# Artificial Regressions and $C(\alpha)$ Tests

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@qed.econ.queensu.ca**

## Abstract

Any artificial regression that can be used to compute Lagrange Multiplier tests can just as easily be used to compute $C(\alpha)$ tests. This also makes it possible to compute Wald-like tests by means of artificial regressions

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada.

Revised, July 3, 1990, TEXed in October 2009

## 1. Introduction

It is now well known that artificial regressions can be used to compute specification tests for a wide variety of econometric models; see for example Engle (1984) and Godfrey (1988). The only tests that are routinely computed in this way, however, are Lagrange Multiplier tests, in which the artificial regression is evaluated at ML (or NLS) estimates under the null hypothesis. In this note we point out that artificial regressions can actually be used to compute other types of test as well. In general, these will be what are called $C(\alpha)$ tests, but in a special case they will be essentially Wald tests. We review some essential material on artificial regressions in the next section, present the main results of the note in section 3, and discuss a specific example, involving tests for serial correlation, in section 4.

## 2. Artificial Regressions

In order to deal with a very general class of artificial regressions, we make use of results established in Davidson and MacKinnon (1990). Suppose there is a fully specified parametrized model characterized by its loglikelihood function, which for a sample of size $n$ can be written as

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

where $\boldsymbol{\theta}$ is a $k$-vector of model parameters. We shall partition $\boldsymbol{\theta}$ as $[\boldsymbol{\theta}_1 \,\vdots\, \boldsymbol{\theta}_2]$ in order to consider the restrictions $\boldsymbol{\theta}_2 = \mathbf{0}$; we focus on this simple case only for expositional simplicity. Here $\boldsymbol{\theta}_i$ is a $k_i$-vector $(i = 1, 2)$ with $k = k_1 + k_2$. The $k$-vector of scores, with typical element $\partial\mathcal{L}(\boldsymbol{\theta})/\partial\theta_i$, will be denoted $\boldsymbol{g}(\boldsymbol{\theta})$. We shall assume that the data were generated by a data-generating process (DGP) characterized by the loglikelihood (1)for some true (but unknown) parameter vector $\boldsymbol{\theta}^0 = [\boldsymbol{\theta}_1^0 \,\vdots\, \mathbf{0}]$, and that suitable regularity conditions are satisfied.

Various artificial regressions can be associated with the model (1). Such a regression always involves two things: a regressand, say $\boldsymbol{r}(\boldsymbol{\theta})$, and a matrix of regressors, say $\boldsymbol{R}(\boldsymbol{\theta})$, both of which depend on the parameter vector $\boldsymbol{\theta}$. It may be written as:

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{b} + \text{residuals}, \tag{2}$$

where we use "residuals" as a neutral term to avoid any implication that (2) is a statistical model. The regressand $\boldsymbol{r}(\boldsymbol{\theta})$ and matrix of regressors $\boldsymbol{R}(\boldsymbol{\theta})$ have certain defining properties. These properties are as follows:

(i) under the DGP characterized by $\boldsymbol{\theta}$, $\rho(\boldsymbol{\theta}) \equiv \operatorname{plim}_{n\to\infty} \boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})$ exists and is a finite, smooth, real-valued function of $\boldsymbol{\theta}$;

(ii) $\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta}) = \rho(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})$; and

(iii) if $\boldsymbol{\theta} \to \boldsymbol{\theta}^0$, then $n^{-1}\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\theta}) \to \rho(\boldsymbol{\theta}^0)\mathcal{I}(\boldsymbol{\theta}^0)$.

As we prove in the paper cited above, artificial regressions that satisfy properties (i) through (iii) for all $\boldsymbol{\theta}$ belonging to some compact $k$-dimensional parameter space $\Theta$ have two crucial features. First, if the artificial regression (2) 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 $\acute{\boldsymbol{b}}$ obtained by OLS have the property that

$$n^{1/2}\acute{\boldsymbol{b}} = n^{1/2}(\hat{\boldsymbol{\theta}} - \acute{\boldsymbol{\theta}}) + o_p(1) \text{ as } n \to \infty, \tag{3}$$

where $\hat{\boldsymbol{\theta}}$ is the (asymptotically efficient) ML estimator of the model (1). Thus the artificial regression can be used to compute one-step estimates $(\acute{\boldsymbol{\theta}} + \boldsymbol{b})$ that are asymptotically equivalent to the ML estimates $\hat{\boldsymbol{\theta}}$.

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

$$n^{-1}\boldsymbol{g}^\top(\acute{\boldsymbol{\theta}})\mathcal{I}^{-1}(\boldsymbol{\theta}^0)\boldsymbol{g}(\acute{\boldsymbol{\theta}}), \tag{4}$$

where $\mathcal{I}^{-1}(\boldsymbol{\theta}^0)$ is the inverse of the information matrix evaluated at $\boldsymbol{\theta}^0$. This result is evidently what allows artificial regressions to be used to compute LM tests. If $\acute{\boldsymbol{\theta}}$ were the vector of restricted estimates $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$, the quantity (4) would be an LM statistic, although in practice $\mathcal{I}(\boldsymbol{\theta}^0)$ would be replaced by $\mathcal{I}(\tilde{\boldsymbol{\theta}})$. LM tests may therefore be based on $nR^2$, or any equivalent test statistic, such as an $F$ test for $\boldsymbol{b}_2 = \mathbf{0}$, from the artificial regression (2) evaluated at $\tilde{\boldsymbol{\theta}}$.

## 3. $C(\alpha)$ Tests and Wald-like Tests

Although it is much less well-known than the LM, LR and Wald tests, there is a fourth classical test that is asymptotically equivalent to them. Proposed originally by Neyman (1959), it is called the $C(\alpha)$ test; see Moran (1970), Smith (1987) and Dagenais and Dufour (1989) for more details. The $C(\alpha)$ test statistic can be written as

$$n^{-1}\Big(\acute{\boldsymbol{g}}^\top\acute{\mathcal{I}}^{-1}\acute{\boldsymbol{g}} - \acute{\boldsymbol{g}}_1^\top(\acute{\mathcal{I}}_{11})^{-1}\acute{\boldsymbol{g}}_1\Big), \tag{5}$$

where $\acute{\boldsymbol{g}}_1$ is the subvector of $\acute{\boldsymbol{g}} \equiv \boldsymbol{g}(\acute{\boldsymbol{\theta}})$ associated with $\boldsymbol{\theta}_1$, and $\acute{\mathcal{I}}_{11}$ is the corresponding submatrix of $\acute{\mathcal{I}} \equiv \mathcal{I}(\acute{\boldsymbol{\theta}})$. All quantities here are evaluated at root-$n$ consistent estimates $\acute{\boldsymbol{\theta}} = [\acute{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$.

The test statistic (5) is the difference between two quadratic forms. Using the results reviewed in the previous section, we see that the first of these is asymptotically equal to $nR^2$ from the artificial regression

$$\acute{\boldsymbol{r}} = \acute{\boldsymbol{R}}_1\boldsymbol{b}_1 + \acute{\boldsymbol{R}}_2\boldsymbol{b}_2 + \text{residuals}, \tag{6}$$

where the regressand and regressors are evaluated at $\acute{\boldsymbol{\theta}}$ and the latter have been partitioned into those that correspond to $\boldsymbol{\theta}_1$ and those that correspond to $\boldsymbol{\theta}_2$. Similarly, the second quadratic form in (5) is asymptotically equal to $nR^2$ from the artificial regression

$$\acute{\boldsymbol{r}} = \acute{\boldsymbol{R}}_1 \boldsymbol{b}_1 + \text{residuals}. \tag{7}$$

The $R^2$ from this regression would be zero if $\acute{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}$, by the first order conditions for $\tilde{\boldsymbol{\theta}}$, but will generally not be zero for any other choice of $\acute{\boldsymbol{\theta}}$.

The test statistic (5) is asymptotically equal to $n$ times the difference between the $R^2$s from (6) and (7). Thus we see that LM tests based on artificial regressions are just special cases of $C(\alpha)$ tests. The difference is that, because $nR^2$ from (7) is generally not zero, one cannot simply use $nR^2$ from (6) as the test statistic in the more general case. An equivalent test statistic, which may well have better finite-sample properties, is the $F$ statistic for $\boldsymbol{b}_2 = \mathbf{0}$ in (6). In their $F$ form the $C(\alpha)$ and LM tests are exactly the same. The intuition that underlies these results is very simple. From (3) and the fact that $\boldsymbol{\theta}_2 = \mathbf{0}$, we see that the OLS estimates of $\boldsymbol{b}_2$ in (6) are asymptotically equivalent to the unrestricted ML estimates $\hat{\boldsymbol{\theta}}$. Thus it is intuitively obvious that an $F$ test for $\boldsymbol{b}_2 = \mathbf{0}$ in (6) should be equivalent to a classical test for $\boldsymbol{\theta}_2 = \mathbf{0}$.

There are many cases in which it is easier to find root-$n$ consistent estimates $\acute{\boldsymbol{\theta}}$ than restricted ML estimates $\tilde{\boldsymbol{\theta}}$. For example, in simultaneous equations models it is much easier to obtain unrestricted reduced form estimates than restricted ones. Thus one possibility is that $\acute{\boldsymbol{\theta}}$ may simply be equal to $[\hat{\boldsymbol{\theta}}_1 \,\vdots\, \mathbf{0}]$, where $\hat{\boldsymbol{\theta}}_1$ is the unrestricted ML estimate of $\boldsymbol{\theta}_1$. In this case the $C(\alpha)$ test based on regressions (6) and (7) is actually a sort of Wald test, or at least a Wald-like test, since it will not be numerically equal to other forms of the Wald test. When the restrictions can easily be written as zero restrictions, this may be the easiest way to compute a test based on unrestricted estimates.

## 4. Tests for Serial Correlation

In this section we discuss a simple case in which $C(\alpha)$ tests based on artificial regressions can be convenient. One often wishes to test a regression model for several different orders of serial correlation. Suppose, for example, that the model to be tested is

$$y_t = \boldsymbol{X}_t \boldsymbol{\beta} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{8}$$

where $\boldsymbol{X}_t$ is a row vector of exogenous and/or predetermined variables and $\boldsymbol{\beta}$ is a vector of unknown parameters. Suppose one wishes to test (8) against the alternative that $u_t$ follows an AR($l$) process. When the null hypothesis is that there is no serial correlation, it is very easy to do so. One simply has to run the artificial regression

$$\tilde{u}_t = \boldsymbol{X}_t \boldsymbol{b} + \sum_{j=1}^{l} c_j \tilde{u}_{t-j} + \text{residuals}, \tag{9}$$

where $u_t$ denotes the $t^{\text{th}}$ residual from OLS estimation of (8). One then either computes $nR^2$ or performs an $F$ test of the hypothesis that all the $c_j$ are zero. See Godfrey (1988) for more details.

But suppose, as often happens in practice, that one begins by testing against AR(1) errors and immediately rejects the null hypothesis. There is now no point in testing (8) any further. If one were limited to using LM tests, one would next re-estimate the model allowing for AR(1) errors, and then test that more complicated model for serial correlation (and other forms of misspecification). But re-estimating the model will turn out to have been a waste of time if the AR(1) model is later rejected. By using a $C(\alpha)$ test one can avoid this re-estimation.

When the error terms in the model (8) are assumed to follow an AR(1) process, that model may be rewritten as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \rho y_t - 1 - \rho\boldsymbol{X}_t - 1\boldsymbol{\beta} + \varepsilon_t, \varepsilon_t \sim \text{IID}(0, \sigma^2). \tag{10}$$

When all the elements of $\boldsymbol{X}_t$ are exogenous, it is very easy to obtain root-$n$ consistent estimates of $\boldsymbol{\beta}$ and $\rho$. The OLS estimates $\tilde{\boldsymbol{\beta}}$ are consistent, and a consistent estimate of $\rho$, say $\tilde{\rho}$, may be obtained by regressing $\tilde{u}_t$ on $\tilde{u}_t - 1$. Then the $C(\alpha)$ test regression for testing the null hypothesis that the error terms in (10) are serially uncorrelated against the alternative that they follow an AR($l$) process is just

$$\begin{aligned} &y_t - \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} - \tilde{\rho}y_{t-1} + \tilde{\rho}\boldsymbol{X}_{t-1}\tilde{\boldsymbol{\beta}} = \\ &(\boldsymbol{X}_t - \tilde{\rho}\boldsymbol{X}_{t-1})\boldsymbol{b} + r(y_{t-1} - \boldsymbol{X}_{t-1}\tilde{\boldsymbol{\beta}}) + \sum_{j=1}^{l} c_j(y_{t-j} - \boldsymbol{X}_{t-j}\tilde{\boldsymbol{\beta}} - \tilde{\rho}y_{t-j-1} + \tilde{\rho}\boldsymbol{X}_{t-j-1}\tilde{\boldsymbol{\beta}}) \end{aligned} \tag{11}$$

One can test the hypothesis that all the $c_j$ are zero by means of an ordinary $F$ test, or by calculating the difference between $nR^2$ from (11) and $nR^2$ from the same regression without the last $l$ regressors.

When $\boldsymbol{X}_t$ includes one or more lagged dependent variables, the procedure just outlined will not work, because the OLS estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\rho}$ are not consistent. One must then use one of the numerous procedures for obtaining root-$n$ consistent estimates that are available for this case, such as the well-known one proposed by Hatanaka (1974). Given those estimates, one would proceed as before. Of course, in many cases one of these procedures will not actually be much easier than obtaining NLS estimates of (10), so that the advantages of not having to obtain the latter may be modest.

## 5. Conclusion

Any artificial regression that can be used to compute LM tests can also be used to compute $C(\alpha)$ tests. Indeed, LM tests based on artificial regressions are actually a special case of $C(\alpha)$ tests, in which the regressand and regressors of the artificial regression are evaluated at restricted ML (or NLS) estimates rather than at parameter estimates that are merely root-$n$ consistent and satisfy the null hypothesis. When they are computed as $F$ tests, LM and $C(\alpha)$ tests are formally the same. However, LM tests can also be computed as $nR^2$ from the artificial regression, and $C(\alpha)$ tests cannot.

## References


Dagenais, M. G. and J.-M. Dufour (1989). "Invariance, nonlinear models and asymptotic tests", Université de Montréal C.R.D.E. working paper.

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

Engle, R. F. (1984). "Wald, likelihood ratio and Lagrange multiplier tests in econometrics", Ch. 13 in: Z. Griliches and M.D. Intriligator, eds., *Handbook of Econometrics*, Vol. II (North-Holland, Amsterdam).

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics*, Cambridge University Press, Cambridge.

Hatanaka, M. (1974). "An efficient two-step estimator for the dynamic adjustment model with autoregressive errors", *Journal of Econometrics* **2**, 199–220.

Moran, P. A. P. (1970). "On asymptotically optimal tests of composite hypotheses", *Biometrika* **57**, 47–55.

Neyman, J. (1959). "Optimal asymptotic tests of composite statistical hypotheses]], in: U. Grenander, ed., *Probability and Statistics*, Wiley, New York.

Smith, R. J. (1987). "Alternative asymptotically optimal tests and their application to dynamic specification", *Review of Economic Studies* **54**, 665–680.