Specification Tests Based on Artificial Regressions

Author(s): Russell Davidson and James G. MacKinnon

Source: *Journal of the American Statistical Association*, Mar., 1990, Vol. 85, No. 409 (Mar., 1990), pp. 220-227

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2289548

Specification Tests Based on Artificial Regressions

RUSSELL DAVIDSON and JAMES G. MacKINNON*

Many specification tests can be computed with artificial linear regressions designed to be used as calculating devices to obtain test statistics and other quantities of interest. This article discusses the general principles that underlie all artificial regressions, and the use of such regressions to compute Lagrange multiplier and other specification tests based on estimates under the null hypothesis. The generality and power of artificial regressions as a means of computing test statistics is demonstrated; how Durbin–Wu–Hausman, conditional moment, and other tests that are not explicitly Lagrange multiplier tests may be computed is shown; and several special cases that illustrate the general results and can be useful in practice are discussed. These include tests of parameter restrictions in nonlinear regression models and tests of binary-choice models such as the logit and probit models.

KEY WORDS: Artificial regression; Binary-choice model; Conditional moment test; Durbin–Wu–Hausman test; Lagrange multiplier test; Nonlinear regression model.

1. INTRODUCTION

Numerous specification tests have been proposed that can be computed with artificial linear regressions. These regressions are artificial in the sense that they are designed to be used solely as calculating devices; the regressand and regressors are constructed so that the desired test statistic is equal to or can easily be computed from one of the quantities normally calculated by an ordinary least squares (OLS) regression program. Artificial regressions essentially the same as those used to calculate test statistics can be used for other purposes, such as calculating consistent estimates of the asymptotic covariance matrix of a vector of parameter estimates and computing one-step efficient estimates from an initial consistent estimate.

The test statistic from an artificial regression is often n (the sample size) times the R^2, sometimes the explained sum of squares, or sometimes an ordinary t test or F test based on the artificial regression. These tests are in some cases derived explicitly as Lagrange multiplier (LM) tests in their score form, and in other cases they are equivalent to such tests. Although many of the tests based on artificial regressions are well known, there has not to our knowledge been an exposition of the general principles that underlie them that may be used to develop new tests and extend existing ones. This article aims to provide such a general exposition, to demonstrate the generality and power of artificial regressions as a means of computing test statistics, and to discuss several special cases that illustrate the general results and can be useful in practice.

2. SOME EXAMPLES

In this section we informally present three well-known artificial regressions and indicate how they may be used for the calculation of test statistics. In Section 3 we discuss matters more formally and in greater generality.

The best-known artificial regression is almost certainly the Gauss–Newton (GN) regression, which was originally derived as a way to calculate least squares estimates for nonlinear regression models (Hartley 1961). Engle (1982, 1984) discussed its use for testing restrictions on nonlinear regression models, dealing with multivariate as well as univariate models. The simplest form of the GN regression applies to the class of univariate nonlinear regression models:

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \tag{1}$$

where $\boldsymbol{\beta} \equiv [\boldsymbol{\beta}_1^T \; \boldsymbol{\beta}_2^T]^T$ is a k vector of parameters (with $k = k_1 + k_2$) and $\mathbf{x}(\boldsymbol{\beta})$ is an n vector of nonlinear functions, which would usually depend on exogenous variables and perhaps lagged values of the dependent variable. The parameter vector $\boldsymbol{\beta}$ has been partitioned because we wish to consider testing the hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$.

The GN regression may be obtained as a first-order Taylor-series approximation to (1) around some value of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^*$. Its general form is

$$\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^*) = \mathbf{X}(\boldsymbol{\beta}^*)\mathbf{b} + \text{residuals}. \tag{2}$$

Here and elsewhere, *residuals* means simply the difference between the regressand and the rest of the right side of the regression. It indicates that no statistical meaning is intended: We merely have a linear regression, which when run may yield useful results. The matrix of derivatives $\mathbf{X}(\boldsymbol{\beta}) \equiv [\mathbf{X}_1(\boldsymbol{\beta}) \; \mathbf{X}_2(\boldsymbol{\beta})]$ is an $n \times k$ matrix with the tith element the derivative of $x_t(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_i$.

If we estimate (1) subject to the restriction that $\boldsymbol{\beta}_2 = \mathbf{0}$ to obtain restricted estimates $\tilde{\boldsymbol{\beta}} \equiv [\tilde{\boldsymbol{\beta}}_1^T \mathbf{0}^T]^T$, the GN regression (2) becomes

$$\tilde{\mathbf{u}} = \tilde{\mathbf{X}}\mathbf{b} + \text{residuals} = \tilde{\mathbf{X}}_1\mathbf{b}_1 + \tilde{\mathbf{X}}_2\mathbf{b}_2 + \text{residuals}, \tag{3}$$

where $\tilde{\mathbf{u}} \equiv \mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}})$ and $\tilde{\mathbf{X}} \equiv \mathbf{X}(\tilde{\boldsymbol{\beta}})$. Here and elsewhere we employ a useful notation whereby functions of parameter vectors (in this case $\boldsymbol{\beta}$) that are evaluated at particular values such as $\tilde{\boldsymbol{\beta}}$ may be written without making the argument explicit. The regressand of (3) is simply the vector of residuals from restricted estimation of (1), and there are k regressors, each of which is a vector of the derivatives of $\mathbf{x}(\boldsymbol{\beta})$ with respect to one of the elements of $\boldsymbol{\beta}$, evaluated

* Russell Davidson is Professor, Department of Economics, Queen's University, Kingston, Ontario K7L 3N6, Canada, and Professor, GREQE, Ecoles des Hautes Etudes en Sciences Sociales, Universités d'Aix-Marseille II et III, Centre de la Vieille Charité, 2 Rue de la Charité, 13002 Marseille, France. James G. MacKinnon is Professor, Department of Economics, Queen's University, Kingston, Ontario K7L 3N6, Canada. This research was supported in part by grants from the Social Sciences and Humanities Research Council of Canada. The authors are grateful to an anonymous referee for comments on earlier versions.

Journal of the American Statistical Association
March 1990, Vol. 85, No. 409, Theory and Methods

at $\tilde{\boldsymbol{\beta}}$. If $\mathbf{x}(\boldsymbol{\beta})$ were a linear regression function with $\mathbf{Z}$ the matrix of independent variables, $\tilde{\mathbf{X}}$ would simply be equal to $\mathbf{Z}$. When the artificial regression (3) is run, nR^2 is asymptotically equivalent to any asymptotically efficient chi-squared or F test of the restrictions $\boldsymbol{\beta}_2 = \mathbf{0}$. When the restrictions are valid, the statistic will under appropriate regularity conditions be distributed as central $\chi^2(k_2)$. An ordinary F test for $\mathbf{b}_2 = \mathbf{0}$ would be asymptotically equivalent to the nR^2 test.

A second well-known example of an artificial regression is provided by the so-called outer-product-of-the-gradient (OPG) regression. Unlike the GN regression, which applies only to nonlinear least squares models, the OPG regression is almost universally applicable to models that can be estimated by maximum likelihood. Suppose that there is a sample of size n that gives rise to a log-likelihood function

$$\mathscr{L}(\boldsymbol{\theta}) \equiv \mathscr{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \tag{4}$$

where $\boldsymbol{\theta}_1$ is a k_1 vector and $\boldsymbol{\theta}_2$ is a k_2 vector of parameters with $k = k_1 + k_2$. We define a matrix $\mathbf{G}(\boldsymbol{\theta})$ with typical element $G_{jt}(\boldsymbol{\theta}) = \partial l_t(\boldsymbol{\theta})/\partial\theta_j$. This is the contribution to the gradient of the log-likelihood function with respect to the jth parameter made by the tth observation. The vector of scores associated with $\mathscr{L}(\boldsymbol{\theta})$, that is, the gradient of the log-likelihood function (4), is denoted by $\mathbf{g}(\boldsymbol{\theta}) \equiv \mathbf{G}^T(\boldsymbol{\theta})\boldsymbol{\iota}$, where $\boldsymbol{\iota}$ is an n vector of 1s.

As is well known, the matrix $n^{-1}\mathbf{G}^T(\ddot{\boldsymbol{\theta}})\mathbf{G}(\ddot{\boldsymbol{\theta}})$ consistently estimates the expectation of the outer product of the gradient, the information matrix $\mathscr{I}(\boldsymbol{\theta})$, whenever $\ddot{\boldsymbol{\theta}}$ consistently estimates $\boldsymbol{\theta}$. Then, if we let $\tilde{\boldsymbol{\theta}}$ denote the maximum likelihood (ML) estimates obtained by maximizing the log-likelihood function (4) subject to the restrictions $\boldsymbol{\theta}_2 = \mathbf{0}$, the OPG artificial regression used for testing these restrictions is

$$\boldsymbol{\iota} = \mathbf{G}(\tilde{\boldsymbol{\theta}})\mathbf{c} + \text{residuals}. \tag{5}$$

The quantity nR^2 from Regression (5), which in this case is equal to the explained sum of squares from the regression, is a test statistic asymptotically distributed as $\chi^2(k_2)$ under the null hypothesis. An early application of this procedure was given by Godfrey and Wickens (1981).

Since the OPG regression is almost always available, one may ask why other artificial regressions are needed at all. They are needed because the finite-sample properties of the OPG regression are often poor. The explained sum of squares from the OPG regression (5) with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ is

$$\boldsymbol{\iota}^T\tilde{\mathbf{G}}(\tilde{\mathbf{G}}^T\tilde{\mathbf{G}})^{-1}\tilde{\mathbf{G}}^T\boldsymbol{\iota} = \tilde{\mathbf{g}}^T(\tilde{\mathbf{G}}^T\tilde{\mathbf{G}})^{-1}\tilde{\mathbf{g}}. \tag{6}$$

The only difference between this and any other form of the LM statistic is that the matrix $n^{-1}\tilde{\mathbf{G}}^T\tilde{\mathbf{G}}$ is used to estimate the information matrix $\mathscr{I}(\boldsymbol{\theta})$. Presumably because $n^{-1}\tilde{\mathbf{G}}^T\tilde{\mathbf{G}}$ often provides a poor estimate of $\mathscr{I}(\boldsymbol{\theta})$ when n is not very large, test statistics based on the OPG regression often have finite-sample distributions that are poorly approximated by their asymptotic distributions. Monte Carlo evidence on this point has been provided by Davidson and MacKinnon (1984a, 1985a), Bera and McKenzie (1986), and Godfrey, McAleer, and McKenzie (1988), among others. All of these articles found that variants of the LM statistic (6) were much more prone to reject incorrectly the null hypothesis than alternative forms of the LM test, many of which were based on different (and less generally applicable) artificial regressions.

The last artificial regression we consider in this section is associated with the simplest type of binary-choice model. The dependent variable y_t may be either 0 or 1, and it is assumed that $\Pr(y_t = 1) = \Psi(\mathbf{X}_t\boldsymbol{\beta})$, where $\Psi(x)$ is a thrice continuously differentiable function that maps from the real line to the 0–1 interval, is weakly increasing in x, and has the properties $\Psi(x) \geq 0$, $\Psi(-\infty) = 0$, $\Psi(\infty) = 1$, and $\Psi(-x) = 1 - \Psi(x)$. Two examples are the probit model, where $\Psi(x)$ is the cumulative standard normal distribution function, and the logit model, where $\Psi(x)$ is the logistic function $(1 + \exp(-x))^{-1}$.

For binary-choice models of this type the artificial regression uses as regressand a vector $\mathbf{r}(\boldsymbol{\beta})$ with typical element

$$r_t(\boldsymbol{\beta}) = (\Psi(-\mathbf{X}_t\boldsymbol{\beta})\Psi(\mathbf{X}_t\boldsymbol{\beta}))^{-1/2}(y_t - \Psi(\mathbf{X}_t\boldsymbol{\beta})), \tag{7}$$

and as regressors a matrix $\mathbf{R}(\boldsymbol{\beta})$ with typical element

$$R_{ti}(\boldsymbol{\beta}) = (\Psi(-\mathbf{X}_t\boldsymbol{\beta})\Psi(\mathbf{X}_t\boldsymbol{\beta}))^{-1/2}\psi(\mathbf{X}_t\boldsymbol{\beta})X_{ti}, \tag{8}$$

where $\psi(x)$ is the first derivative of $\Psi(x)$. This artificial regression can be derived as a variant of the GN regression. We can write a binary-choice model as

$$y_t = \Psi(\mathbf{X}_t\boldsymbol{\beta}) + u_t, \tag{9}$$

where u_t equals either $1 - \Psi(\mathbf{X}_t\boldsymbol{\beta})$ or $-\Psi(\mathbf{X}_t\boldsymbol{\beta})$ and can easily be shown to have variance $\Psi(-\mathbf{X}_t\boldsymbol{\beta})\Psi(\mathbf{X}_t\boldsymbol{\beta})$. Taylor-expanding (9) as if it were an ordinary nonlinear regression model and correcting for the heteroscedasticity of the u_t's yields the artificial regression defined by (7) and (8).

If the parameter vector $\boldsymbol{\beta}$ is partitioned as $\beta^T = [\boldsymbol{\beta}_1^T\ \boldsymbol{\beta}_2^T]^T$ and the binary-choice model is estimated by maximum likelihood subject to the restrictions $\boldsymbol{\beta}_2 = \mathbf{0}$, with restricted ML estimates $\tilde{\boldsymbol{\beta}}^T \equiv [\tilde{\boldsymbol{\beta}}_1^T\ \mathbf{0}]$, then running the artificial regression $\mathbf{r}(\tilde{\boldsymbol{\beta}}) = \mathbf{R}(\tilde{\boldsymbol{\beta}})\mathbf{b} + \text{residuals}$ yields an nR^2 or an explained sum of squares, either of which can serve as a chi-squared test of the restrictions. The two are not numerically equal, and there is reason to prefer the explained sum of squares in finite samples. For more details, and extensions, see Engle (1984) and Davidson and MacKinnon (1984a).

3. THE GENERAL CASE

All of the artificial regressions discussed in Section 2 (and many others) can be understood in terms of a general framework of artificial regressions that share a set of basic properties and can therefore be used for a wide variety of purposes. We deal with the following general case. There is a fully specified parameterized model characterized by its log-likelihood function, which for a sample of size n

can be written as

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}), \tag{10}$$

where $\boldsymbol{\theta}$ is a k vector of model parameters. We frequently wish to partition $\boldsymbol{\theta}$ as $[\boldsymbol{\theta}_1^T \; \boldsymbol{\theta}_2^T]^T$ to consider the restrictions $\boldsymbol{\theta}_2 = \mathbf{0}$. In such cases, $\boldsymbol{\theta}_i$ is a k_i vector ($i = 1, 2$) with $k = k_1 + k_2$. As before, $\mathbf{g}(\boldsymbol{\theta})$ denotes the k vector of scores and $\mathbf{G}(\boldsymbol{\theta})$ denotes the $n \times k$ matrix of the derivatives of $l_t(\boldsymbol{\theta})$ with respect to the elements of $\boldsymbol{\theta}$.

We assume that the data were generated by a data-generating process (DGP) characterized by the log-likelihood (10) for some true (but unknown) parameter vector $\boldsymbol{\theta}^0$ such that

$$\boldsymbol{\theta}^0 \equiv \begin{bmatrix} \boldsymbol{\theta}_1^0 \\ \boldsymbol{\theta}_2^0 \end{bmatrix}.$$

Often we additionally assume that $\boldsymbol{\theta}_2^0 = \mathbf{0}$. The model represented by (10) is assumed to satisfy all of the usual conditions for ML estimation and inference to be asymptotically valid (e.g., see Amemiya 1985, chap. 4). In particular, we assume that the true parameter vector $\boldsymbol{\theta}^0$ is interior to a compact parameter space Θ, and that the information matrix $\mathcal{I}(\boldsymbol{\theta}) \equiv \lim(E(n^{-1}\mathbf{g}(\boldsymbol{\theta})\mathbf{g}^T(\boldsymbol{\theta})))$, which in this case is $k \times k$, is a finite, nonsingular matrix for all $\boldsymbol{\theta}$ in Θ.

Various artificial regressions can be associated with the model (10). They always involve two things: a regressand, say $\mathbf{r}(\boldsymbol{\theta})$, and a matrix of regressors, say $\mathbf{R}(\boldsymbol{\theta})$. The artificial regression can be evaluated at any point $\boldsymbol{\theta} \in \Theta$. It may be written as

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{R}(\boldsymbol{\theta})\mathbf{b} + \text{residuals}. \tag{11}$$

Note again that we use residuals as a neutral term to avoid any implication that (11) is a statistical model.

We develop our theory of artificial regressions under the assumption that $\mathbf{r}(\boldsymbol{\theta})$ and $\mathbf{R}(\boldsymbol{\theta})$ have certain defining properties. These properties are as follows, where all plim's are calculated with DGP any process characterized by the log-likelihood (10) for some set of parameters in Θ:

Property 1. Under the DGP characterized by $\boldsymbol{\theta}$, $\operatorname{plim}_{n\to\infty} n^{-1}\mathbf{r}^T(\boldsymbol{\theta})\mathbf{r}(\boldsymbol{\theta})$ exists and is a finite, smooth, real-valued function of $\boldsymbol{\theta}$, the value of which is denoted by $\rho(\boldsymbol{\theta})$.

Property 2. $\mathbf{R}^T(\boldsymbol{\theta})\mathbf{r}(\boldsymbol{\theta}) = \rho(\boldsymbol{\theta})\mathbf{g}(\boldsymbol{\theta})$.

Property 3. If $\ddot{\boldsymbol{\theta}} \to \boldsymbol{\theta}^0$, then $n^{-1}\mathbf{R}^T(\ddot{\boldsymbol{\theta}})\mathbf{R}(\ddot{\boldsymbol{\theta}}) \to \rho(\boldsymbol{\theta}^0)\mathcal{I}(\boldsymbol{\theta}^0)$.

These properties are not shared by every linear regression used solely to calculate quantities of interest, which we call auxiliary regressions. Nevertheless, in this article any regression termed artificial will satisfy these three properties.

The two crucial features of artificial regressions satisfying Properties 1–3 are expressed in the following two theorems.

Theorem 1. Suppose that the artificial regression (11) is associated with the fully specified parameterized model (10), in the sense that Properties 1–3 are satisfied for all $\boldsymbol{\theta} \in \text{int } \Theta$, a compact k-dimensional parameter space. Suppose further that the model (10) satisfies the regularity conditions of Amemiya (1985, chap. 4). Then, if the artificial regression is evaluated at some $\acute{\boldsymbol{\theta}} \in \Theta$ such that $\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = O(n^{-1/2})$, the artificial parameter estimates $\acute{\mathbf{b}}$ obtained by OLS have the property that

$$n^{1/2}\acute{\mathbf{b}} = n^{1/2}(\hat{\boldsymbol{\theta}} - \acute{\boldsymbol{\theta}}) + o_p(1) \quad \text{as } n \to \infty, \tag{12}$$

where $\hat{\boldsymbol{\theta}}$ is the (asymptotically efficient) ML estimator of the model (10).

Proof. The proof of Theorem 1 is both simple and illuminating. A Taylor expansion of the gradient $\mathbf{g}(\acute{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}^0$ yields

$$\mathbf{g}(\acute{\boldsymbol{\theta}}) \cong \mathbf{g}(\boldsymbol{\theta}^0) + \mathbf{H}(\boldsymbol{\theta}^0)(\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = \mathbf{g}^0 + \mathbf{H}^0(\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}^0), \tag{13}$$

where $\mathbf{H}^0 \equiv \mathbf{H}(\boldsymbol{\theta}^0)$ denotes the Hessian matrix of the log-likelihood function $\mathcal{L}(\boldsymbol{\theta})$ and $\cong$ denotes asymptotic equivalence. Multiplying all quantities in (13) by appropriate powers of n so that they are $O(1)$, and using Properties 2 and 3 and the fact that $\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = O(n^{-1/2})$, we see that

$$\begin{aligned} n^{-1/2}\acute{\mathbf{R}}^T\acute{\mathbf{r}} &= n^{-1/2}\rho^0\mathbf{g}(\acute{\boldsymbol{\theta}}) \\ &\cong n^{-1/2}\rho^0\mathbf{g}^0 - (n^{-1}\mathbf{R}^{0T}\mathbf{R}^0)n^{1/2}(\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}^0). \end{aligned} \tag{14}$$

If (14) is now evaluated at the ML estimator $\hat{\boldsymbol{\theta}}$ instead of at $\acute{\boldsymbol{\theta}}$, the left side of the equation is 0 by the first-order conditions for the maximum of the log-likelihood function, so

$$\mathbf{0} \cong n^{-1/2}\rho^0\mathbf{g}^0 - (n^{-1}\mathbf{R}^{0T}\mathbf{R}^0)n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0). \tag{15}$$

Subtracting (15) from (14) then yields $(n^{-1}\mathbf{R}^{0T}\mathbf{R}^0)n^{1/2}(\hat{\boldsymbol{\theta}} - \acute{\boldsymbol{\theta}}) = n^{-1/2}\acute{\mathbf{R}}^T\acute{\mathbf{r}}$. Using $\mathcal{I}(\boldsymbol{\theta}^0) \cong n^{-1}\mathbf{R}^{0T}\mathbf{R}^0 \cong n^{-1}\acute{\mathbf{R}}^T\acute{\mathbf{R}}$, we obtain (12) and thus prove the theorem.

Theorem 2. Under the regularity conditions of Theorem 1, nR^2 from the artificial regression (11) evaluated at any $\acute{\boldsymbol{\theta}} \in \Theta$ such that $\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = O(n^{-1/2})$ is asymptotically equal to

$$n^{-1}\acute{\mathbf{g}}^T\mathcal{I}^{-1}(\boldsymbol{\theta}^0)\acute{\mathbf{g}}. \tag{16}$$

Proof. The R^2 from (11) is equal to the ratio of the explained sum of squares to the total sum of squares. The total sum of squares, divided by the sample size n, tends to ρ^0 as $n \to \infty$, by Property 1 and $\acute{\boldsymbol{\theta}} \to \boldsymbol{\theta}^0$. The explained sum of squares is $\acute{\mathbf{r}}^T\acute{\mathbf{R}}(\acute{\mathbf{R}}^T\acute{\mathbf{R}})^{-1}\acute{\mathbf{R}}^T\acute{\mathbf{r}}$. By Property 2 this becomes $n^{-1}\acute{\mathbf{r}}^T\acute{\mathbf{R}}(n^{-1}\acute{\mathbf{R}}^T\acute{\mathbf{R}})^{-1}\acute{\mathbf{R}}^T\acute{\mathbf{r}} = n^{-1}\acute{\rho}^2\acute{\mathbf{g}}^T(n^{-1}\acute{\mathbf{R}}^T\acute{\mathbf{R}})^{-1}\acute{\mathbf{g}}$, which by Property 3 is asymptotically equal to $n^{-1}\rho^0\acute{\mathbf{g}}^T\mathcal{I}(\boldsymbol{\theta}^0)^{-1}\acute{\mathbf{g}}$. The asymptotic equivalence of (16) and nR^2 from (11) then follows at once.

By itself, Theorem 2 does not say anything about the distribution of nR^2, but it underlies the use of artificial regressions for LM tests. Let $\tilde{\boldsymbol{\theta}}$ denote a vector of ML estimates subject to the restriction that $\boldsymbol{\theta}_2 = \mathbf{0}$. The score form of the LM test statistic for testing the hypothesis $\boldsymbol{\theta}_2 = \mathbf{0}$ against the alternative $\boldsymbol{\theta}_2 \neq \mathbf{0}$ is

$$n^{-1}\tilde{\mathbf{g}}^T\mathcal{I}(\theta^0)^{-1}\tilde{\mathbf{g}}. \tag{17}$$

If $\boldsymbol{\theta}_2^0 = \mathbf{0}$, it is a familiar result (e.g., see Cox and Hinkley 1974, chap. 9) that (17) is asymptotically distributed as $\chi^2(k_2)$. Moreover, any test statistic of the form of (17) but with the matrix $\mathcal{I}(\boldsymbol{\theta}^0)$ replaced by any matrix that estimates $\mathcal{I}(\boldsymbol{\theta}^0)$ consistently under the null hypothesis is asymptotically equivalent to (17). It follows directly from Theorem 2 that if we are given $\tilde{\mathbf{r}} \equiv \mathbf{r}(\tilde{\boldsymbol{\theta}})$ and $\tilde{\mathbf{R}} \equiv \mathbf{R}(\tilde{\boldsymbol{\theta}})$ that satisfy Properties 1–3, the LM statistic (17) (or an asymptotically equivalent test statistic) may be computed as nR^2 from the artificial regression

$$\tilde{\mathbf{r}} = \tilde{\mathbf{R}}\mathbf{b} + \text{residuals}. \tag{18}$$

In many cases $\rho(\boldsymbol{\theta})$ is equal to 1, perhaps after rescaling of the artificial variables $\mathbf{r}$ and $\mathbf{R}$, and in some cases $n^{-1}\tilde{\mathbf{r}}^T\tilde{\mathbf{r}}$ is equal to 1. In all such cases the explained sum of squares from (18) is asymptotically equal to the test statistic (17). Other expressions asymptotically equal to (17) can be found. It is easy to show that k_2 times the F statistic for the (artificial) hypothesis $\mathbf{b}_2 = \mathbf{0}$ is such an expression. When $k_2 = 1$, so only one restriction is being tested, the square of the t test for $\mathbf{b}_2 = \mathbf{0}$ is asymptotically equal to (17), and the t test itself is asymptotically valid. Which of these variants of the LM test statistic will in finite samples have a distribution closest to the nominal asymptotic one depends on the details of the model being tested.

It is sometimes useful to make explicit the distinction between $\tilde{\mathbf{R}}_1$ and $\tilde{\mathbf{R}}_2$. Regression (18) can be written as

$$\tilde{\mathbf{r}} = \tilde{\mathbf{R}}_1\mathbf{b}_1 + \tilde{\mathbf{R}}_2\mathbf{b}_2 + \text{residuals}. \tag{19}$$

The first-order conditions for $\tilde{\boldsymbol{\theta}}$ imply that $\tilde{\mathbf{r}}^T\tilde{\mathbf{R}}_1 = \mathbf{0}$, so the explained and total sums of squares from (19) must be identical to those from the regression $\tilde{\mathbf{r}} = \tilde{\mathbf{M}}_1\tilde{\mathbf{R}}_2\mathbf{b}_2 +$ residuals, where $\tilde{\mathbf{M}}_1 \equiv \mathbf{I} - \tilde{\mathbf{R}}_1(\tilde{\mathbf{R}}_1^T\tilde{\mathbf{R}}_1)^{-1}\tilde{\mathbf{R}}_1^T$ is the matrix that projects orthogonally onto the orthogonal complement of the subspace spanned by $\tilde{\mathbf{R}}_1$. Thus we see that the numerator of the nR^2 form of the test can be written as

$$\tilde{\mathbf{r}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{R}}_2(\tilde{\mathbf{R}}_2^T\tilde{\mathbf{M}}_1\tilde{\mathbf{R}}_2)^{-1}\tilde{\mathbf{R}}_2^T\tilde{\mathbf{M}}_1\tilde{\mathbf{r}}$$
$$= (n^{-1/2}\tilde{\mathbf{r}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{R}}_2)(n^{-1}\tilde{\mathbf{R}}_2^T\tilde{\mathbf{M}}_1\tilde{\mathbf{R}}_2)^{-1}(n^{-1/2}\tilde{\mathbf{R}}_2^T\tilde{\mathbf{M}}_1\tilde{\mathbf{r}}). \tag{20}$$

Clearly, when the test statistic is written in this form it must have k_2 df, since $n^{-1/2}\tilde{\mathbf{R}}_2^T\tilde{\mathbf{M}}_1\tilde{\mathbf{r}}$ is a k_2 vector.

In some cases of practical interest, some of the parameters $\boldsymbol{\theta}_1$ that may vary freely under the null hypothesis can be treated as nuisance parameters and the artificial regressors corresponding to them dropped from the artificial regression. This situation arises if the information matrix is block-diagonal between these nuisance parameters and all of the other parameters of the model, so those columns of $\mathbf{R}(\boldsymbol{\theta})$ that correspond to the nuisance parameters will be asymptotically orthogonal to all of the remaining columns of $\mathbf{R}(\boldsymbol{\theta})$. This situation arises in the case of the nonlinear regression model with normal errors and allows us to show that the GN regression (3) can be regarded as a special case of the general class of artificial regressions discussed before.

If we assume that the vector $\mathbf{u}$ in (1) is normally distributed, the contribution from the tth observation to the log-likelihood function is $l_t(\boldsymbol{\beta}, \sigma) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - (y_t - x_t(\boldsymbol{\beta}))^2/(2\sigma^2)$. The derivatives of $l_t(\boldsymbol{\beta}, \sigma)$ are

$$\partial l_t/\partial\beta_i = \sigma^{-2}X_{ti}(\boldsymbol{\beta})(y_t - x_t(\boldsymbol{\beta})) \tag{21}$$

and

$$\partial l_t/\partial\sigma = -\sigma^{-1} + \sigma^{-3}(y_t - x_t(\boldsymbol{\beta}))^2. \tag{22}$$

When (21) is multiplied by (22) the expectation of the resulting product is 0, which establishes the familiar result that the information matrix for nonlinear regression models is block-diagonal between $\boldsymbol{\beta}$ and σ. If we are only interested in restrictions on $\boldsymbol{\beta}$, we can construct an artificial regression in which σ is treated as a nuisance parameter. The block-diagonality property implies that the artificial regression need not include a regressor corresponding to σ.

If we make the following definitions corresponding to the artificial variables used in the GN regression (2), $r_t(\boldsymbol{\beta}) = y_t - x_t(\boldsymbol{\beta})$, and $R_{ti}(\boldsymbol{\beta}) = X_{ti}(\boldsymbol{\beta})$ for $i = 1, \ldots, k$, we see that these artificial variables satisfy the defining Properties 1–3. In particular, $\rho(\boldsymbol{\beta}) = \sigma^2$ and $\mathbf{R}^T(\boldsymbol{\beta})\mathbf{r}(\boldsymbol{\beta}) = \mathbf{X}^T(\boldsymbol{\beta})(\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}))$, which is the gradient of the normal log-likelihood with respect to $\boldsymbol{\beta}$, times $\rho(\boldsymbol{\beta})$ $(= \sigma^2)$, and $n^{-1}\mathbf{R}^T(\boldsymbol{\beta})\mathbf{R}(\boldsymbol{\beta}) = n^{-1}\mathbf{X}^T(\boldsymbol{\beta})\mathbf{X}(\boldsymbol{\beta})$, which is $\rho(\boldsymbol{\beta})$ times the $\boldsymbol{\beta}$–$\boldsymbol{\beta}$ block of the information matrix.

It is well known that the assumption of normality is unnecessary for the asymptotic theory of nonlinear regression models estimated by least squares. The conventional theory of the GN regression shows that it too can be used validly with models in which the errors are nonnormal. It may be possible to extend the general theory of artificial regressions to a semiparametric context; the use of the GN regression with nonnormal errors would then be covered. Such an extension is beyond the scope of this article, however.

In contrast, the assumption of homoscedasticity is essential if we are to make asymptotic inferences based on the usual estimated covariance matrix $s^2(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}$. But the work of Eicker (1963) and White (1980), among others, has shown that it is possible to make valid inferences asymptotically even in the presence of heteroscedasticity of unknown form. Auxiliary regressions exist that permit the straightforward calculation of test statistics robust to the presence of such heteroscedasticity (e.g., see Davidson and MacKinnon 1985b; Wooldridge 1987). Nevertheless, these auxiliary regressions have only k_2 rather than k regressors, and thus cannot satisfy Properties 1–3. Although they can be useful, they are therefore not artificial regressions in our sense.

Artificial regressions can be useful for a variety of purposes besides calculating test statistics. For example, suppose that we evaluate $\mathbf{r}(\boldsymbol{\theta})$ and $\mathbf{R}(\boldsymbol{\theta})$ at the unrestricted estimates $\hat{\boldsymbol{\theta}}$ and run the regression

$$\hat{\mathbf{r}} = \hat{\mathbf{R}}\mathbf{b} + \text{residuals}. \tag{23}$$

It is obvious by the first-order conditions that the OLS estimate $\hat{\mathbf{b}}$ must be identically 0, so running (23) is an easy way to verify that $\hat{\boldsymbol{\theta}}$ satisfies the first-order conditions.

Moreover, the OLS covariance matrix estimate from (23) is

$$(\hat{\mathbf{r}}^T\hat{\mathbf{r}}/(n - k))(\hat{\mathbf{R}}^T\hat{\mathbf{R}})^{-1},$$

and by Properties 1 and 3 this is evidently a valid estimate of the inverse of the information matrix. Whether it is an estimate that we would actually want to use in practice will depend in part on whether $n^{-1}\hat{\mathbf{r}}^T\hat{\mathbf{r}}$ equals 1 identically, and whether the degrees-of-freedom correction is deemed appropriate.

Of considerably greater interest to the following results on testing is the fact that artificial regressions can be used to calculate one-step efficient estimates. Suppose that we are somehow able to obtain a consistent but asymptotically inefficient estimate $\acute{\boldsymbol{\theta}}$. The manner in which $\acute{\boldsymbol{\theta}}$ is obtained is unimportant; all we require is that $\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = O(n^{-1/2})$. If we evaluate $\mathbf{r}(\boldsymbol{\theta})$ and $\mathbf{R}(\boldsymbol{\theta})$ at $\acute{\boldsymbol{\theta}}$ and run the artificial regression $\acute{\mathbf{r}} = \acute{\mathbf{R}}\mathbf{b} +$ residuals, then by Theorem 1 the resulting OLS estimates $\acute{\mathbf{b}}$ have the property (12). Consequently, the one-step estimator $\grave{\boldsymbol{\theta}} \equiv \acute{\boldsymbol{\theta}} + \acute{\mathbf{b}} = \acute{\boldsymbol{\theta}} + (\acute{\mathbf{R}}^T\acute{\mathbf{R}})^{-1}\acute{\mathbf{R}}^T\acute{\mathbf{r}}$ is asymptotically equivalent to $\hat{\boldsymbol{\theta}}$. In situations where $\hat{\boldsymbol{\theta}}$ is difficult or expensive to obtain but $\acute{\boldsymbol{\theta}}$ is readily obtainable, this can be a valuable result.

More specifically, Theorem 1 allows one to go in one step from the estimates $\tilde{\boldsymbol{\theta}}$ [obtained by ML estimation of Model (10) under the null hypothesis that $\boldsymbol{\theta}_2 = 0$] to estimates asymptotically equivalent to the unrestricted ML estimates $\hat{\boldsymbol{\theta}}$ of the same model, where the asymptotic equivalence is good whether $\boldsymbol{\theta}_2 = \mathbf{0}$ or not, provided only that $\boldsymbol{\theta}_2 = O(n^{-1/2})$. Furthermore, in one step one can go backward as well, starting from the unrestricted estimates $\hat{\boldsymbol{\theta}}$ and obtaining estimates asymptotically equivalent under the null hypothesis to the restricted estimates $\tilde{\boldsymbol{\theta}}$. The trick is to use in the former case what we may call the artificial regression for the unrestricted model (i.e., the regression with the full set of artificial regressors $[\tilde{\mathbf{R}}_1\ \hat{\mathbf{R}}_2]$) and in the latter case the artificial regression for the restricted model (where the regressors $\hat{\mathbf{R}}_2$ are absent). Formally, the $\tilde{\mathbf{b}}$ got by OLS from Regression (19), when added to $\tilde{\boldsymbol{\theta}} \equiv [\tilde{\boldsymbol{\theta}}_1, \mathbf{0}]$, give an answer equal through order $n^{-1/2}$ to $\hat{\boldsymbol{\theta}}$, whereas the $\hat{\mathbf{b}}_1$ got by OLS from $\hat{\mathbf{r}} = \hat{\mathbf{R}}_1\mathbf{b}_1 +$ residuals, when added to $\hat{\boldsymbol{\theta}}_1$, give an answer equivalent to $\tilde{\boldsymbol{\theta}}_1$ good to the same order of approximation. These convenient results follow from Theorem 1 and that $\tilde{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}^0$ if $\boldsymbol{\theta}_2 = O(n^{-1/2})$ and $\hat{\boldsymbol{\theta}}_1$ is consistent for $\boldsymbol{\theta}_1^0$ if $\boldsymbol{\theta}_2 = \mathbf{0}$.

4. TESTS WHERE THE ALTERNATIVE IS IMPLICIT

Up to this point all of the tests we have discussed have involved an explicit alternative hypothesis. The null is a special case of the alternative, and the restrictions being tested are (or can be reformulated as) zero restrictions. In this section we show that artificial regressions can be used to perform tests where there is no explicit alternative hypothesis. Thus we can write an artificial regression analogous to (19), with $\tilde{\mathbf{R}}_2$ replaced by an $n \times l$ matrix $\tilde{\mathbf{Z}} \equiv \mathbf{Z}(\tilde{\boldsymbol{\theta}})$:

$$\tilde{\mathbf{r}} = \tilde{\mathbf{R}}_1\mathbf{c}_1 + \tilde{\mathbf{Z}}\mathbf{c}_2 + \text{residuals}. \tag{24}$$

Provided that the matrix $\tilde{\mathbf{Z}}$ satisfies certain conditions (discussed in the following) that essentially give it the same properties as $\tilde{\mathbf{R}}_2$ and assuming that the matrix $[\tilde{\mathbf{R}}_1\ \tilde{\mathbf{Z}}]$ has full rank, nR^2 from (24) and other asymptotically equivalent statistics are asymptotically distributed as $\chi^2(l)$ when the data are actually generated by (10) with $\boldsymbol{\theta}_2 = \mathbf{0}$. Thus (24) provides a way to compute a wide variety of test statistics, which need not necessarily be derived explicitly as LM statistics.

We now briefly indicate how to prove the aforementioned proposition. Since the proof is similar to standard proofs for LM tests based on artificial regressions, many details are omitted. As noted previously, it is necessary that $\tilde{\mathbf{Z}}$ satisfy certain conditions for it to have essentially the same properties as $\tilde{\mathbf{R}}_2$. First, we require that plim $(n^{-1}\tilde{\mathbf{r}}^T\tilde{\mathbf{Z}}) = \mathbf{0}$ under the null hypothesis; if this condition were not satisfied, we obviously could not expect the plim of $\mathbf{c}_2$ in (24) to be 0. Second, we require that

$$\text{plim}(n^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{Z}}) = \rho^0\ \text{plim}(n^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}) \tag{25}$$

and

$$\text{plim}(n^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{R}}_1) = \rho^0\ \text{plim}(n^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{R}}_1), \tag{26}$$

conditions that are similar to the requirement that

$$\text{plim}(n^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{R}}_1) = \rho^0\ \text{plim}(n^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{R}}_1). \tag{27}$$

This last requirement does not have to be assumed separately because it is a consequence of Properties 2 and 3, the definition of the information matrix $\mathcal{I}(\boldsymbol{\theta})$, and the consistency of $\tilde{\boldsymbol{\theta}}$. Third, we require that laws of large numbers be applicable to the quantities whose probability limits appear on the right sides of (25)–(27). Finally, we require that a central limit theorem be applicable to the vector

$$n^{-1/2}\tilde{\mathbf{Z}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{r}}. \tag{28}$$

Of course, all of these conditions and assumptions must be verified in individual cases. Since this article is concerned with the general properties of artificial regressions, it seems inappropriate to consider any particular case in detail. In some instances it is not too difficult to find sufficient conditions that guarantee the needed regularity, whereas in others, especially in time-series contexts, elaborate arguments may be necessary, especially for a central limit theorem to apply to (28). There is no doubt, however, that in numerous cases of interest sufficient regularity is present.

Now, consider the vector (28). Asymptotically, it has mean 0 under the null hypothesis, and its asymptotic covariance matrix is plim$(n^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{Z}})$, which is equal to

$$\begin{aligned}\text{plim}(n^{-1}(&\tilde{\mathbf{Z}}^T\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{Z}} - \tilde{\mathbf{Z}}^T\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{R}}_1(\tilde{\mathbf{R}}_1^T\tilde{\mathbf{R}}_1)^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{Z}} \\ &- \tilde{\mathbf{Z}}^T\tilde{\mathbf{R}}_1(\tilde{\mathbf{R}}_1^T\tilde{\mathbf{R}}_1)^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{Z}} \\ &+ \tilde{\mathbf{Z}}^T\tilde{\mathbf{R}}_1(\tilde{\mathbf{R}}_1^T\tilde{\mathbf{R}}_1)^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{R}}_1(\tilde{\mathbf{R}}_1^T\tilde{\mathbf{R}}_1)^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{Z}})).\end{aligned} \tag{29}$$

Rewriting (29) so that each term is a product of $O(1)$ probability limits, using (25)–(27), and simplifying, we find that

$$\text{plim}(n^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{Z}}) = \rho^0\ \text{plim}(n^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{Z}}). \tag{30}$$

This plus the asymptotic normality of (28) implies that the expression $(n^{-1/2}\tilde{\mathbf{r}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{Z}})(\text{plim}(n^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{Z}}))^{-1}(n^{-1/2}\tilde{\mathbf{Z}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{r}})$ is asymptotically distributed as ρ^0 times $\chi^2(l)$. The numerator of nR^2 from Regression (24) tends to this expression as $n \to \infty$. By Property 1 the denominator of nR^2 tends to ρ^0. Thus we conclude that nR^2 from Regression (24) is asymptotically distributed as $\chi^2(l)$ if the DGP satisfies the null hypothesis.

There are numerous examples of tests not designed against explicit alternatives that can be based on artificial regressions. One example is the class of tests called Durbin–Wu–Hausman tests (considered in Sec. 5), which can be based on any artificial regression. Other examples are provided by several tests based on the OPG regression, which we consider in the remainder of this section.

Newey (1985) suggested using the OPG regression to calculate what he called "conditional moment tests"; see Tauchen (1985), who used a related auxiliary regression for similar purposes. The basic idea of conditional moment tests is that parametric statistical models are generally based on assumptions that imply that certain moment conditions must hold. For example, suppose that a model depends on underlying error terms u_t that are assumed to be NID(0, σ^2). Estimation of the model generally yields observable counterparts of these, say $\hat{u}_t$, that are functions of the parameter estimates $\hat{\boldsymbol{\theta}}$. Then, in large samples we would expect that

$$n^{-1}\sum_{t=1}^{n}\hat{u}_t^3 \quad \text{and} \quad n^{-1}\sum_{t=1}^{n}(\hat{u}_t^4 - 3\hat{\sigma}^4), \tag{31}$$

the empirical counterparts of conditions on the third and fourth moments of the u_t's, should both be approximately equal to 0. Newey demonstrated that moment conditions such as these can be tested by running the artificial regression

$$\boldsymbol{\iota} = \hat{\mathbf{G}}\mathbf{c}_1 + \hat{\mathbf{Z}}\mathbf{c}_2 + \text{residuals}, \tag{32}$$

where $\hat{\mathbf{G}} \equiv \mathbf{G}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{Z}}$ is chosen so that $\boldsymbol{\iota}^T\hat{\mathbf{Z}}$ generates the moment conditions to be tested. In the case of (31), $\hat{\mathbf{Z}}$ would consist of two vectors, with typical elements $\hat{u}_t^3$ and $\hat{u}_t^4 - 3\hat{\sigma}^4$.

Moment conditions may arise from the theory of the phenomenon being modeled. For example, in the context of models of rational behavior by economic agents, error terms are often supposed to be orthogonal to everything in agents' information sets. Thus if $\mathbf{W}_t$ denotes an l vector of variables that should be orthogonal to u_t, we could define $\hat{\mathbf{Z}}_t$ as the l vector $\mathbf{W}_t\hat{u}_t$ and then test the l orthogonality conditions by regressing $\boldsymbol{\iota}$ on $\hat{\mathbf{G}}$ and $\hat{\mathbf{Z}}$. Nevertheless, because of the poor finite-sample properties of tests based on the OPG regression, conditional moment tests should be used with great caution when the sample size is not very large.

5. DURBIN–WU–HAUSMAN TESTS

Hausman (1978), following Durbin (1954) and Wu (1973), suggested that it may often be useful to test whether there is any significant difference between two sets of estimates, one consistent and efficient under relatively strong conditions and one consistent under weaker conditions. The original application was comparing estimates obtained by least squares with estimates obtained by instrumental variables. It might be natural as well to compare the vector of restricted estimates $\tilde{\boldsymbol{\theta}}_1$ with the vector of unrestricted estimates $\hat{\boldsymbol{\theta}}_1$ from Model (10). Because of the possibility of a one-step artificial regression, it is not actually necessary to obtain $\hat{\boldsymbol{\theta}}_1$ to do so, and this observation provides the easiest way to see how to use an artificial regression to perform a Durbin–Wu–Hausman (DWH) test.

The estimate of $\mathbf{b}_1$ from the artificial regression (19) is $\tilde{\mathbf{b}}_1 = (\tilde{\mathbf{R}}_1^T\tilde{\mathbf{M}}_2\tilde{\mathbf{R}}_1)^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{M}}_2\tilde{\mathbf{r}}$, where $\tilde{\mathbf{M}}_2 = \mathbf{I} - \tilde{\mathbf{R}}_2(\tilde{\mathbf{R}}_2^T\tilde{\mathbf{R}}_2)^{-1}\tilde{\mathbf{R}}_2^T$. Adding this quantity to $\tilde{\boldsymbol{\theta}}_1$ yields a one-step estimator that is asymptotically equivalent to $\hat{\boldsymbol{\theta}}_1$ in the sense discussed in Section 4. Hence a test based on a comparison of $\tilde{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_1 + \tilde{\mathbf{b}}_1$ is equivalent to one based on a comparison of $\tilde{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_1$.

There is no need to restrict attention to Regression (19); Regression (24), of which the former is a special case, can also be used to obtain one-step estimates of $\boldsymbol{\theta}_1$. The estimate of $\mathbf{c}_1$ from (24) is

$$\begin{aligned}\tilde{\mathbf{c}}_1 &= (\tilde{\mathbf{R}}_1^T\tilde{\mathbf{M}}_Z\tilde{\mathbf{R}}_1)^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{M}}_Z\tilde{\mathbf{r}} \\ &= -(\tilde{\mathbf{R}}_1^T\tilde{\mathbf{M}}_Z\tilde{\mathbf{R}}_1)^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{r}},\end{aligned} \tag{33}$$

where $\tilde{\mathbf{M}}_Z \equiv \mathbf{I} - \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T$ and $\tilde{\mathbf{P}}_Z \equiv \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T = \mathbf{I} - \tilde{\mathbf{M}}_Z$. The corresponding one-step estimate of $\boldsymbol{\theta}_1$ is simply

$$\tilde{\boldsymbol{\theta}}_1 + \tilde{\mathbf{c}}_1 = \tilde{\boldsymbol{\theta}}_1 + (\tilde{\mathbf{R}}_1^T\tilde{\mathbf{M}}_Z\tilde{\mathbf{R}}_1)^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{M}}_Z\tilde{\mathbf{r}}. \tag{34}$$

These one-step estimates are *less* efficient than $\tilde{\boldsymbol{\theta}}_1$ if $\boldsymbol{\theta}_2 = \mathbf{0}$, since [as (34) makes clear] they are equal to $\tilde{\boldsymbol{\theta}}_1$ plus something that should be random noise when the model is correctly specified. If the model were not correctly specified, however, the second term in (34) would not be random noise, and $\tilde{\mathbf{c}}_1$ would differ systematically from 0. Thus the DWH test simply asks whether the second term in (34) is random noise.

We have seen that the difference between the one-step estimate and the restricted ML estimate $\tilde{\boldsymbol{\theta}}_1$ is $\tilde{\mathbf{c}}_1$, which is given in (33). The DWH test is thus concerned with whether the vector

$$n^{-1/2}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{r}} = n^{-1/2}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{M}}_1\tilde{\mathbf{r}} \tag{35}$$

has mean 0 asymptotically. Note that the equalities in (33) and (35) both follow from $\tilde{\mathbf{R}}_1^T\tilde{\mathbf{r}} = \mathbf{0}$. The vector on the right side of (35) looks just like the vector (28), with $\tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1$ playing the role of $\tilde{\mathbf{Z}}$. Hence the result (30) implies that

$$\begin{aligned}&\text{plim}(n^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{M}}_1\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\tilde{\mathbf{M}}_1\tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1) \\ &\qquad = \rho^0\,\text{plim}(n^{-1}\tilde{\mathbf{R}}_1^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{M}}_1\tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1).\end{aligned}$$

It is now evident that we may test the hypothesis that Expression (35) has mean 0 asymptotically by using the test statistic

$$\tilde{\mathbf{r}}^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1(\tilde{\mathbf{R}}_1^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{M}}_1\tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1)^+\tilde{\mathbf{R}}_1^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{r}}/(n^{-1}\tilde{\mathbf{r}}^T\tilde{\mathbf{r}}), \tag{36}$$

where $(\cdot)^+$ denotes a generalized inverse. We must use a

generalized inverse here because the matrix $\tilde{\mathbf{R}}_1^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{M}}_1\tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1$ may not have full rank k_1; in fact, it can have rank at most equal to $\min(k_1, l)$.

Of course, the test statistic (36) may be calculated with an artificial regression, namely

$$\tilde{\mathbf{r}} = \tilde{\mathbf{R}}_1\mathbf{d}_1 + \tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1^*\mathbf{d}_2 + \text{residuals}, \tag{37}$$

where $\tilde{\mathbf{R}}_1^*$ is a matrix that consists of as many columns of $\tilde{\mathbf{R}}_1$ as possible, subject to the constraint that the matrix $[\tilde{\mathbf{R}}_1\ \tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1^*]$ must have full rank. Note that nR^2 from both variants of regression (37), $\tilde{\mathbf{r}}^T\tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1^*(\tilde{\mathbf{R}}_1^{*T}\tilde{\mathbf{P}}_Z\tilde{\mathbf{M}}_1\tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1^*)^{-1}\tilde{\mathbf{R}}_1^{*T}\tilde{\mathbf{P}}_Z\tilde{\mathbf{r}}/(n^{-1}\tilde{\mathbf{r}}^T\tilde{\mathbf{r}})$, must be numerically equal to the test statistic (36). Thus we have provided a general procedure for performing a DWH test with an artificial regression. DWH tests are potentially of interest when either the less efficient set of estimates is not explicitly obtained by relaxing a set of restrictions [although as Davidson and MacKinnon (1987) prove, this is always implicitly the case] or when k_1 is substantially smaller than k_2, so the DWH test involves substantially fewer degrees of freedom than a classical test would.

Ruud (1982, 1984) and Newey (1985) showed that tests asymptotically equivalent to DWH tests can be computed as score tests, so various artificial regressions can be used for computation. Nevertheless, the only artificial regression that has been explicitly suggested for this purpose is the OPG regression discussed in Section 2. The aforementioned results are much more general. They show that for *any* test statistic that can be computed with an artificial regression, whether it is explicitly an LM test or not, there is a DWH version that can be computed by a similar artificial regression. One simply replaces the $n \times l$ matrix $\tilde{\mathbf{Z}}$, whatever it may be, with the matrix $\tilde{\mathbf{P}}_Z\tilde{\mathbf{R}}_1^*$, which in regular cases will be $n \times \min(k_1, l)$. For more on the interpretation of DWH tests, see Davidson and MacKinnon (1989).

6. DOUBLE-LENGTH REGRESSIONS

In Section 2 we discussed three widely used artificial regressions. There are many others, most of which we suspect have yet to be discovered. In this section we discuss one useful but not yet widely used class of artificial regressions, the double-length artificial regressions proposed by Davidson and MacKinnon (1984b). These apply to any model of the form

$$f_t(y_t, \bar{\mathbf{y}}_t, \boldsymbol{\theta}) = \varepsilon_t, \qquad \varepsilon_t \sim \text{NID}(0, 1), \tag{38}$$

where $f_t(\cdot)$ is a nonlinear function that may depend on exogenous variables (hence the t subscript), y_t is an observation on the dependent variable, $\bar{\mathbf{y}}_t$ is a vector of observations on lagged values of y, and $\boldsymbol{\theta}$ is a vector of parameters. Suitable regularity conditions must be assumed, of course; see Davidson and MacKinnon (1984b) for details. Since the nonlinear function $f_t(\cdot)$ may include a transformation to the standard normal, (38) is actually a rather general class of models; univariate and multivariate nonlinear regression models with normal errors are both special cases of it, for example. For a model of this class the contribution to the log-likelihood from the tth observation is $l_t = -\frac{1}{2}\log(2\pi) - \frac{1}{2}f_t^2 + k_t$, where $k_t(y_t, \bar{\mathbf{y}}_t, \boldsymbol{\theta}) \equiv \log|\partial f_t(y_t, \bar{\mathbf{y}}_t, \boldsymbol{\theta})/\partial y_t|$ is a Jacobian term. Now, let us define $F_{ti}(y_t, \bar{\mathbf{y}}_t, \boldsymbol{\theta}) \equiv \partial f_t(y_t, \bar{\mathbf{y}}_t, \boldsymbol{\theta})/\partial\theta_i$ and $K_{ti}(y_t, \bar{\mathbf{y}}_t, \boldsymbol{\theta}) \equiv \partial k_t(y_t, \bar{\mathbf{y}}_t, \boldsymbol{\theta})/\partial\theta_i$, and define $\mathbf{F}(\boldsymbol{\theta})$ and $\mathbf{K}(\boldsymbol{\theta})$ as the $n \times k$ matrices with typical elements $F_{ti}(y_t, \bar{\mathbf{y}}_t, \boldsymbol{\theta})$ and $K_{ti}(y_t, \bar{\mathbf{y}}_t, \boldsymbol{\theta})$. It is easy to see that the gradient is

$$\mathbf{g}(\boldsymbol{\theta}) = -\mathbf{F}^T(\boldsymbol{\theta})\mathbf{f}(\boldsymbol{\theta}) + \mathbf{K}^T(\boldsymbol{\theta})\boldsymbol{\iota}. \tag{39}$$

The fundamental result is that for this class of models the information matrix $\mathcal{I}(\boldsymbol{\theta})$ satisfies

$$\operatorname*{plim}_{n\to\infty}(n^{-1}(\mathbf{F}^T(\boldsymbol{\theta})\mathbf{F}(\boldsymbol{\theta}) + \mathbf{K}^T(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta}))) = \mathcal{I}(\boldsymbol{\theta}), \tag{40}$$

and so can be consistently estimated by the matrix $n^{-1}(\mathbf{F}^T(\ddot{\boldsymbol{\theta}})\mathbf{F}(\ddot{\boldsymbol{\theta}}) + \mathbf{K}^T(\ddot{\boldsymbol{\theta}})\mathbf{K}(\ddot{\boldsymbol{\theta}}))$, where $\ddot{\boldsymbol{\theta}}$ is any consistent estimate of $\boldsymbol{\theta}$. Hence one valid form of the LM statistic for testing hypotheses about $\boldsymbol{\theta}$ is

$$(-\tilde{\mathbf{f}}^T\tilde{\mathbf{F}} + \boldsymbol{\iota}^T\tilde{\mathbf{K}})(\tilde{\mathbf{F}}^T\tilde{\mathbf{F}} + \tilde{\mathbf{K}}^T\tilde{\mathbf{K}})^{-1}(-\tilde{\mathbf{F}}^T\tilde{\mathbf{f}} + \tilde{\mathbf{K}}^T\boldsymbol{\iota}), \tag{41}$$

where as usual $\tilde{\boldsymbol{\theta}}$ denotes ML estimates of $\boldsymbol{\theta}$ subject to the restrictions to be tested. The test statistic (41) is evidently just the explained sum of squares from the double-length artificial regression

$$\begin{bmatrix}\tilde{\mathbf{f}}\\ \boldsymbol{\iota}\end{bmatrix} = \begin{bmatrix}-\tilde{\mathbf{F}}\\ \tilde{\mathbf{K}}\end{bmatrix}\mathbf{b} + \text{residuals}. \tag{42}$$

This artificial regression has $2n$ observations. The regressand is $\tilde{f}_t$ for observation t and unity for observation $t + n$, and the regressors corresponding to $\boldsymbol{\theta}$ are $-\tilde{\mathbf{F}}_t$ for observation t and $\tilde{\mathbf{K}}_t$ for observation $t + n$, with $\tilde{\mathbf{F}}_t$ and $\tilde{\mathbf{K}}_t$ denoting the tth rows of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{K}}$, respectively. It is clear from (39) and (40) that Regression (42) satisfies Properties 1–3, provided that in Property 1 n is replaced by $2n$, corresponding to the double length of the artificial regression. Similarly, all results mentioning nR^2 from the regression should in this context be read as $2nR^2$.

Double-length artificial regressions are particularly useful for testing whether the dependent variable in a regression model should be transformed in some way. For example, they may be used to test both the hypothesis that $\lambda = 0$ and the hypothesis that $\lambda = 1$ in the Box–Cox regression

$$(y_t^\lambda - 1)/\lambda = x_t(\boldsymbol{\beta}) + u_t, \qquad u_t \sim \text{NID}(0, \sigma^2); \tag{43}$$

see Box and Cox (1964). Under the null hypothesis that $\lambda = 0$ the regressand of (43) is simply $\log y_t$, whereas under the null hypothesis that $\lambda = 1$ [and provided that $x_t(\cdot)$ includes the equivalent of a constant term] the regressand is effectively just y_t. LM tests for $\lambda = 0$ and $\lambda = 1$ based on double-length regressions were derived by Davidson and MacKinnon (1985a). They found that these test statistics have finite-sample distributions much closer to their asymptotic distributions than similar LM test statistics (based on the OPG regression) proposed by Godfrey and Wickens (1981), a result confirmed by Godfrey et al. (1988). They also found, analytically, that except when σ^2 is quite small, both forms of the LM statistic have much

greater power than the well-known test proposed by Andrews (1971). Further applications of double-length regressions are discussed in Davidson and MacKinnon (1988).

7. CONCLUSION

In this article we have discussed several aspects of the general theory of artificial linear regressions. We have shown that whenever it is possible to construct an artificial regression that satisfies our Properties 1–3, one can evaluate that regression at restricted estimates and use n times the R^2 from the regression as a test statistic. Moreover, it is possible in many cases to calculate tests based on artificial regressions without ever explicitly specifying an alternative hypothesis. One simply has to construct test regressors so that they satisfy certain properties. This opens the door to a wide range of tests. One-step efficient estimates are readily calculated with artificial regressions, and we have used this fact to show that it is always possible to calculate a Durbin–Wu–Hausman variant of any test based on an artificial regression. Finally, we have discussed several specific artificial regressions that can be very useful in practice.

[*Received July 1987. Revised February 1989.*]

REFERENCES

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

Andrews, D. F. (1971), "A Note on the Selection of Data Transformations," *Biometrika*, 58, 249–254.

Bera, A. K., and McKenzie, C. R. (1986), "Alternative Forms and Properties of the Score Test," *Journal of Applied Statistics*, 13, 13–25.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.

Davidson, R., and MacKinnon, J. G. (1984a), "Convenient Specification Tests for Logit and Probit Models," *Journal of Econometrics*, 25, 241–262.

——— (1984b), "Model Specification Tests Based on Artificial Linear Regressions," *International Economic Review*, 25, 485–502.

——— (1985a), "Testing Linear and Loglinear Regressions Against Box–Cox Alternatives," *Canadian Journal of Economics*, 25, 499–517.

——— (1985b), "Heteroskedasticity-Robust Tests in Regression Directions," *Annales de l'INSÉE*, 59, 183–218.

——— (1987), "Implicit Alternatives and the Local Power of Test Statistics," *Econometrica*, 55, 1305–1329.

——— (1988), "Double-Length Artificial Regressions," *Oxford Bulletin of Economics and Statistics*, 50, 203–217.

——— (1989), "Testing for Consistency Using Artificial Regressions," *Econometric Theory*, 5, 363–384.

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

Eicker, F. (1963), "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *The Annals of Mathematical Statistics*, 34, 447–456.

Engle, R. F. (1982), "A General Approach to Lagrange Multiplier Model Diagnostics," *Journal of Econometrics*, 20, 83–104.

——— (1984), "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics," in *Handbook of Econometrics*, eds. Z. Griliches and M. Intriligator, Amsterdam: North-Holland, pp. 775–826.

Godfrey, L. G., McAleer, M., and McKenzie, C. R. (1988), "Variable Addition and Lagrange Multiplier Tests for Linear and Logarithmic Regression Models," *Review of Economics and Statistics*, 70, 492–503.

Godfrey, L. G., and Wickens, M. R. (1981), "Testing Linear and Log-Linear Regressions for Functional Form," *Review of Economic Studies*, 48, 487–496.

Hartley, H. O. (1961), "The Modified Gauss–Newton Method for the Fitting of Nonlinear Regressions by Least Squares," *Technometrics*, 3, 269–280.

Hausman, J. A. (1978), "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1272.

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

Ruud, P. A. (1982), "Score Tests of Consistency," mimeo, University of California, Berkeley, Dept. of Economics.

——— (1984), "Tests of Specification in Econometrics," *Econometric Reviews*, 3, 211–242.

Tauchen, G. E. (1985), "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics*, 30, 415–443.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

Wooldridge, J. M. (1987), "Specification Testing and Quasi-Maximum Likelihood Estimation," Working Paper 479, Massachusetts Institute of Technology, Dept. of Economics.

Wu, D. (1973), "Alternative Tests of Independence Between Stochastic Regressors and Disturbances," *Econometrica*, 41, 733–750.