COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 741

TWO STAGE AND RELATED ESTIMATORS AND THEIR APPLICATIONS

Adrian Pagan

April 8, 1985

TWO STAGE AND RELATED ESTIMATORS AND THEIR APPLICATIONS*

Adrian Pagan

"We shall not cease from our exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time".

T.S. Elliot, *Little Gidding*.

1. Introduction

One of the most consistently troublesome problems faced by applied econometricians arises when, in order to estimate the parameters they are ultimately interested in, it becomes necessary to quantify a number of incidental or nuisance parameters. Moreover, in many of these instances, it is the presence of these incidental parameters which converts a relatively simple computational problem into a very complex one. Examples in the early history of econometrics would be the presence of serial correlation, heteroscedasticity, seemingly unrelated equations and simultaneity. More recent examples would be the construction of expectations through auxiliary regressions as in Barro (1977), unobservable variables such as market clearing prices by Lewis (1983) and the estimation of mixtures of limited dependent and continuous random variable models e.g. Heckman (1979).

Because estimation would generally be easy if the individual parameters were known, a very common strategy for dealing with them has emerged: They are replaced by a nominated value which may or may not be estimated from the data. It would not be too inaccurate to assert that most issues of journals

*This is a revision of a paper done at the Australian National University, August 1984.

featuring applied econometric work contain at least one application of this approach, making it a very widespread practice. Such frequency of appearance suggests that the procedure is well understood and, by its practice, seemingly ratified. What is less well understood is that it can be very misleading. Most researchers recognise the potential that it creates for inconsistencies in estimators, and the same normally draw attention to its lack of efficiency. A smaller number observe that it can also lead to invalid inferences e.g. Durbin (1970), Sims (1977), Johnston (1972) and there is probably a growing awareness of this point. For particular types of models such a point has been emphasised by Heckman (1978), Amemiya (1978), Hsiao and Mountain (1982), McCallum (1979), Pagan (1984) and Newey (1984).

What is lacking in most of these contributions is the development and use of a general framework for two-stage estimators, a deficiency which is unfortunate as it serves to mask a common structure that enables the unification of what are seemingly diverse results. By eschewing a specific treatment it also becomes possible to identify relationships that tend to be obscured in particular instances, and to appreciate the limits of some conclusions that have been reached. Accordingly, this paper attempts to discuss the issues surrounding two-stage (and related estimators) from a very general perspective, before applying the principles established to special cases.

To be more specific we will be concerned with the generic problem characterized by a log likelihood L and three sets of parameters (α, β, γ) of dimension $(p \times 1)$, $(q \times 1)$ and $(s \times 1)$ respectively. Throughout β will be the parameters of interest and (α, γ) incidental parameters. It is necessary to partition the incidental parameters in such a way, as there are two strategies for dealing with them. The first involves replacement

by a consistent estimator, $\tilde{\alpha}$ say; the second is to prescribe them e.g. at $\gamma = \bar{\gamma} = 0$. Any combination of the two options may, and frequently is, used for a given set of incidental parameters. Hence, in what follows, α will be those parameters replaced by consistent estimators $\tilde{\alpha}$, while γ will be prescribed at $\bar{\gamma}$. Two-stage estimators then solve the program $\{\max_{\beta} L(\beta,\tilde{\alpha},\bar{\gamma})\}$.

Section 2 of this paper presents a series of theorems dealing with the properties of two-stage estimators under these various scenarios. Conditions are presented for consistency and efficiency of the estimators as well as general expressions for their covariance matrices[1]. Some mention is also made of related estimators that are used in later sections to improve on the efficiency of two-stage procedures. Section 3 proceeds to apply these theorems to derive results established in Pagan (1984), Wickens (1982) and Turkington (1984) concerning models with expectations. Each of these papers presents a series of disparate results and it is not easy to appreciate where the diversity comes from. Furthermore, some conclusions are puzzling. When looked at from a unified perspective however the outcomes are easily apprehended and are intuitively reasonable.

Section 4 moves on to an area where the complexity of estimation is very much bound up with the presence of incidental parameters; that of rational expectations. These nuisance parameters have meant difficulties in maximizing the likelihood e.g. see Eckstein (1984, f.n.18) and Miskin (1982, p.26), and it is natural to look at two-stage estimators as an

---

[1] Strictly speaking, it is not necessary that a likelihood be maximized, provided the function used acts like a likelihood in the sense that conditions (a) - (d) of section 2 hold for it. Newey (1984) for example works with a general function as does Liang (1984) and Pierce (1982). Each of these articles contains some of the theorems given later, as does Durbin (1970), although without the presence of the parameters γ.

alternative. Problems with a loss of efficiency and difficulties in computing the covariance matrix for two-stage methods motivate the development of an estimator asymptotically equivalent to the maximum likelihood estimator (MLE), but which can be computed as a double length regression.

Section 5 uses the framework established in section 2 to set up diagnostic test statistics for certain mis-specifications, and to show how many of these test statistics are effectively based upon two-stage estimators. Consequently, they are not unique, and this section explores some of the choices for α and γ that appear in the literature.

Section 6 of the paper turns to applications of two-stage estimators that arise in the analysis of censored data. It is not uncommon in that area for models to exhibit a variable expressed as a "normal" or "expected" value, and therefore two-stage estimators are frequently advocated. Even when anticipated quantities are not present however, a staged approach is sometimes advanced as a simple solution to the problem of consistent estimation in the presence of incidental parameters; most notably by Heckman (1976). Accordingly, section 6 provides an analysis of how Heckman's solution relates to the more conventional two-stage procedure discussed elsewhere in the paper.

2. Two-Stage Estimator Theory

Denote $\theta = (\alpha'\beta'\gamma')'$ and let $\hat{\theta}$ be the MLE of θ. Standard MLE properties are assumed to hold (including global identifiability), and we will particularly require

(a) $T^{-1}\,\partial L/\partial\theta = T^{-1}d_\theta \xrightarrow{p} 0$

(b) $T^{-\frac{1}{2}}d_\theta \xrightarrow{\mathcal{D}} N(0, I_{\theta\theta})$ where $I_{\theta\theta} = \underset{T\to\infty}{\text{plim}}\, T^{-1}(-\partial^2 L/\partial\theta\partial\theta') = \underset{T\to\infty}{\text{plim}}\, -T^{-1}H_{\theta\theta}$

(c) $I_{\theta\theta}$ is non-singular

(d) $T^{\frac{1}{2}}(\hat{\theta}-\theta_o) \xrightarrow{\mathcal{D}} N(0, I_{\theta\theta}^{-1})$ where θ_o is the true value of θ and is an interior point of a compact set.

From the definitions above H is the Hessian of the log likelihood and $I_{\theta\theta}$ will be the asymptotic information matrix. It will also be convenient to assume that, for any estimator θ^* of θ converging in probability to $\underline{\theta}$, $-T^{-1}H(\theta^*) \xrightarrow{p} I(\underline{\theta})$. This last restriction enables the employment of the mean value theorem and certainly holds for the models considered in this paper.

Consistency of the two-stage estimator is established as Theorem 1.

Theorem 1: *With* γ *prescribed to* $\bar{\gamma}$, $\tilde{\alpha}$ *a root-*β *consistent estimator of* α, *and* $\tilde{\beta} = \max_\beta [L(\beta,\tilde{\alpha},\bar{\gamma})]$, $\tilde{\beta} \xrightarrow{p} \beta_o$ *iff* $I_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma})(\bar{\gamma}-\gamma_o) = 0$.

Proof: $\tilde{\beta}$ is the solution to the first order conditions $d_\beta(\tilde{\beta},\tilde{\alpha},\bar{\gamma}) = 0$. Expanding $d_\beta(\beta_o,\alpha_o,\gamma_o)$ around $\beta = \tilde{\beta}$, $\alpha = \tilde{\alpha}$ and $\gamma = \bar{\gamma}$ by the mean value theorem gives

$$d_\beta(\theta_o) = d_\beta(\tilde{\beta},\tilde{\alpha},\bar{\gamma}) + H_{\beta\beta}(\theta^*)(\beta_o-\tilde{\beta}) + H_{\beta\gamma}(\theta^*)(\gamma_o-\bar{\gamma}) + H_{\beta\alpha}(\theta^*)(\alpha_o-\tilde{\alpha}) \quad (1)$$

where θ^* lies between θ_o and $(\tilde{\beta},\tilde{\alpha},\bar{\gamma})$.

$$T^{-1}d_\beta(\theta_o) = I_{\beta\beta}(\underline{\theta})(\tilde{\beta}-\beta_o) + I_{\beta\gamma}(\underline{\theta})(\bar{\gamma}-\gamma_o) + I_{\beta\alpha}(\underline{\theta})(\tilde{\alpha}-\alpha_o) + o_p(1). \quad (2)$$

Necessity is shown as follows. Suppose $\tilde{\beta}$ is consistent. Then $\beta^* \xrightarrow{p} \beta_o$, $\alpha^* \xrightarrow{p} \alpha_o$ and therefore $I_{\beta\gamma}(\underline{\theta}) = I_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma})$. Since $I_{\beta\beta}$ is non-singular, $\tilde{\alpha} \xrightarrow{p} \alpha_o$, and $T^{-1}d_\beta(\theta_o) \xrightarrow{p} 0$ by (a), $I_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma})(\bar{\gamma}-\gamma_o)$ must be zero. Sufficiency is obvious. □

Theorem 1 shows that a fairly weak set of conditions are needed for consistency of a two-stage estimator, and by careful attention to design can generally be made to hold. But, even if $\tilde{\alpha}$ is not consistent, it is sometimes possible for $\tilde{\beta}$ to be consistent, and an interesting extension of Theorem 1 is provided in Theorem 2.

Theorem 2: *If* γ *is prescribed to* $\bar{\gamma}$ *and* $\tilde{\alpha}-\alpha_o \xrightarrow{p} \phi$, $\tilde{\beta}$ *is a consistent estimator of* β *if*

$$I_{\beta\alpha}(\underline{\beta},\alpha_o+\phi,\bar{\gamma})\phi + I_{\beta\gamma}(\underline{\beta},\alpha_o+\phi,\bar{\gamma})(\bar{\gamma}-\gamma_o) = 0 \quad (3)$$

for any $\underline{\beta}$.

Proof: Follows directly from (2) by observing that $\tilde{\alpha} \xrightarrow{p} \alpha_o+\phi$ and $\underline{\theta}' = (\underline{\beta},\alpha_o+\phi,\bar{\gamma})$. Independence of (3) from $\underline{\beta}$ is important to allow $\tilde{\beta}$ to potentially have any limit. □

The requirements of Theorem 2 are quite severe and unlikely to hold very widely. Nevertheless, the expectations literature contains some models for which (3) is valid, and the theorem is therefore invoked in the next section.

Turning to the efficiency of the two-stage estimator of β, we wish

to isolate conditions under which it is fully efficient relative to the MLE estimator $\hat{\beta}$. Once again, the requirements depend upon the ways of eliminating nuisance parameters. Suppose initially that a consistent estimator is always used i.e. there are no γ. Theorem 3 provides some guidance on that case.

Theorem 3: *Assuming* $\tilde{\beta}$ *is consistent it is efficient relative to the MLE iff* $I_{\beta\alpha}(\beta_o,\alpha_o)T^{\frac{1}{2}}(\tilde{\alpha}-\hat{\alpha})$ is $o_p(1)$.

Proof: Expanding $d_\beta(\beta_o,\alpha_o)$ around $\hat{\beta}$, $\hat{\alpha}$ gives

$$T^{-\frac{1}{2}}d_\beta(\beta_o,\alpha_o) = T^{-\frac{1}{2}}d_\beta(\hat{\beta},\hat{\alpha}) + I_{\beta\beta}(\theta_o)T^{\frac{1}{2}}(\hat{\beta}-\beta_o) + I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\hat{\alpha}-\alpha_o) + o_p(1). \tag{4}$$

Similarly, expansion around $\tilde{\beta}$, $\tilde{\alpha}$ provides

$$T^{-\frac{1}{2}}d_\beta(\beta_o,\alpha_o) = T^{-\frac{1}{2}}d_\beta(\tilde{\beta},\tilde{\alpha}) + I_{\beta\beta}(\theta_o)T^{\frac{1}{2}}(\tilde{\beta}-\beta_o) + I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o) + o_p(1). \tag{5}$$

Subtracting (5) from (4)

$$I_{\beta\beta}(\theta_o)T^{\frac{1}{2}}(\hat{\beta}-\tilde{\beta}) + I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\hat{\alpha}-\tilde{\alpha}) + o_p(1) = 0. \tag{6}$$

Since $I_{\beta\beta}(\theta_o)$ is non-singular $T^{\frac{1}{2}}(\hat{\beta}-\tilde{\beta})$ is $o_p(1)$ iff $I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\tilde{\alpha}-\hat{\alpha})$ is $o_p(1)$. □

Theorem 3 throws up a number of sufficient conditions for $\tilde{\beta}$ to be efficient. Most importantly $I_{\beta\alpha}(\beta_o,\alpha_o) = 0$, which provides the justification for feasible GLS in the presence of exogenous regressors as well as Zellner's SUR estimator. This condition was emphasized by Stein (1956) and, more recently, by Manski (1985), although the latter also cites it as necessary. A less interesting situation, emphasized by Pierce (1982), is when $\tilde{\alpha}$ is fully efficient. But both of these are too severe, and Theorem 4 gives a weaker sufficient condition that is used in the following section.

Theorem 4: *If* $T^{\frac{1}{2}}(\tilde{\alpha}-\alpha) \xrightarrow{d} N(0,V_{\tilde{\alpha}})$ *with* $\tilde{\alpha}$ *and* $\hat{\alpha}$ *jointly normally distributed as* $T \to \infty$, *a sufficient condition for* $\tilde{\beta}$ *to be efficient is that* $A = V_{\tilde{\alpha}} I_{\alpha\beta} I^{-1}_{\beta\beta} - I^{-1}_{\alpha\alpha} I_{\alpha\beta} I^{\beta\beta}$ *equal zero, where* $I^{\beta\beta}$ *is the block matrix corresponding to* β *in* I^{-1}. *A necessary condition is that* $I_{\beta\alpha} A = 0$.

Proof: From Theorem 3 we require $I_{\beta\alpha} T^{\frac{1}{2}}(\tilde{\alpha}-\hat{\alpha}) \xrightarrow{P} 0$. With $\tilde{\alpha}$ and $\hat{\alpha}$ asymptotically jointly normally distributed, this holds iff $I_{\beta\alpha}(V_{\tilde{\alpha}}-I^{\alpha\alpha})I_{\alpha\beta} = 0$, using the result that the variance of $T^{\frac{1}{2}}(\tilde{\alpha}-\hat{\alpha})$ is the difference of the variances because $\hat{\alpha}$ is fully efficient - Hausman (1978).

$$\therefore \quad I_{\beta\alpha}(V_{\tilde{\alpha}}-I^{\alpha\alpha})I_{\alpha\beta} = I_{\beta\alpha}V_{\tilde{\alpha}}I_{\alpha\beta} - I_{\beta\alpha}I^{\alpha\alpha}I_{\alpha\beta}. \tag{7}$$

From $I^{-1}I = I$, $I^{\alpha\alpha}I_{\alpha\beta} + I^{\alpha\beta}I_{\beta\beta} = 0$ and (7) reduces to

$$I_{\beta\alpha}(V_{\tilde{\alpha}}I_{\alpha\beta}+I^{\alpha\beta}I_{\beta\beta}). \tag{8}$$

$$= I_{\beta\alpha}(V_{\tilde{\alpha}}I_{\alpha\beta}I^{-1}_{\beta\beta}+I^{\alpha\beta})I_{\beta\beta}. \tag{9}$$

From $II^{-1} = I$, $I_{\alpha\beta}I^{\beta\beta} + I_{\alpha\alpha}I^{\alpha\beta} = 0$ simplifying (9) to

$$I_{\beta\alpha}(V_{\tilde{\alpha}}I_{\alpha\beta}I^{-1}_{\beta\beta}-I^{-1}_{\alpha\alpha}I_{\alpha\beta}I^{\beta\beta})I_{\beta\beta} \tag{10}$$

$$= I_{\beta\alpha} A I_{\beta\beta}. \tag{11}$$

Sufficiency follows immediately from (11) as does necessity since $I_{\beta\beta}$ is strictly positive definite. □

Unfortunately, the introduction of a prescribed set of parameters renders simple extensions of Theorems 3 and 4 impossible. To see why, observe that $\hat{\beta}$ is based upon $\hat{\gamma}$, which is a consistent estimator of γ_o, whereas $\tilde{\beta}$ derives from $\bar{\gamma}$ which need not be. This means that the expansions corresponding to (4) and (5) differ both in the points at which I is evaluated and the fact that the two terms $\bar{\gamma} - \gamma_o$ and $\hat{\gamma} - \gamma_o$ do not differ

from one another by terms of order one in probability. A special case which is of some importance to expectations models is analysed in Theorem 5.

Theorem 5: *Let* $I_{\beta\beta}$ *be independent of* γ, $I_{\beta\gamma}(\theta_o) = 0$ *and assume that* $[I_{\beta\alpha}(\beta_o,\alpha_o,\bar{\gamma})-I_{\beta\alpha}(\theta_o)]T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o) - T^{-\frac{1}{2}}H_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma})(\bar{\gamma}-\gamma_o)$ *is* $o_p(1)$. *Then* $\tilde{\beta}$ *is efficient relative to the MLE iff* $I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\hat{\alpha}-\tilde{\alpha})$ *is* $o_p(1)$.

Proof: Under the assumptions the analogues of (4) and (5) in this expanded problem would be (ignoring terms of $o_p(1)$)

$$T^{-\frac{1}{2}}d_\beta(\theta_o) = I_{\beta\beta}(\theta_o)T^{\frac{1}{2}}(\hat{\beta}-\beta_o) + I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\hat{\alpha}-\alpha_o) \tag{12}$$

$$T^{-\frac{1}{2}}d_\beta(\theta_o) = I_{\beta\beta}(\theta_o)T^{\frac{1}{2}}(\tilde{\beta}-\beta_o) + I_{\beta\alpha}(\beta_o,\alpha_o,\bar{\gamma})T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o) - T^{-\frac{1}{2}}H_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma})(\bar{\gamma}-\gamma_o). \tag{13}$$

Subtracting (13) from (12) and using the non-singularity of $I_{\beta\beta}(\theta_o)$ shows that $\hat{\beta}$ is efficient relative to the MLE iff

$$I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\hat{\alpha}-\alpha_o) - I_{\beta\alpha}(\beta_o,\alpha_o,\bar{\gamma})T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o) + T^{-\frac{1}{2}}H_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma})(\bar{\gamma}-\gamma_o) \tag{14}$$

is $o_p(1)$. (14) can be re-arranged as

$$I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\hat{\alpha}-\tilde{\alpha}) + T^{-\frac{1}{2}}H_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma})(\bar{\gamma}-\gamma_o) + [I_{\beta\alpha}(\theta_o) - I_{\beta\alpha}(\beta_o,\alpha_o,\bar{\gamma})]$$
$$T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o). \tag{15}$$

With the sum of the last two terms being $o_p(1)$, $\hat{\beta}$ is efficient iff $I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\hat{\alpha}-\tilde{\alpha})$ is $o_p(1)$.

□

Theorem 5 therefore identifies a context in which Theorem 3 extends. One item of importance to note from Theorem 5 is that, whenever $\bar{\gamma} \neq \gamma_o$, it will be necessary that $T^{-\frac{1}{2}}H_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma})$ possess a limiting distribution if the limit distribution of $\tilde{\beta}$ is to be centred at β_o. In other cases the mis-specification involved in pre-setting γ at $\bar{\gamma} \neq \gamma_o$ will shift the mean of this limiting distribution away from β_o, as in the formal analysis provided by White (1982).

Two-stage estimators have a positive appeal for their computational simplicity but a negative aspect if the efficiency loss is too great. If the balance of the two forces swings too much in favour of the latter, it becomes worth while to consider a fully efficient estimator constructed in two steps. Theorem 6 provides a statement on this well-known estimator.

Theorem 6: *Define* $\hat{\theta}_{2SML} = \bar{\theta} + \bar{C}^{-1}\bar{c}$ *where* $\bar{C} \xrightarrow{p} I_{\theta\theta}$, $T^{-1}(\bar{c} - \bar{d}_\theta)$ *are* $o_p(T^{-\frac{1}{2}})$ *and* "-" *indicates evaluation at a root-T consistent estimator,* $\bar{\theta}$, *of* θ. *Then* $T^{\frac{1}{2}}(\hat{\theta}_{2SML} - \theta_o)$ *has the same limiting distribution as* $T^{\frac{1}{2}}(\hat{\theta} - \theta_o)$.

Proof: From the proof of Rothenberg and Leenders (1964), the asymptotically efficient two step estimator of θ is $\dot{\theta} = \bar{I}_{\theta\theta}^{-1}(T^{-1}\bar{d}_\theta) + \bar{\theta}$. Since $\bar{C} \xrightarrow{p} I_{\theta\theta}$ and $T^{-1}(\bar{d}_\theta - \bar{c})$ are $o_p(T^{-\frac{1}{2}})$, it is clear that $T^{\frac{1}{2}}(\hat{\theta}_{2SML} - \dot{\theta}) \xrightarrow{p} 0$. □

Perhaps the simplest way to compute $\hat{\theta}_{2SML}$ is to take one iteration of the Newton-Raphson algorithm for maximizing L, in which case $\bar{C} = -T^{-1}\bar{H}_{\theta\theta}$ and $\bar{c} = T^{-1}\bar{d}_\theta$, but sometimes substantial simplification is achieved through replacing $-T^{-1}\bar{H}_{\theta\theta}$ by $\bar{I}_{\theta\theta}$ or some other consistent estimator of it. Similarly, it is not always the case that the scores $\bar{d}_\theta$ are best. A good illustration of this last point occurs with Three Stage Least Squares, where the derivatives are replaced by quantities not affecting the limiting distribution but which produce an estimator that is much more easily computed.

Although no recourse is made to it in this paper, there may be some circumstances in which joint estimation of α and β is difficult, but estimation of each individually is relatively easy. For this purpose, Theorem 7 shows how an efficient estimator of β can be constructed in a staged fashion.

Theorem 7: *Let* $\bar{\alpha}$, $\bar{\beta}$ *be root-T consistent estimators of* α *and* β. *Define* $\alpha^{(1)} = \bar{\alpha} + \bar{I}^{-1}_{\alpha\alpha}\bar{d}_{\alpha}$, $\beta^{(1)} = I^{-1}_{\beta\beta}(\bar{\beta},\alpha^{(1)})d_{\beta}(\bar{\beta},\alpha^{(1)}) + \bar{\beta}$ *and* $\beta^{(2)} = \bar{\beta} + \bar{I}^{\beta\beta}\bar{I}_{\alpha\alpha}(\bar{\beta}-\beta^{(1)})$. *Then* $T^{\frac{1}{2}}(\beta^{(2)}-\beta_o)$ *has the same limiting distribution as* $T^{\frac{1}{2}}(\hat{\beta}-\beta_o)$.

Proof: The proof follows the same lines as Kohn (1978, p.623). Although he is concerned with ARMA processes in fact the result applies more generally.

□

Whether Theorem 7 is of great interest is hard to say; the fact that $\bar{I}^{\beta\beta}$ is required to weight together $\bar{\beta}$ and $\beta^{(1)}$ necessitates the (partitioned) inversion of $I_{\theta\theta}$, and therefore one might as well have taken one step of the Newton-Raphson algorithm. However, it is worth stating the result, as a number of applications of this staged approach appear in the time series literature; a fact made clear by Kohn.

The last issue to be dealt with concerns valid inferences. To make these about β it is necessary to establish the covariance matrix of $\tilde{\beta}$. But it is this need which tends to be ignored in most applications of two-stage estimators. Consistency of $\tilde{\beta}$ is frequently invoked as an implicit (or explicit) justification for the use of $I^{-1}_{\beta\beta}$ as the covariance matrix of $T^{\frac{1}{2}}(\tilde{\beta}-\beta_o)$. As evident from Theorem 1 however, consistency of $\tilde{\beta}$ follows almost automatically (if $\bar{\gamma} = \gamma_o$), yet the limiting distribution of $\tilde{\beta}$ was affected by the distribution of $\tilde{\alpha}$. Hence, it is rarely going to be the case that $I^{-1}_{\beta\beta}$ is the correct covariance matrix for $T^{\frac{1}{2}}(\tilde{\beta}-\beta)$, leading to invalid inferences if investigators act as if it were.

It turns out to be quite difficult to provide general expressions for the covariance matrix of a two-stage estimator without knowing how $\tilde{\alpha}$ was generated. Theorem 8 gives a result on the situation when the incidental parameters are all consistently estimated.

Theorem 8: *Suppose* $\begin{bmatrix} T^{\frac{1}{2}}(\tilde{\alpha}-\alpha) \\ T^{\frac{1}{2}}d_\beta \end{bmatrix} \xrightarrow{d} N\left(0, \begin{bmatrix} V_{\tilde{\alpha}} & V_1 \\ V_1' & I_{\beta\beta} \end{bmatrix}\right)$ *and that no nuisance parameters are prescribed. Then*

$$T^{\frac{1}{2}}(\tilde{\beta}-\beta) \xrightarrow{d} N(0, I_{\beta\beta}^{-1} - 2I_{\beta\beta}^{-1}I_{\beta\alpha}V_1 I_{\beta\beta}^{-1} + I_{\beta\beta}^{-1}I_{\beta\alpha}V_{\tilde{\alpha}}I_{\alpha\beta}I_{\beta\beta}^{-1}).$$

Proof: From (2) with $\bar{\gamma} = \gamma_o$,

$$T^{\frac{1}{2}}(\tilde{\beta}-\beta_o) = I_{\beta\beta}^{-1}(\theta_o)[T^{-\frac{1}{2}}d_\beta(\theta_o) - I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o)] + o_p(1) \qquad (16)$$

and the result follows directly. □

A combination of Theorems 3 and 8 highlights the fact that block diagonality of the information matrix provides a sufficient condition for both efficiency in estimation and valid inferences (more precisely diagonality under the null hypothesis for the latter). Additionally, it demonstrates that even when efficiency obtains, as in the context of Theorem 4, valid inferences are not assured, and the need for separate investigation of both questions is stressed.

Theorem 8 also reveals the importance of $\tilde{\alpha}$ to the covariance matrix of $\tilde{\beta}$. Depending on the covariance of $\tilde{\alpha}$ and d_β, V_1, a wide range of estimates of the covariance matrix of $\tilde{\beta}$ would be appropriate. When $V_1 = 0$, $I_{\beta\beta}^{-1}$ differs from the true covariance matrix by a positive semi-definite matrix $I_{\beta\beta}^{-1}I_{\beta\alpha}V_{\tilde{\alpha}}I_{\alpha\beta}I_{\beta\beta}^{-1}$, and so a directional result is possible concerning the error caused by incorrectly using $I_{\beta\beta}^{-1}$ as the covariance matrix of $T^{\frac{1}{2}}(\tilde{\beta}-\beta)$. One example of this is Theorem 8 in Pagan (1984, p.234) which arose out of Barro's (1977) use of a two-stage estimator. When $V_1 \neq 0$, few specific predictions can be made, a notable exception however being given as a corollary to Theorem 8.

Corollary 8.1: If $V_1 = I^{-1}_{\alpha\alpha}I_{\alpha\beta} + o_p(1)$ while $V_{\tilde{\alpha}} = I^{-1}_{\alpha\alpha}$, the asymptotic covariance matrix of $T^{\frac{1}{2}}(\tilde{\beta}-\beta)$ is $I^{-1}_{\beta\beta} - I^{-1}_{\beta\beta}I_{\beta\alpha}I^{-1}_{\alpha\alpha}I_{\alpha\beta}I^{-1}_{\beta\beta}$. □

This corollary is interesting mainly because the conditions occur in the derivation of Durbin's h-statistic for serial correlation in the presence of lagged dependent variables, and illustrates the possibility that $I^{-1}_{\beta\beta}$ may actually overstate standard errors.

Any difficulties faced by an investigator in getting the covariance matrix of $\tilde{\beta}$ becomes magnified once some of the incidental parameters are prescribed. Now it is almost impossible to present a general result. An important one for later development however is

Theorem 9: *Under the conditions of Theorem 5 the asymptotic covariance matrix of* $T^{\frac{1}{2}}(\tilde{\beta}-\beta)$ *is as set out in Theorem 8.*

Proof: From (15) and the statement of Theorem 5,

$$T^{-\frac{1}{2}}d_\beta(\theta_o) = I_{\beta\beta}(\theta_o)T^{\frac{1}{2}}(\tilde{\beta}-\beta_o) + I_{\beta\alpha}(\theta_o)T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o). \tag{17}$$

□

No doubt some other situations could be isolated where simplifications to the covariance matrix of $\tilde{\beta}$ eventuate, for example if $T^{-\frac{1}{2}}H_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma})$ had zero covariance with $T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o)$ and $T^{-\frac{1}{2}}d_\alpha$, but it scarcely seems worth setting these out in detail. On balance, one cannot help wondering if the nomination of values for γ is wise. There is advantage, in terms of an easier computation of the covariance matrix of $\tilde{\beta}$, in always replacing γ with a consistent estimator, although the effect of a switch from consistent estimation of γ to prescription of it can be to lose efficiency, and that may be regarded as undesirable.

3. Some Simple Models with Unobservables and Expectations

A very widespread application of two-stage estimators is to models containing unobserved variables, the most popular examples being those involving expectations. Normally the expectations (z_t^e) are assumed related to some set of observed variables. (18) and (19) below constitute the simplest representation of such models.

$$y_t = z_t^e\delta + e_t = (w_t\alpha)\delta + e_t \tag{18}$$

$$z_t = z_t^e + \eta_t = w_t\alpha + \eta_t \tag{19}$$

with e_t and η_t being bivariate normally distributed with zero mean and covariance matrix, $\text{diag}\{\sigma_e^2\sigma_\eta^2\}$. (18) and (19) have the obvious matrix form

$$y = (W\alpha)\delta + e \tag{20}$$

$$z = W\alpha + \eta \tag{21}$$

and W will be assumed to have characteristics such that the requirements (a) - (d) at the beginning of section 2 are met.

The log likelihood of (20) and (21) is

$$L = -T\log 2\pi - {}^T/_2\log\sigma_e^2 - {}^T/_2\log\sigma_\eta^2 - \sigma_e^{-2}e'e - \sigma_\eta^{-2}\eta'\eta. \tag{22}$$

Identifying β as δ, γ' as $(\sigma_e^2\sigma_\eta^2)$, we observe that, with $\alpha = \tilde{\alpha}$, (22) becomes (ignoring constants)

$$-{}^T/_2\log\sigma_e^2 - {}^T/_2\log\sigma_\eta^2 - \sigma_e^{-2}(y-W\tilde{\alpha}\delta)'(y-W\tilde{\alpha}\delta) - \sigma_\eta^{-2}\tilde{\eta}'\tilde{\eta}, \tag{23}$$

and so the two-stage estimator of $\beta = \delta$ is $\tilde{\beta} = (\tilde{\alpha}'W'W\tilde{\alpha})^{-1}\tilde{\alpha}'W'y$, showing it is invariant to σ_η^2 and σ_e^2. Accordingly, without loss of generality, γ can be prescribed to γ_o, the true value.

The two-stage estimator of β involves the regression of y against $\tilde{z}^e = W\tilde{\alpha}$, and the theorems in section 2 can be immediately employed to state and prove some propositions about it.

Proposition 3.1 (Pagan, 1984, Theorem 2): *The two-stage estimator of* δ *in* (20) *is asymptotically efficient if* $\tilde{\alpha}$ *is the OLS estimator formed by the regression of* z *against* W.

Proof: As $\gamma = \gamma_o$ Theorem 4 can be applied and a sufficient condition for the proposition to be valid is that

$$A = V_{\tilde{\alpha}} I_{\alpha\beta} I^{-1}_{\beta\beta} - I^{-1}_{\alpha\alpha} I_{\alpha\beta} I^{\beta\beta} \tag{24}$$

be zero. Pagan (1984, eq.(14)) gives $I^{-1}_{\alpha\alpha} = (\sigma_e^{-2}\delta^2+\sigma_\eta^{-2})^{-1}(W'W)^{-1}$ and $V_{\tilde{\alpha}} = \underset{T\to\infty}{\text{plim}}\ \sigma_\eta^2(T^{-1}W'W)^{-1} = \psi I^{-1}_{\alpha\alpha}$ where $\psi = \sigma_\eta^2(\sigma_e^{-2}\delta^2+\sigma_\eta^{-2}) = \sigma_e^{-2}(\sigma_\eta^2\delta^2+\sigma_e^2)$. Substituting into (24)

$$A = I^{-1}_{\alpha\alpha} I_{\alpha\beta}(\psi I^{-1}_{\beta\beta} - I^{\beta\beta}). \tag{25}$$

From Pagan (1984, eq.(12) and (17)), $I^{-1}_{\beta\beta} = \underset{T\to\infty}{\text{plim}}\ T\sigma_e^2(\alpha'W'W\alpha)^{-1}$ and $I^{\beta\beta} = \underset{T\to\infty}{\text{plim}}\ T(\sigma_\eta^2\delta^2+\sigma_e^2)(\alpha'W'W\alpha)^{-1}$ so that A is zero when $\psi\sigma_e^2 - (\sigma_\eta^2\delta^2+\sigma_e^2)$ is zero. But this follows directly from the definition of ψ above. □

As might be expected this efficiency result is restricted to a special set of circumstances. Proposition 2 identifies a simple variant of (20) for which it fails, namely when the equation is augmented by an extra set of regressors X which are neither orthogonal to W nor linear combinations of it.

Proposition 3.2 (Pagan, 1984, Theorem 4): *If* (20) *contains other regressors* X *such that* $T^{-1}X'W \neq 0$ *as* $T \to \infty$ *and* X *is not a linear combination of* W, *the two-stage estimator of* δ *and the coefficients of* X, $\tilde{\beta}$, *will not be efficient if* $\tilde{\alpha}$ *is generated as in Proposition 1.*

Proof: $V_{\tilde{\alpha}}$ and $I_{\alpha\alpha}$ remain the same as in Proposition 3.1 since neither is changed by the presence of extra regressors in (20). Therefore A is defined by (25) and, from Theorem 4, a necessary condition for efficiency is that $I_{\beta\alpha}A = 0$

$$\text{i.e.} \quad I_{\beta\alpha}I_{\alpha\alpha}^{-1}I_{\alpha\beta}(\psi I_{\beta\beta}^{-1} - I^{\beta\beta}) \tag{26}$$

is zero. When $T^{-1}X'W \xrightarrow{p} 0$, $I_{\beta\alpha} \neq 0$ and $I_{\beta\alpha}I_{\alpha\alpha}^{-1}I_{\alpha\beta} > 0$ because $I_{\alpha\alpha} > 0$. Hence (26) is zero only if $\psi I_{\beta\beta}^{-1} - I^{\beta\beta}$ equals zero. From the definitions of $(I^{\beta\beta})^{-1}$ and $I_{\beta\beta}$ in Pagan (1984, Appendix, eq.(1) - (4) and (7) - (9))

$$\psi^{-1}I_{\beta\beta} - (I^{\beta\beta})^{-1} = \text{diag}\{0, \sigma_e^{-4}(\sigma_e^{-2}\delta^2+\sigma_\eta^{-2})^{-1}(\operatorname{plim}_{T\to\infty} T^{-1}X'M_WX)\}$$

which equals zero only if $\operatorname{plim}_{T\to\infty} T^{-1}X'M_WX = 0$ i.e. $T^{-1}X'W \xrightarrow{p} 0$ or X a linear combination of W. □

Thus it is rare to find a fully efficient two-stage estimator, motivating the search for alternatives that are efficient and yet resemble the two-stage procedure in their computational aspects. Suppose (21) is substituted into (20), augmented with $X\beta_1$, to give

$$y = z\delta + X\beta_1 + u \tag{27}$$

and so obtain a two-equation (triangular) system of simultaneous equations composed of (21) and (27). One is then naturally drawn to the two-stage least squares estimator of $\delta, \tilde{\delta}_{2SLS}$. Obviously, whenever X does not appear in (27), $\tilde{\delta}_{2SLS} = \tilde{\delta}$ as the set of predetermined variables are just W. Otherwise, we have the following proposition.

Proposition 3.3 (Turkington (1984)): *When* $I_{\beta\alpha} \neq 0$, *the two-stage least squares estimator of* $\beta' = (\delta'\beta_1')$ *from* (21) *and* (27) *is less efficient than the two-stage estimator of Proposition 2.*

Proof: As the estimator of α used to form the two stage estimator of $\beta(\tilde{\alpha})$ is from the regression of z against W, while that to form $\tilde{\beta}_{2SLS}(\tilde{\alpha}_{2SLS})$ is from the regression of z against W and X, $V_{\tilde{\alpha}} < V_{\tilde{\alpha}_{2SLS}}$, where V represents the covariance matrix. As $V_1 = 0$ in both cases, Theorem 8 shows that

$$V_{\tilde{\beta}_{2SLS}} - V_{\tilde{\beta}} = I^{-1}_{\beta\beta} I_{\beta\alpha} (V_{\tilde{\alpha}_{2SLS}} - V_{\tilde{\alpha}}) I_{\alpha\beta} I^{-1}_{\beta\beta}. \tag{28}$$

Now $I_{\beta\alpha} \neq 0$, $V_{\tilde{\alpha}_{2SLS}} - V_{\tilde{\alpha}} > 0$, meaning that (28) is strictly positive definite

□

Having shown that 2SLS is less efficient than the two-stage estimator for the model (21) and (27), it is interesting to speculate on what would happen if three-stage least squares (3SLS) was employed. Proposition 3.4 states that 3SLS is asymptotically efficient, a conclusion reached by Wickens (1982, p. 60) and which he describes as somewhat surprising.

Proposition 3.4 (Turkington (1984), Wickens (1982)): *The 3SLS estimator of* β *in* (21) *and* (27) *is asymptotically efficient. It is a member of the class of asymptotically efficient two step MLE's of Theorem 6.*

Proof: Although tedious to do, it can be shown that the 3SLS estimator follows from the two step MLE of Theorem 6 by using a consistent estimator of the information matrix and replacing the scores $\bar{d}_\theta$ with terms that differ by $o_p(T^{\frac{1}{2}})$. Instead, more insight is available by demonstrating that the asymptotic covariance matrix of $\tilde{\beta}_{3SLS}$ is $I^{\beta\beta}$. To do this, define $G = (W\alpha \vdots X)$, and let $\Sigma = \{\sigma^{ij}\}$ be the covariance matrix of the errors in (21) and (27). Standard theory then gives $V_{\tilde{\beta}_{3SLS}}$ as

$$V^{-1}_{\tilde{\beta}_{3SLS}} = \operatorname*{plim}_{T\to\infty} T^{-1}[\sigma^{11} G'G - (\sigma^{12} G'W)[\sigma^{22} W'W]^{-1}(\sigma^{12} W'G)] \tag{29}$$

where σ^{ij} are the elements of Σ^{-1}. From inspection of u and η

$$\Sigma = \begin{pmatrix} \delta^2\sigma^2_\eta + \sigma^2_e & -\delta\sigma^2_\eta \\ -\delta\sigma^2_\eta & \sigma^2_\eta \end{pmatrix}$$

producing $\sigma^{11} = \sigma_e^{-2}$, $\sigma^{12} = \delta\sigma_e^{-2}$ and $\sigma^{22} = \sigma_\eta^{-2}\sigma_e^{-2}(\sigma_\eta^2\delta^2+\sigma_e^2)$. From partitioned inversion theory

$$(I^{\beta\beta})^{-1} = I_{\beta\beta} - I_{\beta\alpha}I_{\alpha\alpha}^{-1}I_{\alpha\beta} \tag{30}$$

and, substituting for the components in (30) from Pagan (1984, Appendix, eq.(1) - (6)),

$$(I^{\beta\beta})^{-1} = \operatorname*{plim}_{T\to\infty} T^{-1}(\sigma_e^{-2}G'G - G'W(W'W)^{-1}W'G\delta^2\sigma_e^{-4}(\delta^2\sigma_e^{-2}+\sigma_\eta^{-2})^{-1}). \tag{31}$$

Accordingly, $\tilde{V}_{3SLS}^{-1} = (I^{\beta\beta})^{-1}$ if

$$(\sigma^{12})^2(\sigma^{22})^{-1} = \delta^2\sigma_e^{-4}(\delta^2\sigma_e^{-2}+\sigma_\eta^{-2})^{-1}$$

which holds from the expressions for σ^{ij}. □

The 3SLS estimator is therefore perfectly efficient because it correctly weights together the information from both equations. But it is a result which is very specialized. Comparing (29) and (30) it is apparent that $I_{\alpha\alpha}$ must be a scalar multiple of $W'W$. If (say) z_{t-1}^e appeared in (20), $I_{\alpha\alpha} = \operatorname*{plim}_{T\to\infty} T^{-1}\sigma_e^{-2}[(W\delta_1+W_{-1}\delta_2)'(W\delta_1+W_{-1}\delta_2)] + \sigma_\eta^{-2}W'W$, where δ_1 and δ_2 are the coefficients of z_t^e and z_{t-1}^e respectively, and the cross term $W_{-1}'W$ means that $I_{\alpha\alpha}$ cannot be written in terms of $W'W$ unless $T^{-1}W_{-1}'W \xrightarrow{p} 0$ and stationarity of w_t is assumed. Consequently, as a method of achieving efficient estimation in a general context, 3SLS is inadequate[2]. The fact that it is an approximation to the efficient two step MLE does suggest however that it may be profitable to concentrate upon this latter option, and this is done in the following section.

---

[2] For the same reason, Proposition 3.1 fails to hold if z_t^e entered into (18) in a non-linear function. With the function as $h(z_t^e)$, $I_{\alpha\alpha}$ will involve $\delta^2\Sigma h_{1,t}w_t'w_t$ where $h_{1,t} = \partial h/\partial z_t^e$, and this becomes a multiple of $W'W$ only if $h_{1,t}$ is not a function of t.

A final proposition concerning this simple model is of interest.

Proposition 3.5: *If* $\delta \neq 0$, *the asymptotic "t-statistics" for* $\tilde{\delta}$ *obtained from the output provided with the regression of* y *against* $W\tilde{\alpha}$ *are higher than those based upon the true standard errors associated with* $\tilde{\delta}$.

Proof: $d_\delta = \alpha'W'e$, $\tilde{\alpha} - \alpha = (W'W)^{-1}W'\eta$ so that $V_1 = 0$. Applying Theorem 8, the standard errors found from $I^{-1}_{\beta\beta}$ are smaller than the true values whenever $I_{\delta\alpha} \neq 0$. But $I_{\delta\alpha} = \underset{T\to\infty}{\text{plim}}\, T^{-1}\alpha'W'W\delta \neq 0$ if $\delta \neq 0$. □

Proposition 3.5 is mentioned explicitly in Pagan (1984) and Newey (1984) but has been obviously recognised by a number of other researchers in the area of expectations e.g. McCallum (1979).

In all of the discussion above, the error terms e_t and η_t were assumed uncorrelated, and some attention needs to be paid to how the preceding propositions are modified once this restriction is relaxed. Let the covariance matrix of $[e_t \vdots \eta_t]'$ be Ω with elements $\{\omega_{ij}\}$. Its inverse Ω^{-1} has elements $\{\omega^{ij}\}$. Under the new specification the log likelihood corresponding to (22) is

$$L = -T/2 \log 2\pi - T/2 \log |\Omega| - \tfrac{1}{2}(e'\eta')\Omega^{-1}\begin{pmatrix} e \\ \eta \end{pmatrix} \qquad (32)$$

and now features a cross term between e and η. As the two-stage estimator minimizes

$$-T/2 \log(\omega^{11}) - \tfrac{1}{2}\omega_{11}^{-1}(y-W\tilde{\alpha}\delta)'(y-W\tilde{\alpha}\delta), \qquad (33)$$

it is clear that ω^{11} and ω^{22} can be assumed equal to their true value while ω^{12} has been set to zero. Two elements of γ are therefore γ_o while one is $\bar{\gamma} = 0 \neq \gamma_o$. Because of this fact Theorem 3 is not immediately appropriate for analysing questions of efficiency; recourse must be had to its extension in Theorem 5. Proposition 3.6 elicits the implications of this specification error for Proposition 3.1.

Proposition 3.6: *Proposition 3.1 remains valid when* $\omega^{12} \neq 0$.

Proof: First the condition set out in Theorem 5 is verified. That is

$$[I_{\delta\alpha}(\alpha_o,\beta_o,\bar{\gamma}) - I_{\delta\alpha}(\theta_o)]T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o) - T^{-\frac{1}{2}}H_{\delta\gamma}(\beta_o,\alpha_o,\bar{\gamma})(\bar{\gamma}-\gamma_o) \qquad (34)$$

is $o_p(1)$. As there is only one non-zero element in $\bar{\gamma}-\gamma_o$, and this corresponds to ω^{12}, (34) is evaluated directly only for that parameter. This is done by substituting $I_{\delta\alpha}(\theta_o) = \omega^{12}\alpha_o'C$, $I_{\delta\alpha}(\alpha_o,\beta_o,\bar{\gamma}) = \bar{\omega}^{12}\alpha_o'C = 0$, $T^{\frac{1}{2}}(\tilde{\alpha}-\alpha_o) = C^{-1}T^{-\frac{1}{2}}W'\eta$, $T^{-\frac{1}{2}}H_{\delta\gamma}(\beta_o,\alpha_o,\bar{\gamma}) = T^{-\frac{1}{2}}\alpha_o'W'\eta$ and $\bar{\gamma}-\gamma_o = -\omega^{12}(C = \operatorname{plim}_{T\to\infty} T^{-1}W'W)$ into (34) to produce

$$(-\omega^{12}\alpha_o'C)(C^{-1}T^{-\frac{1}{2}}W'\eta) + T^{-\frac{1}{2}}\omega^{12}\alpha_o'W'\eta = 0. \qquad (35)$$

Because of (35) Theorem 4 may now be invoked. Using the same argument as in Proposition 3.1 it is sufficient to show that $\psi I_{\beta\beta}^{-1} - I^{\beta\beta} = 0$, where $V_{\tilde{\alpha}} = \psi I_{\alpha\alpha}^{-1}$. From Turkington (1984, Appendix),

$$I_{\alpha\alpha}^{-1} = \sigma_u^{-2}r^2C^{-1}, \quad I_{\beta\beta}^{-1} = \sigma_\eta^{-2}r^2(\alpha_o'C\alpha_o)^{-1} \quad \text{and} \quad I^{\beta\beta} = \sigma_u^2(\alpha_o'C\alpha_o)^{-1}$$

where $r = \sigma_\eta^2\sigma_e^2 - \sigma_{e\eta}^2$ and $\sigma_u^2 = \sigma_\eta^2\delta^2 + \sigma_e^2$. Thus $\psi = \sigma_\eta^2\sigma_u^2/r^2$ and $\psi I_{\beta\beta}^{-1} - I^{\beta\beta} = 0$. □

Pagan (1984) comments that Proposition 3.1 holds even if the errors were correlated and here a formal proof has been provided. Furthermore, because the condition in Theorem 5 holds, it immediately follows that the covariance matrix of the two-stage estimator is as stated in Theorem 8.

A number of propositions are now invalidated. Because $d_\beta = \omega^{11}G'e$ and $\tilde{\alpha}-\alpha_o = (W'W)^{-1}W'\eta$, it is clear that $V_1 \neq 0$ whenever $\omega_{12} \neq 0$; a feature which demands the reconsideration of Proposition 3.3 in Proposition 3.7.

Proposition 3.7 (Turkington (1984): *Under the same conditions as in Proposition 3.3, but with the disturbances of* e *and* η *being correlated, the two-stage least squares estimator may be as efficient as the two-stage estimator.*

Proof: The correlation between d_β and either $(\tilde{\alpha}_{2SLS}-\alpha_o)$ or $(\tilde{\alpha}-\alpha_o)$ is now non-zero. As $V_{\tilde{\beta}_{2SLS}}$ and $V_{\tilde{\beta}}$ depend upon this correlation (Theorem 8) the 2SLS estimator can be more efficient than the two stage estimator. Turkington gives an exact relation for Proposition 3.3 to hold. □

Proposition 3.5 suffers a similar fate, for the same reason. Pagan's (1984) directional result for the t-statistics in Barro's (1977) model, which effectively has η and η_{-1} as regressors in X, loses its force when η and e are uncorrelated. But, a lack of correlation is needed in such a model to preserve consistency of estimators. To see this, observe that the covariance between e and η, γ_o, has been invalidly prescribed at $\bar{\gamma} = 0$. From Theorem 1 $\tilde{\beta}$ is consistent only if $I_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma}) = 0$. For that element of β corresponding to the regressor $\hat{\eta} = M_W z$, $I_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma}) = \sigma_e^{-2}E(\eta e') = \sigma_e^{-2}\sigma_{e\eta} \neq 0$. As Pudney (1982) pointed out, this lack of diagonality in the information matrix also characterized Sargent's (1976) model.

One final proposition finishes this section, and it concerns the consistency of $\tilde{\beta}$ when (21) is mis-specified, in the sense that extra terms $\bar{W}$ should appear in it[3].

[3] In this proposition it is convenient to assume, without loss of generality, that σ_η^2 and σ_e^2 are known so that γ can be identified exclusively with $\bar{W}$.

Proposition 3.8 (Attfield (1984), Wickens (1982)): *If* $z^e = W\alpha + \bar{W}\gamma$ *in* (20) *and* (21), *but* γ *is prescribed to* $\bar{\gamma} = 0$ *in estimation, the two-stage estimator of* $\beta = \delta$ *formed by solving* $\max_{\beta} L(\beta, \tilde{\alpha}, \bar{\gamma})$, *where* $\tilde{\alpha} = (W'W)^{-1}W'z$, is consistent.

Proof: As $\tilde{\alpha}$ is inconsistent, Theorem 2 needs to be utilized. It is easily checked that $I_{\beta\alpha} = T^{-1}\sigma_e^{-2}(W\alpha+\bar{W}\gamma)'W\beta + o_p(1)$ and $I_{\beta\gamma} = T^{-1}\sigma_e^{-2}(W\alpha+\bar{W}\gamma)'\bar{W}\beta + o_p(1)$ yielding the equivalences $I_{\beta\alpha}(\alpha_o+\phi, \bar{\gamma}, \underline{\beta}) = T^{-1}\sigma_e^{-2}(\alpha_o+\phi)'W'W\underline{\beta}$ and $I_{\beta\gamma}(\alpha_o+\phi, \bar{\gamma}, \underline{\beta}) = T^{-1}\sigma_e^{-2}(\alpha_o+\phi)'W'\bar{W}\underline{\beta} + o_p(1)$. Now $\phi = (W'W)^{-1}W'\bar{W}\gamma_o + o_p(1)$ so that

$$T^{-1}W'\bar{W}\gamma_o = (T^{-1}W'W)\phi + o_p(1). \tag{36}$$

Using (36),

$$\begin{aligned} I_{\beta\gamma}(\alpha_o+\phi, \bar{\gamma}, \underline{\beta})(\bar{\gamma}-\gamma_o) &= -\sigma_e^{-2}(\alpha_o+\phi)'(T^{-1}W'\bar{W})\gamma_o\underline{\beta} + o_p(1) \\ &= -\sigma_e^{-2}(\alpha_o+\phi)'(T^{-1}W'W)\phi\underline{\beta} + o_p(1), \end{aligned} \tag{37}$$

while

$$I_{\beta\alpha}(\alpha_o+\phi, \bar{\gamma}, \underline{\beta})\phi = \sigma_e^{-2}(\alpha_o+\phi)'(T^{-1}W'W)\phi\underline{\beta} + o_p(1). \tag{38}$$

Inspection shows that (37) + (38) equals zero for any $\underline{\beta}$, and Theorem 2 gives this as the requirement for $\tilde{\beta}$ to be consistent. □

Wickens described Proposition 3.8 by observing that $\tilde{\beta}$ is the 2SLS estimator with a truncated set of predetermined variables, and truncation does not affect consistency in this special model. Extension to other models is possible. Provided a stationarity assumption is made about w_t, $\tilde{\beta}$ will remain consistent if (20) had z^e_{t-j} in place of z_t; the crucial element in the proof of Proposition 3.8 being that $\phi = \operatorname{plim}_{T\to\infty}(W'_{-j}W_{-j})^{-1}W'_{-j}\bar{W}_{-j}\gamma_o$. One must have some doubts about the importance of Proposition 3.8 however. Exclusive concentration upon the first moment hides the fact that

misspecification will normally invalidate inferences, just as in Theorem 8, and Attfield's enthusiasm for the estimator seems a bit misplaced. Nevertheless, there are instances where inferences would be correct, notably when testing if $\beta = 0$, and in these circumstances it can be a useful result.

4. More Complex Models with Expectations

In many applications models are more complex than those analysed in the preceding section. This complexity primarily stems from the presence of leads and lags in expectations; that is, from the presence of terms such as z_t^e or z_{t-1}^e. Many of the propositions established in section 3 relied heavily upon the fact that only "current" anticipations appeared in the model, particularly those that managed to derive an efficient estimator of β using an inefficient estimator of α. When one departs from these simple structures, it will rarely be possible to obtain efficient two-stage estimators of β without a corresponding estimator of α. Furthermore, the covariance matrix of any two stage estimator will need to be computed separately as in Theorem 8; a fact highlighted for Barro's model by Mishkin (1982), Newey (1984) and Pagan (1984).

Both of these features emphasise the need for an efficient estimator whose covariance matrix is easily computed. It has not proved easy to maximize the likelihood for some of these models, and frequently the "maximum likelihood" estimates are presented without standard errors - see Sargent (1978, p. 1023) or Eckstein (1984, p. 14) for example. Because the two-step MLE of Theorem 6 is asymptotically efficient, it appeals as a potentially profitable approach, and is exploited below.

(21) and (20) are now replaced by

$$y = g(X;\beta,\alpha) + e \tag{39}$$

$$z = W\alpha + \eta \tag{40}$$

where $g(X;\beta,\alpha)$ indicates some function of X involving both β and α. X may or may not include W. (39) specializes to (20) when $g(X;\beta,\alpha) = (W\alpha)\delta$; to (27) when $g(X;\beta,\alpha) = (W\alpha)\delta + X\beta$; to models with a lagged expectation when $g(X;\beta,\alpha) = (W_{-1}\alpha)\delta$; and to future expectations when $g(X;\beta,\alpha) = W_{-1}\psi(\alpha,\beta)$, with ψ a vector valued non-linear function of α and β - see Eckstein (1984, p. 13) for an example of ψ. It is worth indicating that (39) and (40) have even wider application. One obvious use is to model cases when the logarithm of expectations are involved. Then $g(X;\beta,\alpha) = \log((W\alpha)\delta)$.

Proposition 4.1 below describes the computation of an asymptotically efficient two-step estimator of β and α. Initial consistent estimators of β and α are available from the two-stage estimators described earlier. For convenience it is assumed that $\gamma' = (\sigma_\eta^2 \sigma_e^2)$ are set at their true values. This is a costless assumption for the same reason as in Proposition 3.1. Initially e and η are assumed uncorrelated.

Proposition 4.1: *Let* $\theta' = (\beta'\alpha')$ *be a* $K \times 1$ *vector,* G_θ *be the* $T \times K$ *matrix of derivatives of* $g(X;\theta)$ *with respect to* θ *and define* $M_\theta = (0 \vdots W)$. *With* $\tilde{\theta}$ *the consistent two-stage estimator of* θ, *the asymptotically efficient two step estimator of* θ *is* $\tilde{\theta}_{2SML} = \tilde{\theta} + \Delta\theta$, *where* $\Delta\theta$ *are the coefficient estimates from the double-length regression that has*

$$\begin{bmatrix} \tilde{\sigma}_e^{-1} \tilde{G}_\theta \\ \tilde{\sigma}_\eta^{-1} \tilde{M}_\theta \end{bmatrix}$$

as independent and

$$\begin{bmatrix} \tilde{\sigma}_e^{-1} \tilde{e} \\ \tilde{\sigma}_\eta^{-1} \tilde{\eta} \end{bmatrix}$$

as dependent variables (a tilde indicating evaluation at $\tilde{\theta}$*).*

Proof: The log likelihood is (omitting constants)

$$-T/2 \log \sigma_e^2 - T/2 \log \sigma_\eta^2 - \tfrac{1}{2}\sigma_e^{-2}(y-g(X;\theta))'(y-g(X;\theta)) - \tfrac{1}{2}\sigma_\eta^{-2}(z-W\alpha)'(z-W\alpha), \tag{41}$$

from which

$$I_{\theta\theta} = T^{-1}[\sigma_e^{-2}G_\theta'G_\theta + \sigma_\eta^{-2}M_\theta'M_\theta] \quad \text{and}$$

$$d_\theta = \sigma_e^{-2}G_\theta'e + \sigma_\eta^{-2}M_\theta'\eta.$$

Applying Theorem 6 with $\tilde{I}_{\theta\theta}$ and $\tilde{d}_\theta$ as $\bar{C}$ and $\bar{c}$ respectively,

$$\tilde{\theta}_{2SML} = \tilde{\theta} + (\tilde{\sigma}_e^{-2}\tilde{G}_\theta'\tilde{G}_\theta + \tilde{\sigma}_\eta^{-2}\tilde{M}_\theta'\tilde{M}_\theta)^{-1}(\tilde{\sigma}_e^{-2}\tilde{G}_\theta'\tilde{e} + \tilde{\sigma}_\eta^{-2}\tilde{M}_\theta'\tilde{\eta}). \tag{42}$$

Defining $\Phi' = (\tilde{\sigma}_e^{-1}\tilde{G}_\theta' \vdots \tilde{\sigma}_\eta^{-1}\tilde{M}_\theta')$, $\phi' = (\tilde{e}'\tilde{\eta}')$,

$$\tilde{\theta}_{2SML} = \tilde{\theta} + (\Phi'\Phi)^{-1}\Phi'\phi \tag{43}$$

showing that $\Delta\theta = \tilde{\theta}_{2SML} - \tilde{\theta}$ comes from a double-length regression. □

Proposition 4.1 demonstrates how the asymptotically efficient estimator of θ is to be obtained. Its actual covariance matrix is I^{-1}, whereas a regression program would provide the estimate $\hat{\sigma}^2\tilde{I}^{-1}$, $\hat{\sigma}^2$ being formed out of residuals from the double length regression. $\hat{\sigma}^2 \xrightarrow{P} 1$ but, if desired, one could re-scale this output by the factor $\hat{\sigma}^{-2}$.

Although many applications are characterized by an assumption of uncorrelated errors in the two equations, this restriction may well be too strong when models contain only anticipated values. Nevertheless, Proposition 4.1 extends to this more complex case, albeit with slightly greater difficulty in implementation. The extension relies upon the fact that the covariance matrix of $(e'\eta')'$, Ω, can be factored via the Cholesky decomposition as $\Omega^{-1} = P'P$.

Proposition 4.2: *With the same definitions as in Proposition 4.1, but with* $E(e\eta') \neq 0$ *and* $P = \{P_{ij}\}$ $i,j = 1,2$, *the asymptotically efficient estimator of* θ *is* $\tilde{\theta}_{2SML} = \tilde{\theta} + \Delta\theta$, *where* $\Delta\theta$ *are the coefficient estimates from the double-length regression that has*

$$\begin{pmatrix} \tilde{P}_{11}\tilde{G}_\theta + \tilde{P}_{12}\tilde{M}_\theta \\ \tilde{P}_{21}\tilde{G}_\theta + \tilde{P}_{22}\tilde{M}_\theta \end{pmatrix}$$

as independent and

$$\begin{pmatrix} \tilde{P}_{11}\tilde{e} + \tilde{P}_{12}\tilde{\eta} \\ \tilde{P}_{21}\tilde{e} + \tilde{P}_{22}\tilde{\eta} \end{pmatrix}$$

as dependent variable.

Proof: The log-likelihood is (omitting constants)

$$-T/2 \log\Omega - \tfrac{1}{2}[y-G(X;\theta) \vdots z-W\alpha]'\Omega^{-1}\begin{bmatrix} y-g(X;\theta) \\ z-W\alpha \end{bmatrix} \qquad (44)$$

from which, using $\Omega^{-1} = P'P$ and $\tilde{A}' = (\tilde{G}_\theta' \vdots \tilde{M}_\theta')$,

$$T.\tilde{I}_{\theta\theta} = \tilde{A}'\tilde{\Omega}^{-1}\tilde{A} = \tilde{A}'\tilde{P}'\tilde{P}\tilde{A}$$

$$\tilde{d}_\theta = \tilde{A}'\tilde{\Omega}^{-1}\begin{pmatrix} \tilde{e} \\ \tilde{\eta} \end{pmatrix} = \tilde{A}'\tilde{P}'\tilde{P}\begin{pmatrix} \tilde{e} \\ \tilde{\eta} \end{pmatrix}.$$

It follows immediately from the definitions of $\tilde{P}$ and $\tilde{A}$ that $\Delta\theta$ will be the coefficients from the regression described in the proposition. □

Computation of the Cholesky decomposition of Ω^{-1} is a routine task on most computers. When only two equations are involved it can be done easily with a hand calculator using the formulae in (say) Graybill (1969, p.299). Initial consistent estimators of Ω are available from the two-stage estimators of β and α; obvious extensions of lemma 1 in Pagan (1984) demonstrate the consistency of these estimators.

A better appreciation of the two propositions above may be obtained by noting that (39) and (40), linearized around $\tilde{\theta}$, are

$$\tilde{e} \simeq \tilde{G}_{\alpha}(\alpha-\tilde{\alpha}) + \tilde{G}_{\beta}(\beta-\tilde{\beta}) + e \qquad (39a)$$

$$\eta \simeq W(\alpha-\tilde{\alpha}) + \eta \ , \qquad (40a)$$

where the approximation involves omitting terms that do not affect the limiting distribution of $T^{\frac{1}{2}}(\tilde{\theta}_{2SML} - \theta_o)$. One possible way of generating $\Delta\theta$ would be to apply Zellner's SUR estimator to (39a) and (40a), but there are numerical difficulties owing to the fact that OLS on (39a) fails as $(\tilde{G}_{\alpha}\ \tilde{G}_{\beta})$ is not of full column rank. Thus, the traditional first step of the SUR procedure (designed to consistently estimate the covariance matrix of the errors e and η), cannot be performed. However, if (39a) and (40a) are divided by $\tilde{\sigma}_e^{-1}$ and $\tilde{\sigma}_{\eta}^{-1}$ respectively, the errors in these normalized equations would have the same variance of unity. Consequently, once (39a) and (40a) have been re-formatted, OLS may be applied by pooling both equations together. This is the rationale for the two propositions advanced above, and it makes clear how extensions to a multi-equation environment are possible.

5. Diagnostic Tests

Perhaps in no other area of econometrics is the need to eliminate nuisance parameters more pressing than in the construction of diagnostic tests; this literature even giving rise, in the form of Durbin (1970), to what is arguably the best treatment of the topic available in the econometric literature. Diagnostic tests are frequently based upon two-stage estimators, and their variety can be substantially accounted for by the different ways in which the two-stage estimator is generated.

Some general results can be given first.

Proposition 5.1: *Let* $\tilde{\alpha}$ *be a root-T consistent of* α *under the null hypothesis* H_o: $\beta = \beta_o$ *and define* $\beta_{(n)}$ *as the estimate of* β *from the n'th iteration of the scoring algorithm applied to solve the program* $\max_{\beta} L(\beta,\tilde{\alpha})$. *When* $\beta_{(o)} = \beta_o$, *a test of the null hypothesis based on* $\beta_{(1)}$ *is asymptotically equivalent to testing if the scores* $d_\beta(\beta_o,\tilde{\alpha})$ *are zero.*

Proof:
$$T^{\frac{1}{2}}(\beta_{(n)}-\beta_{(n-1)}) = I^{-1}_{\beta\beta}(\beta_{(n-1)},\tilde{\alpha})T^{-\frac{1}{2}}d_\beta(\beta_{(n-1)},\tilde{\alpha}) \tag{45}$$

Let $n = 1$ and note that $\tilde{\alpha} \xrightarrow{p} \alpha_o$ under H_o making $I(\beta_{(o)},\tilde{\alpha}) \xrightarrow{p} I(\beta_o,\alpha_o)$. Accordingly (45) becomes

$$T^{\frac{1}{2}}(\beta_{(1)}-\beta_o) = I^{-1}_{\beta\beta}(\beta_o,\alpha_o)T^{-\frac{1}{2}}d_\beta(\beta_o,\tilde{\alpha}) + o_p(1) . \tag{46}$$

As $I_{\beta\beta}(\beta_o,\alpha_o)$ is non-singular the proposition follows directly. □

Perhaps the most important implication of Proposition 5.1 arises when $\tilde{\alpha}$ is set at $\bar{\alpha} = \max_{\alpha} L(\beta_o,\alpha)$. The scores become $d_\beta(\beta_o,\bar{\alpha})$ and the test statistic based on this quantity is the score or Lagrange Multiplier (LM) test statistic for H_o: $\beta = \beta_o$. Thus the LM test may be interpreted as comparing the estimate of β obtained from one iteration of the scoring algorithm (or Newton-Raphson) to the log likelihood $L(\beta,\bar{\alpha})$[4]. Following the same method of proof as in Durbin (1970), and under the sequence of local alternatives $\beta_n = \beta_o + \psi/\sqrt{T}$, it is easily seen that $T^{\frac{1}{2}}(\beta_{(1)}-\beta_o)$ is asymptotically normally distributed with mean $I^{-1}_{\beta\beta}D\psi$ and variance $I^{-1}_{\beta\beta}D\,I^{-1}_{\beta\beta}$, where $D = I_{\beta\beta} - I_{\beta\alpha}I^{-1}_{\alpha\alpha}I_{\alpha\beta}$ (see Corollary 8.1 for the variance). Consequently, the non-centrality parameter of the associated χ^2 statistic is given by $\psi' D\, I^{-1}_{\beta\beta}I_{\beta\beta}\, D^{-1}\, I_{\beta\beta}I^{-1}_{\beta\beta}\, D\psi = \psi' D\psi$, which satisfies Neyman's (1959)

---

[4] Of course, there is no reason why further iterations should not be performed, as the limiting distribution of $T^{\frac{1}{2}}(\beta_{(n)}-\beta_o)$ is the same as $T^{\frac{1}{2}}(\beta_{(1)}-\beta_{(o)})$. When $L(\beta,\bar{\alpha})$ is quadratic in β, one iteration immediately yields the maximum.

criterion for a locally optimal test. What is most striking about this result is that a test statistic based on $\beta_{(1)}$ is asymptotically locally most powerful, despite the fact that $\beta_{(1)}$ is an inefficient estimator of β (unless $I_{\beta\alpha} = 0$). This contention can be easily verified by comparing $I_{\beta\beta}^{-1} D I_{\beta\beta}^{-1}$ with the variance of $T^{\frac{1}{2}}(\hat{\beta}-\beta_o)$, namely D^{-1}.

One might be interested in examining test statistics formed from efficient estimators of β. By far the best known of these would be the Wald test statistic $T(\hat{\beta}-\beta_o)'(I^{\beta\beta})^{-1}(\hat{\beta}-\beta_o)$ but, as a class of asymptotically efficient estimators of β was defined in Theorem 6, it is of some interest that, for a consistent $\tilde{\alpha}$, the test statistic based on $\hat{\beta}_{2SML}$ is identical to one known in the literature as the $C(\alpha)$ test. Proposition 5.2 demonstrates this last outcome.

Proposition 5.2: *Define* $\theta_{(n)} = \theta_{(n-1)} + I^{-1}(\theta_{(n-1)})T^{-1}d_\theta(\theta_{(n-1)})$ *and set* $\theta'_{(o)} = (\tilde{\alpha}'\beta'_o) = \tilde{\theta}$, *where* $\tilde{\alpha}-\alpha_o$ *is* $O_p(T^{-\frac{1}{2}})$ *under* H_o. *With this choice for* α, *the* $C(\alpha)$ *test of Neyman (1959) is* $T.C(\alpha) = (\tilde{d}_\beta - \tilde{I}_{\beta\alpha}\tilde{I}_{\alpha\alpha}^{-1}\tilde{d}_\alpha)'\tilde{I}^{\beta\beta}(\tilde{d}_\beta - \tilde{I}_{\beta\alpha}\tilde{I}_{\alpha\alpha}^{-1}\tilde{d}_\alpha)$, *where the tilde indicates evaluation at* $\tilde{\theta}$. *Then*

$$T.C(\alpha) = (\beta_{(1)}-\beta_o)'(\tilde{I}^{\beta\beta})^{-1}(\beta_{(1)}-\beta_o) .$$

Proof:

$$T.\begin{bmatrix} \beta_{(1)}-\beta_o \\ \alpha_{(1)}-\tilde{\alpha} \end{bmatrix} = \tilde{I}^{-1}\tilde{d}_\theta = \begin{bmatrix} \tilde{I}^{\beta\beta}\tilde{I}^{\beta\alpha} \\ \tilde{I}^{\alpha\beta}\tilde{I}^{\alpha\alpha} \end{bmatrix}\begin{bmatrix} \tilde{d}_\beta \\ \tilde{d}_\alpha \end{bmatrix} \tag{47}$$

$$\therefore \quad T.(\beta_{(1)}-\beta_o) = \tilde{I}^{\beta\beta}\tilde{d}_\beta + \tilde{I}^{\beta\alpha}\tilde{d}_\alpha \tag{48}$$

or

$$T.(\tilde{I}^{\beta\beta})^{-1}(\beta_{(1)}-\beta_o) = \tilde{d}_\beta - \tilde{I}_{\beta\alpha}\tilde{I}_{\alpha\alpha}^{-1}\tilde{d}_\alpha , \tag{49}$$

using the fact that $(\tilde{I}^{\beta\beta})^{-1}\tilde{I}^{\beta\alpha} + \tilde{I}_{\beta\alpha}\tilde{I}_{\alpha\alpha}^{-1} = 0$. (49) yields the desired result. □

The test statistic based on the estimator of β from the first step of the scoring algorithm above was termed the "indirect LM test" in Breusch and Pagan (1980), and reflected a failure to recognise it as identical to the asymptotically efficient $C(\alpha)$ test described elsewhere in that paper. In most instances it would probably be easier to compute the $C(\alpha)$ test by performing the iteration to obtain $\beta_{(1)}$ rather than explicitly computing it from the right hand side of (49).

Some general points about diagnostic tests have been sketched above, from which it should be apparent that a central issue in their construction is how to eliminate incidental parameters before making inferences about those of interest. It is worth using the framework of this paper to look at some of the vast literature on diagnostic tests for particular model deficiencies, so as to show how a general approach can order that literature. For brevity, we concentrate solely upon testing for serial correlation in the linear regression model, but the approach is much more widely applicable.

The linear regression model is

$$y_t = x_t\alpha + u_t \tag{50}$$

and it is desired to test if $\beta = 0$ in $u_t = \beta u_{t-1} + e_t$, where $e_t \sim$ n.i.i.d.$(0,\sigma^2)$. Because $I_{\alpha\sigma^2} = 0$, $I_{\beta\sigma^2} = 0$, σ^2 can be taken as known without any loss of generality. For a given $\tilde{\alpha}$ therefore, $\max_{\beta} L(\beta,\tilde{\alpha})$ produces $\tilde{\beta} = (\Sigma\tilde{u}_{t-1}^2)^{-1}\Sigma\tilde{u}_{t-1}\tilde{u}_t + o_p(1)$, where $\tilde{u}_t = y_t - x_t\tilde{\alpha}$. Clearly there are many test statistics for $\beta = 0$, depending upon the choice of α, and we proceed to classify some below.

As mentioned previously the best choice for $\tilde{\alpha}$, in terms of asymptotic local power, is $\bar{\alpha} = \max_{\alpha} L(\beta = 0,\alpha)$. For (50) this implies that $\bar{\alpha} = (\Sigma x_t'x_t)^{-1}\Sigma x_t y_t = \hat{\alpha}_{OLS}$. The LM test for serial correlation is therefore

based upon $\tilde{\beta}$ computed from the regression of the least squares residuals $\hat{u}_t = y_t - x_t\hat{\alpha}_{OLS}$ against their lagged values $\hat{u}_{t-1}$, using as covariance matrix that given in Theorem 8. But there are circumstances in which the same asymptotic power is available even if $\bar{\alpha}$ is not adopted. Theorem 3 showed that the limiting distribution of $\tilde{\beta}$ is independent of the choice of $\tilde{\alpha}$ if $I_{\beta\alpha} = 0$.[5] Obviously, if x_t is strictly exogenous, $I_{\beta\alpha} = 0$ and any choice for $\tilde{\alpha}$ (subject to the requirement of root-T consistency) suffices.

Apart from $\bar{\alpha}$ two other choices are featured in the literature when regressors are exogenous. The first derives from Berenblutt and Webb (1973). Consider first differencing (50) to produce $\dot{y}_t = \dot{x}_t\alpha + \dot{u}_t$ where $\dot{y}_t = y_t - y_{t-1}$ and $\dot{x}_t = x_t - x_{t-1}$. A regression of $\dot{y}_t$ against $\dot{x}_t$ provides a consistent estimator of α. The Berenblutt-Webb statistic is then formed from the first order serial correlation coefficient $(\Sigma\tilde{u}^2_{t-1})^{-1}\Sigma\tilde{u}_{t-1}\tilde{u}_t$ in which the residuals $\tilde{u}_t$ are not the OLS ones but are derived from the estimate of $\tilde{\alpha}$ coming from the differenced-data regression. The second is a generalization of this idea by King (1985). It relies upon quasi-differenced data $\dot{y}_t = y_t - \lambda y_{t-1}$, $\dot{x}_t = x_t - \lambda x_{t-1}$ to be the inputs to the regression generating $\tilde{\alpha}$. Clearly, since the asymptotic distribution of $\tilde{\beta}$ does not depend on $\tilde{\alpha}$, it is also independent of λ, and all test statistics constructed with differing (finite) values of λ have identical asymptotic power. Small sample considerations must therefore

---

[5] In the development of the theorems of section 2 (β_o,α_o) were the true values of β and α. Under a sequence of local alternatives these are replaced by β_T and α_o, where $\beta_T = \beta_o + T^{-\frac{1}{2}}\psi$, as the properties (a) and (b) of section 2 apply to $d_\theta(\beta_T,\alpha_o)$ rather than $d_\theta(\beta_o,\alpha_o)$.

dictate the selection of λ. King opts for a value of .5 or .75.[6]

Once x_t consists of variables that are not strictly exogenous, the situation changes; it is no longer a matter of indifference about which $\tilde{\alpha}$ is to be used. The score-test variant with $\bar{\alpha} = (\Sigma x_t' x_t)^{-1}\Sigma x_t' y_t$ is equivalent to Durbin's h statistic, and represents a benchmark for any power comparisons. Few alternative $\tilde{\alpha}$'s have been proposed. An interesting one has, however, been recently given by Inder (1984) for models in which x_t contains y_{t-1} and some z_t that are strictly exogenous. Partitioning $\tilde{\alpha}' = (\tilde{\alpha}_1'\tilde{\alpha}_2')$ to correspond to $x_t = (y_{t-1} \vdots z_t)$, he makes the following selections. For $\tilde{\alpha}_1$ set it to $\bar{\alpha}_1$ which is the relevant sub-vector of $\bar{\alpha}$. For $\tilde{\alpha}_2$, let it be the coefficient estimates from the regression of $y_t - y_{t-1}\bar{\alpha}_1$ against z_t.

Now the h-statistic employs an estimator of β, β_h, derived from $\bar{\alpha}' = (\bar{\alpha}_1'\bar{\alpha}_2')$, while Inder's works with β_I based on $(\bar{\alpha}_1'\tilde{\alpha}_2')$. Applying (6) gives

$$I_{\beta\beta}T^{\frac{1}{2}}(\beta_h-\beta_I) = I_{\beta\alpha_1}T^{\frac{1}{2}}(\bar{\alpha}_1-\tilde{\alpha}_1) + I_{\beta\alpha_2}T^{\frac{1}{2}}(\bar{\alpha}_2-\tilde{\alpha}_2). \qquad (51)$$

But $I_{\beta\alpha_2} = 0$ because of the exogeneity of z_t, so that it immediately follows that $T^{\frac{1}{2}}(\beta_h-\beta_I)$ is $o_p(1)$, i.e. the h-statistic and Inder's statistic have the same limiting distribution. What is crucial to this

---

[6] It may not be immediately obvious that King's test statistic is asymptotically equivalent to that described above. Ignoring end effects, he uses $(\Sigma\hat{u}_t^2)^{-1}\Sigma(\tilde{u}_t-\lambda\tilde{u}_{t-1})^2$, where $\hat{u}_t$ are the OLS residuals while $\tilde{u}_t$ are the residuals from the quasi-differenced model. After centering and normalizing, the test statistic $T^{\frac{1}{2}}S(\lambda) = T^{\frac{1}{2}}[(1+\lambda^2)^{-1}(\Sigma\hat{u}_t^2)^{-1}(\Sigma(\tilde{u}_t-\lambda\tilde{u}_{t-1})^2)-1]$ is asymptotically a standard normal deviate. It is easily shown that $T^{\frac{1}{2}}S(\lambda)$ is distributed as $(1+\lambda^2)^{-1}\sigma^{-2}T^{-\frac{1}{2}}(\Sigma(\tilde{u}_t-\lambda\tilde{u}_{t-1})^2-\hat{u}_t^2)$ under H_o. But $T^{-\frac{1}{2}}\Sigma(\tilde{u}_t-\lambda\tilde{u}_{t-1})^2-\hat{u}_t^2 = -2\lambda T^{-\frac{1}{2}}\Sigma\tilde{u}_t\tilde{u}_{t-1} + o_p(1)$ and therefore King's test is essentially based upon the first order serial correlation coefficient of $\tilde{u}_t$.

equivalence is the common strategy for dealing with the incidental parameter α_1. Without this, there is unlikely to be any equivalence with the h-statistic.

6. Some Censored Variable Applications

Although the most likely models to exhibit anticipated values are those based upon time series, there are nevertheless a significant number of studies employing data on individuals in which a "permanent" or "potential variable" appears as a regressor, and some proxy is constructed for it from an auxiliary regression. These "potential" or "permanent" variables can be viewed as conditional expectations and so (19) applies directly with the subscript representing individuals. There are many examples of this type of model, particularly in labour economics and demography. Thus Nickell (1979) makes the probability of leaving the unemployment pool a function of the expected wage available to the individual given his personal characteristics. Similarly, Long and Jones (1980) model the participation rate of women as a function of the potential wage. In each of these cases the unobserved expected wage is replaced by predictions from an auxiliary linear regression.

It does not seem possible to provide a complete account of the problems raised in such applications, as there is great diversity in the way in which the decisions taken by economic agents are modelled. For this reason our analysis is limited to the procedure adopted by Long and Jones, where the decision to participate is modelled through probit analysis. Accordingly, equation (19) is retained but (18) is modified to reflect the existence of a latent variable y_t^*, with the observations y_t being binary.

$$\therefore \qquad y_t^* = (w_t\alpha)\delta + x_t\beta_1 + e_t \ . \tag{52}$$

It is assumed that e_t and η_t in (19) and (52) are bivariate normal with zero mean and covariance matrix with elements $\{\sigma_{ij}\}$ $i = 1, 2$, $j = 1, 2$. Very few existing studies are specific about whether $\sigma_{12} = 0$ or not, but it is hard to see why such a restriction would be appropriate. Nevertheless, it is invariably the case that σ_{12} is prescribed to zero. For the equivalent continuous variable case of Section 3, such an action only raised questions regarding efficiency and the nature of inferences (Propositions 3.6, 3.7), and the two-stage estimator remained consistent. Such a convenient result is less obvious when data is censored, and to demonstrate this fact, as well as providing a simple adjustment to obviate any difficulties created by the invalid prescription, it is useful to modify (52) so as to highlight the parameter σ_{12} .

The modification employed originates in Telser (1964), and has been exploited recently by Heckman (1978) and Smith and Blundell (1983). Because e_t and η_t are bivariate normal, it is possible to express e_t as $\eta_t\gamma + \varepsilon_t$, where $\gamma = \sigma_{22}^{-1}\sigma_{21}$ and $E(\varepsilon_t\eta_t) = 0$. Thus (52) becomes

$$y_t^* = (w_t\alpha)\delta + x_t\beta_1 + \eta_t\gamma + \varepsilon_t = q_t\beta + \eta_t\gamma + \varepsilon_t \ , \tag{53}$$

making the log-likelihood of the binary data and z_t

$$-\frac{T}{2}\log \sigma_{22} + \sum_{t=1}^{T}\{y_t \log(1-F_t) - (y_t-1)\log F_t - \frac{1}{2}\sigma_{22}^{-1}\eta_t^2\} \ . \tag{54}$$

where $F_t = \text{prob}(y_t \leq 0) = \int_{-\infty}^{q_t\beta+\eta_t\gamma} (2\pi)^{-1/2}\exp\left\{-\frac{1}{2}\lambda^2\right\}d\lambda$ (imposing the identifying condition that $\text{var}(\varepsilon_t) = 1$) .

Now the two-stage estimator used in empirical studies involves fitting a probit model to (53) with $\gamma = 0$ and $w_t\alpha$ replaced by $w_t\tilde{\alpha}$ i.e., $\tilde{\beta}$ is the solution to $\tilde{\beta} = \max_{\beta} L(\beta, \tilde{\alpha}, \gamma = 0)$. This raises the question of whether $\tilde{\beta}$ is a consistent estimator; clearly from Theorem 1 it will be necessary that $I_{\beta\gamma}(\beta_0, \alpha_0, \overline{\gamma}) = 0$, and one might proceed to evaluate this element in the information matrix. However, such an operation is quite complex, owing to the presence of random variables η_t in the limits of the integrals involved in F_t . Some progress is possible by taking the expectation conditional upon η_t first, as this allows one to exploit the independence of the scores d_β and d_γ.

$$\therefore \quad I_{\beta\gamma} = \lim_{T\to\infty} E_\eta[T^{-1}E(\Sigma d_{\beta,t}d_{\gamma,t}|\eta_t)] \tag{55}$$

$$= \lim_{T\to\infty} E_\eta[T^{-1}\Sigma F_t^{-1}(1-F_t)^{-1}q_t\eta_t f_t^2] \tag{56}$$

using the standard results for the information matrix of a probit model --see Davidson and MacKinnon (1984, p. 244)--where $f_t = (2\pi)^{-1/2}\exp\left(-\frac{1}{2}(q_t\beta+\eta_t\gamma)^2\right)$. Unfortunately, getting the unconditional expectation of (56) proves difficult, But, as interest only centers on whether $I_{\beta\gamma}(\beta_0, \alpha_0, 0) = 0$, Proposition 6.1 shows that the two stage probit estimator is consistent.

Proposition 6.1: *The two-stage probit estimator of* β *in (53), constructed by assuming that* $\gamma = 0$, *is consistent.*

Proof: $I_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma}=0)$ can be evaluated by setting $\gamma = 0$ inside the expectation in (56). This makes f_t and F_t non-stochastic, and therefore

$$I_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma}=0) = \lim_{T\to\infty} T^{-1}\Sigma F_t^{-1}(1-F_t)^{-1}E(q_t\eta_t)f_t^2 = 0 \ . \qquad (57)$$

□

That the two-stage probit estimator of β is consistent follows readily from (53) by writing it as

$$y_t^* = q_t\beta + v_t \ , \qquad (58)$$

and observing that v_t is a normally distributed random variable. Hence, the variance of v_t rather than ε_t might be normalized to unity, and this will have no effect upon the consistency of the probit estimator applied to (58). Of course, this property is one peculiar to the normal distribution, and it therefore seems unlikely that Proposition 6.1 would extend to other cases in the literature e.g. where e_t has the logit rather than normal density.

Even when the two stage estimator is consistent, the problem of valid inferences remains. In general it will be necessary to compute the covariance matrix using Theorem 8, with $\sigma_{22}^{-1}\sigma_{12}$ included amongst β as one of the parameters to be estimated i.e. the two-stage probit estimator is applied to (53) with η_t replaced by $\tilde{\eta}_t = z_t - w_t\tilde{\alpha}$. Of course this modification does not mean that $I_{\beta\beta}^{-1}$ will be a consistent estimator of the covariance matrix of $\tilde{\beta}' = (\tilde{\delta}\tilde{\beta}_1'\tilde{\gamma}')$, but a directional result relating $I_{\beta\beta}^{-1}$ to the actual covariance matrix of $T^{\frac{1}{2}}(\tilde{\beta}-\beta_o)$ is provided in Proposition 6.2.

Proposition 6.2: *The standard errors from a two-stage probit analysis applied to (53) with* $\eta_t = \tilde{\eta}_t$ *understate the true standard errors.*

Proof: From Theorem 8, this follows provided $V_1 = 0$. Now the scores are $d_\beta = -\Sigma y_t(1-F_t)^{-1}f_t q_t - \Sigma(y_t-1)F_t^{-1}f_t q_t$ while $\tilde{\alpha} = (W'W)^{-1}W'\eta + \alpha_0$

$$\therefore \quad E(d_\beta(\tilde{\alpha}-\alpha_0)') = V_1 = E_\eta[E(d_\beta(\tilde{\alpha}-\alpha_0)'|\eta)] = E_\eta[E(d_\beta|\eta)(\tilde{\alpha}-\alpha_0)'] \tag{59}$$

and $E(d_\beta|\eta) = 0$ since $E(y_t|\eta_t) = 1 - F_t$, $E((y_t-1)|\eta_t) = F_t$. □

The fact that σ_{12} may not be zero is a source of problems in many censored variable applications. Perhaps the best known of these is the Gronau/Lewis model, analyzed extensively by Heckman (1976). Heckman proposed a two-stage estimator of the parameters of interest in this model, and it is therefore of interest to determine its location within the framework of the current paper.

The Gronau/Lewis model consists of two equations

$$y_t = x_t\beta_1 + e_t \tag{60}$$

$$z_t = w_t\alpha + \eta_t \tag{61}$$

where e_t and η_t are bivariate normal with covariance matrix elements $\{\sigma_{ij}\}$. Data is observed on x_t and w_t for observations on T individuals, but y_t is observed only if z_t exceeds zero. Thus the variable y_t is censored and it is assumed that there are only T_1 observations on y_t . Two strategies for the estimation of β are now possible; one utilizes all T observations, the other works with the *truncated* data set composed only of the T_1 observed values for y_t . Heckman chooses this latter option.

The log likelihood of the truncated data set is available in Bloom and Killingworth (1985, p. 133) as

$$-\frac{T_1}{2}\log\sigma_{11}-\frac{1}{2}\sum_{t=1}^{T_1}\sigma_{11}^{-1}e_t^2+\sum_{t=1}^{T_1}\log F(K_t|\sigma_3^{1/2})-\sum_{t=1}^{T_1}\log F(J_t) \qquad (62)$$

using the identifying restriction $\sigma_{22}=1$ and defining $\sigma_3=1-\sigma_{11}^{-1}\sigma_{12}^2$, $K_t=w_t\alpha+\sigma_{12}\sigma_{11}^{-1}(y_t-x_t\beta_1)$, $J_t=w_t\alpha$. F_t is the standard normal distribution function. It is apparent from (62) that replacing α by a consistent estimator $\tilde{\alpha}$ eliminates the last term from the log likelihood in (62), but the presence of σ_{12} and β_1 in K_t means that any "two-stage estimator" formed by maximizing $L(\beta_1, \sigma_{12}, \sigma_{11}, \tilde{\alpha})$ with respect to β_1, σ_{12} and σ_{11} must be iterative. Setting $\sigma_{12}=0$ would obviate the need for iteration, but the resulting estimator of β_1 will generally be inconsistent as $I_{\beta\gamma}(\beta_o,\alpha_o,\bar{\gamma}=0)\neq 0$.
Of course, if an initial consistent estimator of σ_{12}, $\tilde{\sigma}_{12}$, was available an efficient two-step estimator could be constructed as in Theorem 6, but obtaining $\tilde{\sigma}_{12}$ would require consistent estimators of β and α.

Now the whole point of a two-stage estimator is that both stages be simple to compute, preferably with a regression program, and that the estimator be consistent. This suggests isolating the requirements for consistency in such a way that two-stage estimators of the Heckman type can be encompassed in our general framework. Proposition 6.3 therefore presents a condition for consistency that underlay the two-stage and two-step efficient estimators analyzed in Section 2, but which is capable of generating a broader class of consistent estimators.

Proposition 6.3: *Let* $\theta' = (\beta'\alpha')$, $\bar{\theta}$ *be some value of* θ , *and define* $\tilde{\beta} = A^{-1}[T^{-1}d_\beta(\bar{\theta}) + A\bar{\beta}] + o_p(1)$. *If* A *is chosen such that*

$$T^{-1}d_\beta(\bar{\theta}) + A(\bar{\beta} - \beta_0) \xrightarrow{p} 0 , \tag{63}$$

$\tilde{\beta}$ *is a consistent estimator of* β_0 .

Proof: $\tilde{\beta} - \beta_0 = A^{-1}[T^{-1}d_\beta(\bar{\theta}) + A(\bar{\beta} - \beta_0)]$ and the R.H.S. tends to zero by (63). □

Proposition 6.2 emphasizes that the class of consistent estimators may be very wide indeed, and the corresponding set of "two-stage" estimators will not be limited to those which solve $\max_\beta L(\beta,\tilde{\alpha})$. This last estimator is $\tilde{\beta} = I_{\beta\beta}(\tilde{\alpha}, \beta_0)T^{-1}d_\beta(\beta_0, \tilde{\alpha}) + o_p(1)$, and is therefore consistent if (63) holds, or if $T^{-1}d_\beta(\beta_0, \tilde{\alpha}) \xrightarrow{p} 0$, an event that occurs when $\tilde{\alpha} \xrightarrow{p} \alpha_0$. In similar fashion, the two-step efficient MLE of Theorem 8 has $\beta = \theta$ (all parameters are estimated jointly), defines $\tilde{\theta}_{2SML} - \tilde{\theta} = I_{\theta\theta}(\tilde{\theta})^{-1}T^{-1}d_\theta(\tilde{\theta})$ and achieves consistency because $T^{-1}d(\tilde{\theta}) + I_{\theta\theta}(\tilde{\theta})(\tilde{\theta} - \theta_0) \xrightarrow{p} 0$ as $\tilde{\theta}$ is consistent.

Reverting to the Gronau/Lewis example, it is natural to seek a two-stage estimator by studying the structure of $d_\beta(\theta)$ at different values for θ . Differentiating (62) with respect to β_1 and σ_{12} yields

$$d_{\beta_1} = \sigma_{11}^{-1}\sum_{t=1}^{T_1} x_t'(y_t - x_t\beta_1) - \sum_{t=1}^{T_1} F(K_t|\sigma_3^{1/2})^{-1}x_t'\sigma_{12}\sigma_{11}^{-1}f(K_t|\sigma_3^{1/2}) \tag{64}$$

$$d_{\sigma_{12}} = \sum_{t=1}^{T_1} F(K_t|\sigma_3^{1/2})^{-1}\sigma_3^{-1}(w_t\alpha + \sigma_{12}\sigma_{11}^{-1}e_t)f(K_t|\sigma_3^{1/2})\sigma_{12}$$

$$+ \sum_{t=1}^{T_1} F(K_t|\sigma_3^{1/2})e_t\sigma_{11}^{-1}f(K_t|\sigma_3^{1/2}) , \tag{65}$$

where $f(\psi_t) = (2\pi)^{-\frac{1}{2}}\exp(-\psi_t^2)$. Setting $\bar{\theta} = (\bar{\sigma}_{12}\ \bar{\beta}_1')$ $= (0,0)$ gives

$$d_{\beta_1}(\bar{\theta}) = \sigma_{11}^{-1}\sum_{t=1}^{T_1} x_t y_t \tag{66}$$

$$d_{\sigma_{12}}(\bar{\theta}) = \sigma_{11}^{-1}\sum_{t=1}^{T_1} F(w_t\alpha)^{-1} f(w_t\alpha) y_t \ , \tag{67}$$

and the structure of these derivatives immediately establishes Heckman's estimator as a special case of the class of estimators distinguished in Proposition 6.4.

Proposition 6.4: *Heckman's (1976) two-stage estimator is a member of the class of estimators in Proposition 6.3 with* $\bar{\theta}' = (0\ \tilde{\alpha}')$, $\tilde{\alpha}$ *being the probit estimator of* α *applied to (61) using all the data,* $A = \tilde{\Phi}'\tilde{\Phi}$ *where* $\tilde{\Phi}$ *has* t^{th} *row* $\phi_t = \sigma_{11}^{-1/2}(x_t\ \tilde{F}_t^{-1}\tilde{f}_t)$, *and the likelihood is that for the truncated data set* $t \in T_1$.

Proof: When $\bar{\theta}' = (0\ \tilde{\alpha}')$, then derivatives of the likelihood are (66) and (67) with F_t and f_t becoming $\tilde{F}_t$ and $\tilde{f}_t$ as they depend upon α . Hence, with $\beta' = (\beta_1'\ \sigma_{12})$,

$$d_\theta(\bar{\theta}) = \sigma_{11}^{-\frac{1}{2}}\tilde{\Phi}'y \ . \tag{68}$$

Choosing $T \cdot A$ as $\tilde{\Phi}'\tilde{\Phi}$, Proposition 6.3 defines the two-stage estimator as $(\tilde{\Phi}'\tilde{\Phi})^{-1}\tilde{\Phi}'y$, which is Heckman's version.[7] As $\tilde{\alpha}$ is consistent,

---

[7]Notice that the factor σ_{11} cancels in this expression, so that it can be set to unity in the definition of ϕ_t .

(63) is satisfied if $E(d_\beta(0, \alpha_0)) = -A\beta_0$ i.e. if $E[(\Phi'y) - \sigma_{11}^{\frac{1}{2}}\Phi'\Phi\beta_0] = 0$ or $E(y - \sigma_{11}^{\frac{1}{2}}\Phi\beta_0) = 0$. But for observations in the truncated data set, $E(y_t) = x_t\beta_1 + \sigma_{12}f_t/F_t = \sigma_{11}^{\frac{1}{2}}\phi_t\beta_0$ (Amemiya (1984, eq. (12))). □

7. Conclusion

Applied econometric research frequently encounters the difficulty that estimation of the parameters of interest is complex owing to the presence of incidental parameters. It is tempting therefore to try to circumvent the difficulties by proceeding in two stages. In the first, some estimates are made of the incidental parameters. In the second, these estimates are treated as though they were population values, leading to a large reduction in the dimension of the unknown parameter space, possibly even down to that of the parameters of interest only. The properties of such a staged process, particularly as they relate to the trio of issues arising from the consistency and efficiency of the estimator and the provision of reliable inferences, were then the central concern of section 2 of the paper.

Section 2 assembled the core of results used later in analysing a range of applications that have featured two-stage estimators. Thus sections 3 and 4 dealt respectively with models characterized by the presence of current and future anticipations. In section 4, an alternative to the traditional two-stage estimator was proposed. This suggested estimator possesses the attractive properties of asymptotic efficiency and of being computable with a regression program. As such, it would seem to be especially useful for those instances in which Barro's (1977) approach to decomposing series into unanticipated and anticipated components is followed. Section 5 turned to the area of diagnostic tests, re-interpreting

these as based upon a variety of two-stage estimators. Some integration and clarification of that material was achieved by looking at diagnostic tests in this way. Finally, section 6 examined some popular two-step estimators arising in the analysis of censored data.

The applications of sections 3 through 5 constitute only a sub-set of the types of models in which two-stage estimators appear. A few others were mentioned in the introduction, but by no means does that selection exhaust the set of circumstances characterized by incidental parameters. To give just one example, the whole question of exogeneity, as formulated by Engle et al (1979) and Koopmans (1950), is tied up with the efficiency of $\tilde{\beta} = \max_{\beta} L(\beta,\tilde{\alpha},\bar{\gamma})$, where $\bar{\gamma}$ is prescribed and $\tilde{\alpha}$ can assume any value. As might be expected from Theorem 3, conditions for exogeneity are therefore most usefully expressed in terms of the information matrix, capturing as it does the efficiency of the MLE against which comparisons may be made. There are many more such examples, all featuring incidental parameters, and it seems important that the similarities of the problems be emphasised rather than their differences. Only then can the geography be fully appreciated.

References

Amemiya, T. (1978), "On a Two-Step Estimation of a Multivariate Logit Model", *Journal of Econometrics*, 8, 13-21.

Amemiya, T. (1984), "Tobit Models: A Survey", *Journal of Econometrica* 24, 3-61.

Attfield, C.L.F. (1983), "An Analysis of the Implications of Omitting Variables from the Monetary Growth Equation in a Model of Real Output and Unanticipated Money Growth", *European Economic Review*, 23, 261-280.

Barro, R.J. (1977), "Unanticipated Money Growth and Unemployment in the United States", *American Economic Review*, 67, 101-115.

Berenblutt, I.I. and G.I. Webb (1973), "A New Test for Autocorrelated Errors in the Linear Regression Model", *Journal of the Royal Statistical Society* B, 35, 33-50.

Bloom, D. E. and M. R. Killingsworth (1985), "Correcting for Truncation Bias Caused by a Latent Truncation Variable", *Journal of Econometrics*, 27, 131-136.

Breusch, T.S. and A.R. Pagan (1980), "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics", *Review of Economic Studies*, 47, 239-253.

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

Durbin, J. (1970), "Testing for Serial Correlation in Least-Squares Regression when some of the Regressors are Lagged Dependent Variables", *Econometrica*, 38, 410-421.

Eckstein, Z. (1984), "A Rational Expectations Model of Agricultural Supply", *Journal of Political Economy*, 92, 1-19.

Engle, R.F., Hendry, D.F. and J.F. Richard (1983), "Exogeneity", *Econometrica*, 51, 277-304.

Graybill, F.A. (1969), *Introduction to Matrices with Applications in Statistics* (Wadsworth: Belmont, California).

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

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475-492.

Heckman, J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System", *Econometrica*, 46, 931-959.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error", *Econometrica*, 47, 153-161.

Hsiao, C. and D. Mountain (1982), "Estimating the Short-Run Income Elasticity of Demand for Electricity Using Cross-Sectional Categorized Data", (University of Toronto, mimeo).

Inder, B.A. (1984), "A New Test for Autocorrelation in the Disturbances of the Dynamic Linear Regression Model", (paper presented to the Second Australasian Meeting of the Econometric Society, Sydney).

Johnston, J. (1972), *Econometric Methods* (McGraw Hill, 2nd ed., New York).

King, M.L. (1985), "A Point Optimal Test for Autoregressive Disturbances", *Journal of Econometrics*, 27, 21-37.

Kohn, R. (1977), "Note Concerning the Akaike and Hannan Estimation Procedures for an Autoregressive-Moving Average Process", *Biometrika*, 64, 622-625.

Koopmans, T.C. (1950), "When is an Equation System Complete For Statistical Purposes", in T.C. Koopmans (ed.), *Statistical Inference in Dynamic Economic Models* (John Wiley and Sons, New York).

Lewis, P.E.T. (1983), "Disequilibrium in the Australian Aggregate Labour Market", *Economics Letters*, 11, 185-189.

Liang, K.Y. (1984), "The Asymptotic Efficiency of Conditional Likelihood Methods", *Biometrika*, 71, 305-313.

Long, J.E. and E. B. Jones (1980), "Labor Force Entry and Exit by Married Women: A Longitudinal Analysis", *Review of Economics and Statistics*, 62, 1-6.

McCallum, B. (1979), "Topics Concerning the Formulation, Estimation and Use of Macroeconomic Models with Rational Expectations," ASA *Proceedings of the Business and Economic Statistics Section*, 65-72.

Manski, C.F. (1985), "Adaptive Estimation of Non-Linear Regression Models", *Econometric Reviews* (forthcoming).

Mishkin, F. (1982), "Does Anticipated Monetary Policy Matter? An Econometric Investigation", *Journal of Political Economy*, 90, 22-51.

Newey, W. (1984), "A Method of Moments Interpretation of Sequential Estimators", *Economics Letters*, 14, 201-206.

Neyman, J. (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses", in U. Grenander (ed.), *Probability and Statistics* (Wiley, New York), 213-234.

Nickell, S. (1979), Estimating the Probability of Leaving Unemployment", *Econometrica*, 47, 1249-1266.

Pagan, A.R. (1984), "Econometric Issues in the Analysis of Regressions with Generated Regressors", *International Economic Review*, 25, 221-247.

Pierce, D.A. (1982), "The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics", *Annals of Statistics*, 10, 475-478.

Pudney, S.E. (1982), "The Identification of Rational Expectation Models Under Structural Neutrality", *Journal of Economic Dynamics and Control*, 4, 117-121.

Rothenberg, T.J. and C.T. Leenders (1964), "Efficient Estimation of Simultaneous Equation Systems", *Econometrica*, 32, 57-76.

Sargent, T. (1976), "A Classical Macroeconomic Model for the United States", *Journal of Political Economy*, 84, 207-237.

Sargent, T.J. (1978), "Estimation of Dynamic Labor Demand Schedules Under Rational Expectations", *Journal of Political Economy*, 86, 1009-1044.

Sims, C. (1977), "Comment", *Journal of the American Statistical Association*, 72, 23-24.

Smith, R. and R. Blundell (1983), "An Exogeneity Test for the Simultaneous Equation Tobit Model with an Application to Labour Supply", Queen's University Discussion Paper 546.

Stein, C. (1956), "Efficient Nonparametric Testing and Estimation", Proceedings of the Third Berkeley Symposium in Mathematical Statistics and Probability, 1, 187-196 (University of California Press, Berkeley).

Telser, L. (1964), "Iterative Estimation of a Set of Linear Regression Equations", *Journal of the American Statistical Association*, 59, 845-862.

Turkington, D. (1984), "A Note on Two-Stage Least Squares, Three-Stage Least Squares and Maximum Likelihood Estimation in an Expectations Model", *International Economic Review*, (forthcoming).

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50, 1-26.

Wickens, M. (1982), "The Efficient Estimation of Econometric Models with Rational Expectations", *Review of Economic Studies*, 49, 55-68.