Department of Economics

Econometrics Working Paper EWP1107

ISSN 1485-6441

THE OPTIMAL CONSTRUCTION OF INSTRUMENTS IN NONLINEAR REGRESSION: IMPLICATIONS FOR GMM INFERENCE

Kenneth G. Stewart

Department of Economics, University of Victoria,
Victoria, B.C., Canada

May 2011

Abstract

Interpreted as an instrumental variables estimator, nonlinear least squares constructs its instruments optimally from the explanatory variables using the nonlinear specification of the regression function. This has implications for the use of GMM estimators in nonlinear regression models, including systems of nonlinear regressions, where the explanatory variables are exogenous or predetermined and so serve as their own instruments, and where the restrictions under test are the only source of overidentification. In such situations the use of GMM test criteria involves a suboptimal construction of instruments; the use of optimally constructed instruments leads to conventional non-GMM test criteria. These implications are illustrated with two empirical examples, one a classic study of models of the short-term interest rate.

Keywords: optimal instruments, nonlinear regression, generalized method of moments

JEL Classifications: C12; C13

Author Contact:
Kenneth G. Stewart, Dept. of Economics, University of Victoria, P.O. Box 1700, STN CSC, Victoria, B.C., Canada V8W 2Y2; e-mail: kstewart@uvic.ca; Voice: (250) 721-8534; FAX: (250) 721-6214

Generalized method of moments (GMM) test criteria are sometimes applied to nonlinear models in which all the explanatory variables are treated as exogenous or predetermined, the instrument set is specified to consist solely of these regressors, and the maintained model is exactly identified. The only source of overidentification is the restrictions under test. As we shall see, the classic study of models of the short-term interest rate by Chan, Karolyi, Longstaff, and Sanders (1992; henceforth CKLS) is an example.

This paper shows that this practice involves a suboptimal construction of instruments that vitiates the supposed benefits of GMM. Recognizing this turns out to involve nothing more than applying the result that, when nonlinear least squares (NLS) is interpreted as an instrumental variables (IV) estimator, asymptotic efficiency requires that the instruments be constructed optimally using the nonlinear model specification. Although, in contrast to NLS-as-IV, an expanded instrument set that includes the optimally constructed ones can in principle be used to obtain a more efficient GMM estimator, the need for analytic derivatives means that this is unlikely to be implemented in practice, CKLS being a case in point.

We begin by expositing these principles and then turn to their application in two examples of GMM. The first is a single-equation nonlinear regression, the second the CKLS system.

I. NLS, NLIV, and GMM

Consider a nonlinear regression model denoted by, following the notation of Davidson and MacKinnon (1993, 2004),

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}.$$

The nonlinear least squares (NLS) estimator minimizes the sum-of-squares function

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}))'(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})).$$

It is well known that, under typical assumptions on the explanatory variables and disturbance, the NLS estimator is consistent. Furthermore, conditional on the information contained in the model, NLS is asymptotically efficient under disturbance normality.

It is also well known, particularly in the special case of a linear model, that this asymptotic efficiency is robust to additional information in the form of instrumental variables uncorrelated with the disturbance. An easy way to show this redundance for the nonlinear model is to relate NLS to nonlinear instrumental variables estimation and then explore the implications of introducing additional instruments.

Nonlinear Least Squares as an IV Estimator

It is useful to begin by reminding ourselves that, for models in which the regression function $\boldsymbol{x}(\boldsymbol{\beta})$ is differentiable, NLS has a method-of-moments interpretation. Continuing with the Davidson-MacKinnon notation, define the matrix of derivatives $\boldsymbol{X}(\boldsymbol{\beta}) \equiv \partial\boldsymbol{x}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. Then the NLS estimator satisfies the first order necessary conditions

$$\boldsymbol{X}(\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})) = \boldsymbol{0}, \tag{1}$$

which require the NLS residuals to be orthogonal to the derivative matrix. Assuming that the estimator is identified by the model and data set, in situations in which these nonlinear orthogonality conditions yield multiple solutions, so that they are necessary but not sufficient for a solution, direct minimization of the objective function $S(\boldsymbol{\beta})$ seeks the unique solution associated with a global minimum. Of course, in the special case of a linear regression model where $\boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$ the derivative matrix is simply the regressor matrix, $\boldsymbol{X}(\boldsymbol{\beta}) = \boldsymbol{X}$, and the orthogonality conditions reduce to the familiar first order conditions for OLS, which in that case are not only necessary but sufficient and can be solved for the familiar closed-form formula.

A key assumption on which the consistency of NLS rests is that the explanatory variables $\boldsymbol{X}$ be predetermined. In situations in which this is untenable estimation requires an instrument set $\boldsymbol{Z}$ comprising at least as many instruments as there are coefficients in $\boldsymbol{\beta}$. Amemiya (1974) showed that a consistent estimator is obtained by minimizing

$$Q(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}))'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})), \tag{2}$$

which we call the nonlinear instrumental variables (NLIV) estimator (although Amemiya called it nonlinear two-stage least squares).

Like NLS, and with the same qualifications (differentiability of the regression function; identification), NLIV has a method of moments interpretation. The first order necessary conditions for a minimum are

$$\boldsymbol{X}(\boldsymbol{\beta})\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})) = \boldsymbol{0}, \tag{3}$$

which requires the NLIV residuals to be orthogonal to the projection of the derivative matrix on the subspace spanned by the instrument set.

It is illuminating to consider the sense in which NLS is a special case of NLIV. Consider the special case in which the explanatory variables $\boldsymbol{X}$ are predetermined and so qualify to serve as their own instruments. It is significant that NLS is *not* obtained by setting $\boldsymbol{Z} = \boldsymbol{X}$ in the NLIV estimator, as would be true for linear regression. Setting $\boldsymbol{Z} = \boldsymbol{X}$ does not reduce the NLIV orthogonality conditions (3) to those for NLS (1), indicating that the minimization of $Q(\boldsymbol{\beta})$ with $\boldsymbol{Z} = \boldsymbol{X}$ is not equivalent to minimizing $S(\boldsymbol{\beta})$. Since NLS is the optimal estimator in these circumstances, such an NLIV estimator must be suboptimal.[1]

Instead inspection reveals that the minimization of the two objective functions is equivalent only if we set $\boldsymbol{Z} = \boldsymbol{X}(\boldsymbol{\beta})$, which does reduce the NLIV orthogonality conditions to those of NLS. Thus the optimal construction of the instrument set requires not just variables that qualify as instruments, but that they be used optimally. Their optimal employment uses the information embodied in the model—the regression function $\boldsymbol{x}(\boldsymbol{\beta})$—to construct the derivative matrix $\boldsymbol{X}(\boldsymbol{\beta})$ for use in the orthogonality conditions.

The Redundance of Additional Instruments

In general in instrumental variables estimation the availability of additional information in the form of additional instruments contributes to efficiency of the NLIV estimator. Consider an instrument set $\boldsymbol{Z}$ partitioned as $\boldsymbol{Z} = [\boldsymbol{Z}_1; \boldsymbol{Z}_2]$. Then the omission of $\boldsymbol{Z}_2$ from the instruments upon which NLIV is based results in a loss of efficiency. (See Davidson and MacKinnon, 2004, Exercise 8.8.)

However this intuition fails when the regressors $\boldsymbol{X}$ themselves qualify as instruments. Suppose that, in addition to $\boldsymbol{X}$, there are other variables $\boldsymbol{Z}_2$ that are predetermined with respect to the disturbance, and consider expanding the instrument set to include them. That is, consider generalizing the NLIV estimator from the NLS instrument set $\boldsymbol{Z} = \boldsymbol{X}(\boldsymbol{\beta})$ to the seemingly more informative one $\boldsymbol{Z} = [\boldsymbol{X}(\boldsymbol{\beta}); \boldsymbol{Z}_2]$. Define the projection matrices $\boldsymbol{P}_Z \equiv \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ and $\boldsymbol{P}_1 \equiv \boldsymbol{Z}_1(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'$. Under a partition $\boldsymbol{Z} = [\boldsymbol{Z}_1; \boldsymbol{Z}_2]$ it is a well-known (Davidson and MacKinnon, 2004, p. 66) property of projections that $\boldsymbol{P}_1\boldsymbol{P}_Z = \boldsymbol{P}_1$. Writing this out explicitly and premultiplying by $\boldsymbol{Z}_1'$ yields

$$\boldsymbol{Z}_1'\boldsymbol{Z}_1(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' = \boldsymbol{Z}_1'\boldsymbol{Z}_1(\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'$$

or, simplifying both sides,

$$\boldsymbol{Z}_1'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' = \boldsymbol{Z}_1'.$$

Setting $\boldsymbol{Z}_1 = \boldsymbol{X}(\boldsymbol{\beta})$ yields

$$\boldsymbol{X}(\boldsymbol{\beta})'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' = \boldsymbol{X}(\boldsymbol{\beta})'.$$

Thus when additional instruments beyond $\boldsymbol{X}(\boldsymbol{\beta})$ are included in $\boldsymbol{Z}$ the NLIV orthogonality conditions (3) nevertheless reduce to those for NLS, the first order conditions (1), regardless of the nature of the instruments. This establishes that, in such an instance, minimizing $Q(\boldsymbol{\beta})$ is equivalent to minimizing $S(\boldsymbol{\beta})$, even though $Q(\boldsymbol{\beta})$ does not reduce algebraically to $S(\boldsymbol{\beta})$.

In conclusion, the asymptotic efficiency of NLS is robust to the availability of additional information in the form of variables from outside the model that qualify as instruments.

Once the instrument set $\boldsymbol{Z}$ includes the derivatives $\boldsymbol{X}(\boldsymbol{\beta})$, expanding it to include additional instruments $\boldsymbol{Z}_2$ does not alter the NLIV estimator—it still reduces to NLS. In the terminology of Breusch, Qian, Schmidt, and Wyhowski (1999) the additional instruments $\boldsymbol{Z}_2$ are redundant.

Two special cases of this result are of interest. First and most obviously, it specializes immediately to linear regression, revealing that extraneous instruments cannot be used to improve OLS. Second and less obviously, it explains why, in nonlinear regression, it is not possible to construct a more efficient estimator by supplementing the NLS instruments $\boldsymbol{X}(\boldsymbol{\beta})$ with the raw $\boldsymbol{X}$ themselves—setting $\boldsymbol{Z} = [\boldsymbol{X}(\boldsymbol{\beta}); \boldsymbol{X}]$.

GMM

An important qualification to this redundance-of-additional-instruments result is that it does not generalize to GMM. The GMM estimator is defined to minimize a criterion function of the form

$$J(\boldsymbol{\beta}) = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}))' \boldsymbol{Z} \hat{\boldsymbol{W}} \boldsymbol{Z}' (\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})). \tag{4}$$

Assuming efficient GMM estimation, the weighting matrix $\hat{\boldsymbol{W}}$ denotes a consistent estimator for the inverse of the asymptotic variance of $(1/\sqrt{n})\boldsymbol{Z}'\boldsymbol{\varepsilon}$. $J(\boldsymbol{\beta})$ is a generalization of the NLIV criterion (2) in that it reduces to $Q(\boldsymbol{\beta})$ in the special case of a classical disturbance having a scalar covariance matrix.

Another special case in which GMM reduces to NLIV is when the instrument set is exactly identifying, so that an estimator exists that sets the sample moments $\boldsymbol{Z}'(\boldsymbol{y}-\boldsymbol{x}(\boldsymbol{\beta}))$ to zero without reference to the weighting matrix $\hat{\boldsymbol{W}}$. Such a case arises when the explanatory variables all qualify as instruments and we set $\boldsymbol{Z} = \boldsymbol{X}(\boldsymbol{\beta})$, yielding the sample moments (1) that define the NLS estimator. Thus, when only $\boldsymbol{X}(\boldsymbol{\beta})$ are used as instruments, once again NLS is obtained as the asymptotically efficient estimator.

However now this conclusion is not robust to the availability of additional instruments. If instruments $\boldsymbol{Z}_2$ are available to supplement $\boldsymbol{X}(\boldsymbol{\beta})$, then setting $\boldsymbol{Z} = [\boldsymbol{X}(\boldsymbol{\beta}); \boldsymbol{Z}_2]$ in $J(\boldsymbol{\beta})$ yields a GMM estimator that is more efficient asymptotically than the NLS estimator that GMM reduces to when $\boldsymbol{Z} = \boldsymbol{X}(\boldsymbol{\beta})$. Remarkably, this efficiency improvement holds even if we simply set $\boldsymbol{Z}_2 = \boldsymbol{X}$, so $\boldsymbol{Z} = [\boldsymbol{X}(\boldsymbol{\beta}); \boldsymbol{X}]$.[2] Of course, actually implementing such a GMM estimator may be problematic because it requires specifying $J(\boldsymbol{\beta})$ in terms of the analytic derivatives $\boldsymbol{X}(\boldsymbol{\beta})$, the expressions for which will be complex in all but the simplest nonlinear models. Because these analytic derivatives are themselves functions of the coefficient vector, the numerics of minimizing $J(\boldsymbol{\beta})$ will be complicated considerably. It would not be surprising if the promise of efficiency improvements is insufficient to induce empirical researchers to overcome these difficulties, complications that NLS avoids. This is particularly true in view of the well-known feature of GMM that expansion of the instrument set beyond the most relevant instruments, although in principle improving asymptotic efficiency, tends to lead to a deterioration in its finite-sample properties.

In summary, as with IV estimation, in general additional instruments improve the asymptotic efficiency of the GMM estimator, although there is the usual tradeoff with finite-sample bias. Unlike IV estimation, however, this remains true even when the regressors themselves all serve as instruments. Instruments that are redundant to IV estimation may not be to GMM.

II. Implications for GMM Inference

Presented in this manner these background results may seem elementary. That their implications for inference are nevertheless nontrivial in application may be illustrated with two empirical examples. It is useful to begin with a single equation model in which the issues emerge in their starkest form. We then turn to the more interesting and perhaps controversial CKLS application, which shows that similar considerations extend to the systems context.

Empirical Example: A Cobb-Douglas Production Function

Let us begin with a simple textbook example of NLS. Stewart (2005, Chap. 13) estimates a Cobb-Douglas production function with an additive disturbance,

$$Q_i = \gamma K_i^{\beta} L_i^{\alpha} + \varepsilon_i, \tag{5}$$

using cross-section data on 24 industries. The data are from Pyatt and Stone (1964) and were used in studies by Feldstein (1967) and Mizon (1977). In addition to reporting NLS coefficient estimates and standard errors (see his Example 2 on p. 565) Stewart tests the hypothesis of constant returns to scale (CRS), $\alpha+\beta=1$, using the inference procedures that apply most naturally in this context. It will be of interest to contrast Stewart's conventional estimation and testing strategy with an alternative, and so we refer to his as *Inference Strategy 1.*

Insert Table 1 around here

Inference Strategy 1: NLS with likelihood ratio or Wald tests NLS estimation results for the unrestricted and CRS-restricted functions are presented in the left half of Table 1, and reproduce results from Stewart (Example 2, p. 565; Example 6, pp. 577–8), although he does not present the heteroskedasticity-robust standard errors. A likelihood ratio statistic is computed as $LR = 2(\mathscr{L}_U - \mathscr{L}_R)$, where the unrestricted and restricted loglikelihood function values are (Stewart, Table 13.5) $\mathscr{L}_U = -132.123$ and $\mathscr{L}_R = -132.155$. This yields $LR = 0.0648$ (Stewart, Table 13.6), which does not come close to rejecting CRS at conventional significance levels (for example, $\chi^2_{0.10}(1) = 2.71$).

Alternatively, the Wald statistic is $W = 0.0646$ or, in its heteroskedasticity-robust variant, $W = 0.1283$.[3] Thus, although applied econometricians might debate the merits of these alternative test criteria (the LR statistic is invariant to reparameterizaton of the model and restriction while Wald statistics are not; but the Wald statistic can be robustified, giving some indication of the sensitivity of the test decision to heteroskedasticity in the data), in this application the substantive test decision is insensitive to these variations: CRS is not rejected.

The analysis of Section I shows that this textbook approach to estimation can be interpreted as NLIV with instruments set to $\boldsymbol{Z} = \boldsymbol{X}(\boldsymbol{\beta})$. For most purposes this interpretation is little more than a curiosity, particularly given that a benchmark approach to testing (the likelihood ratio test) comes from outside the IV framework. The NLIV interpretation is of interest, however, in comparing Strategy 1 with an alternative that we attribute to a hypothetical analyst.

Inference Strategy 2: GMM with distance or Wald tests Treating the explanatory variables of the Cobb-Douglas function as exogenous, the hypothetical analyst advocates GMM estimation using the instruments $\boldsymbol{X}_i = [1, K_i, L_i]$. (The instrument set includes the unit vector because, even though the model does not include an intercept, the disturbance has zero mean and so is orthogonal to the unit vector in the population.) Formally, this minimizes the GMM criterion (4) with $\boldsymbol{Z} = \boldsymbol{X}$.

The analyst acknowledges that, for the unrestricted model (5) with three coefficients, an estimator $\hat{\boldsymbol{\beta}}$ generally exists that will set the expression $\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}))$ to zero. That is, the three instruments are exactly identifying and so the weighting matrix is irrelevant to the minimization of $J(\boldsymbol{\beta})$; GMM reduces to NLIV based on $\boldsymbol{Z} = \boldsymbol{X}$ and the GMM criterion must be identically zero. Under the CRS restriction, however, one of the coefficients is eliminated, the instrument set becomes overidentifying, and GMM differs from NLIV given some nonscalar covariance specification—in this application presumably one of heteroskedasticity given that the data are cross-sectional.

Even in the exactly identified maintained model, because this implementation of NLIV is based on $\boldsymbol{Z} = \boldsymbol{X}$ rather than $\boldsymbol{Z} = \boldsymbol{X}(\boldsymbol{\beta})$, the GMM estimates differ from those of Strategy 1, as do Wald tests. The GMM analog to the LR statistic is the distance statistic of Newey and West (1987),

$$D = J(\tilde{\boldsymbol{\beta}}) - J(\hat{\boldsymbol{\beta}}) \xrightarrow{d} \chi^2(g). \tag{6}$$

Here $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ denote the restricted and unrestricted GMM estimators and g is the number of restrictions. It is well known that, to ensure that the statistic is nonnegative, the two values of the objective function must be computed using a common weighting matrix. The weighting matrix for either $\tilde{\boldsymbol{\beta}}$ or $\hat{\boldsymbol{\beta}}$ may be used, giving rise to two ways of computing the statistic:

$$D_1 = \hat{J}(\tilde{\boldsymbol{\beta}}) - \hat{J}(\hat{\boldsymbol{\beta}}) \qquad \text{(uses } \hat{\boldsymbol{W}} \text{ of unrestricted model)} \tag{7a}$$

$$D_2 = \tilde{J}(\tilde{\boldsymbol{\beta}}) - \tilde{J}(\hat{\boldsymbol{\beta}}) \qquad \text{(uses } \tilde{\boldsymbol{W}} \text{ of restricted model).} \tag{7b}$$

Practices vary regarding the choice of weighting matrix to hold constant. Applied researchers often use D_1, presumably on the intuition that it seems inappropriate to impose on the weighting matrix the restrictions that are under test. However Hansen (2006) presents analytical and simulation evidence supporting D_2.

In the present application in which the unrestricted model is exactly identified and so $J(\hat{\boldsymbol{\beta}}) = 0$ regardless of the weighting matrix, these minimum distance statistics simplify to

$$D_1 = \hat{J}(\tilde{\boldsymbol{\beta}}), \quad D_2 = \tilde{J}(\tilde{\boldsymbol{\beta}}). \tag{8}$$

The latter, $J(\boldsymbol{\beta})$ for the restricted model, is familiar as the Hansen-Sargan test for overidentification.

The results of Strategy 2 are presented in the right half of Table 1, and are similar to those of Strategy 1. The point estimates suggest a roughly 16%/84% division of factor payments between capital and labor, only slightly different from the NLS estimates of 14%/86%; two-standard deviation confidence bounds on either set of estimates easily include those of the other strategy. CRS is not rejected by either Wald or distance tests, which yield almost identical p-values. Interestingly, in this example the distance statistic is almost unaffected by the choice of weighting matrix to hold constant.

In terms of the formal differences between the two strategies, the hypothetical analyst might assert several advantages of Strategy 2 over Strategy 1. First, GMM makes no assumption about the form of the population distribution, in contrast to likelihood-based methods. As well, GMM estimation of the restricted model, because it is overidentified, uses the nonscalar covariance structure, in principle improving the efficiency of those estimates. Another byproduct of the overidentification of the restricted model is that it yields the Hansen-Sargan test for overidentification, a useful model diagnostic.

However it is unlikely that most econometricians would be persuaded by these arguments. Instead they would point out that the entire GMM exercise is predicated on the inefficient instrument set $\boldsymbol{Z} = \boldsymbol{X}$. The use of the nonscalar covariance structure in the estimation of the restricted model, and the availability of the Hansen-Sargan test, are merely artifacts of this initial inefficient instrument choice. Furthermore, estimation of the restricted model then becomes dependent on the specification of that nonscalar covariance structure, which may introduce the possibility of specification error. That GMM does not require an assumption on the form of the population distribution is shared by NLS. NLS also shares the ability to obtain standard errors and Wald tests that are robust to heteroskedasticity. It is only the LR test that uses the additional assumption of normality. But this assumption yields benefits that may be worth its cost. First, the LR statistic involves no issues of holding the covariance matrix constant across estimations, a simplification that becomes increasingly valuable with more complex nested testing structures. Second, the small-sample behavior of LR tests and the quality of the finite-sample approximation they provide is better understood and perhaps more reliable than for distance tests. Presumably the performance of the LR test is enhanced by the fact that it is based on optimally-constructed instruments.

In this example this debate between the two strategies might be dismissed as immaterial given the lack of sensitivity of the substantive results of the analysis to them. However this is not always the case, as the next example illustrates.

Empirical Example: Models of the Short-Term Interest Rate

As a second example illustrating the practical importance of the considerations highlighted by the analysis of Section I, consider the model of the short-term interest rate estimated by Chan, Karolyi, Longstaff, and Sanders (1992). The CKLS model is of particular interest because it nests several classic models of the interest rate as special cases, so these can be tested as restrictions on their system. Consequently their analysis has often been cited or replicated; see, for example, Bibby, Jacobsen, and Sørensen (2006, Example 5.5), Mills and Markellos (2008, Example 4.4), and Zivot and Wang (2006, Sec. 21.7.5).[4] Many papers extend the CKLS analysis in various directions. Bliss and Smith (1998) find the results to be sensitive to the treatment of structural breaks, while Treepongkaruna and Gray (2003) study their robustness to different data sets and sampling frequencies. Brenner, Harjes, and Kroner (1996) and Koedijk, Nissen, Schotman, and Wolff (1997) nest the CKLS model within more general frameworks that permit GARCH volatility. As well, the CKLS model has become a canonical application for illustrating alternative approaches to the estimation of continuous time models: see Jiang and Knight (1997), Nowman (1997), and Yu and Phillips (2001).

The CKLS estimation strategy uses a discrete-time approximation to an underlying continuous-time process for the interest rate. This discrete-time approximation specifies the interest rate as evolving according to a first-order autoregression with a disturbance variance that depends on the interest rate itself:

$$r_{t+1} - r_t = \alpha + \beta r_t + \varepsilon_{t+1} \tag{9a}$$

$$\mathrm{E}(\varepsilon_{t+1}) = 0 \tag{9b}$$

$$\mathrm{E}(\varepsilon_{t+1}^2) = \sigma^2 r_t^{2\gamma}. \tag{9c}$$

A distinguishing feature of this model is that, although the variance specification (9c) permits systematic variation in volatility, this variation depends only on the level of the interest rate. By contrast, the GARCH alternatives investigated by Brenner et al. (1996) and Koedijk et al. (1997) specify an autoregressive conditional volatility that depends directly on information shocks.

The various interest rate models encompassed as special cases of the CKLS model, and the associated parameter restrictions, are summarized in Table 2, which reproduces Table I of CKLS.

Insert Table 2 around here

Turning to the population moments implicit in the model, the zero-mean disturbance (9b) implies

$$\mathrm{E}(\Delta r_{t+1} - \alpha - \beta r_t) = 0$$

while the variance specification (9c) implies

$$\mathrm{E}[(\Delta r_{t+1} - \alpha - \beta r_t)^2] - \sigma^2 r_t^{2\gamma} = 0.$$

In terms of its empirical content, therefore, the CKLS model may be represented as a two-equation system of seemingly unrelated regressions (SUR).[5]

$$\Delta r_{t+1} = \alpha + \beta r_t + u_{1,t+1}$$
$$(\Delta r_{t+1} - \alpha - \beta r_t)^2 = \sigma^2 r_t^{2\gamma} + u_{2,t+1}$$

The lagged interest rate r_t is predetermined in relation to the period $t+1$ shocks generating the disturbances $u_{1,t+1}$, $u_{2,t+1}$; if these disturbances are serially uncorrelated then the model is a true SUR rather than simultaneous system. This is implicitly the assumption adopted by CKLS because they treat r_t as satisfying the requirements for an instrumental variable. The nonlinearity-in-parameters of the second equation makes this system nonlinear as a whole, with the implications for the optimal construction of instruments revealed in Section I. Paralleling the Cobb-Douglas example, two strategies for estimation and testing may be identified.

Inference Strategy 1: Nonlinear GLS Under a suitable specification for the disturbances u_{1t}, u_{2t} the nonlinear SUR system may be estimated by feasible generalized least squares (often called Zellner estimation in the case of a classical SUR covariance structure). For a disturbance covariance matrix satisfying the Oberhofer-Kmenta (1974) conditions, iterating on the covariance matrix yields maximum likelihood estimators, and so likelihood-based inference becomes available, as with iterative Zellner estimation. Of course, Wald tests may also be used, the comparative advantages of Wald versus LR tests being as described in the Cobb-Douglas example.

However Strategy 1 is not that adopted by CKLS, who instead use:

Inference Strategy 2: GMM Were the autoregression (9a) to be estimated as an OLS regression the instrument set is effectively $[1, r_t]$, and this is the instrument set used by CKLS for the system as a whole; the relevant moments are therefore

$$\begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+1}^2 - \sigma^2 r_t^{2\gamma} \end{bmatrix} \otimes [1, r_t]' = \begin{bmatrix} \Delta r_{t+1} - \alpha - \beta r_t \\ (\Delta r_{t+1} - \alpha - \beta r_t)^2 - \sigma^2 r_t^{2\gamma} \end{bmatrix} \otimes [1, r_t]', \qquad (10)$$

which corresponds to equation (4) of CKLS. These four moments serve to exactly identify the four parameters α, β, γ, and σ^2 of the maintained model, but are overidentifying under any of the restrictions of Table 2. In this respect inference in the model is, therefore, analogous to that of the Cobb-Douglas example, although the nested testing structure implied by the restrictions of Table 2 is more complex. In general Newey-West distance statistics are computed as (6); CKLS indicate that they use the weighting matrix from the unrestricted model, and thus the statistic D_1 in our notation. As in the Cobb-Douglas example, when the unrestricted model is the maintained model its exact identification yields $J(\hat{\boldsymbol{\beta}}) = 0$ and so the CKLS Newey-West test statistic is simply $D_1 = \hat{J}(\tilde{\boldsymbol{\beta}})$.

In order to compare the two strategies we begin by replicating CKLS. Table 4 reports the results of my replication of their GMM estimates, as reported in their Table III, supplemented with Wald tests. In order to gauge the precision of the replication all values are reported to an accuracy one digit greater than in the CKLS table. The replication is exact or very close. The maintained model is replicated exactly, perhaps because, under exact identification, the estimated covariance matrix does not play a role in the coefficient point estimates. (That is, for the maintained model GMM reduces to NLIV, albeit based on the inefficient instrument set $\boldsymbol{Z} = \boldsymbol{X}$.) The coefficient estimates of the maintained model are therefore not sensitive to variations across software in the numerics of the estimation of the weighting matrix, aiding replication.[6]

Insert Tables 3 and 4 around here

Contrasted with these are the Strategy 1 nonlinear GLS results of Table 3. As in our Cobb-Douglas example, qualitatively the coefficient estimates and their t statistics are broadly similar across the two strategies. Focusing on the maintained model to illustrate, note that under both strategies the coefficient estimates have similar degrees of statistical significance. Both estimates of β imply that the autoregressive process for the interest rate (9a) is stable (because $1 + \hat{\beta} = 1 - 0.5921 = 0.4079 < 1$ in the case of Strategy 2, while $1 + \hat{\beta} = -0.262$ in the case of strategy 1). This implies long run convergence to an interest rate of $r^* = -\hat{\alpha}/\hat{\beta} = 0.0408/0.5921 = 0.0689$ in the case of Strategy 2, or the nearly identical $r^* = 0.0859/1.2620 = 0.0681$ in the case of Strategy 1, a plausible long run annual rate for the sample period in question (June 1964–December 1989). Even so, both strategies yield estimates of α and β that are not highly significant, so the hypothesis that they are zero—as specified in some of the special-case models—cannot necessarily be rejected. CKLS observe (p. 1217) that "...there appears to be only weak evidence of mean reversion in the short-term rate; the parameter β is insignificant in the unrestricted model ...", a finding that emerges from both estimation strategies.

Another important respect in which the two strategies yield consistent results is with respect to the role of γ. CKLS remark that "...the conditional volatility of the process is highly sensitive to the level of the short-term yield; the unconstrained estimate of γ is 1.499. This result is important since this is much higher than the values used in most of the models." Strategy 1 similarly yields $\hat{\gamma} = 1.3871$, also highly significant.

Despite these similarities, there is an important difference in the the results yielded by the two strategies. The LR tests of Strategy 1 provide more decisive rejections of the nested models, rejecting all but Model 6 (Brennan-Schwartz) at a 1% level of significance. By contrast, the distance tests of Strategy 2 do not reject Models 4 (Dothan), 5 (GBM), 6 (Brennan-Schwartz), or 7 (CIR-VR) at conventional significance levels. The Wald tests, on the other hand, tend to be more favorable to the special case models under Strategy 1 than under Strategy 2. The GLS Wald tests do not reject models 4, 5, and 6 at conventional significance levels, whereas the GMM Wald tests reject all the special-case models at 10%, although not necessarily at more stringent significance levels.

Conclusions

Our analysis suggests general implications for the optimal construction of instruments in nonlinear regression, including systems of nonlinear regressions—that is, situations in which all the explanatory variables are exogenous or predetermined and so qualify as instruments. In such situations nonlinear least squares (or, in the systems context, nonlinear feasible generalized least squares) constructs instruments optimally from the regressors using the nonlinear specification of the model. In contrast, GMM does not. Tests of restrictions on the maintained model should surely therefore be based on the NLS (or nonlinear FGLS) rather than GMM results. This is particularly in view of the advantages of the generally simpler likelihood-based inference methods afforded by the least squares estimators, which do not involve issues of holding the GMM weighting matrix constant across restricted and unrestricted models that complicate the application of Newey-West tests. The advantages of likelihood-based inference include not just its simplicity of implementation but also, in many applications, better-understood properties of asymptotic approximation.

The practical importance of this conclusion has been illustrated with two empirical examples, one elementary, the other a classic and widely-cited comparison of models of the short-term interest rate. In the latter, important test decisions are altered by the use of optimally constructed instruments.

Of course, GMM continues to be an appropriate estimator in situations where some regressors are not exogenous or predetermined, so that consistent estimation requires instruments from outside the specification of the estimating equations. In such situations any nonlinearity of the model specification does not provide information relevant to the optimal employment of the instruments.

GMM would also be an appropriate estimator, at least in principle, in situations where the regressors all qualify as instruments and the researcher is willing to supplement $\boldsymbol{X}(\boldsymbol{\beta})$ with additional instruments, $\boldsymbol{X}$ or otherwise. In this case, as discussed in Section I, the additional instruments improve the asymptotic efficiency of the GMM estimator whereas they do not improve NLS/NLIV. Here the maintained model is overidentified, the restrictions under test are not the only source of overidentification, and the GMM distance statistics take the full form (7) rather than reducing to (8). However most empirical researchers are unlikely to find this route to efficiency improvements appealing. In addition to requiring the formulation of the GMM criterion in terms of potentially complex expressions for analytic derivatives and the accompanying numerical complexities, it is well known that expanding the instrument set in GMM tends to come at the cost of finite-sample bias. Thus one cannot fault CKLS for not using the instrument set $\boldsymbol{Z} = [\boldsymbol{X}(\boldsymbol{\beta}); \boldsymbol{X}]$ rather than the $\boldsymbol{Z} = \boldsymbol{X}$ that they did use; our argument is instead that they should have used $\boldsymbol{Z} = \boldsymbol{X}(\boldsymbol{\beta})$, i.e. nonlinear FGLS in their systems context, or NLS in the simpler single-equation context.

Notes

[1] This is another way of saying that the NLIV estimator cannot be obtained via a two-step application of least squares, nonlinear or otherwise, in contrast to the linear case. It is for this reason that Amemiya's name *nonlinear two-stage least squares* is misleading. Attempts to arrive at an estimator by such a two-step process invariably lead to an inconsistent estimator; see the remarks to this effect in Davidson and MacKinnon (1993, p. 225). Historically, this is why Amemiya's demonstration that the minimization of $Q(\boldsymbol{\beta})$ yields a consistent NLIV estimator was of landmark significance.

[2] This conclusion is reminiscent of the early contribution to the GMM literature by Cragg (1983). He showed that, in the context of linear regression with heteroskedasticity, an efficiency improvement over OLS could be achieved with a GMM estimator based on an instrument set that supplements the exogenous regressors with nonlinear functions of them such as powers and cross-products.

[3] Stewart reports size-corrected values of $W = 0.0565$ (p. 596) and 0.112 (p. 634), respectively. Here we focus on the non-size-corrected versions in order to facilitate comparison with the GMM results.

[4] The CKLS paper is reprinted in Hughston (2001). Their results have also been (approximately) replicated in unpublished papers by Christensen and Poulsen (1999) and Christensen, Poulsen, and Sørensen (2001). For surveys of interest rate modeling with some discussion of CKLS see Campbell, Lo, and MacKinlay (1997, Chaps. 10, 11; especially pp. 449–451) or James and Webber (2000, Chap. 17).

[5] For the short-term interest rate r_t CKLS used the one-month Treasury bill yield over the period June 1964–December 1989. In their data set this appears as a continuously compounded monthly return, and so must be multiplied by 12 to be expressed conventionally as an annualized return. Because the model is estimated with annualized returns sampled monthly, for technical reasons related to the discrete time approximation of a continuous time process a factor 1/12 must be introduced in estimation; the SUR system is modified as

$$\begin{aligned}\Delta r_{t+1} &= (\alpha + \beta r_t)/12 + u_{1,t+1} \\ [\Delta r_{t+1} - (\alpha + \beta r_t)/12]^2 &= \sigma^2 r_t^{2\gamma}/12 + u_{2,t+1},\end{aligned}$$

and similarly for the moments (10).

[6] All estimation results reported in this paper were obtained using the econometrics package TSP. The numerics of TSP's nonlinear estimation routines have been favorably evaluated by McCullough (1999).

References

Amemiya, T. (1974) The nonlinear two-stage least squares estimator. *Journal of Econometrics* 2, 105–110.

Bibby, B.M., Jacobsen, M., and M. Sørensen (2006) Estimating functions for discretely sample diffusion-type models. In Y. Aït-Sahalia and L.P. Hansen (eds.), *Handbook of Financial Econometrics.* North-Holland.

Bliss, R.R., and D.C. Smith (1998) The elasticity of interest rate volatility: Chan, Karolyi, Longstaff, and Sanders revisited. *Journal of Risk* 1, 21–46.

Brenner, R.J., Harjes, R.H., and K.F. Kroner (1996) Another look at models of the short-term interest rate. *Journal of Financial and Quantitative Analysis* 31, 85–107.

Breusch, T., Qian, H., Schmidt, P., and D. Wyhowski (1999) "Redundancy of moment conditions," *Journal of Econometrics* 91, 89–111.

Campbell, J.Y., Lo, A.W., and A.C. MacKinlay (1997) *The Econometrics of Financial Markets.* Princeton University Press.

Chan, K.C., Karolyi, G.A., Longstaff, F.A., and A.B. Sanders (1992) An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance* 47, 1209–1227.

Christensen, B.J., and R. Poulson (1999) Optimal martingale and likelihood methods for models of the short rate of interest, with Monte Carlo evidence for the CKLS specification and applications to nonlinear drift models. Working Paper, University of Aarhus.

Christensen, B.J., Poulson, R., and M. Sørensen (2001) Optimal inference in diffusion models of the short rate of interest. Working Paper 102, University of Aarhus Centre for Analytical Finance.

Cragg, J.G. (1983) More efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* 51, 751–763.

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

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

Feldstein, M.S. (1967) Alternative Methods of Estimating a CES Production Function for Britain. *Economica* 34, 384–394.

Hansen, B.E. (2006) Edgeworth expansions for the Wald and GMM statistics for nonlinear restrictions. In D. Corbae, S.N. Durlauf, and B.E. Hansen (eds.), *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, pp. 9–35. Cambridge University Press.

Hughston, L. (2001) *The New Interest Rate Models: Recent Developments in the Theory and Application of Yield Curve Dynamics.* Risk Publications.

James, J., and N. Webber (2000) *Interest Rate Modelling.* Wiley.

Koedijk, K.G., Nissen, F.G.J.A., Schotman, P.C., and C.C.P. Wolff (1997) The dynamics of short-term interest rate volatility reconsidered. *European Finance Review* 1, 105–130.

Jiang, G.J., and J.L. Knight (1997) A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model. *Econometric Theory* 13, 615–645.

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

Mills, T.C., and R.N. Markellos (2008) *The Econometric Modelling of Financial Time Series.* Cambridge University Press.

Mizon, G.E. (1977) Inferential Procedures in Nonlinear Models: An Application in a UK Industrial Cross Section Study of Factor Substitution and Returns to Scale. *Econometrica* 45, 1221–1242.

Newey, W.K., and K.D. West (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.

Nowman, K.B. (1997) Gaussian estimation of single-factor continuous time models of the term structure of interest rates. *Journal of Finance* 52, 1695–1706.

Oberhofer, J., and J. Kmenta (1974) A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica* 42, 570–590.

Pyatt, G., and R. Stone (1964) *Capital, Output and Employment 1948–60, A Programme for Growth,* Vol. 4. Chapman and Hall.

Stewart, K.G. (2005) *Introduction to Applied Econometrics.* Brooks/Cole Thomson Learning.

Treepongkaruna, S., and S. Gray (2003) On the robustness of short-term interest rate models. *Accounting and Finance* 43, 87–121.

Yu, J., and P.C.B. Phillips (2001) A Gaussian approach for continuous time models of the short-term interest rate. *Econometrics Journal* 4, 210–224.

Zivot, E., and J. Wang (2006) *Modeling Financial Time Series with S-Plus,* 2nd ed. Springer.

Table 1: Estimation Results for a Cobb-Douglas Production Function

	Strategy 1: NLS		Strategy 2: GMM	
	Maintained	CRS-restricted	Maintained	CRS-restricted
Coefficients[a]				
γ	1.6212	1.7318	1.6992	1.7379
	(0.4524)	(0.0465)	(0.2944)	(0.0600)
	[0.3141]	[0.0644]		
β	0.1476	0.1427	0.1659	0.1640
	(0.0462)	(0.0395)	(0.0305)	(0.0268)
	[0.0542]	[0.0489]		
α	0.8531	0.8573	0.8379	0.8360
	(0.0465)	(0.0395)	(0.0302)	(0.0268)
	[0.0488]	[0.0489]		
Criterion function value[b]	$\mathscr{L}_{\mathrm{U}} = -132.123$	$\mathscr{L}_{\mathrm{R}} = -132.155$	$J(\hat{\boldsymbol{\beta}}) = 0$	$\tilde{J}(\tilde{\boldsymbol{\beta}}) = 0.01722$
Wald test of CRS (p-value)[c]			$W = 0.0177$ (0.894)	
non-heteroskedasticity robust	$W = 0.0646$ (0.799)			
heteroskedasticity robust	$W = 0.1283$ (0.720)			
Likelihood ratio test of CRS (p-value)	$LR = 0.0648$ (0.799)			
Minimum distance test of CRS (p-value)				
using unrestricted model covariance matrix			$D_1 = \hat{J}(\tilde{\boldsymbol{\beta}}) = 0.01712$ (0.896)	
using restricted model covariance matrix			$D_2 = \tilde{J}(\tilde{\boldsymbol{\beta}}) = 0.01722$ (0.896)	

[a] Conventionally computed standard errors in parentheses. NLS heterosdedasticity-robust standard errors in brackets.

[b] Covariance matrix iterated to convergence.

[c] Wald statistics are non-size-corrected; to obtain size-corrected values multiply by $(n-K)/n = 21/24$.

Table 2: Alternative Models of the Short Term Interest Rate as Restrictions on the CKLS Model

	Model	α	β	σ^2	γ
1.	Merton		0		0
2.	Vasicek				0
3.	Cox-Ingersoll-Ross square root (CIR-SR)				0.5
4.	Dothan	0	0		1
5.	Geometric Brownian motion (GBM)	0			1
6.	Brennan-Schwartz (BS)				1
7.	Cox-Ingersoll-Ross variable rate (CIR-VR)	0	0		1.5
8.	Constant elasticity of variance (CEV)	0			

Table 3: *Strategy 1:* Nonlinear GLS Results for Alternative Models of the Short-Term Interest Rate

	Model	α	β	σ^2	γ	d.f.	LR test	Wald test
							(p-value)	(p-value)
0.	Maintained	0.0859	−1.2620	1.0295	1.3871			
		(1.8968)	(−1.7337)	(0.6482)	(4.4569)			
1.	Merton	0.0011	0.0	0.0008	0.0	2	121.3051	21.1511
		(0.0650)		(3.2174)			(0.0000)	(0.0000)
2.	Vasicek	0.0942	−1.3865	0.0008	0.0	1	69.3923	19.8643
		(1.7578)	(−1.6152)	(4.4818)			(0.0000)	(0.0000)
3.	CIR-SR	0.0922	−1.3489	0.0155	0.5	1	35.2251	8.1248
		(1.7853)	(−1.6348)	(4.2156)			(0.0000)	(0.0044)
4.	Dothan	0.0	0.0	0.1894	1.0	3	69.5244	4.6756
				(3.6899)			(0.0000)	(0.1972)
5.	GBM	0.0	−0.2644	0.1983	1.0	2	56.4569	3.6178
			(−0.8811)	(1.5426)			(0.0000)	(0.1638)
6.	BS	0.0886	−1.3003	0.1892	1.0	1	6.3896	1.5472
		(1.8277)	(−1.6762)	(3.8359)			(0.0115)	(0.2135)
7.	CIR-VR	0.0	0.0	1.6839	1.5	3	61.6634	7.9883
				(3.1225)			(0.0000)	(0.0462)
8.	CEV	0.0	−0.2577	1.3067	1.4332	1	50.2852	3.5979
			(−0.9252)	(0.6783)	(4.8755)		(0.0000)	(0.0578)

Notes: Coefficient t-ratios are in parentheses; t statistics and Wald tests are heteroskedasticity-robust.

LR and Wald tests are of the restricted model against the alternative of the maintained model.

Table 4: *Strategy 2:* GMM Results for Alternative Models of the Short-Term Interest Rate

	Model	α	β	σ^2	γ	d.f.	Distance test	Wald test
							(p-value)	(p-value)
0.	Maintained	0.04082	−0.59214	1.67038	1.49990			
		(1.855)	(−1.552)	(0.773)	(5.948)			
1.	Merton	0.00550	0.0	0.00042	0.0	2	6.75330	35.5386
		(1.437)		(7.272)			(0.03416)	(0.0000)
2.	Vasicek	0.01540	−0.17763	0.00042	0.0	1	8.84288	35.3766
		(0.793)	(−0.522)	(7.115)			(0.00294)	(0.0000)
3.	CIR-SR	0.01884	−0.23155	0.00732	0.5	1	6.12651	15.7219
		(0.973)	(−0.683)	(7.546)			(0.01332)	(0.0001)
4.	Dothan	0.0	0.0	0.11729	1.0	3	5.62694	7.8324
				(7.973)			(0.13124)	(0.0496)
5.	GBM	0.0	0.10113	0.11848	1.0	2	3.15430	5.5320
			(1.504)	(8.036)			(0.20656)	(0.0629)
6.	BS	0.02421	−0.31366	0.11857	1.0	1	2.21241	3.9297
		(1.237)	(−0.917)	(8.091)			(0.13690)	(0.0474)
7.	CIR-VR	0.0	0.0	1.5778	1.5	3	6.2067	6.3042
				(8.00)			(0.1019)	(0.0977)
8.	CEV	0.0	0.10300	0.43240	1.24438	1	2.98565	3.4399
			(1.528)	(0.615)	(4.000)		(0.08401)	(0.0636)

Note: Coefficient t-ratios are in parentheses. Distance and Wald tests are of the restricted model against the alternative of the maintained model.