally be calculated using slightly different versions of the same artificial regression. The regressand and the regressors that correspond to parameters estimated under the null remain the same, while the test regressors that correspond to parameters not estimated under the null vary from one artificial regression to the next. Once one has set up the basic artificial regression, it is generally very easy to use it to calculate a variety of test statistics. Of course, interpreting the results of a number of specification tests of the same model requires a good deal of care, as the discussion of Section IV.B makes
clear. It is entirely possible that several tests against different alternatives may reject the hypothesis that the model is correctly specified, even though only one type of misspecification is actually present.

References


MacKinnon: Model Specification Tests and Artificial Regressions


KRÄGER, WALTER, ed. Econometrics of Structural Change, Heidelberg: Physica-Verlag, 1986. [also Empirical Econ., 13(3).]


- “Heteroskedasticity/Robust Tests for Structural -
null