Davidson, Russell; MacKinnon, James

Working Paper

Artificial regressions

Queen's Economics Department Working Paper, No. 1038

Provided in Cooperation with:
Queen’s University, Department of Economics (QED)

Suggested Citation: Davidson, Russell; MacKinnon, James (2001) : Artificial regressions, Queen's Economics Department Working Paper, No. 1038, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
http://hdl.handle.net/10419/67825

Queen's Economics Department Working Paper No. 1038

Artificial Regressions

Russell Davidson
McGill University

James MacKinnon
Queen's University

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

1-2001

Artificial Regressions

by

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

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

russell@ehess.cnrs-mrs.fr

and

James G. MacKinnon

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

jgm@econ.queensu.ca

Abstract

Associated with every popular nonlinear estimation method is at least one "artificial" linear regression. We define an artificial regression in terms of three conditions that it must satisfy. Then we show how artificial regressions can be useful for numerical optimization, testing hypotheses, and computing parameter estimates. Several existing artificial regressions are discussed and are shown to satisfy the defining conditions, and a new artificial regression for regression models with heteroskedasticity of unknown form is introduced.

A slightly earlier version of this paper appeared as GREQAM Document de Travail 99a04. This research was supported, in part, by the Social Sciences and Humanities Research Council of Canada.

January, 2001

1. Introduction

All popular nonlinear estimation methods, including nonlinear least squares (NLS), maximum likelihood (ML), and the generalized method of moments (GMM), yield estimators which are asymptotically linear. Provided the sample size is large enough, the behavior of these nonlinear estimators in the neighborhood of the true parameter values closely resembles the behavior of the ordinary least squares (OLS) estimator. A particularly illuminating way to see the relationship between any nonlinear estimation method and OLS is to formulate the **artificial regression** that corresponds to the nonlinear estimator.

An artificial regression is a linear regression in which the regressand and regressors are constructed as functions of the data and parameters of the nonlinear model that is really of interest. In addition to helping us understand the asymptotic properties of nonlinear estimators, artificial regressions are often extremely useful as calculating devices. Among other things, they can be used to estimate covariance matrices, as key ingredients of nonlinear optimization methods, to compute one-step efficient estimators, and to calculate test statistics.

In the next section, we discuss the defining properties of an artificial regression. In the subsequent section, we introduce the Gauss-Newton regression, which is probably the most popular artificial regression. Then, in Section 4, we illustrate a number of uses of artificial regressions, using the Gauss-Newton regression as an example. In Section 5, we develop the most important use of artificial regressions, namely, hypothesis testing. We go beyond the Gauss-Newton regression in Sections 6 and 7, in which we introduce two quite generally applicable artificial regressions, one for models estimated by maximum likelihood, and one for models estimated by the generalized method of moments. Section 8 shows how artificial regressions may be modified to take account of the presence of heteroskedasticity of unknown form. Then, in Sections 9 and 10, we discuss double-length regressions and artificial regressions for binary response models, respectively.

2. The Concept of an Artificial Regression

Consider a fully parametric nonlinear model that is characterized by a parameter vector $\boldsymbol{\theta}$ which belongs to a parameter space $\Theta \subseteq \mathbb{R}^k$ and which can be estimated by minimizing a criterion function $Q(\boldsymbol{\theta})$ using n observations. In the case of a nonlinear regression model estimated by nonlinear least squares, $Q(\boldsymbol{\theta})$ would be one half the sum of squared residuals, and in the case of a model estimated by maximum likelihood, $Q(\boldsymbol{\theta})$ would be minus the loglikelihood function.

If an artificial regression exists for such a model, it always involves two things: a regressand, $\boldsymbol{r}(\boldsymbol{\theta})$, and a matrix of regressors, $\boldsymbol{R}(\boldsymbol{\theta})$. The number of regressors for the artificial regression is equal to k, the number of parameters. The number of "observations" for the artificial regression is often equal to n, but it may also be equal to a small integer, such as 2 or 3, times n. We can write a generic artificial

regression as

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{b} + \text{residuals}, \tag{1}$$

where $\boldsymbol{b}$ is a k–vector of coefficients. "Residuals" is used here as a neutral term to avoid any implication that (1) is a statistical model. The regressand and regressors in (1) can be evaluated at any point $\boldsymbol{\theta} \in \Theta$, and the properties of the artificial regression will depend on the point at which they are evaluated. In many cases, we will want to evaluate (1) at a vector of estimates $\acute{\boldsymbol{\theta}}$ that is root-n consistent. This means that, if the true parameter vector is $\boldsymbol{\theta}_0 \in \Theta$, then $\acute{\boldsymbol{\theta}}$ approaches $\boldsymbol{\theta}_0$ at a rate proportional to $n^{-1/2}$. One such vector that is of particular interest is $\hat{\boldsymbol{\theta}}$, the vector of estimates which minimizes the criterion function $Q(\boldsymbol{\theta})$.

For (1) to constitute an artificial regression, the vector $\boldsymbol{r}(\boldsymbol{\theta})$ and the matrix $\boldsymbol{R}(\boldsymbol{\theta})$ must satisfy certain defining properties. These may be stated in a variety of ways, which depend on the class of models to which the artificial regression is intended to apply. For the purposes of this paper, we will say that (1) is an artificial regression if it satisfies the following three conditions:

(i) The estimator $\hat{\boldsymbol{\theta}}$ is defined, uniquely in a neighborhood in Θ, by the k equations $\boldsymbol{R}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{r}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$;

(ii) for any root-n consistent $\acute{\boldsymbol{\theta}}$, a consistent estimate of $\operatorname{Var}(\operatorname{plim} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0))$ is given by the inverse of $n^{-1}\boldsymbol{R}^\top(\acute{\boldsymbol{\theta}})\boldsymbol{R}(\acute{\boldsymbol{\theta}})$. Formally,

$$\operatorname{Var}\left(\operatorname*{plim}_{n\to\infty} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right) = \operatorname*{plim}_{n\to\infty}\left(n^{-1}\boldsymbol{R}^\top(\acute{\boldsymbol{\theta}})\boldsymbol{R}(\acute{\boldsymbol{\theta}})\right)^{-1};$$

(iii) if $\acute{\boldsymbol{b}}$ denotes the vector of estimates from the artificial regression (1) with regressand and regressors evaluated at $\acute{\boldsymbol{\theta}}$, then

$$\acute{\boldsymbol{\theta}} + \acute{\boldsymbol{b}} = \hat{\boldsymbol{\theta}} + o_p(n^{-1/2}).$$

Many artificial regressions actually satisfy a stronger version of condition (i):

(i′) $\boldsymbol{g}(\boldsymbol{\theta}) = -\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})$,

where $\boldsymbol{g}(\boldsymbol{\theta})$ denotes the gradient of the criterion function $Q(\boldsymbol{\theta})$. Clearly, condition (i′) implies condition (i), but not *vice versa*. The minus sign in (i′) is due to the arbitrary choice that the estimator is defined by minimizing $Q(\boldsymbol{\theta})$ rather than maximizing it.

Condition (ii) has been written in a particularly simple form, and some nonstandard artificial regressions do not actually satisfy it. However, as we will see, this does not prevent them from having essentially the same properties as artificial regressions that do satisfy it.

Condition (iii), which is perhaps the most interesting of the three conditions, will be referred to as the **one-step property**. It says that, if we take one step from

an initial consistent estimator $\acute{\boldsymbol{\theta}}$, where the step is given by the coefficients $\acute{\boldsymbol{b}}$ from the artificial regression, we will obtain an estimator that is asymptotically equivalent to $\hat{\boldsymbol{\theta}}$.

The implications of these three conditions will become clearer when we study specific artificial regressions in the remainder of this paper. These conditions differ substantially from the conditions used to define an artificial regression in Davidson and MacKinnon (1990), because that paper was concerned solely with artificial regressions for models estimated by maximum likelihood.

3. The Gauss-Newton Regression

Associated with every nonlinear regression model is a somewhat nonstandard artificial regression which is probably more widely used than any other. Consider the univariate, nonlinear regression model

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \quad t = 1, \dots, n, \tag{2}$$

where y_t is the t^{th} observation on the dependent variable, and $\boldsymbol{\beta}$ is a k–vector of parameters to be estimated. The scalar function $x_t(\boldsymbol{\beta})$ is a nonlinear regression function. It determines the mean value of y_t as a function of unknown parameters $\boldsymbol{\beta}$ and, usually, of explanatory variables, which may include lagged dependent variables. The explanatory variables are not shown explicitly in (2), but the t subscript on $x_t(\boldsymbol{\beta})$ reminds us that they are present. The model (2) may also be written as

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \tag{3}$$

where $\boldsymbol{y}$ is an n–vector with typical element y_t, and $\boldsymbol{x}(\boldsymbol{\beta})$ is an n–vector of which the t^{th} element is $x_t(\boldsymbol{\beta})$.

The nonlinear least squares (NLS) estimator $\hat{\boldsymbol{\beta}}$ for model (3) minimizes the sum of squared residuals. It is convenient to use for the criterion function to be minimized this sum divided by 2. Thus we define

$$Q(\boldsymbol{\beta}) = \tfrac{1}{2}\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big). \tag{4}$$

The Gauss-Newton regression can be derived as an approximation to **Newton's Method** for the minimization of $Q(\boldsymbol{\beta})$. In this case, Newton's Method consists of the following iterative procedure. One starts from some suitably chosen starting value, $\boldsymbol{\beta}_{(0)}$. At step m of the procedure, $\boldsymbol{\beta}_{(m)}$ is updated by the formula

$$\boldsymbol{\beta}_{(m+1)} = \boldsymbol{\beta}_{(m)} - \boldsymbol{H}_{(m)}^{-1}\boldsymbol{g}_{(m)},$$

where the $k \times 1$ vector $\boldsymbol{g}_{(m)}$ and the $k \times k$ matrix $\boldsymbol{H}_{(m)}$ are, respectively, the gradient and the Hessian of $Q(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, evaluated at $\boldsymbol{\beta}_{(m)}$. For general $\boldsymbol{\beta}$, we have

$$\boldsymbol{g}(\boldsymbol{\beta}) = -\boldsymbol{X}^\top(\boldsymbol{\beta})\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big),$$

where the matrix $\boldsymbol{X}(\boldsymbol{\beta})$ is an $n \times k$ matrix with ti^{th} element the derivative of $x_t(\boldsymbol{\beta})$ with respect to β_i, the i^{th} component of $\boldsymbol{\beta}$. A typical element of the Hessian $\boldsymbol{H}(\boldsymbol{\beta})$ is

$$H_{ij}(\boldsymbol{\beta}) = -\sum_{t=1}^{n}\Big(\big(y_t - x_t(\boldsymbol{\beta})\big)\frac{\partial X_{ti}(\boldsymbol{\beta})}{\partial \beta_j} - X_{ti}(\boldsymbol{\beta})X_{tj}(\boldsymbol{\beta})\Big), \quad i,j = 1,\ldots,k. \tag{5}$$

The Gauss-Newton procedure is one of the set of so-called **quasi-Newton** procedures, in which the exact Hessian is replaced by an approximation. Here, only the second term in (5) is used, so that the $\boldsymbol{H}(\boldsymbol{\beta})$ of Newton's method is replaced by the matrix $\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$. Thus the Gauss-Newton updating formula is

$$\boldsymbol{\beta}_{(m+1)} = \boldsymbol{\beta}_{(m)} + \big(\boldsymbol{X}_{(m)}^\top \boldsymbol{X}_{(m)}\big)^{-1}\boldsymbol{X}_{(m)}^\top\big(\boldsymbol{y} - \boldsymbol{x}_{(m)}\big), \tag{6}$$

where we write $\boldsymbol{X}_{(m)} = \boldsymbol{X}(\boldsymbol{\beta}_{(m)})$ and $\boldsymbol{x}_{(m)} = \boldsymbol{x}(\boldsymbol{\beta}_{(m)})$. The updating term on the right-hand side of (6) is the set of OLS parameter estimates from the **Gauss-Newton regression**, or **GNR**,

$$\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{X}(\boldsymbol{\beta})\boldsymbol{b} + \text{residuals}, \tag{7}$$

where the variables $\boldsymbol{r}(\boldsymbol{\beta}) \equiv \boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})$ and $\boldsymbol{R}(\boldsymbol{\beta}) \equiv \boldsymbol{X}(\boldsymbol{\beta})$ are evaluated at $\boldsymbol{\beta}_{(m)}$. Notice that there is no regressor in (7) corresponding to the parameter σ^2, because the criterion function $Q(\boldsymbol{\beta})$ does not depend on σ^2. This is one of the features of the GNR that makes it a nonstandard artificial regression.

The GNR is clearly a linearization of the nonlinear regression model (3) around the point $\boldsymbol{\beta}$. In the special case in which the original model is linear, $\boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{X}$ is the matrix of independent variables. Since $\boldsymbol{X}(\boldsymbol{\beta})$ is equal to $\boldsymbol{X}$ for all $\boldsymbol{\beta}$ in this special case, the GNR will simply be a regression of the vector $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ on the matrix $\boldsymbol{X}$.

An example is provided by the nonlinear regression model

$$y_t = \beta_1 Z_{t1}^{\beta_2} Z_{t2}^{1-\beta_2} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{8}$$

where Z_{t1} and Z_{t2} are independent variables. The regression function here is nonlinear and has the form of a Cobb-Douglas production function. In many cases, of course, it would be reasonable to assume that the error term is multiplicative, and it would then be possible to take logarithms of both sides and use ordinary least squares. But if we wish to estimate (8) as it stands, we must use nonlinear least squares. The GNR that corresponds to (8) is

$$y_t - \beta_1 Z_{t1}^{\beta_2} Z_{t2}^{1-\beta_2} = b_1 Z_{t1}^{\beta_2} Z_{t2}^{1-\beta_2} + b_2 \beta_1 Z_{t2}\left(\frac{Z_{t1}}{Z_{t2}}\right)^{\beta_2} \log\left(\frac{Z_{t1}}{Z_{t2}}\right) + \text{residual}.$$

The regressand is y_t minus the regression function, the first regressor is the derivative of the regression function with respect to β_1, and the second regressor is the derivative of the regression function with respect to β_2

Now consider the defining conditions of an artificial regression. We have

$$\boldsymbol{R}^\top(\boldsymbol{\theta})\,\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{X}^\top(\boldsymbol{\beta})\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big), \tag{9}$$

which is just minus the gradient of $Q(\boldsymbol{\beta})$. Thus condition (i′) is satisfied.

Next, consider condition (iii). Let $\acute{\boldsymbol{\beta}}$ denote a vector of initial estimates, which are assumed to be root-n consistent. The GNR (7) evaluated at these estimates is

$$\boldsymbol{y} - \acute{\boldsymbol{x}} = \acute{\boldsymbol{X}}\boldsymbol{b} + \text{residuals},$$

where $\acute{\boldsymbol{x}} \equiv \boldsymbol{x}(\acute{\boldsymbol{\beta}})$ and $\acute{\boldsymbol{X}} \equiv \boldsymbol{X}(\acute{\boldsymbol{\beta}})$. The estimate of $\boldsymbol{b}$ from this regression is

$$\acute{\boldsymbol{b}} = (\acute{\boldsymbol{X}}^\top\acute{\boldsymbol{X}})^{-1}\acute{\boldsymbol{X}}^\top(\boldsymbol{y} - \acute{\boldsymbol{x}}). \tag{10}$$

The **one-step efficient estimator** is then defined to be

$$\grave{\boldsymbol{\beta}} \equiv \acute{\boldsymbol{\beta}} + \acute{\boldsymbol{b}}. \tag{11}$$

By Taylor expanding the expression $n^{-1/2}\acute{\boldsymbol{X}}^\top(\boldsymbol{y} - \acute{\boldsymbol{x}})$ around $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ is the true parameter vector, and using standard asymptotic arguments, it can be shown that, to leading order,

$$n^{-1/2}\acute{\boldsymbol{X}}^\top(\boldsymbol{y} - \acute{\boldsymbol{x}}) = n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u} - n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0\,n^{1/2}(\acute{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

where $\boldsymbol{X}_0 \equiv \boldsymbol{X}(\boldsymbol{\beta}_0)$. This relation can be solved to yield

$$n^{1/2}(\acute{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = (n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0)^{-1}\big(n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u} - n^{-1/2}\acute{\boldsymbol{X}}^\top(\boldsymbol{y} - \acute{\boldsymbol{x}})\big). \tag{12}$$

Now it is a standard result that, asymptotically,

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = (n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0)^{-1}(n^{-1/2}\boldsymbol{X}_0^\top\boldsymbol{u}); \tag{13}$$

see, for example, Davidson and MacKinnon (1993, Section 5.4). By (10), the second term on the right-hand side of (12) is asymptotically equivalent to $-n^{1/2}\acute{\boldsymbol{b}}$. Thus (12) implies that

$$n^{1/2}(\acute{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - n^{1/2}\acute{\boldsymbol{b}}.$$

Rearranging this and using the definition (11), we see that, to leading order asymptotically,

$$n^{1/2}(\grave{\boldsymbol{\beta}} - \boldsymbol{\beta}) = n^{1/2}(\acute{\boldsymbol{\beta}} + \acute{\boldsymbol{b}} - \boldsymbol{\beta}_0) = n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

In other words, after both are centered and multiplied by $n^{1/2}$, the one-step estimator $\grave{\boldsymbol{\beta}}$ and the NLS estimator $\hat{\boldsymbol{\beta}}$ tend to the same random variable asymptotically. This is just another way of writing condition (iii) for model (3).

Finally, consider condition (ii). Since $\boldsymbol{X}(\boldsymbol{\beta})$ plays the role of $\boldsymbol{R}(\boldsymbol{\theta})$, we see that

$$\frac{1}{n}\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta}). \tag{14}$$

If the right-hand side of (14) is evaluated at any root-n consistent estimator $\acute{\boldsymbol{\beta}}$, it must tend to the same probability limit as $n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0$. It is a standard result, following straightforwardly from (13), that, if $\hat{\boldsymbol{\beta}}$ denotes the NLS estimator for the model (3), then

$$\lim_{n\to\infty} \operatorname{Var}\bigl(n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\bigr) = \sigma_0^2 \operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{X}_0^\top\boldsymbol{X}_0)^{-1}, \tag{15}$$

where σ_0^2 is the true variance of the error terms; see, for example, Davidson and MacKinnon (1993, Chapter 5). Thus the GNR would satisfy condition (ii) except that there is a factor of σ_0^2 missing. However, this factor is automatically supplied by the regression package. The estimated covariance matrix will be

$$\widehat{\operatorname{Var}}(\acute{\boldsymbol{b}}) = \acute{s}^2(\acute{\boldsymbol{X}}^\top\acute{\boldsymbol{X}})^{-1}, \tag{16}$$

where $\acute{s}^2 = \text{SSR}/(n-k)$ is the estimate of σ^2 from the artificial regression. It is not hard to show that $\acute{s}^2$ estimates σ_0^2 consistently, and so it is clear from (15) that (16) provides a reasonable way to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$.

It is easy to modify the GNR so that it actually satisfies condition (ii). We just need to divide both the regressand and the regressors by s, the standard error from the original, nonlinear regression. When this is done, (14) becomes

$$\frac{1}{n}\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\theta}) = \frac{1}{ns^2}\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta}),$$

and condition (ii) is seen to be satisfied. However, there is rarely any reason to do this in practice.

Although the GNR is the most commonly encountered artificial regression, it differs from most artificial regressions in one key respect: There is one parameter, σ^2, for which there is no regressor. This happens because the criterion function, $Q(\boldsymbol{\beta})$, depends only on $\boldsymbol{\beta}$. The GNR therefore has only as many regressors as $\boldsymbol{\beta}$ has components. This feature of the GNR is responsible for the fact that it does not quite satisfy condition (ii). The fact that $Q(\boldsymbol{\beta})$ does not depend on σ^2 also causes the asymptotic covariance matrix to be block diagonal between the $k \times k$ block that corresponds to $\boldsymbol{\beta}$ and the 1×1 block that corresponds to σ^2.

4. Uses of the GNR

The GNR, like other artificial regressions, has several uses, depending on the parameter values at which the regressand and regressors are evaluated. If we evaluate them at $\hat{\boldsymbol{\beta}}$, the vector of NLS parameter estimates, regression (7) becomes

$$\boldsymbol{y} - \hat{\boldsymbol{x}} = \hat{\boldsymbol{X}}\boldsymbol{b} + \text{residuals}, \tag{17}$$

where $\hat{\boldsymbol{x}} \equiv \boldsymbol{x}(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{X}} \equiv \boldsymbol{X}(\hat{\boldsymbol{\beta}})$. By condition (i), which follows from the first-order conditions for NLS estimation, the OLS estimate $\hat{\boldsymbol{b}}$ from this regression is a zero vector. In consequence, the explained sum of squares, or ESS, from regression (17) will be 0, and the SSR will be equal to

$$\|\boldsymbol{y} - \hat{\boldsymbol{x}}\|^2 = (\boldsymbol{y} - \hat{\boldsymbol{x}})^\top (\boldsymbol{y} - \hat{\boldsymbol{x}}),$$

which is the SSR from the original nonlinear regression.

Although it may seem curious to run an artificial regression all the coefficients of which are known in advance to be zero, there can be two very good reasons for doing so. The first reason is to check that the vector $\hat{\boldsymbol{\beta}}$ reported by a program for NLS estimation really does satisfy the first-order conditions. Computer programs for calculating NLS estimates do not yield reliable answers in every case; see McCullough (1999). The GNR provides an easy way to see whether the first-order conditions are actually satisfied. If all the t statistics for the GNR are not less than about 10^{-4}, and the R^2 is not less than about 10^{-8}, then the value of $\hat{\boldsymbol{\beta}}$ reported by the program should be regarded with suspicion.

The second reason to run the GNR (17) is to calculate an estimate of $\text{Var}(\hat{\boldsymbol{\beta}})$, the covariance matrix of the NLS estimates. The usual OLS covariance matrix from regression (17) is

$$\widehat{\text{Var}}(\hat{\boldsymbol{b}}) = s^2(\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}})^{-1}, \tag{18}$$

which is similar to (16) except that everything is now evaluated at $\hat{\boldsymbol{\beta}}$. Thus running the GNR (17) provides an easy way to calculate what is arguably the best estimate of $\text{Var}(\hat{\boldsymbol{\beta}})$. Of course, for (18) to provide an asymptotically valid covariance matrix estimate, it is essential that the error terms in (2) be independent and identically distributed, as we have assumed so far. We will discuss ways to drop this assumption in Section 7.

Since the GNR satisfies the one-step property, it and other artificial regressions can evidently be used to obtain one-step efficient estimates. However, although one-step estimation is of considerable theoretical interest, it is generally of modest practical interest, for two reasons. Firstly, we often do not have a root-n consistent estimator to start from and, secondly, modern computers are so fast that the savings from stopping after just one step are rarely substantial.

What is often of great practical interest is the use of the GNR as part of a numerical minimization algorithm to find the NLS estimates $\hat{\boldsymbol{\beta}}$ themselves. In

practice, the classical Gauss-Newton updating procedure (6) should generally be replaced by

$$\boldsymbol{\beta}_{(m)} = \boldsymbol{\beta}_{(m-1)} + \alpha_{(m)} \boldsymbol{b}_{(m)},$$

where $\alpha_{(m)}$ is a scalar that is chosen in various ways by different algorithms, but always in such a way that $Q(\boldsymbol{\beta}_{(m+1)}) < Q(\boldsymbol{\beta}_{(m)})$. Numerical optimization methods are discussed by Press et al. (1992), among many others. Artificial regressions other than the GNR allow these methods to be used more widely than just in the least squares context.

5. Hypothesis Testing with Artificial Regressions

Artificial regressions like the GNR are probably employed most frequently for hypothesis testing. Suppose we wish to test a set of r equality restrictions on $\boldsymbol{\theta}$. Without loss of generality, we can assume that these are zero restrictions. This allows us to partition $\boldsymbol{\theta}$ into two subvectors, $\boldsymbol{\theta}_1$ of length $k-r$, and $\boldsymbol{\theta}_2$ of length r, the restrictions being that $\boldsymbol{\theta}_2 = \mathbf{0}$. If the estimator $\hat{\boldsymbol{\theta}}$ is not only root-n consistent but also asymptotically normal, an appropriate statistic for testing these restrictions is

$$\hat{\boldsymbol{\theta}}_2^\top \big(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_2)\big)^{-1} \hat{\boldsymbol{\theta}}_2, \tag{19}$$

which will be asymptotically distributed as $\chi^2(r)$ under the null if $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_2)$ is a suitable estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}_2$.

Suppose that $\boldsymbol{r}(\boldsymbol{\theta})$ and $\boldsymbol{R}(\boldsymbol{\theta})$ define an artificial regression for the estimator $\hat{\boldsymbol{\theta}}$. Let $\acute{\boldsymbol{\theta}} \equiv [\acute{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$ be a vector of root-n consistent estimates under the null. Then, if the variables of the artificial regression are evaluated at $\acute{\boldsymbol{\theta}}$, the regression can be expressed as

$$\boldsymbol{r}(\acute{\boldsymbol{\theta}}_1, \mathbf{0}) = \boldsymbol{R}_1(\acute{\boldsymbol{\theta}}_1, \mathbf{0})\boldsymbol{b}_1 + \boldsymbol{R}_2(\acute{\boldsymbol{\theta}}_2, \mathbf{0})\boldsymbol{b}_2 + \text{residuals}, \tag{20}$$

where the partitioning of $\boldsymbol{R} = [\boldsymbol{R}_1 \;\; \boldsymbol{R}_2]$ corresponds to the partitioning of $\boldsymbol{\theta}$ as $[\boldsymbol{\theta}_1 \vdots \boldsymbol{\theta}_2]$. Regression (20) will usually be written simply as

$$\acute{\boldsymbol{r}} = \acute{\boldsymbol{R}}_1 \boldsymbol{b}_1 + \acute{\boldsymbol{R}}_2 \boldsymbol{b}_2 + \text{residuals},$$

although this notation hides the fact that $\acute{\boldsymbol{\theta}}$ satisfies the null hypothesis.

By the one-step property, $\acute{\boldsymbol{b}}_2$ from (20) is asymptotically equivalent under the null to the estimator $\hat{\boldsymbol{\theta}}_2$, since under the null the true value of $\boldsymbol{\theta}_2$ is zero. This suggests that we may replace $\hat{\boldsymbol{\theta}}_2$ in (19) by $\acute{\boldsymbol{b}}_2$. By property (ii), the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is estimated by $(n^{-1}\acute{\boldsymbol{R}}^\top \acute{\boldsymbol{R}})^{-1}$. A suitable estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}_2$ can be obtained from this by use of the Frisch-Waugh-Lovell (FWL) theorem: See Davidson and MacKinnon (1993, Chapter 1) for a full treatment of the FWL Theorem. The estimate is $(\acute{\boldsymbol{R}}_2^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{R}}_2)^{-1}$, where the orthogonal projection matrix $\acute{\boldsymbol{M}}_1$ is defined by

$$\acute{\boldsymbol{M}}_1 = \mathbf{I} - \acute{\boldsymbol{R}}_1 (\acute{\boldsymbol{R}}_1^\top \acute{\boldsymbol{R}}_1)^{-1} \acute{\boldsymbol{R}}_1^\top. \tag{21}$$

By the same theorem, we have that

$$\acute{\boldsymbol{b}}_2 = (\acute{\boldsymbol{R}}_2^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{R}}_2)^{-1} \acute{\boldsymbol{R}}_2^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{r}}. \tag{22}$$

Thus the artificial regression version of the test statistic (19) is

$$\acute{\boldsymbol{b}}_2^\top \acute{\boldsymbol{R}}_2^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{R}}_2 \acute{\boldsymbol{b}}_2 = \acute{\boldsymbol{r}}^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{R}}_2 (\acute{\boldsymbol{R}}_2^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{R}}_2)^{-1} \acute{\boldsymbol{R}}_2^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{r}}. \tag{23}$$

The following theorem demonstrates the asymptotic validity of (23).

> **Theorem 1:** If the regressand $\boldsymbol{r}(\boldsymbol{\theta})$ and the regressor matrix $\boldsymbol{R}(\boldsymbol{\theta})$ define an artificial regression for the root-n consistent, asymptotically normal, estimator $\hat{\boldsymbol{\theta}}$, and if the partition $\boldsymbol{R} = [\boldsymbol{R}_1 \;\; \boldsymbol{R}_2]$ corresponds to the partition $\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \vdots \boldsymbol{\theta}_2]$, then the statistic (23), computed at any root-n consistent $\acute{\boldsymbol{\theta}} = [\acute{\boldsymbol{\theta}}_1 \vdots \boldsymbol{0}]$, is asymptotically distributed as $\chi^2(r)$ under the null hypothesis that $\boldsymbol{\theta}_2 = \boldsymbol{0}$, and is asymptotically equivalent to the generic statistic (19).

Proof: To prove this theorem, we need to show two things. The first is that

$$n^{-1} \acute{\boldsymbol{R}}_2^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{R}}_2 = n^{-1} \boldsymbol{R}_2(\boldsymbol{\theta}_0) \boldsymbol{M}_1(\boldsymbol{\theta}_0) \boldsymbol{R}_2(\boldsymbol{\theta}_0) + o_p(1),$$

where $\boldsymbol{\theta}_0$ is the true parameter vector, and $\boldsymbol{M}_1(\boldsymbol{\theta}_0)$ is defined analogously to (21). This result follows by standard asymptotic arguments based on the one-step property. The second is that the vector

$$n^{-1/2} \acute{\boldsymbol{R}}_2^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{r}} = n^{-1/2} \boldsymbol{R}_2^\top(\boldsymbol{\theta}_0) \boldsymbol{M}_1(\boldsymbol{\theta}_0) \boldsymbol{r}(\boldsymbol{\theta}_0) + o_p(1)$$

is asymptotically normally distributed. The equality here also follows by standard asymptotic arguments. The asymptotic normality of $\hat{\boldsymbol{\theta}}$ implies that $\acute{\boldsymbol{b}}$ is asymptotically normally distributed. Therefore, by (22), $n^{-1/2} \acute{\boldsymbol{R}}_2^\top \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{r}}$ must also be asymptotically normally distributed. These two results imply that, asymptotically under the null hypothesis, the test statistic (23) is a quadratic form in a normally distributed r–vector, the mean of which is zero, and the inverse of its covariance matrix. Such a quadratic form follows the $\chi^2(r)$ distribution. ■

Remarks: The statistic (23) can be computed as the difference between the sums of squared residuals (SSR) from the regressions

$$\acute{\boldsymbol{r}} = \acute{\boldsymbol{R}}_1 \boldsymbol{b}_1 + \text{residuals, and} \tag{24}$$

$$\acute{\boldsymbol{r}} = \acute{\boldsymbol{R}}_1 \boldsymbol{b}_1 + \acute{\boldsymbol{R}}_2 \boldsymbol{b}_2 + \text{residuals.} \tag{25}$$

Equivalently, it can be computed as the difference between the explained sums of squares (ESS), with the opposite sign, or as the ESS from the FWL regression corresponding to (25):

$$\acute{\boldsymbol{M}}_1 \acute{\boldsymbol{r}} = \acute{\boldsymbol{M}}_1 \acute{\boldsymbol{R}}_2 \boldsymbol{b}_2 + \text{residuals.}$$

If $\operatorname{plim} n^{-1}\acute{\boldsymbol{r}}^\top\acute{\boldsymbol{r}} = 1$ for all root-n consistent $\acute{\boldsymbol{\theta}}$, there are other convenient ways of computing (23), or statistics asymptotically equivalent to it. One is the ordinary F statistic for $\boldsymbol{b}_2 = \boldsymbol{0}$ in regression (25):

$$F = \frac{\acute{\boldsymbol{r}}^\top\acute{\boldsymbol{M}}_1\acute{\boldsymbol{R}}_2(\acute{\boldsymbol{R}}_2^\top\acute{\boldsymbol{M}}_1\acute{\boldsymbol{R}}_2)^{-1}\acute{\boldsymbol{R}}_2^\top\acute{\boldsymbol{M}}_1\acute{\boldsymbol{r}}/r}{\acute{\boldsymbol{r}}^\top\acute{\boldsymbol{M}}_1\acute{\boldsymbol{r}}/(n-k)}, \tag{26}$$

which works because the denominator tends to a probability limit of 1 as $n \to \infty$. This statistic is, of course, in F rather than χ^2 form.

Another frequently used test statistic is available if $\acute{\boldsymbol{\theta}}$ is actually the vector of restricted estimates, that is, the estimator that minimizes the criterion function when the restriction that $\boldsymbol{\theta}_2 = \boldsymbol{0}$ is imposed. In this case, n times the uncentered R^2 from (25) is a valid test statistic. With this choice of $\acute{\boldsymbol{\theta}}$, the ESS from (24) is zero, by property (i). Thus (23) is just the ESS from (25). Since $nR^2 = \text{ESS}/(\text{TSS}/n)$, where TSS denotes the total sum of squares, and since $\text{TSS}/n \to 1$ as $n \to \infty$, it follows that this statistic is asymptotically equivalent to (23).

Even though the GNR does not satisfy condition (ii) when it is expressed in its usual form with all variables not divided by the standard error s, the F statistic (26) and the nR^2 statistic are still valid test statistics, because they are both ratios. In fact, variants of the GNR are routinely used to perform many types of specification tests. These include tests for serial correlation similar to the ones proposed by Godfrey (1978), nonnested hypothesis tests where both models are parametric (Davidson and MacKinnon, 1981), and nonnested hypothesis tests where the alternative model is nonparametric (Delgado and Stengos, 1994). They also include several Durbin-Wu-Hausman, or DWH, tests, in which an efficient estimator is compared with an inefficient estimator that is consistent under weaker conditions; see Sections 7.9 and 11.4 of Davidson and MacKinnon (1993).

6. The OPG Regression

By no means all interesting econometric models are regression models. It is therefore useful to see if artificial regressions other than the GNR exist for wide classes of models. One of these is the **outer-product-of-the-gradient regression**, or **OPG regression**, a particularly simple artificial regression that can be used with most models that are estimated by maximum likelihood. Suppose we are interested in a model of which the loglikelihood function can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta}), \tag{27}$$

where $\ell_t(\cdot)$ denotes the contribution to the loglikelihood function associated with observation t. This is the log of the density of the dependent variable(s) for

observation t, conditional on observations $1, \ldots, t-1$. Thus lags of the dependent variable(s) are allowed. The key feature of (27) is that $\ell(\boldsymbol{\theta})$ is a sum of contributions from each of the n observations.

Now let $\boldsymbol{G}(\boldsymbol{\theta})$ be the matrix with typical element

$$G_{ti}(\boldsymbol{\theta}) \equiv \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_i}; \quad t = 1, \ldots, n, \;\; i = 1, \ldots, k.$$

The matrix $\boldsymbol{G}(\boldsymbol{\theta})$ is called the **matrix of contributions to the gradient**, or the **CG matrix**, because the derivative of the sample loglikelihood (27) with respect to θ_i, the i^{th} component of $\boldsymbol{\theta}$, is the sum of the elements of column i of $\boldsymbol{G}(\boldsymbol{\theta})$. The OPG regression associated with (27) can be written as

$$\boldsymbol{\iota} = \boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{b} + \text{residuals}, \tag{28}$$

where $\boldsymbol{\iota}$ denotes an n–vector of 1s.

It is easy to see that the OPG regression (28) satisfies the conditions for it to be an artificial regression. Condition (i′) is evidently satisfied, since $\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{G}^\top(\boldsymbol{\theta})\boldsymbol{\iota}$, the components of which are the derivatives of $\ell(\boldsymbol{\theta})$ with respect to each of the θ_i. Condition (ii) is also satisfied, because, under standard regularity conditions, if $\boldsymbol{\theta}$ is the true parameter vector,

$$\operatorname*{plim}_{n\to\infty}\big(n^{-1}\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\theta})\big) = \operatorname*{plim}_{n\to\infty}\big(n^{-1}\boldsymbol{G}^\top(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})\big) = \mathcal{I}(\boldsymbol{\theta}).$$

Here $\mathcal{I}(\boldsymbol{\theta})$ denotes the information matrix, defined as

$$\mathcal{I}(\boldsymbol{\theta}) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E\big(\boldsymbol{G}_t^\top(\boldsymbol{\theta})\boldsymbol{G}_t(\boldsymbol{\theta})\big),$$

where $\boldsymbol{G}_t(\cdot)$ is the t^{th} row of $\boldsymbol{G}(\cdot)$. Since, as is well-known, the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is given by the inverse of the information matrix, condition (ii) is satisfied under the further weak regularity condition that $\mathcal{I}(\boldsymbol{\theta})$ should be continuous in $\boldsymbol{\theta}$. Condition (iii) is also satisfied, since it can be shown that one-step estimates from the OPG regression are asymptotically equivalent to maximum likelihood estimates. The proof is quite similar to the one for the GNR given in Section 3.

It is particularly easy to compute an LM test by using the OPG regression. Let $\tilde{\boldsymbol{\theta}}$ denote the constrained ML estimates obtained by imposing r restrictions when maximizing the loglikelihood. Then the ESS from the OPG regression

$$\boldsymbol{\iota} = \boldsymbol{G}(\tilde{\boldsymbol{\theta}})\boldsymbol{b} + \text{residuals}, \tag{29}$$

which is equal to n times the uncentered R^2, is the OPG form of the LM statistic.

Like the GNR, the OPG regression can be used for many purposes. The use of what is essentially the OPG regression for obtaining maximum likelihood estimates and computing covariance matrices was advocated by Berndt, Hall, Hall, and Hausman (1974). Using it to compute Lagrange Multiplier, or LM, tests was suggested by Godfrey and Wickens (1981), and using it to compute information matrix tests was proposed by Chesher (1983) and Lancaster (1984). The OPG regression is appealing for all these uses because it applies to a very wide variety of models and requires only first derivatives. In general, however, both estimated covariance matrices and test statistics based on the OPG regression are not very reliable in finite samples. In particular, a large number of papers, including Chesher and Spady (1991), Davidson and MacKinnon (1985a, 1992), and Godfrey, McAleer, and McKenzie (1988), have shown that, in finite samples, LM tests based on the OPG regression tend to overreject, often very severely.

Despite this drawback, the OPG regression provides a particularly convenient way to obtain various theoretical results. For example, suppose that we are interested in the variance of $\hat{\theta}_2$, the last element of $\hat{\boldsymbol{\theta}}$. If $\boldsymbol{\theta}_1$ denotes a vector of the remaining $k-1$ elements, and $\boldsymbol{G}(\boldsymbol{\theta})$ and $\boldsymbol{b}$ are partitioned in the same way as $\boldsymbol{\theta}$, the OPG regression becomes

$$\boldsymbol{\iota} = \boldsymbol{G}_1(\boldsymbol{\theta})\boldsymbol{b}_1 + \boldsymbol{G}_2(\boldsymbol{\theta})b_2 + \text{residuals},$$

and the FWL regression derived from this by retaining only the last regressor is

$$\boldsymbol{M}_1\boldsymbol{\iota} = \boldsymbol{M}_1\boldsymbol{G}_2 b_2 + \text{residuals},$$

where $\boldsymbol{M}_1 \equiv \mathbf{I} - \boldsymbol{G}_1(\boldsymbol{G}_1^\top\boldsymbol{G}_1)^{-1}\boldsymbol{G}_1^\top$, and the dependence on $\boldsymbol{\theta}$ has been suppressed for notational convenience. The covariance matrix estimate from this is just

$$(\boldsymbol{G}_2^\top\boldsymbol{M}_1\boldsymbol{G}_2)^{-1} = \big(\boldsymbol{G}_2^\top\boldsymbol{G}_2 - \boldsymbol{G}_2^\top\boldsymbol{G}_1(\boldsymbol{G}_1^\top\boldsymbol{G}_1)^{-1}\boldsymbol{G}_1^\top\boldsymbol{G}_2\big)^{-1}. \tag{30}$$

If we divide each of the components of (30) by n and take their probability limits, we find that

$$\lim_{n\to\infty} \operatorname{Var}\big(n^{1/2}(\hat{\theta}_2 - \theta_{20})\big) = (\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12})^{-1},$$

where θ_{20} is the true value of θ_2. This is a very well-known result, but, since its relation to the FWL theorem is not obvious without appeal to the OPG regression, it is not usually obtained in such a convenient or illuminating way.

7. An Artificial Regression for GMM Estimation

Another useful artificial regression, much less well known than the OPG regression, is available for a class of models estimated by the generalized method of moments (GMM). Many such models can be formulated in terms of functions $f_t(\boldsymbol{\theta})$ of the model parameters and the data, such that, when they are evaluated at the true $\boldsymbol{\theta}$, their expectations conditional on corresponding information sets, Ω_t, vanish. The Ω_t usually contain all information available prior to the time of observation t, and so, as with the GNR and the OPG regression, lags of dependent variables are allowed.

Let the $n \times l$ matrix $\boldsymbol{W}$ denote the instruments used to obtain the GMM estimates. The t^{th} row of $\boldsymbol{W}$, denoted $\boldsymbol{W}_t$, must contain variables in Ω_t only. The dimension of $\boldsymbol{\theta}$ is k, as before, and, for $\boldsymbol{\theta}$ to be identified, we need $l \geq k$. The GMM estimates with $l \times l$ weighting matrix $\boldsymbol{A}$ are obtained by minimizing the criterion function

$$Q(\boldsymbol{\theta}) = \frac{1}{2} \boldsymbol{f}^\top(\boldsymbol{\theta}) \boldsymbol{W} \boldsymbol{A} \boldsymbol{W}^\top \boldsymbol{f}(\boldsymbol{\theta}) \tag{31}$$

with respect to $\boldsymbol{\theta}$. Here $\boldsymbol{f}(\boldsymbol{\theta})$ is the n–vector with typical element $f_t(\boldsymbol{\theta})$. For the procedure known as efficient GMM, the weighting matrix $\boldsymbol{A}$ is chosen so as to be proportional, asymptotically at least, to the inverse of the covariance matrix of $\boldsymbol{W}^\top \boldsymbol{f}(\boldsymbol{\theta})$. In the simplest case, the $f_t(\boldsymbol{\theta})$ are serially uncorrelated and homoskedastic with variance 1, and so an appropriate choice is $\boldsymbol{A} = (\boldsymbol{W}^\top \boldsymbol{W})^{-1}$. With this choice, the criterion function (31) becomes

$$Q(\boldsymbol{\theta}) = \frac{1}{2} \boldsymbol{f}^\top(\boldsymbol{\theta}) \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{f}(\boldsymbol{\theta}), \tag{32}$$

where $\boldsymbol{P}_{\boldsymbol{W}}$ is the orthogonal projection on to the columns of $\boldsymbol{W}$.

Let $\boldsymbol{J}(\boldsymbol{\theta})$ be the negative of the $n \times k$ Jacobian matrix of $\boldsymbol{f}(\boldsymbol{\theta})$, so that the ti^{th} element of $\boldsymbol{J}(\boldsymbol{\theta})$ is $-\partial f_t / \partial \theta_i(\boldsymbol{\theta})$. The first-order conditions for minimizing (32) are

$$\boldsymbol{J}^\top(\boldsymbol{\theta}) \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}. \tag{33}$$

By standard arguments, it can be seen that the vector $\hat{\boldsymbol{\theta}}$ that solves (33) is asymptotically normal and asymptotically satisfies the equation

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = (n^{-1} \boldsymbol{J}_0^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{J}_0)^{-1} n^{-1/2} \boldsymbol{J}_0^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{f}_0, \tag{34}$$

with $\boldsymbol{J}_0 = \boldsymbol{J}(\boldsymbol{\theta}_0)$ and $\boldsymbol{f}_0 = \boldsymbol{f}(\boldsymbol{\theta}_0)$. See Davidson and MacKinnon (1993, Chapter 17), for a full discussion of GMM estimation.

Now consider the artificial regression

$$\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{J}(\boldsymbol{\theta}) \boldsymbol{b} + \text{residuals}. \tag{35}$$

By the first-order conditions (33) for $\boldsymbol{\theta}$, this equation clearly satisfies condition (i), and in fact it also satisfies condition (i′) for the criterion function $Q(\boldsymbol{\theta})$

of (32). Since the covariance matrix of $\boldsymbol{f}(\boldsymbol{\theta}_0)$ is just the identity matrix, it follows from (34) that condition (ii) is also satisfied. Arguments just like those presented in Section 3 for the GNR can be used to show that condition (iii), the one-step property, is also satisfied by (35).

If the $f_t(\boldsymbol{\theta}_0)$ are homoskedastic but with unknown variance σ^2, regression (35) can be used in exactly the same way as the GNR. Either the regressand and regressors can be divided by a suitable consistent estimate of σ, or else all test statistics can be computed as ratios, in F or nR^2 form, as appropriate.

An important special case of (35) is provided by the class of regression models, linear or nonlinear, estimated with instrumental variables (IV). Such a model can be written in the form (3), but it will be estimated by minimizing, not the criterion function (4) related to the sum of squared residuals, but rather

$$Q(\boldsymbol{\beta}) \equiv \tfrac{1}{2}\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top \boldsymbol{P}_{\boldsymbol{W}} \big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big),$$

where $\boldsymbol{W}$ is an $n \times l$ matrix of instrumental variables. This criterion function has exactly the same form as (32), with $\boldsymbol{\beta}$ instead of $\boldsymbol{\theta}$, and with $\boldsymbol{f}(\boldsymbol{\beta}) = \boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})$. In addition, $\boldsymbol{J}(\boldsymbol{\beta}) = \boldsymbol{X}(\boldsymbol{\beta})$, where $\boldsymbol{X}(\boldsymbol{\beta})$ is defined, exactly as for the GNR, to have ti^{th} element $\partial x_t / \partial \beta_i(\boldsymbol{\beta})$. The resulting artificial regression for the IV model, which takes the form

$$\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}(\boldsymbol{\beta}) \boldsymbol{b} + \text{residuals}, \tag{36}$$

is often referred to as a GNR, because, except for the projection matrix $\boldsymbol{P}_{\boldsymbol{W}}$, it is identical to (7): See Davidson and MacKinnon (1993, Chapter 7).

8. Artificial Regressions and Heteroskedasticity

Covariance matrices and test statistics calculated via the GNR (7), or via artificial regressions such as (35) and (36), are not asymptotically valid when the assumption that the error terms are IID is violated. Consider a modified version of the nonlinear regression model (3), in which $E(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is an $n \times n$ diagonal matrix with t^{th} diagonal element ω_t^2. Let $\hat{\boldsymbol{\Omega}}$ denote an $n \times n$ diagonal matrix with the squared residual $\hat{u}_t^2$ as the t^{th} diagonal element. It has been known since the work of White (1980) that the matrix

$$(\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}})^{-1} \hat{\boldsymbol{X}}^\top \hat{\boldsymbol{\Omega}} \hat{\boldsymbol{X}} (\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}})^{-1} \tag{37}$$

provides an estimator of $\text{Var}(\hat{\boldsymbol{\beta}})$, which can be used in place of the usual estimator, $s^2(\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}})^{-1}$. Like the latter, this **heteroskedasticity-consistent covariance matrix estimator**, or **HCCME**, can be computed by means of an artificial regression. We will refer to this regression as the **heteroskedasticity-robust Gauss-Newton Regression**, or **HRGNR**.

In order to derive the HRGNR, it is convenient to begin with a linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, and to consider the criterion function

$$Q(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The negative of the gradient of this function with respect to $\boldsymbol{\beta}$ is

$$\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{38}$$

and its Hessian is the matrix

$$\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}, \tag{39}$$

of which the inverse is the HCCME if we replace $\boldsymbol{\Omega}$ by $\hat{\boldsymbol{\Omega}}$. Equating the gradient to zero just yields the OLS estimator, since $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}$ are $k \times k$ nonsingular matrices.

Let $\boldsymbol{V}$ be an $n \times n$ diagonal matrix with t^{th} diagonal element equal to ω_t; thus $\boldsymbol{V}^2 = \boldsymbol{\Omega}$. Consider the $n \times k$ regressor matrix $\boldsymbol{R}$ defined by

$$\boldsymbol{R} = \boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{V}^2 \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{P}_{\boldsymbol{V}\boldsymbol{X}} \boldsymbol{V}^{-1} \boldsymbol{X}, \tag{40}$$

where $\boldsymbol{P}_{\boldsymbol{V}\boldsymbol{X}}$ projects orthogonally on to the columns of $\boldsymbol{V}\boldsymbol{X}$. We have

$$\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}, \tag{41}$$

which is just the Hessian (39). Let $\boldsymbol{U}(\boldsymbol{\beta})$ be a diagonal matrix with t^{th} diagonal element equal to $y_t - \boldsymbol{X}_t\boldsymbol{\beta}$. Then, if we define $\boldsymbol{R}(\boldsymbol{\beta})$ as in (40) but with $\boldsymbol{V}$ replaced by $\boldsymbol{U}(\boldsymbol{\beta})$, we find that $\hat{\boldsymbol{R}}^\top \hat{\boldsymbol{R}}$ is the HCCME (37).

In order to derive the regressand $\boldsymbol{r}(\boldsymbol{\beta})$, note that, for condition (i′) to be satisfied, we require

$$\boldsymbol{R}^\top(\boldsymbol{\beta}) \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{U}^2(\boldsymbol{\beta}) \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta});$$

recall (38). Since the t^{th} element of $\boldsymbol{U}(\boldsymbol{\beta})$ is $y_t - \boldsymbol{X}_t\boldsymbol{\beta}$, this implies that

$$\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{U}^{-1}(\boldsymbol{\beta})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\iota}.$$

In the general nonlinear case, $\boldsymbol{X}$ becomes $\boldsymbol{X}(\boldsymbol{\beta})$, and the HRGNR has the form

$$\boldsymbol{\iota} = \boldsymbol{P}_{\boldsymbol{U}(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})} \boldsymbol{U}^{-1}(\boldsymbol{\beta}) \boldsymbol{X}(\boldsymbol{\beta}) \boldsymbol{b} + \text{residuals}, \tag{42}$$

where now the t^{th} diagonal element of $\boldsymbol{U}(\boldsymbol{\beta})$ is $y_t - x_t(\boldsymbol{\beta})$. When $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, the vector of NLS estimates,

$$\begin{aligned} \hat{\boldsymbol{r}}^\top \hat{\boldsymbol{R}} &= \boldsymbol{\iota}^\top \boldsymbol{P}_{\hat{\boldsymbol{U}}\hat{\boldsymbol{X}}} \hat{\boldsymbol{U}}^{-1} \hat{\boldsymbol{X}} \\ &= \boldsymbol{\iota}^\top \hat{\boldsymbol{U}}\hat{\boldsymbol{X}}(\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}\hat{\boldsymbol{X}})^{-1} \hat{\boldsymbol{X}}^\top \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^{-1} \hat{\boldsymbol{X}} \\ &= \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{X}}(\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{\Omega}} \hat{\boldsymbol{X}})^{-1} \hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}} = \boldsymbol{0}, \end{aligned} \tag{43}$$

because the NLS first-order conditions give $\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{u}} = \boldsymbol{0}$. Thus condition (i) is satisfied for the nonlinear case. Condition (ii) is satisfied by construction, as can be seen by putting hats on everything in (41).

For condition (iii) to hold, regression (42) must satisfy the one-step property. We will only show that this property holds for linear models. Extending the argument to nonlinear models would be tedious but not difficult. In the linear case, evaluating (42) at an arbitrary $\acute{\boldsymbol{\beta}}$ gives

$$\acute{\boldsymbol{b}} = (\boldsymbol{X}^\top \acute{\boldsymbol{U}}^{-1} \boldsymbol{P}_{\acute{\boldsymbol{U}}\boldsymbol{X}} \acute{\boldsymbol{U}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \acute{\boldsymbol{U}}^{-1} \boldsymbol{P}_{\acute{\boldsymbol{U}}\boldsymbol{X}} \boldsymbol{\iota}.$$

With a little algebra, it can be shown that this reduces to

$$\acute{\boldsymbol{b}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \acute{\boldsymbol{u}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\acute{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}} - \acute{\boldsymbol{\beta}}, \tag{44}$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator. It follows that the one-step estimator $\acute{\boldsymbol{\beta}} + \acute{\boldsymbol{b}}$ is equal to $\hat{\boldsymbol{\beta}}$, as we wished to show. In the nonlinear case, of course, we obtain an asymptotic equality rather than an exact equality.

As with the ordinary GNR, the HRGNR is particularly useful for hypothesis testing. If we partition $\boldsymbol{\beta}$ as $[\boldsymbol{\beta}_1 \vdots \boldsymbol{\beta}_2]$ and wish to test the r zero restrictions $\boldsymbol{\beta}_2 = \boldsymbol{0}$, we need to run two versions of the regression and compute the difference between the two SSRs or SSEs. The two regressions are:

$$\boldsymbol{\iota} = \boldsymbol{P}_{\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}} \tilde{\boldsymbol{U}}^{-1} \tilde{\boldsymbol{X}}_1 \boldsymbol{b}_1 + \text{residuals, and} \tag{45}$$

$$\boldsymbol{\iota} = \boldsymbol{P}_{\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}} \tilde{\boldsymbol{U}}^{-1} \tilde{\boldsymbol{X}}_1 \boldsymbol{b}_1 + \boldsymbol{P}_{\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}} \tilde{\boldsymbol{U}}^{-1} \tilde{\boldsymbol{X}}_2 \boldsymbol{b}_2 + \text{residuals.} \tag{46}$$

It is important to note that the first regression is *not* the HRGNR for the restricted model, because it uses the matrix $\boldsymbol{P}_{\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}}$ rather than the matrix $\boldsymbol{P}_{\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}_1}$. In consequence, the regressand in (45) will not be orthogonal to the regressors. This is why we need to run two artificial regressions. We could compute an ordinary F statistic instead of the difference between the SSRs from (45) and (46), but there would be no advantage to doing so, since the F form of the test merely divides by a stochastic quantity that tends to 1 asymptotically.

The HRGNR appears to be new. The trick of multiplying $\boldsymbol{X}(\boldsymbol{\beta})$ by $\boldsymbol{U}^{-1}(\boldsymbol{\beta})$ in order to obtain an HCCME by means of an OLS regression was used, in a different context, by Messer and White (1984). This trick does cause a problem in some cases. If any element on the diagonal of the matrix $\boldsymbol{U}(\boldsymbol{\beta})$ is equal to 0, the inverse of that element cannot be computed. Therefore, it is necessary to replace any such element by a small, positive number before computing $\boldsymbol{U}^{-1}(\boldsymbol{\beta})$.

A different, and considerably more limited, type of heteroskedasticity-robust GNR, which is applicable only to hypothesis testing, was first proposed by Davidson and MacKinnon (1985b). It was later rediscovered by Wooldridge (1990, 1991) and extended to handle other cases, including regression models with error terms that have autocorrelation as well as heteroskedasticity of unknown form.

It is possible to construct a variety of artificial regressions that provide different covariance matrix estimators for regression models. From (43) and (44), it follows that any artificial regression with regressand

$$\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{U}^{-1}(\boldsymbol{\beta})\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)$$

and regressors

$$\boldsymbol{R}(\boldsymbol{\beta}) = \boldsymbol{P}_{\boldsymbol{U}(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})}\boldsymbol{U}^{-1}(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})$$

satisfies properties (i) and (iii) for the least-squares estimator, for any nonsingular matrix $\boldsymbol{U}(\boldsymbol{\beta})$. Thus any sandwich covariance matrix estimator can be computed by choosing $\boldsymbol{U}(\boldsymbol{\beta})$ appropriately; the estimator (37) is just one example. In fact, it is possible to develop artificial regressions that allow testing not only with a variety of different HCCMEs, but also with some sorts of heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimators. It is also a simple matter to use such estimators with modified versions of the artificial regression (35) used with models estimated by GMM.

9. Double-Length Regressions

Up to this point, the number of observations for all the artificial regressions we have studied has been equal to n, the number of observations in the data. In some cases, however, artificial regressions may have $2n$ or even $3n$ observations. This can happen whenever each observation makes two or more contributions to the criterion function.

The first **double-length artificial regression**, or **DLR**, was proposed by Davidson and MacKinnon (1984a). We will refer to it as *the* DLR, even though it is no longer the only artificial regression with $2n$ observations. The class of models to which the DLR applies is a subclass of the one used for GMM estimation. Such models may be written as

$$f_t(y_t, \boldsymbol{\theta}) = \varepsilon_t, \quad t = 1, \ldots, n, \quad \varepsilon_t \sim \text{NID}(0, 1), \tag{47}$$

where, as before, each $f_t(\cdot)$ is a smooth function that depends on the data and on a k–vector of parameters $\boldsymbol{\theta}$. Here, however, the f_t are assumed to be normally distributed conditional on the information sets Ω_t, as well as being of mean zero, serially uncorrelated, and homoskedastic with variance 1. Further, f_t may depend only on a scalar dependent variable y_t, although lagged dependent variables are allowed as explanatory variables.

The class of models (47) is much less restrictive than it may at first appear to be. In particular, it is not essential that the error terms follow the normal distribution, although it is essential that they follow some specified, continuous distribution, which can be transformed into the standard normal distribution, so as to allow the model to be written in the form of (47). A great many models

that involve transformations of the dependent variable can be put into the form of (47). For example, consider the Box-Cox regression model

$$\tau(y_t, \lambda) = \sum_{i=1}^{k} \beta_i \tau(X_{ti}, \lambda) + \sum_{j=1}^{l} \gamma_j Z_{tj} + u_t, \quad u_t \sim N(0, \sigma^2), \tag{48}$$

where $\tau(x, \lambda) = (x^\lambda - 1)/\lambda$ is the Box-Cox transformation (Box and Cox, 1964), y_t is the dependent variable, the X_{ti} are independent variables that are always positive, and the Z_{tj} are additional independent variables. We can rewrite (48) in the form of (47) by making the definition

$$f_t(y_t, \boldsymbol{\theta}) = \frac{1}{\sigma}\Big(\tau(y_t, \lambda) - \sum_{i=1}^{k} \beta_i \tau(X_{ti}, \lambda) - \sum_{j=1}^{l} \gamma_j Z_{tj}\Big).$$

For the model (47), the contribution of the t^{th} observation to the loglikelihood function $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ is

$$\ell_t(y_t, \boldsymbol{\theta}) = -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2} f_t^2(y_t, \boldsymbol{\theta}) + k_t(y_t, \boldsymbol{\theta}),$$

where

$$k_t(y_t, \boldsymbol{\theta}) \equiv \log\left|\frac{\partial f_t(y_t, \boldsymbol{\theta})}{\partial y_t}\right|$$

is a Jacobian term. Now let us make the definitions

$$F_{ti}(y_t, \boldsymbol{\theta}) \equiv \frac{\partial f_t(y_t, \boldsymbol{\theta})}{\partial \theta_i} \quad \text{and} \quad K_{ti}(y_t, \boldsymbol{\theta}) \equiv \frac{\partial k_t(y_t, \boldsymbol{\theta})}{\partial \theta_i}$$

and define $\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{\theta})$ and $\boldsymbol{K}(\boldsymbol{y}, \boldsymbol{\theta})$ as the $n \times k$ matrices with typical elements $F_{ti}(y_t, \boldsymbol{\theta})$ and $K_{ti}(y_t, \boldsymbol{\theta})$ and typical rows $\boldsymbol{F}_t(\boldsymbol{y}, \boldsymbol{\theta})$ and $\boldsymbol{K}_t(\boldsymbol{y}, \boldsymbol{\theta})$. Similarly, let $\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{\theta})$ be the n–vector with typical element $f_t(y_t, \boldsymbol{\theta})$.

The DLR, which has $2n$ artificial observations, may be written as

$$\begin{bmatrix} \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{\theta}) \\ \boldsymbol{\iota} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y}, \boldsymbol{\theta}) \end{bmatrix} \boldsymbol{b} + \text{residuals}. \tag{49}$$

Since the gradient of $\ell(\boldsymbol{y}, \boldsymbol{\theta})$ is

$$\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\theta}) = -\boldsymbol{F}^\top(\boldsymbol{y}, \boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{\theta}) + \boldsymbol{K}^\top(\boldsymbol{y}, \boldsymbol{\theta}) \boldsymbol{\iota}, \tag{50}$$

we see that regression (49) satisfies condition (i′). It can also be shown that it satisfies conditions (ii) and (iii), and thus it has all the properties of an artificial regression.

The DLR can be used for many purposes, including nonnested hypothesis tests of models with different functional forms (Davidson and MacKinnon, 1984a), tests of functional form (MacKinnon and Magee, 1990), and tests of linear and loglinear regressions against Box-Cox alternatives like (48) (Davidson and MacKinnon, 1985a). The latter application has recently been extended to models with AR(1) errors by Baltagi (1999). An accessible discussion of the DLR may be found in Davidson and MacKinnon (1988). When both the OPG regression and the DLR are available, the finite-sample performance of the latter always seems to be very much better than that of the former.

As we remarked earlier, the DLR is not the only artificial regression with $2n$ artificial observations. In particular, Orme (1995) showed how to construct such a regression for the widely-used tobit model, and Davidson and MacKinnon (1999) provided evidence that Orme's regression generally works very well. It makes sense that a double-length regression should be needed in this case, because the tobit loglikelihood is the sum of two summations, which are quite different in form. One summation involves all the observations for which the dependent variable is equal to zero, and the other involves all the observations for which it takes on a positive value.

10. 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 Gauss-Newton regression. It was independently suggested by Engle (1984) and Davidson and MacKinnon (1984b).

The object of a binary response model is to predict the probability that the binary dependent variable, y_t, is equal to 1 conditional on some information set Ω_t. A useful class of binary response models can be written as

$$E(y_t \,|\, \Omega_t) = \Pr(y_t = 1) = F(\boldsymbol{Z}_t\boldsymbol{\beta}). \tag{51}$$

Here $\boldsymbol{Z}_t$ is a row vector of explanatory variables that belong to Ω_t, $\boldsymbol{\beta}$ is the vector of parameters to be estimated, and $F(x)$ is the differentiable cumulative distribution function (CDF) of some scalar probability distribution. For the probit model, $F(x)$ is the standard normal CDF. For the logit model, $F(x)$ is the logistic function

$$\frac{\exp(x)}{1+\exp(x)} = \big(1+\exp(-x)\big)^{-1}.$$

The loglikelihood function for this class of binary response models is

$$\ell(\boldsymbol{\beta}) = \sum_{t=1}^{n} \Big((1-y_t)\log\big(1-F(\boldsymbol{Z}_t\boldsymbol{\beta})\big) + y_t \log\big(F(\boldsymbol{Z}_t\boldsymbol{\beta})\big)\Big), \tag{52}$$

If $f(x) = F'(x)$ is the density corresponding for the CDF $F(x)$, the first-order conditions for maximizing (52) are

$$\sum_{t=1}^{n} \frac{(y_t - \hat{F}_t)\hat{f}_t Z_{ti}}{\hat{F}_t(1 - \hat{F}_t)} = 0, \;\; i = 1, \ldots, k, \tag{53}$$

where Z_{ti} is the ti^{th} component of $\boldsymbol{Z}_t$, $\hat{f}_t \equiv f(\boldsymbol{Z}_t\hat{\boldsymbol{\beta}})$ and $\hat{F}_t \equiv F(\boldsymbol{Z}_t\hat{\boldsymbol{\beta}})$.

There is more than one way to derive the artificial regression that corresponds to the model (51). The easiest is to rewrite it in the form of the nonlinear regression model

$$y_t = F(\boldsymbol{Z}_t\boldsymbol{\beta}) + u_t. \tag{54}$$

The error term u_t here is evidently non-normal and heteroskedastic. Because y_t is like a Bernoulli trial with probability p given by $F(\boldsymbol{Z}_t\boldsymbol{\beta})$, and the variance of a Bernoulli trial is $p(1-p)$, the variance of u_t is

$$v_t(\boldsymbol{\beta}) \equiv F(\boldsymbol{Z}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{Z}_t\boldsymbol{\beta})\big). \tag{55}$$

The ordinary GNR for (54) would be

$$y_t - F(\boldsymbol{Z}_t\boldsymbol{\beta}) = f(\boldsymbol{Z}_t\boldsymbol{\beta})\boldsymbol{Z}_t\boldsymbol{b} + \text{residual},$$

but the ordinary GNR is not appropriate because of the heteroskedasticity of the u_t. Multiplying both sides by the square root of the inverse of (55) yields the artificial regression

$$v_t^{-1/2}(\boldsymbol{\beta})\big(y_t - F(\boldsymbol{Z}_t\boldsymbol{\beta})\big) = v_t^{-1/2}(\boldsymbol{\beta})f(\boldsymbol{Z}_t\boldsymbol{\beta})\boldsymbol{Z}_t\boldsymbol{b} + \text{residual}. \tag{56}$$

This regression has all the usual properties of artificial regressions. It can be seen from (53) that it satisfies condition (i′). Because a typical element of the information matrix corresponding to (52) is

$$\mathcal{I}_{ij}(\boldsymbol{\beta}) = \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n} Z_{ti}Z_{tj}\,\frac{f(\boldsymbol{Z}_t\boldsymbol{\beta})^2}{F(\boldsymbol{Z}_t\boldsymbol{\beta})\big(1 - F(\boldsymbol{Z}_t\boldsymbol{\beta})\big)}\right),$$

it is not difficult to show that regression (56) satisfies condition (ii). Finally, since (56) has the structure of a GNR, the arguments used in Section 3 show that it also satisfies condition (iii), the one-step property.

As an artificial regression, (56) can be used for all the things that other artificial regressions can be used for. In particular, when it is evaluated at restricted estimates $\tilde{\boldsymbol{\beta}}$, the explained sum of squares is an LM test statistic for testing the restrictions. The normalization of the regressand by its standard error means that other test statistics, such as nR^2 and the ordinary F statistic for the coefficients on the regressors that correspond to the restricted parameters to be zero, are also asymptotically valid. However, they seem to have slightly poorer finite-sample properties than the ESS (Davidson and MacKinnon, 1984b). It is, of course, possible to extend regression (56) in various ways. For example, it has been extended to tests of the functional form of $F(x)$ by Thomas (1993) and to tests of ordered logit models by Murphy (1996).

11. Conclusion

In this paper, we have introduced the concept of an artificial regression and discussed several examples. We have seen that artificial regressions can be useful for minimizing criterion functions, computing one-step estimates, calculating covariance matrix estimates, and computing test statistics. The last of these is probably the most common application. There is a close connection between the artificial regression for a given model and the asymptotic theory for that model. Therefore, as we saw in Section 6, artificial regressions can also be very useful for obtaining theoretical results.

Most of the artificial regressions we have discussed are quite well-known. This is true of the Gauss-Newton regression discussed in Sections 3 and 4, the OPG regression discussed in Section 6, the double-length regression discussed in Section 9, and the regression for binary response models discussed in Section 10. However, the artificial regression for GMM estimation discussed in Section 7 does not appear to have been treated previously in published work, and we believe that the heteroskedasticity-robust GNR discussed in Section 8 is new.

References

Baltagi, B. (1999). "Double length regressions for linear and log-linear regressions with AR(1) disturbances," *Statistical Papers*, **40**, 199–209.

Berndt, E. R., B. H. Hall, R. E. Hall, and J. A. Hausman (1974). "Estimation and inference in nonlinear structural models," *Annals of Economic and Social Measurement*, **3**, 653–665.

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

Chesher, A. (1983). "The information matrix test: simplified calculation via a score test interpretation," *Economics Letters*, **13**, 45–48.

Chesher, A., and R. Spady (1991). "Asymptotic expansions of the information matrix test statistic," *Econometrica*, **59**, 787–815.

Davidson, R., and J. G. MacKinnon (1981). "Several tests for model specification in the presence of alternative hypotheses," *Econometrica*, **49**, 781–793.

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

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

Davidson, R., and J. G. MacKinnon (1985a). "Testing linear and loglinear regressions against Box-Cox alternatives," *Canadian Journal of Economics*, **18**, 499–517.

Davidson, R., and J. G. MacKinnon (1985b). "Heteroskedasticity-robust tests in regression directions," *Annales de l'INSEE*, **59/60**, 183–218.

Davidson, R., and J. G. MacKinnon (1988). "Double-length artificial regressions," *Oxford Bulletin of Economics and Statistics*, **50**, 203–217.

Davidson, R., and J. G. MacKinnon (1990). "Specification tests based on artificial regressions," *Journal of the American Statistical Association*, **85**, 220–227.

Davidson, R., and J. G. MacKinnon (1992). "A new form of the information matrix test," *Econometrica*, **60**, 145–157.

Davidson, R., and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York: Oxford University Press.

Davidson, R., and J. G. MacKinnon (1999). "Bootstrap testing in nonlinear models," *International Economic Review*, **40**, 487–508.

Delgado, M. A., and T. Stengos (1994). "Semiparametric specification testing of non-nested econometric models," *Review of Economic Studies*, **61**, 291–303.

Engle, R. F. (1984). "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics," Ch. 13 in *Handbook of Econometrics*, Vol. II. Eds.: Zvi Griliches and Michael D. Intriligator. Amsterdam: North-Holland.

Godfrey, L. G. (1978). "Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables," *Econometrica*, **46**, 1293–1301.

Godfrey, L. G., M. McAleer, and C. R. McKenzie (1988). "Variable addition and Lagrange Multiplier tests for linear and logarithmic regression models," *Review of Economics and Statistics*, **70**, 492–503.

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

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

MacKinnon, J. G., and L. Magee (1990). "Transforming the dependent variable in regression models," *International Economic Review*, **31**, 315–339.

McCullough, B. D. (1999). "Econometric software reliability: EViews, LIMDEP, SHAZAM, and TSP," *Journal of Applied Econometrics*, **14**, 191–202.

Messer, K., and H. White (1984). "A note on computing the heteroskedasticity consistent covariance matrix using instrumental variable techniques," *Oxford Bulletin of Economics and Statistics*, **46**, 181–184.

Murphy, A. (1996). "Simple LM tests of mis-specification for ordered logit models," *Economics Letters*, **52**, 137–141.

Orme, C. (1995). "On the use of artificial regressions in certain microeconometric models," *Econometric Theory*, **11**, 290–305.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, (1992). *Numerical Recipes in C*, Second edition, Cambridge, Cambridge University Press.

Thomas, J. (1993). "On testing the logistic assumption in binary dependent variable models," *Empirical Economics*, **18**, 381–392.

White, H. (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, **48**, 817–38.

Wooldridge, J. M. (1990). "A unified approach to robust, regression-based specification tests," *Econometric Theory*, **6**, 17–43.

Wooldridge, J. M. (1991). "On the application of robust, regression-based diagnostics to models of conditional means and conditional variances," *Journal of Econometrics*, **47**, 5–46.