Maximum Likelihood Estimation by Artificial Regression

by

Russell Davidson

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

Department of Economics
McGill University
Montreal, Quebec, Canada
H3A 2T7

email: **Russell.Davidson@mcgill.ca**

Abstract

Artificial regressions are developed, based on elementary zero functions, that exploit the fact that the normal distribution is completely characterised by its first two moments. These artificial regressions can be used as the basis of numerical algorithms for the maximum likelihood estimation of models with normally distributed random elements, and other estimation techniques based on the optimisation of criterion functions. The proposed algorithms are often simpler to program than many conventional algorithms for the optimisation of functions, and they have the advantage that an asymptotically correct estimate of the covariance matrix of the parameter estimates is computed as a by-product. Specific examples discussed include regression models with ARMA or (G)ARCH errors.

Keywords: Artificial regression, elementary zero function, estimating equation, maximum likelihood.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. I am greatly indebted to James MacKinnon for numerous helpful suggestions and comments, and to participants at the 2003 Econometric Study Group Annual Conference in Bristol.

November 2003

1. Introduction

The point of view taken in this paper is that the basic unit of statistical information is an *elementary zero function*, that is, a function of the data associated with one observation of the sample and of parameters. The expectation of an elementary zero function is zero when the parameters are those associated with the data-generating process (DGP) that generated the sample. A zero function may be informative or not about any given parameter. If the function does not depend on a parameter, then it cannot be informative about it. Even if it does depend on a parameter, it may still not be informative about it. Whether a zero function is informative about a parameter, and, if so, to what extent, is measured by a quantity that, in some circumstances, can be described as the *optimal instrument* for that zero function and that parameter. If this instrument is zero, then the function is not informative about the parameter.

In this context, instrumental variables are viewed as coefficients by which a set of zero functions are weighted in the construction of the *estimating equations*, which are equations to be solved to find an estimator of the parameters. Estimating equations are said to be (asymptotically) optimal if the estimator they define has minimum (asymptotic) variance in the class of estimators defined using the given set of elementary zero functions. Optimal instruments are then the instruments used in optimal estimating equations.

The theory of estimating functions was originally developed by Godambe (1960); see also Godambe and Thompson (1978). In Section 2, we recall this theory, and explore the role of elementary zero functions in formulating optimal estimating equations. It is particularly simple to find the optimal instruments associated with a set of elementary zero functions for given parameters if the zero functions are homoskedastic and serially uncorrelated when evaluated at the "true" parameters, that is, those of the true DGP.

It is well known that the multivariate normal distribution is characterised completely by its first two moments. This suggests that efficient estimation of the parameters of models in which the random elements are multivariate normal can be performed by basing the estimating equations on elementary zero functions associated with the first two moments of these random elements. In Section 3, we show that this is indeed the case, and that the asymptotically optimal estimating equations are just the likelihood equations obtained by formulating a loglikelihood function under the assumption of normality.

This result makes it possible to develop *artificial regressions* that correspond to models with normal random elements estimated by maximum likelihood. The theory of artificial regressions is set out for models estimated by ML in Davidson and MacKinnon (1990), and in a more complete form in Davidson and MacKinnon (2001); see also Davidson and MacKinnon (2004), Chapter 15. Artificial regressions have numerous advantages; they may be used as the basis of an algorithm for computing estimators, for computing estimates of the covariance matrices of estimators, and for hypothesis testing. Many artificial regressions exist already for models estimated by ML. The best known, if not the best behaved, is no doubt the OPG artificial regression introduced by Godfrey and Wickens (1981). The

artificial regressions proposed in Section 4, based on elementary zero functions, have much better properties than almost all of those found in the existing literature, because the covariance matrix estimator they provide is the efficient score estimator, that is, the inverse of the information matrix evaluated at the MLE.

It is common practice to estimate a wide variety of time-series models by maximum likelihood, under the assumption of normally distributed errors. In many cases, it is quite difficult to formulate the loglikelihood function correctly, and in some cases, it appears to be impossible to do so analytically. In Section 5, it is seen that some of these difficulties can be overcome by use of the artificial regressions developed in Section 4. In particular, these regressions allow one to do full-information ML estimation of regression models with ARMA errors, with no analytical effort beyond the formulation of the stationary covariance matrix for the ARMA process under consideration, and of its derivatives with respect to the ARMA parameters.

In Section 6, artificial regressions are developed for (G)ARCH models. These are a little more complicated to set up than those for ARMA models, but no more so than setting up the loglikelihood function itself along with its derivatives. Since ML estimation of GARCH models is numerically very delicate, these artificial regressions offer the hope of greater numerical stability than what is provided by most currently available packages. Finally, Section 7 offers a few concluding remarks.

2. Elementary Zero Functions

Most parameter estimators can be defined by a set of estimating equations, by which a corresponding set of *estimating functions*, which depend on observed data and the parameters to be estimated, are set equal to zero. Estimating equations are often referred to in the econometrics literature as *moment conditions*. In many cases, the estimating equations are the first-order conditions for the maximisation or minimisation of a criterion function; obvious examples are least squares, maximum likelihood, and GMM. Estimating functions are usually zero functions, that is, functions which, when evaluated at the true parameter values, yield random variables of expectation zero. Estimating functions with this property are called unbiased.

Elementary zero functions, of which estimating functions are usually linear combinations, play a role analogous to that of residuals in a regression model. They depend on observed variables and on a vector of parameters, say $\boldsymbol{\theta}$. Let $f_t(\boldsymbol{\theta}, y_t)$ denote an elementary zero function for observation t. In general, there may well be more than one elementary zero function for each observation.

The parameters $\boldsymbol{\theta}$ are defined in the context of a model, denoted by $\mathbb{M}$, which is defined as a set of DGPs. A parameter-defining mapping associates to each DGP $\mu \in \mathbb{M}$ a unique parameter $\boldsymbol{\theta}_\mu$, often called the "true" value of $\boldsymbol{\theta}$ for that DGP. It is important to note that the uniqueness goes just one way here: A given parameter vector $\boldsymbol{\theta}$ may correspond to many DGPs, but each DGP corresponds to just one parameter vector. The existence of a parameter-defining mapping thus assumes that there are no problems of unidentified parameters left unsolved.

The key property of elementary zero functions can be written as

$$\mathrm{E}_\mu\big(f_t(\boldsymbol{\theta}_\mu, y_t)\big) = 0, \tag{1}$$

where $\mathrm{E}_\mu(\cdot)$ denotes the expectation under the DGP μ, and $\boldsymbol{\theta}_\mu$ is the (unique) parameter vector associated with μ. It is assumed that property (1) holds for all t and for all $\mu \in \mathbb{M}$.

If there are k parameters to be estimated, we need k estimating equations, and so k estimating functions. In general, these are linear combinations of the elementary zero functions, formed using instrumental variables as the coefficients of the linear combinations. If there are N elementary zero functions, let the $N \times k$ matrix $\boldsymbol{W}$ be the matrix of instruments. The estimating functions are thus the k components of $\boldsymbol{W}^\top \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$. A condition that ensures that they are unbiased is the following predeterminedness condition:

$$\mathrm{E}_\mu(f_t(\boldsymbol{\theta}_\mu, \boldsymbol{y}) \,|\, \boldsymbol{W}_t) = 0, \quad t = 1, \ldots, N, \tag{2}$$

where $\boldsymbol{W}_t$ is the t^{th} row of $\boldsymbol{W}$. The estimating equations

$$\boldsymbol{W}^\top \boldsymbol{f}(\hat{\boldsymbol{\theta}}, \boldsymbol{y}) = \boldsymbol{0} \tag{3}$$

implicitly define the estimator $\hat{\boldsymbol{\theta}}$.

For $\hat{\boldsymbol{\theta}}$ to be consistent, we require that

$$\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu) \equiv \operatorname*{plim}_{n\to\infty \,\mu} \frac{1}{n} \boldsymbol{W}^\top \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$$

exists and satisfies an asymptotic identification condition. Here n denotes the sample size. Condition (2) implies that $\boldsymbol{\alpha}(\boldsymbol{\theta}_\mu; \mu) = \boldsymbol{0}$ for all $\mu \in \mathbb{M}$, and the needed condition is that $\boldsymbol{\alpha}(\boldsymbol{\theta}; \mu) \neq \boldsymbol{0}$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_\mu$. For asymptotic normality, we assume that the f_t are continuously differentiable in a neighbourhood of the true parameters $\boldsymbol{\theta}_\mu$. Performing a first-order Taylor expansion of (3) around $\boldsymbol{\theta}_\mu$ and introducing some appropriate factors of powers of n, we obtain the result that

$$n^{-1/2} \boldsymbol{W}^\top \boldsymbol{f}(\boldsymbol{\theta}_\mu) + n^{-1} \boldsymbol{W}^\top \boldsymbol{F}(\bar{\boldsymbol{\theta}}) n^{1/2} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu) = \boldsymbol{0}, \tag{4}$$

where the $N \times k$ matrix $\boldsymbol{F}(\boldsymbol{\theta})$ has typical element

$$F_{ti}(\boldsymbol{\theta}) \equiv \frac{\partial f_t(\boldsymbol{\theta})}{\partial \theta_i},$$

where θ_i is the i^{th} element of $\boldsymbol{\theta}$. The notation $\boldsymbol{F}(\bar{\boldsymbol{\theta}})$ in (4) is a convenient shorthand: Row t of the matrix is the corresponding row of $\boldsymbol{F}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}_t$, where the $\bar{\boldsymbol{\theta}}_t$ all satisfy the inequality

$$\big\|\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\big\| \leq \big\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\big\|.$$

The consistency of $\hat{\boldsymbol{\theta}}$ then implies that the $\bar{\boldsymbol{\theta}}_t$ also tend to $\boldsymbol{\theta}_\mu$ as $n \to \infty$.

Asymptotic normality also needs the stronger asymptotic identification condition that $\operatorname{plim} n^{-1}\boldsymbol{W}^\top\boldsymbol{F}(\boldsymbol{\theta}_\mu)$ should be nonsingular. This allows us to solve (4) to obtain

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu) \stackrel{a}{=} -\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{F}(\boldsymbol{\theta}_\mu)\Big)^{-1} n^{-1/2}\boldsymbol{W}^\top\boldsymbol{f}(\boldsymbol{\theta}_\mu). \tag{5}$$

It follows that, assuming that a central limit theorem applies to $n^{-1/2}\boldsymbol{W}^\top\boldsymbol{f}(\boldsymbol{\theta}_\mu)$, the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu)$ is normal, with mean zero, and finite covariance matrix.

If we assume that the elementary zero functions are homoskedastic and serially uncorrelated, at least for a subset of interest of the DGPs in $\mathbb{M}$, it is easy enough to determine the asymptotic covariance matrix. Let $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})$ denote the vector of all the elementary zero functions. Then we assume that

$$\mathrm{E}\big(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{y})\boldsymbol{f}^\top(\boldsymbol{\theta}, \boldsymbol{y})\big) = \sigma^2\mathbf{I}$$

for the DGPs of interest. The covariance matrix of the limit of the right-hand side of (5) can then be seem to be

$$\begin{aligned}\sigma^2\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{F}(\boldsymbol{\theta}_\mu)\Big)^{-1}\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W}\Big)\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{F}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{W}\Big)^{-1}& \\ = \sigma^2 \operatorname*{plim}_{n\to\infty}\Big(\frac{1}{n}\boldsymbol{F}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{F}(\boldsymbol{\theta}_\mu)\Big)^{-1},&\end{aligned} \tag{6}$$

where $\boldsymbol{P}_{\boldsymbol{W}} \equiv \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$ is the orthogonal projection on to the space spanned by the instruments.

If $\boldsymbol{F}(\boldsymbol{\theta}_\mu)$ is predetermined in the sense of (2), then it is clear from (6) that the asymptotic covariance matrix is minimised by setting $\boldsymbol{W} = \boldsymbol{F}(\boldsymbol{\theta}_\mu)$. Since $\boldsymbol{\theta}_\mu$ is unknown, it is necessary in practice to replace it by a consistent estimate for practical purposes. Usually, however, $\boldsymbol{F}(\boldsymbol{\theta}_\mu)$ is not predetermined. In that case, define the matrix $\bar{\boldsymbol{F}}$ in terms of its typical row $\bar{\boldsymbol{F}}_t$, and an $N \times k$ matrix $\boldsymbol{V}$, as follows:

$$\bar{\boldsymbol{F}}_t \equiv \mathrm{E}\big(\boldsymbol{F}_t(\boldsymbol{\theta}_\mu) \,|\, \Omega_t\big) \quad \text{and} \quad \boldsymbol{V} \equiv \boldsymbol{F}(\boldsymbol{\theta}_\mu) - \bar{\boldsymbol{F}}, \tag{7}$$

where Ω_t denotes information that is predetermined with respect to the elementary zero function f_t. This implies that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\bar{\boldsymbol{F}}^\top\boldsymbol{F}(\boldsymbol{\theta}_\mu) = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\bar{\boldsymbol{F}}^\top(\bar{\boldsymbol{F}} + \boldsymbol{V}) = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\bar{\boldsymbol{F}}^\top\bar{\boldsymbol{F}}.$$

The term $\operatorname{plim} n^{-1}\bar{\boldsymbol{F}}^\top\boldsymbol{V}$ equals $\mathbf{O}$ because $\mathrm{E}(\boldsymbol{V}_t \,|\, \Omega_t) = \mathbf{0}$, and the conditional expectation $\bar{\boldsymbol{F}}_t$ belongs to the information set Ω_t. A straightforward asymptotic argument can then be used to show that setting $\boldsymbol{W} = \bar{\boldsymbol{F}}$ minimises the asymptotic covariance matrix over the set of admissible instruments. Thus the optimal instruments associated with the homoskedastic and serially uncorrelated set of zero functions f_t are the columns of the matrix $\bar{\boldsymbol{F}}$, or, equivalently, of $-\bar{\boldsymbol{F}}$.

3. Maximum Likelihood

If the parameters $\boldsymbol{\theta}$ of a model $\mathbb{M}$ are to be estimated by maximum likelihood, one defines a loglikelihood function as follows:

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

where the contribution $\ell_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ is the log of the density of observation y_t conditional on observations $y_1, \ldots, y_{t-1}$ for the DGP characterised by $\boldsymbol{\theta}$. The notation $\boldsymbol{y}^t$ signifies the vector of observations $y_1, \ldots, y_t$.

In regular cases, the estimating equations for ML estimation are the likelihood equations

$$\frac{\partial \ell(\boldsymbol{y}, \boldsymbol{\theta})}{\partial \theta_i} = \sum_{t=1}^{n} \frac{\partial \ell_t(\boldsymbol{y}^t, \boldsymbol{\theta})}{\partial \theta_i} = 0, \quad i = 1, \ldots, k.$$

From this, it is clear that the elementary zero functions implicitly used by ML are the derivatives $\partial \ell_t / \partial \theta_i$ of the loglikelihood contributions, k of them per observation. These functions are uncorrelated for different observation indices s and t, but are contemporaneously correlated within each observation. In addition, they are heteroskedastic. On the other hand, the instruments are just vectors with each element equal to 1. In this section, we will see that, for a certain class of models, these same likelihood equations can be constructed along the lines of the last section with homoskedastic and uncorrelated elementary zero functions, combined with optimal instruments defined as in (7).

Suppose that the (scalar) observations y_t, $t = 1, \ldots, n$, are jointly normally distributed with expectations $x_t(\boldsymbol{\theta})$ and $n \times n$ covariance matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$. The regression functions $x_t(\boldsymbol{\theta})$ may depend on exogenous explanatory variables and lagged dependent variables. Then the loglikelihood function for the sample of n observations can be written as

$$\ell(\boldsymbol{y}, \boldsymbol{\theta}) \equiv -\frac{n}{2} \log 2\pi - \frac{1}{2} \log \det \boldsymbol{\Omega}(\boldsymbol{\theta}) - \frac{1}{2} \boldsymbol{u}^\top(\boldsymbol{\theta}) \boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}) \boldsymbol{u}(\boldsymbol{\theta}), \tag{8}$$

where $\boldsymbol{u}(\boldsymbol{\theta})$ is the n–vector with typical element $y_t - x_t(\boldsymbol{\theta})$. Let the lower-triangular matrix $\boldsymbol{A}(\boldsymbol{\theta})$ be such that $\boldsymbol{A}^\top(\boldsymbol{\theta}) \boldsymbol{A}(\boldsymbol{\theta}) = \boldsymbol{\Omega}^{-1}(\boldsymbol{\theta})$. Then (8) can be rewritten as the sum of contributions $\ell_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ defined as

$$\ell_t(\boldsymbol{y}^t, \boldsymbol{\theta}) = -\frac{1}{2} \log 2\pi + \log a_{tt}(\boldsymbol{\theta}) - \frac{1}{2} \Big(\sum_{s=1}^{t} a_{ts}(\boldsymbol{\theta}) u_s(\boldsymbol{\theta}) \Big)^2, \tag{9}$$

where $a_{ts}(\boldsymbol{\theta})$ is element (t, s) of $\boldsymbol{A}(\boldsymbol{\theta})$.

The likelihood equations for estimating $\boldsymbol{\theta}$ are the first-order conditions for the maximisation of (8) with respect to $\boldsymbol{\theta}$, and they can be written in terms of the derivatives of the contributions (9):

$$\frac{\partial \ell_t}{\partial \theta_i} = \sum_{s=1}^{t} \Big(\frac{\partial \ell_t}{\partial u_s} \frac{\partial u_s}{\partial \theta_i} + \frac{\partial \ell_t}{\partial a_{ts}} \frac{\partial a_{ts}}{\partial \theta_i} \Big), \quad i = 1, \ldots, k. \tag{10}$$

The derivatives $\partial u_s/\partial\theta_i$ and $\partial a_{ts}/\partial\theta_i$ are model specific, but the derivatives of the ℓ_t can always be written as

$$\frac{\partial \ell_t}{\partial u_s} = -a_{ts}v_t, \quad s \le t, \text{ and}$$
$$\frac{\partial \ell_t}{\partial a_{ts}} = \frac{\delta_{ts}}{a_{tt}} - u_s v_t,$$

where δ_{ts} is the Kronecker delta and

$$v_t(\boldsymbol{\theta}) = \sum_{s=1}^{t} a_{ts}(\boldsymbol{\theta}) u_s(\boldsymbol{\theta}), \tag{11}$$

or $\boldsymbol{v}(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{\theta})$ in vector notation.

Note that the $v_t(\boldsymbol{\theta})$ form a set of homoskedastic, uncorrelated, elementary zero functions. The normal distribution is completely determined by its first two moments, and so the obvious set of elementary zero functions to use for estimation in the present case is constituted by the functions $v_t(\boldsymbol{\theta})$ and $(v_t^2(\boldsymbol{\theta}) - 1)/\sqrt{2}$, for $t = 1, \ldots, n$. It is easy to check that these functions are homoskedastic and uncorrelated, and, since the y_t are assumed to be normal, it follows that, evaluated at the true $\boldsymbol{\theta}$, v_t is independent of v_s for $s \neq t$.

The optimal instrument to use in conjunction with v_t for θ_i is the expectation of

$$\frac{\partial v_t}{\partial \theta_i} = \sum_{s=1}^{t} \Big(a_{ts} \frac{\partial u_s}{\partial \theta_i} + u_s \frac{\partial a_{ts}}{\partial \theta_i} \Big) \tag{12}$$

conditional on the information set Ω_t, which, in addition to all exogenous variables, may be assumed to contain the v_s for $s < t$. In order to be able to deal with ARCH phenomena, we wish to allow the a_{ts} to be random, but we assume that $a_{ts} \in \Omega_t$, so that any random variables on which a_{ts} may depend must be predetermined at t. Because the v_t are independent, we have $\mathrm{E}(v_t \,|\, \Omega_t) = 0$. Further, $\partial u_t/\partial\theta_i \in \Omega_t$, since this derivative is $-\partial x_t/\partial\theta_i$.

We have $\boldsymbol{u} = \boldsymbol{A}^{-1}\boldsymbol{v}$, where $\boldsymbol{A}^{-1}$ is lower triangular because $\boldsymbol{A}$ is. Write this relation as $u_t = \sum_{s=1}^{t} a^{ts} v_s$. It follows that $\mathrm{E}(u_s \,|\, \Omega_t) = u_s$ for $s < t$, and $\mathrm{E}(u_t \,|\, \Omega_t) = \sum_{s=1}^{t-1} a^{ts} v_s = u_t - a^{tt} v_t = u_t - v_t/a_{tt}$, since $a^{tt} = 1/a_{tt}$ by the lower triangularity property. The conditional expectation of (12) becomes

$$\mathrm{E}\Big(\frac{\partial v_t}{\partial \theta_i} \;\Big|\; \Omega_t \Big) = \sum_{s=1}^{t} a_{ts} \frac{\partial u_s}{\partial \theta_i} + \sum_{s=1}^{t} u_s \frac{\partial a_{ts}}{\partial \theta_i} - \frac{1}{a_{tt}} \frac{\partial a_{tt}}{\partial \theta_i} v_t. \tag{13}$$

The optimal instrument for $(v_t^2 - 1)/\sqrt{2}$ is

$$\frac{1}{\sqrt{2}} \mathrm{E}\Big(\frac{\partial v_t^2}{\partial \theta_i} \;\Big|\; \Omega_t \Big) = \sqrt{2}\, \mathrm{E}\Big(v_t \frac{\partial v_t}{\partial \theta_i} \;\Big|\; \Omega_t \Big). \tag{14}$$

The only term of (12) that is not in Ω_t is $u_t \partial a_{tt}/\partial\theta_i$, and so this is the only term to contribute to (14), which, on noting that $\mathrm{E}(v_t u_t) = a^{tt}$, can be seen to be equal to

$$\sqrt{2}\,\frac{\partial a_{tt}}{\partial \theta_i}\mathrm{E}(v_t u_t) = \sqrt{2}\,\frac{1}{a_{tt}}\frac{\partial a_{tt}}{\partial \theta_i}. \tag{15}$$

Adding v_t times (13) and $(v_t^2 - 1)/\sqrt{2}$ times (15) gives the contribution from observation t to the estimating function for θ_i. This contribution is

$$\begin{aligned}
&\sum_{s=1}^{t} a_{ts} v_t \frac{\partial u_s}{\partial \theta_i} + \sum_{s=1}^{t} v_t u_s \frac{\partial a_{ts}}{\partial \theta_i} - \frac{1}{a_{tt}}\frac{\partial a_{tt}}{\partial \theta_i} \\
&= \sum_{s=1}^{t} a_{ts} v_t \frac{\partial u_s}{\partial \theta_i} + \sum_{s=1}^{t} \Big(v_t u_s - \frac{\delta_{ts}}{a_{tt}} \Big) \frac{\partial a_{ts}}{\partial \theta_i},
\end{aligned}$$

which is exactly the negative of (10). We have proved the following theorem.

Theorem 1

Consider the model defined by the set of loglikelihood contributions (9), where $u_t(\boldsymbol{\theta}) = y_t - x_t(\boldsymbol{\theta})$ and $x_t(\boldsymbol{\theta})$ and the $a_{ts}(\boldsymbol{\theta})$, $s \leq t$, are in the information set defined by the y_s, $s = 1, \ldots, t-1$, and the exogenous explanatory variables. The likelihood equations for this model are identical to the estimating equations based on the set of homoskedastic and uncorrelated elementary zero functions $v_t(\boldsymbol{\theta})$ and $(v_t(\boldsymbol{\theta}) - 1)^2/\sqrt{2}$, where $v_t(\boldsymbol{\theta})$ is defined by (11), in conjunction with the optimal instruments for these zero functions, as given by (13) and (15). ■

4. An Artificial Regression

The result of Theorem 1 may seem rather academic, but it can be put to excellent practical use by means of an artificial regression. This artificial regression can lead to much simpler ML estimation than by the usual maximisation of the loglikelihood function, and can furnish the efficient score estimate of the information matrix, and thus an efficient estimate of the asymptotic covariance matrix of the ML estimator.

An artificial regression is a linear regression, for which the regressand and the regressors are functions of both data and parameters. Such a regression, which can be written as

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{b} + \text{residuals}, \tag{16}$$

corresponds to a parametrised model $\mathbb{M}$ and to an estimator $\hat{\boldsymbol{\theta}}$ of the parameters of the model if the following three conditions are satisfied.

- The artificial regressand and the artificial regressors are orthogonal when evaluated at $\hat{\boldsymbol{\theta}}$, that is,

$$\boldsymbol{R}^{\top}(\hat{\boldsymbol{\theta}})\boldsymbol{r}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}. \tag{17}$$

  Equations (17) are therefore estimating equations for $\hat{\boldsymbol{\theta}}$.

- Under any DGP $\mu \in \mathbb{M}$, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by

$$\mathrm{Var}\big(\operatorname*{plim}_{n\to\infty,\mu} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu)\big) = \operatorname*{plim}_{n\to\infty,\mu}\big(n^{-1}\boldsymbol{R}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{R}(\hat{\boldsymbol{\theta}})\big)^{-1}, \tag{18}$$

  where $\boldsymbol{\theta}_\mu$ is the true parameter vector for the DGP μ and n is the sample size.

- The artificial regression allows for one-step estimation, in the sense that, if $\acute{\boldsymbol{\theta}}$ is any root-n consistent estimator of the model parameters, and $\acute{\boldsymbol{b}}$ denotes the vector of OLS parameter estimates obtained by running the artificial regression with $\boldsymbol{\theta} = \acute{\boldsymbol{\theta}}$, then, under any DGP $\mu \in \mathbb{M}$,

$$\acute{\boldsymbol{\theta}} + \acute{\boldsymbol{b}} = \hat{\boldsymbol{\theta}} + O_p(n^{-1}). \tag{19}$$

An artificial regression can be used as the basis for an algorithm of nonlinear estimation, for the estimation of covariance matrices, and for hypothesis testing. Specifically, the estimator $\hat{\boldsymbol{\theta}}$ can be computed by an algorithm that uses the artificial regression as the Gauss-Newton regression is used in the computation of nonlinear least squares estimates. A starting value is chosen for $\boldsymbol{\theta}$, the artificial variables are evaluated at this starting value, and $\boldsymbol{\theta}$ is updated by adding to the starting value the vector of OLS parameter estimates from the artificial regression[1]. It is clear that, if an iterative procedure based on this updating step converges, the value of $\boldsymbol{\theta}$ at convergence satisfies the estimating equations (17).

The sort of artificial regression that interests us here has a regressand which is a vector of homoskedastic, serially uncorrelated, zero functions, and regressors that are the negatives of the optimal instruments for those zero functions and the model parameters. For such artificial regressions, we can prove the following Lemma, which is a slight specialisation of a result given in Davidson and MacKinnon (2001).

Lemma 1

Let an artificial regression of the form (16) be constructed with regressand $\boldsymbol{r}(\boldsymbol{\theta})$, with typical element $r_t(\boldsymbol{\theta})$, such that $\mathrm{E}_\mu(r_t(\boldsymbol{\theta}_\mu)) = 0$ and $\mathrm{E}_\mu(r_t(\boldsymbol{\theta}_\mu)r_s(\boldsymbol{\theta}_\mu)) = \delta_{ts}$ for all $\mu \in \mathbb{M}$. The regressor corresponding to the zero function r_t and the parameter θ_i is the negative of the optimal instrument

$$R_{ti}(\boldsymbol{\theta}) \equiv -\mathrm{E}\Big(\frac{\partial r_t}{\partial \theta_i}(\boldsymbol{\theta}) \;\Big|\; \Omega_t\Big),$$

where the expectation is computed under a DGP with associated parameter vector $\boldsymbol{\theta}$. Let the estimator $\hat{\boldsymbol{\theta}}$ be defined by the estimating equations $\boldsymbol{R}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{r}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$. Then such an artificial regression satisfies the three conditions (17), (18), and (19) for the estimator $\hat{\boldsymbol{\theta}}$.

Proof: In Appendix. ■

[1] In practice, the OLS parameter estimates should be used to give the direction of the update in the parameter space, but not necessarily the magnitude, which can be found by a one-dimensional optimisation algorithm.

For the models considered in the previous section and the ML estimator, we construct an artificial regression with $N = 2n$ artificial observations, one for each of the elementary zero functions in the statement of Theorem 1. For each real observation there are two artificial observations, for which the elements of the regressand are $v_t(\boldsymbol{\theta})$ and $(v_t^2(\boldsymbol{\theta}) - 1)/\sqrt{2}$, where $v_t(\boldsymbol{\theta})$ is defined in terms of data and parameters by (11). There are k regressors, one for each parameter to be estimated. For observation t and parameter θ_i, the two elements of the regressor are the negatives of the optimal instruments (13) and (15), respectively. We may write the regressors as

$$\begin{aligned} R_{ti}^{(1)} &= -\sum_{s=1}^{t} a_{ts} \frac{\partial u_s}{\partial \theta_i} - \sum_{s=1}^{t} u_s \frac{\partial a_{ts}}{\partial \theta_i} + \frac{1}{a_{tt}} \frac{\partial a_{tt}}{\partial \theta_i} v_t, \text{ and} \\ R_{ti}^{(2)} &= -\sqrt{2}\, \frac{1}{a_{tt}} \frac{\partial a_{tt}}{\partial \theta_i}. \end{aligned} \tag{20}$$

It is then convenient to represent the artificial regression schematically as follows:

$$\begin{bmatrix} v_t \\ (v_t^2 - 1)/\sqrt{2} \end{bmatrix} = \sum_{i=1}^{k} \begin{bmatrix} R_{ti}^{(1)} \\ R_{ti}^{(2)} \end{bmatrix} b_i + \text{residuals}. \tag{21}$$

Here it is understood that everything depends on the parameter vector $\boldsymbol{\theta}$, except for the artificial parameters b_i.

By Lemma 1, this artificial regression satisfies the three required conditions, and by Theorem 1 the estimator that corresponds to it is the ML estimator $\hat{\boldsymbol{\theta}}$. Regarding condition (18), it turns out that the covariance matrix estimator $\boldsymbol{R}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{R}(\hat{\boldsymbol{\theta}})$ provided by the artificial regression (21) is the inverse of the information matrix evaluated at $\hat{\boldsymbol{\theta}}$. This property implies that, in the neighbourhood of $\hat{\boldsymbol{\theta}}$, the matrix $\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\theta})$ is asymptotically equal to the negative of the Hessian of the loglikelihood function.

The particular properties of the artificial regression (21) are collected in the following theorem.

Theorem 2

> The artificial regression (21) corresponds to the model defined by the set of loglikelihood contributions (9) and to the maximum likelihood estimator of that model. In addition, the matrix of cross-products of the regressors evaluated at $\hat{\boldsymbol{\theta}}$ is the efficient score estimator of the information matrix.

Proof: In Appendix. ■

In order to implement the artificial regression (21), we need to be able to evaluate the elements $a_{ts}(\boldsymbol{\theta})$, for $t = 1, \ldots, n$ and $s = 1, \ldots, t$, and their derivatives with respect to the model parameters. We assume that the matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is available to us as a function of $\boldsymbol{\theta}$ in analytic form. Then computing the matrix $\boldsymbol{A}(\boldsymbol{\theta})$ for any given $\boldsymbol{\theta}$ is just a question of matrix inversion and the Cholesky decomposition. For the derivatives, notice that, for $i = 1, \ldots, k$,

$$\frac{\partial}{\partial \theta_i}(\boldsymbol{A}^\top \boldsymbol{A}) = \frac{\partial \boldsymbol{A}^\top}{\partial \theta_i} \boldsymbol{A} + \boldsymbol{A}^\top \frac{\partial \boldsymbol{A}}{\partial \theta_i} = \frac{\partial \boldsymbol{\Omega}^{-1}}{\partial \theta_i} = -\boldsymbol{\Omega}^{-1} \frac{\partial \boldsymbol{\Omega}}{\partial \theta_i} \boldsymbol{\Omega}^{-1}.$$

From this, it follows that

$$(\boldsymbol{A}^\top)^{-1}\frac{\partial \boldsymbol{A}^\top}{\partial \theta_i} + \frac{\partial \boldsymbol{A}}{\partial \theta_i}\boldsymbol{A}^{-1} = -\boldsymbol{A}\frac{\partial \boldsymbol{\Omega}}{\partial \theta_i}\boldsymbol{A}^\top. \tag{22}$$

If $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is available in analytic form, then $\partial\boldsymbol{\Omega}/\partial\theta_i$ can be found analytically as well. It follows that the right-hand side of (22) can be computed once $\boldsymbol{A}$ has been computed. The second term on the left-hand side is lower triangular, and the first term is its transpose. The second term can thus be constructed with its principal diagonal equal to half that of the right-hand side, and with the triangle below the principal diagonal equal to that of the right-hand side. Finally, $\partial\boldsymbol{A}/\partial\theta$ is the result of this computation times $\boldsymbol{A}$. Thus analytic differentiation of $\boldsymbol{\Omega}$ with respect to the parameters, along with some standard matrix manipulations, is enough to implement (21).

5. Regression Models with ARMA errors

It is usually not very hard to write down the covariance matrix $\boldsymbol{\Omega}$ of an ARMA(p, q) process as an analytic function of the autoregressive and moving average parameters. It is on the other hand not at all simple to perform maximum likelihood estimation of regression models with ARMA errors without conditioning on the early observations in the sample, and, when there is an MA component, making some assumptions about unobserved error terms. The artificial regression described in the previous section makes it feasible to dispense with conditioning and arbitrary assumptions, and to do full-information ML estimation. In addition to formulating the (stationary) ARMA covariance matrix, the only other analytical effort required is differentiating this matrix with respect to the parameters on which it depends.

Our first illustration is the estimation of a model with MA(1) errors. Even for this simple case, full-information ML estimation is usually a complicated business; see for instance Hamilton (1994), Chapter 5. A general nonlinear regression model can be written as

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \tag{23}$$

where $\boldsymbol{x}(\boldsymbol{\beta})$ denotes an n–vector of regression functions that depend on exogenous or predetermined explanatory variables, and, possibly nonlinearly, on a k–vector $\boldsymbol{\beta}$ of parameters. The MA(1) error process can be written as

$$u_t = \varepsilon_t - \alpha\varepsilon_{t-1}, \quad \varepsilon_t \sim \text{NID}(0, \sigma^2).$$

With this sort of model, it turns out that it is not necessary to treat the variance parameter σ^2 as an element of the overall parameter vector $\boldsymbol{\theta}$, since its ML estimate can be computed explicitly as a function of observed data and the other parameters. It would of course be possible to treat σ^2 in the same way as $\boldsymbol{\beta}$ and α, but it would be a little more complicated for no gain.

It is well known that the covariance matrix of the error terms u_t is σ^2 times a matrix with all diagonal elements equal to $1+\alpha^2$, the elements on the first super- and sub-diagonal equal to $-\alpha$, and all other elements 0:

$$\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} 1+\alpha^2 & -\alpha & 0 & \dots & 0 \\ -\alpha & 1+\alpha^2 & -\alpha & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1+\alpha^2 \end{bmatrix}. \tag{24}$$

The only derivative of this matrix we need is that with respect to α. It is clear that this derivative is σ^2 times a matrix that has 2α on the principal diagonal, -1 on the two adjacent diagonals, and zero elsewhere.

The steps required to set up the artificial regression for given values of $\boldsymbol{\beta}$, α, and σ^2 are then as follows:

- Form the matrix $\boldsymbol{\Omega}$ for the given α and σ^2 according to (24), and compute $\boldsymbol{A}$ by the Cholesky decomposition so that $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{\Omega}^{-1}$.
- Form the matrix $\partial\boldsymbol{\Omega}/\partial\alpha$ as described above, and then form the matrix $-\boldsymbol{A}(\partial\boldsymbol{\Omega}/\partial\alpha)\boldsymbol{A}^\top$, which is the right-hand side of (22) for $\theta_i = \alpha$.
- Compute $\partial\boldsymbol{A}/\partial\alpha$ as described at the end of the previous section.
- Form the vector of residuals $u_t(\boldsymbol{\beta}) \equiv y_t - x_t(\boldsymbol{\beta})$ for the given $\boldsymbol{\beta}$, and use them to construct the zero functions v_t by the relation $\boldsymbol{v} = \boldsymbol{A}\boldsymbol{u}$. Form the regressand by stacking the vector of the v_t on top of the vector of the second-moment zero functions $(v_t^2 - 1)/\sqrt{2}$.
- Form the regressors corresponding to the parameters $\boldsymbol{\beta}$. The upper block is the matrix $\boldsymbol{A}\boldsymbol{X}(\boldsymbol{\beta})$, where $\boldsymbol{X}(\boldsymbol{\beta})$ is the Jacobian matrix of the regression functions. The lower block is a zero matrix.
- The regressor corresponding to α is made up of the vector

$$-\frac{\partial\boldsymbol{A}}{\partial\alpha}\boldsymbol{u} + \boldsymbol{w}, \quad \text{where} \quad w_t = \frac{v_t}{a_{tt}}\left(\frac{\partial\boldsymbol{A}}{\partial\alpha}\right)_{tt},$$

  stacked on top of the vector with typical element $-\sqrt{2}a_{tt}(\partial\boldsymbol{A}/\partial\alpha)_{tt}$.

After setting up and running the artificial regression (by OLS), the parameters $\boldsymbol{\beta}$ and α are updated by adding to them the corresponding artificial parameter estimates, while σ^2 is updated by the formula

$$\sigma^2 = \frac{1}{n(1+\alpha^2)}\sum_{t=1}^{n}\big(y_t - x_t(\boldsymbol{\beta})\big)^2.$$

The artificial regression can then be set up again for the new parameter values, and the whole procedure iterated until convergence. For models with MA(q) errors, the procedure is completely analogous, but more complicated on account of there being q MA parameters.

Next we consider models with error terms that follow an AR(p) process, with no MA component, since the general artificial regression (21) can be considerably simplified in this case. As above, we illustrate the simplest case, with a vector $\boldsymbol{u}$ of AR(1) errors defined by the autoregression $u_t = \rho u_{t-1} + \sigma v_t$, where v_t is white noise. The model (23) with AR(1) errors can be estimated in many ways. One traditional method is feasible generalised least squares (GLS), but this method requires that the explanatory variables should be exogenous; lagged dependent variables are not allowed, on pain of inconsistent estimates. A more flexible procedure is to transform the model so that the error term is σv_t rather than u_t, yielding

$$y_t = x_t(\boldsymbol{\beta}) + \rho y_{t-1} - \rho x_{t-1}(\boldsymbol{\beta}) + \sigma v_t. \tag{25}$$

This model can then be estimated by nonlinear least squares, if one is prepared to drop the first observation, for which the lags are not observed. The Gauss-Newton regression (GNR) that corresponds to (25) can be written as

$$\begin{aligned} y_t - x_t(\boldsymbol{\beta}) - \rho y_{t-1} + \rho x_{t-1}(\boldsymbol{\beta}) = \\ (\boldsymbol{X}_t(\boldsymbol{\beta}) - \rho \boldsymbol{X}_{t-1}(\boldsymbol{\beta}))\boldsymbol{b}_{\boldsymbol{\beta}} + (y_{t-1} - x_{t-1}(\boldsymbol{\beta}))b_\rho + \text{residual}, \end{aligned} \tag{26}$$

where $\boldsymbol{X}(\boldsymbol{\beta})$ is the $n \times k$ Jacobian matrix of the regression functions contained in $\boldsymbol{x}(\boldsymbol{\beta})$, and $\boldsymbol{b}_{\boldsymbol{\beta}}$ and b_ρ are artificial parameters that correspond to $\boldsymbol{\beta}$ and ρ respectively.

The GNR (26) can be interpreted as an artificial regression for which the elementary zero functions are the residuals $y_t - x_t(\boldsymbol{\beta}) - \rho y_{t-1} + \rho x_{t-1}(\boldsymbol{\beta})$ that constitute the regressand, and the instruments are the derivatives of these zero functions with respect to $\boldsymbol{\beta}$ and ρ. No second moment information is used, and no information from the first observation. The second of these flaws may be corrected by appending an extra artificial observation to (26), which can be written schematically as

$$(1-\rho^2)^{1/2}\big(y_1 - x_1(\boldsymbol{\beta})\big) = [\,(1-\rho^2)^{1/2}\boldsymbol{X}_1(\boldsymbol{\beta}) \quad 0\,] \begin{bmatrix} \boldsymbol{b}_{\boldsymbol{\beta}} \\ b_\rho \end{bmatrix} + \text{residual}. \tag{27}$$

The scaling by $(1-\rho^2)^{1/2}$ serves to make the variance of the regressand the same as that of the other elements of the regressand, under the assumption that the AR(1) process is stationary. This is because the stationary variance of the u_t process is $\sigma^2/(1-\rho^2)$. The regressors are minus the unconditional expectations of the derivatives of the regressand, and are therefore optimal instruments. Note that, since the derivative with respect to ρ is proportional to the error term, its expectation is zero.

We now examine what information, if any, can be extracted from second moment information. Obviously, such information must be used in order to estimate σ^2. An estimator that uses the information in all the observations is

$$\hat{\sigma}^2 = \frac{1}{n}\Big((1-\hat{\rho}^2)\big(y_1 - x_1(\hat{\boldsymbol{\beta}})\big)^2 + \sum_{t=2}^{n}\big(y_t - x_t(\hat{\boldsymbol{\beta}}) - \hat{\rho}y_{t-1} + \hat{\rho}x_{t-1}(\hat{\boldsymbol{\beta}})\big)^2\Big). \tag{28}$$

Consider the second-moment zero functions $(y_t - x_t(\boldsymbol{\beta}) - \rho y_{t-1} + \rho x_{t-1}(\boldsymbol{\beta}))^2 - \sigma^2$ for $t = 2, \ldots, n$. The derivative with respect to σ^2 is 1, which implies that these functions should be weighted equally in the estimating equation for σ^2, as is the case in (28). The derivatives with respect to $\boldsymbol{\beta}$ and ρ are all proportional to the error terms, and so have expectations of 0. These zero functions are therefore uninformative about $\boldsymbol{\beta}$ and ρ, and it is appropriate to make no use of them in the estimation of these parameters.

The second-moment zero function is $(1-\rho^2)(y_1 - x_1(\boldsymbol{\beta}))^2 - \sigma^2$ for the first observation. The variance of this function is $2\sigma^4$. The derivatives with respect to the elements of $\boldsymbol{\beta}$ still have expectations of 0, but minus the derivative with respect to ρ is $2\rho(y_1 - x_1(\boldsymbol{\beta}))^2$, with expectation $2\rho\sigma^2/(1-\rho^2)$. When we scale the zero function by a factor of $1/(\sigma\sqrt{2})$, the variance of the rescaled function is σ^2, the same as the variance of the first-moment zero functions. Thus if we append another artificial observation to the GNR (26), along with (27), as follows:

$$\frac{(1-\rho^2)\big(y_1 - x_1(\boldsymbol{\beta})\big)^2 - \sigma^2}{\sigma\sqrt{2}} = \begin{bmatrix} \mathbf{0} & \dfrac{\rho\sigma\sqrt{2}}{(1-\rho)^2} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_{\boldsymbol{\beta}} \\ b_\rho \end{bmatrix} + \text{residual},$$

then we take account of the information about ρ in the second-moment zero function for the first observation.

In order to obtain ML estimates from this augmented GNR, at each iteration we evaluate the artificial observations at the current values of $\boldsymbol{\beta}$ and ρ, and run the regression. The artificial parameter estimates are used to update $\boldsymbol{\beta}$ and ρ, and σ^2 is updated using the formula (28). Although this procedure is not based directly on the artificial regression (21), it is easy to show that, since (28) is the likelihood equation for σ^2, the estimates of all the parameters, $\boldsymbol{\beta}$, ρ, and σ^2, are ML estimates when convergence is achieved.

A similarly augmented GNR can be developed to deal with the case of AR(p) errors. In that case, the error terms u_t in the regression model obey the recursion

$$u_t = \sum_{i=1}^{p} \rho_i u_{t-i} + v_t, \quad v_t \text{ white noise.}$$

The information in the observations after the p^{th} can be extracted by dropping the first p observations and then estimating the model

$$y_t = x_t(\boldsymbol{\beta}) + \sum_{i=1}^{p} \rho_i \big(y_{t-i} - x_{t-i}(\boldsymbol{\beta})\big) + v_t$$

by nonlinear least squares. It is easy enough to see that the second moments of these observations are uninformative about $\boldsymbol{\beta}$ and ρ. For the first p observations, it is necessary, both for the loglikelihood function and for an augmented GNR, to formulate the unconditional covariance matrix $\boldsymbol{\Omega}_p$ of the first p error terms. Under the hypothesis of stationarity, this is the covariance matrix of p consecutive elements of the stationary distribution. One way to find this covariance matrix is

to solve the Yule-Walker equations for an AR(p) process. These equations can be written as

$$s_0 - \sum_{i=1}^{p} \rho_i s_i = 1, \quad \text{and}$$

$$\rho_i s_0 - s_i + \sum_{j=1, j\neq i}^{p} \rho_j s_{|i-j|} = 0, \quad i = 1, \ldots, p, \tag{29}$$

where s_0 times σ^2 is the stationary variance, and s_i times σ^2 the stationary covariance of elements of the u_t process separated by i periods. Equations (29) are linear with respect to the s_i, $i = 0, 1, \ldots, p$, and so can be solved without difficulty for the s_i in terms of the ρ_i. The matrix $\boldsymbol{\Omega}_p$ is then given by

$$\boldsymbol{\Omega}_p = \sigma^2 \begin{bmatrix} s_0 & s_1 & \cdots & s_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p-1} & s_{p-2} & \cdots & s_0 \end{bmatrix}. \tag{30}$$

The derivative of $\boldsymbol{\Omega}_p$ with respect to ρ_k, $k = 1, \ldots, p$, can be computed by solving the Yule-Walker equations differentiated with respect to ρ_k. These equations can be written as

$$\frac{\partial s_0}{\partial \rho_k} - \sum_{i=1}^{p} \rho_i \frac{\partial s_i}{\partial \rho_k} = s_k, \quad \text{and}$$

$$\rho_i \frac{\partial s_0}{\partial \rho_k} - \frac{\partial s_i}{\partial \rho_k} + \sum_{j=1, j\neq i}^{p} \rho_j \frac{\partial s_{|i-j|}}{\partial \rho_k} = -s_{|i-k|}, \quad i = 1, \ldots, p.$$

The GNR must now be augmented by $2p$ extra artificial observations for the first- and second-moment zero functions for the first p observations. For p sufficiently small, this can be done analytically, but in all cases the numerical procedure described above allows us to evaluate the elements of the regressand and regressors for these artificial observations. The procedure is likely to take less than the usual time to compute, because the matrices are p–dimensional rather than n–dimensional. The $2p$ artificial observations thus constructed are appended to the ordinary GNR for the observations after the first p, and, when the iterations based on the GNR converge, we have obtained the ML estimates.

The amount of work needed to set up the artificial regressions described above is similar to what is needed to set up the loglikelihood function taking full account of all observations. One may argue that, once the loglikelihood is available, estimation can safely be left to the computer. This may or may not be true. In the past, it was thought worthwhile to devote substantial effort in order to develop numerical methods for specific models that would have better properties than general algorithms for optimising functions. For the case of a linear model with AR(1) errors, the classical example is the paper of Beach and MacKinnon (1978a); see also Beach and MacKinnon (1978b) for the case of AR(2) errors. The artificial regression for these cases is distinctly simpler to program than the Beach-MacKinnon procedures. In addition, estimation by artificial regression has associated advantages – automatic computation of a covariance matrix estimate, ease of performing hypothesis tests and specification tests, *etc*.

6. ARCH and GARCH

Models with (G)ARCH errors are notoriously difficult to estimate accurately. Although most econometrics software packages claim to perform ML estimation of regression models with (G)ARCH errors, they cannot all be doing so correctly, as they can give very different results with the same data; see Brooks, Burke, and Persand (2001) for evidence on this matter. ML estimation *should* be relatively trouble free given the work of Fiorentini, Calzolari, and Panattoni (1996). In their paper, analytical expressions are given for the first and second derivatives of the loglikelihood functions for models of this type. Thus it is possible to use Newton's method with analytic second derivatives, and this method is known to converge faster than all others in the neighbourhood of the ML estimates.

Other than making Newton's method feasible, second derivatives are needed in order to form the Hessian estimate of the information matrix. However, artificial regressions have their advantages. Those we consider in this paper do not need second derivatives, and provide the efficient score information matrix estimator. A few simulations show that they converge quite reliably, although they do not have the built-in safeguard of ML estimation of models with ARMA errors, whereby one cannot leave the stationarity region without crossing a singularity. A technique that seems to work well is to watch for iterations that lead to parameter values outside the stationarity region, and, when they occur, take a shorter step that remains inside.

All models of this class have the form

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \mathrm{N}\big(0, \sigma_t^2(\boldsymbol{\beta}, \boldsymbol{\phi})\big), \tag{31}$$

with a possibly nonlinear regression function $x_t(\boldsymbol{\beta})$ and skedastic function $\sigma_t^2(\boldsymbol{\beta}, \boldsymbol{\phi})$ which depends on the parameters $\boldsymbol{\beta}$ of the regression function through lagged squared residuals and on a set of GARCH parameters $\boldsymbol{\phi}$. Both $x_t(\boldsymbol{\beta})$ and $\sigma_t^2(\boldsymbol{\beta}, \boldsymbol{\phi})$ belong to the information set Ω_t. The simplest case is when the errors follow an ARCH(1) process, for which

$$\sigma_t^2 = \alpha + \gamma u_{t-1}^2(\boldsymbol{\beta}), \quad \text{where} \quad u_t(\boldsymbol{\beta}) \equiv y_t - x_t(\boldsymbol{\beta}), \tag{32}$$

and the vector $\boldsymbol{\phi}$ has the two components α and γ. Note that u_{t-1} is predetermined at time t.

For any model of the form (31), the matrix $\boldsymbol{A}$ is just $\mathrm{diag}\{1/\sigma_t\}$, a diagonal matrix, since (G)ARCH models introduce conditional heteroskedasticity, but not autocorrelation. The elementary zero functions $v_t(\boldsymbol{\beta}, \boldsymbol{\phi})$ are then defined as $u_t(\boldsymbol{\beta})/\sigma_t(\boldsymbol{\beta}, \boldsymbol{\phi})$. From (20), it follows that the regressors corresponding to these zero functions take the form

$$\begin{aligned} R_{ti}^{(1)} &= -\frac{1}{\sigma_t(\boldsymbol{\beta}, \boldsymbol{\phi})} \frac{\partial u_t}{\partial \theta_i}(\boldsymbol{\beta}) - u_t(\boldsymbol{\beta}) \frac{\partial \sigma_t}{\partial \theta_i}(\boldsymbol{\beta}, \boldsymbol{\phi}) + \sigma_t(\boldsymbol{\beta}, \boldsymbol{\phi}) \frac{\partial \sigma_t}{\partial \theta_i}(\boldsymbol{\beta}, \phi) \frac{u_t(\boldsymbol{\beta})}{\sigma_t(\boldsymbol{\beta}, \boldsymbol{\phi})} \\ &= -\frac{1}{\sigma_t(\boldsymbol{\beta}, \boldsymbol{\phi})} \frac{\partial u_t}{\partial \theta_i}(\boldsymbol{\beta}), \end{aligned} \tag{33}$$

where θ_i is any component of either $\boldsymbol{\beta}$ or $\boldsymbol{\phi}$. For components of $\boldsymbol{\phi}$, the regressor (33) is zero, since the residuals do not depend on $\boldsymbol{\phi}$. For components of $\boldsymbol{\beta}$, (33) becomes $\boldsymbol{X}_t(\boldsymbol{\beta})/\sigma_t(\boldsymbol{\beta}, \boldsymbol{\phi})$.

For the zero functions $(v_t^2 - 1)/\sqrt{2}$, the regressors are, again from (20),

$$R_{ti}^{(2)} = \frac{\sqrt{2}}{\sigma_t(\boldsymbol{\beta}, \boldsymbol{\phi})} \frac{\partial \sigma_t}{\partial \theta_i}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \frac{1}{\sigma_t^2(\boldsymbol{\beta}, \boldsymbol{\phi})\sqrt{2}} \frac{\partial \sigma_t^2}{\partial \theta_i}(\boldsymbol{\beta}, \boldsymbol{\phi}). \tag{34}$$

A little calculation using (33) and (34) shows that the regressors for a model with ARCH(1) errors are

$$\begin{aligned}
R_{t\boldsymbol{\beta}}^{(1)} &= \frac{1}{\sigma_t} \boldsymbol{X}_t(\boldsymbol{\beta}) & R_{t\alpha}^{(1)} &= 0 & R_{t\gamma}^{(1)} &= 0 \\
R_{t\boldsymbol{\beta}}^{(2)} &= -\frac{\gamma\sqrt{2}}{\sigma_t^2} \boldsymbol{X}_{t-1}(\boldsymbol{\beta}) u_{t-1}(\boldsymbol{\beta}) & R_{t\alpha}^{(2)} &= \frac{1}{\sigma_t^2\sqrt{2}} & R_{t\gamma}^{(2)} &= \frac{1}{\sigma_t^2\sqrt{2}}\, u_{t-1}^2(\boldsymbol{\beta})
\end{aligned}$$

The artificial regression defined above should be run over observations 2 to n. In order to use the information in the first observation, we can form the zero function

$$v_1(\alpha, \boldsymbol{\beta}, \gamma) \equiv \frac{(1-\gamma)^{1/2} u_1(\boldsymbol{\beta})}{\alpha^{1/2}},$$

using the fact that the unconditional variance of the ARCH(1) process defined in (32) is $\alpha/(1-\gamma)$. The other zero function is, of course, $(v_1^2 - 1)/\sqrt{2}$. The regressors are

$$\begin{aligned}
R_{1\boldsymbol{\beta}}^{(1)} &= \frac{1}{\sigma_1} \boldsymbol{X}_1(\boldsymbol{\beta}) & R_{1\alpha}^{(1)} &= 0 & R_{1\gamma}^{(1)} &= 0 \\
R_{1\boldsymbol{\beta}}^{(2)} &= 0 & R_{1\alpha}^{(2)} &= -\frac{1}{(1-\gamma)\sigma_1^2\sqrt{2}} & R_{1\gamma}^{(2)} &= \frac{1}{(1-\gamma)\sqrt{2}},
\end{aligned}$$

where $\sigma_1^2 \equiv \alpha/(1-\gamma)$.

Although it is perfectly in order to use these two artificial observations in order to take account of the first observation, we do *not* obtain the ML estimator by doing so. This is because the *unconditional* distribution of (G)ARCH errors is not normal. It is usually possible to obtain expressions for all the moments of these distributions, to the extent that they exist, as shown in Engle (1982) for the case of ARCH(1), but there does not seem to exist anywhere in the literature an expression for the density, without which the loglikelihood contribution for the first observation cannot be constructed. One might consider using the information in higher moments, but in many cases these do not exist.

With GARCH models, there is a further complication that we illustrate with the celebrated GARCH(1,1) model proposed by Bollerslev (1986). The conditional variance in this model obeys the following recursion:

$$\sigma_t^2 = \alpha + \gamma u_{t-1}^2 + \delta \sigma_{t-1}^2.$$

Since the σ_t^2 are not observed, even dropping observations is not enough to give us the information needed to initialise this recursion. The various *ad hoc* tricks commonly used to get around this difficulty can all be implemented with our artificial regression, about which we say no more here, since writing it down involves complicated and unenlightening expressions.

7. Concluding Remarks

The approach to estimation based on elementary zero functions is very general; it is at the heart of the Generalised Method of Moments. In contrast, maximum likelihood estimation relies on very precise assumptions about the distributions of the random elements in a model. In this paper, it is shown that, when these random elements are normally distributed, maximum likelihood and the elementary zero function approach yield exactly the same estimating equations. In this case, therefore, the two approaches are not just asymptotically equivalent, but numerically identical.

This fact has consequences useful for econometric practice. Many software packages exist that allow practitioners to perform a vast array of estimating procedures. Many of them require a bare minimum of work on the part of the user, and so are often treated as black boxes that turn out empirical results on demand. In all too many cases, it has been shown that even experienced users may be led astray by such results, since different packages give results that often differ widely with the same data and ostensibly the same estimation method.

Other packages make more demands on their users, requiring them to formulate loglikelihood functions and other criterion functions explicitly, sometimes along with their derivatives with respect to the parameters to be estimated. These packages allow users more control of what they are doing, and thus help to avoid meaningless results. The artificial regressions studied in this paper are typically no harder to set up than a criterion function and its derivatives. In particular, computer algebra is just as helpful with either approach. Since artificial regressions are *linear* regressions, their numerical implementation is much less subject to numerical instability than most optimisation algorithms. In addition, they can be used with profit in bootstrapping; see Davidson and MacKinnon (1999). Thus, in addition to the theoretical insights yielded by the approach based on elementary zero functions, many practical and numerical advantages also accrue from its use.

Appendix

Proof of Lemma 1

Condition (17) is just the definition of the estimator $\hat{\boldsymbol{\theta}}$, and so is satisfied by construction.

Since row t of the matrix $\boldsymbol{R}(\boldsymbol{\theta})$ is by definition contained in the information set Ω_t, we can see, by analogy with (5), that, under the DGP $\mu \in \mathbb{M}$,

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu) \stackrel{a}{=} -\big(n^{-1}\boldsymbol{R}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{F}(\boldsymbol{\theta}_\mu)\big)^{-1} n^{-1/2}\boldsymbol{R}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{r}(\boldsymbol{\theta}_\mu), \tag{35}$$

where element ti of the matrix $\boldsymbol{F}(\boldsymbol{\theta}_\mu)$ is $(\partial r_t/\partial\theta_i)(\boldsymbol{\theta}_\mu)$. By the definition of the regressor $R_{ti}(\theta)$, it follows that

$$\mathrm{E}_\mu\big(F_{ti}(\boldsymbol{\theta}_\mu) \,\big|\, \Omega_t\big) = -R_{ti}(\boldsymbol{\theta}_\mu),$$

and so, by a law of large numbers,

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{R}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{F}(\boldsymbol{\theta}_\mu) = -\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{R}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{R}(\boldsymbol{\theta}_\mu). \tag{36}$$

Given that the elementary zero functions $r_t(\boldsymbol{\theta}_\mu)$ are uncorrelated and homoskedastic with variance 1, we can apply a central limit theorem to the expression $n^{-1/2}\boldsymbol{R}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{r}(\boldsymbol{\theta}_\mu)$, and conclude that

$$\operatorname*{plim}_{n\to\infty} n^{-1/2}\boldsymbol{R}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{r}(\boldsymbol{\theta}_\mu) \sim \mathrm{N}\big(\mathbf{0}, \operatorname*{plim}_{n\to\infty} n^{-1}\boldsymbol{R}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{R}(\boldsymbol{\theta}_\mu)\big).$$

It follows that

$$\operatorname*{plim}_{n\to\infty} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu) \sim \mathrm{N}\big(\mathbf{0}, \operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{R}^\top(\boldsymbol{\theta}_\mu)\boldsymbol{R}(\boldsymbol{\theta}_\mu))^{-1}\big),$$

and so we see that condition (18) is satisfied, since $\operatorname{plim}\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\mu$.

The OLS estimates from running the artificial regression (16) with variables evaluated at a root-n consistent estimate $\acute{\boldsymbol{\theta}}$ is

$$\acute{\boldsymbol{b}} = (\acute{\boldsymbol{R}}^\top\acute{\boldsymbol{R}})^{-1}\acute{\boldsymbol{R}}^\top\acute{\boldsymbol{r}},$$

where $\acute{\boldsymbol{R}} \equiv \boldsymbol{R}(\acute{\boldsymbol{\theta}})$ and $\acute{\boldsymbol{r}} \equiv \boldsymbol{r}(\acute{\boldsymbol{\theta}})$. Because $\acute{\boldsymbol{\theta}}$ is root-n consistent, it follows readily from this that

$$n^{-1}\boldsymbol{R}_0^\top\boldsymbol{R}_0\, n^{1/2}\acute{\boldsymbol{b}} = n^{-1/2}\acute{\boldsymbol{R}}^\top\acute{\boldsymbol{r}} + O_p(n^{-1/2}) \tag{37}$$

with $\boldsymbol{R}_0 \equiv \boldsymbol{R}(\boldsymbol{\theta}_\mu)$. By Taylor expansion, we may write

$$n^{-1/2}\acute{\boldsymbol{R}}^\top\acute{\boldsymbol{r}} = n^{-1/2}\boldsymbol{R}_0^\top\boldsymbol{r}_0 + n^{-1}\boldsymbol{R}_0^\top\boldsymbol{F}_0\, n^{1/2}(\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu) + O_p(n^{-1/2}),$$

with $\boldsymbol{r}_0 \equiv \boldsymbol{r}(\boldsymbol{\theta}_\mu)$ and $\boldsymbol{F}_0 \equiv \boldsymbol{F}(\boldsymbol{\theta}_\mu)$. Thus

$$n^{-1/2}\acute{\boldsymbol{R}}^\top\acute{\boldsymbol{r}} = n^{-1/2}\boldsymbol{R}_0^\top\boldsymbol{r}_0 - n^{-1}\boldsymbol{R}_0^\top\boldsymbol{R}_0\, n^{1/2}(\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu) + O_p(n^{-1/2}). \tag{38}$$

But, from (35) and (36), we know that

$$n^{-1/2}\boldsymbol{R}_0^\top\boldsymbol{r}_0 = n^{-1}\boldsymbol{R}_0^\top\boldsymbol{R}_0\, n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu) + O_p(n^{-1/2}). \tag{39}$$

Assembling (37), (38), and (39) shows that

$$n^{-1}\boldsymbol{R}_0^\top\boldsymbol{R}_0\big(n^{1/2}\acute{\boldsymbol{b}} + n^{1/2}(\acute{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu) - n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu)\big) = O_p(n^{-1/2}).$$

Since $n^{-1}\boldsymbol{R}_0^\top\boldsymbol{R}_0 = O_p(1)$, the above equation is equivalent to (19). ■

Proof of Theorem 2

That the artificial regression (21) corresponds to the model defined by the set of loglikelihood contributions (9) and to the maximum likelihood estimator of that model follows immediately from Theorem 1 and Lemma 1.

By the term "efficient score estimator of the information matrix", we mean the estimator of which element ij is defined by the equation

$$\hat{\boldsymbol{I}}_{ij} = \sum_{t=1}^{n} \boldsymbol{I}_{ij}^{t}(\hat{\boldsymbol{\theta}}) \equiv \sum_{t=1}^{n} \mathrm{E}_{\boldsymbol{\theta}}\Big(\frac{\partial \ell_t}{\partial \theta_i}(\boldsymbol{\theta}) \frac{\partial \ell_t}{\partial \theta_j}(\boldsymbol{\theta}) \;\Big|\; \Omega_t \Big) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{40}$$

Since the regressors $R_{ti}^{(1)}(\boldsymbol{\theta})$ and $R_{ti}^{(2)}(\boldsymbol{\theta})$ are by definition the negatives of the optimal instruments, it follows by the work leading to Theorem 1 that

$$\frac{\partial \ell_t}{\partial \theta_i}(\boldsymbol{\theta}) = v_t(\boldsymbol{\theta}) R_{ti}^{(1)}(\boldsymbol{\theta}) + \frac{v_t^2(\boldsymbol{\theta}) - 1}{\sqrt{2}} R_{ti}^{(2)}(\boldsymbol{\theta}).$$

Since $R_{ti}^{(i)}(\boldsymbol{\theta}) \in \Omega_t$, $i = 1, 2$, and since the elementary zero functions $v_t(\boldsymbol{\theta})$ and $(v_t^2(\boldsymbol{\theta}) - 1)/\sqrt{2}$ all have variance 1 and are mutually uncorrelated, we can calculate that

$$\mathrm{E}_{\boldsymbol{\theta}}\Big(\frac{\partial \ell_t}{\partial \theta_i}(\boldsymbol{\theta}) \frac{\partial \ell_t}{\partial \theta_j}(\boldsymbol{\theta}) \;\Big|\; \Omega_t \Big) = R_{ti}^{(1)}(\boldsymbol{\theta}) R_{tj}^{(1)}(\boldsymbol{\theta}) + R_{ti}^{2)}(\boldsymbol{\theta}) R_{tj}^{(2)}(\boldsymbol{\theta}).$$

Since

$$\sum_{t=1}^{n} R_{ti}^{(1)}(\boldsymbol{\theta}) R_{tj}^{(1)}(\boldsymbol{\theta}) + R_{ti}^{2)}(\boldsymbol{\theta}) R_{tj}^{(2)}(\boldsymbol{\theta}) = \big(\boldsymbol{R}^{\top}(\boldsymbol{\theta}) \boldsymbol{R}(\boldsymbol{\theta}) \big)_{ij},$$

we see that $\boldsymbol{R}^{\top}(\hat{\boldsymbol{\theta}}) \boldsymbol{R}(\hat{\boldsymbol{\theta}})$ is equal to the efficient score estimator defined by (40). ■

References

Beach, C. M., and J. G. MacKinnon (1978a). "A maximum likelihood procedure for regression with autocorrelated errors," *Econometrica*, **46**, 51–58.

Beach, C. M., and J. G. MacKinnon (1978b). "Full maximum likelihood estimation of second-order autoregressive error models," *Journal of Econometrics*, **7**, 187–98.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, **31**, 307–27.

Brooks, C., S. P. Burke, and G. Persand (2001). "Benchmarks and the Accuracy of GARCH Model Estimation," *International Journal of Forecasting*, **17**, 45–56.

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 (1999). "Bootstrap testing in Nonlinear Models," *International Economic Review*, **40**, 487–508.

Davidson, R. and J. G. MacKinnon (2001). "Artificial Regressions," Ch. 1 in *A Companion to Econometric Theory*, ed. B. Baltagi, Oxford, Blackwell Publishers, 16–37.

Davidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*, Oxford University Press, New York.

Engle, R. F. (1982). "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, **50**, 987–1007.

Fiorentini, G., G. Calzolari, and L. Panattoni (1996). "Analytic derivatives and the computation of GARCH estimates," *Journal of Applied Econometrics*, **11**, 399–417.

Godambe, V. P. (1960). "An optimum property of regular maximum likelihood estimation," *Annals of Mathematical Statistics* **31**, 1208–11.

Godambe, V. P., and M. E. Thompson (1978). "Some aspects of the theory of estimating equations," *Journal of Statistical Planning and Inference*, **2**, 95–104.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton, Princeton University Press.

Godfrey, L. G., and M. R. Wickens (1981). "Testing linear and log-linear regressions for functional form," *Review of Economic Studies*, **48**, 487–96.

White, H. (1982). "Maximum likelihood estimation of misspecified models," *Econometrica*, **50**, 1–26.