University of California, Berkeley
Department of Agricultural &
Resource Economics

CUDARE Working Papers

Year 1983 *Paper 281*

On Efficient Estimation of Some Limited Dependent Variables Models

Yacov Tsur

ON EFFICIENT ESTIMATION OF SOME LIMITED DEPENDENT VARIABLES MODELS†

Yacov Tsur‡

INTRODUCTION

In empirical research, it is often the case that the data set under investigation is incomplete. Consequently, a plethora of estimation methods have been developed to deal with the several possible pattern of missing data, such as in Afifi and Elashoff [1966], Griliches, Hall and Hausman [1978], and Dempster, Rubin and Laird [1977]. In general, these methodologies use prior knowledge or information on the process generating the data and on the missing patterns to "complete" the data set. Standard estimation procedures are then employed on this new data set.

The econometric literature distinguishes two types of models corresponding to missing data problems: incomplete data on predetermined or right-hand-side (RHS) variables only; and incomplete data on the dependent, and possibly on the RHS, variables. In the latter models, known as Limited Dependent Variable (LDV) models, the most widely used estimation procedures follow Heckman [1976,1978] and Lee [1978]. These procedures do not use "filled-in" methods but rather employ a two stage procedure where the second stage focuses only on the *observed* part of the sample with additional terms that account for the

† This research was partially supported by Water Resources Center Grant W622 and by the Giannini Foundation of Agricultural Economics.

‡ Department of Agricultural and Resource Economics, University of California, Berkeley.

I wish to thank T. Rothenberg, L. Le Cam, M. Hanemann and P. Ruud for helpful discussions and comments. Any remaining errors are, of course, my responsibility.

selection mechanism (or missing patterns), derived in the first stage.

Maximum likelihood (ML) methods are frequently considered (Amemiya [1973], Hausman and Wise [1978], Heckman [1978], Duncan [1980], Cosslet [1981], among others) but rarely used in practical applications because of computational and implementation difficulties. (See Griliches *et al.* [1978] as an exception.) This is primarily because the log likelihood functions of LDV models is generally highly non-linear in the parameters and often contains multiple roots so that iterations with arbitrary initial values may result in a root that does not correspond to the global maximum of the function (Amemiya [1973]). To overcome the possibility of a "wrong root" requires a "good" initial parameter value. This leads to the consideration of a single iteration procedure of the form $\vartheta^P - R(\vartheta^P)^{-1} L_\vartheta(\vartheta^P)$, where ϑ^P is the preliminary estimate, $R(\vartheta^P)$ is an estimate of the information matrix and $L_\vartheta(\vartheta^P)$ is the gradient of the log likelihood function evaluated at ϑ^P. The outcome of this procedure, denoted as the "Linearized Maximum Likelihood" (LML) procedure, is asymptotically equivalent to the ML estimator provided that ϑ^P approaches the true parameter at an "appropriate" rate as the sample size increases. (See Le Cam [1960, 1969], and Rothenberg and Leenders [1964] for a theoretical discussion, and Berndt, Hall, Hall and Hausman [1974] for implementation). In fact any estimate of the form $\vartheta^P - R(\vartheta^P)^{-1}\omega_\vartheta(\vartheta^P)$, such that $\sqrt{T}(\omega_\vartheta(\vartheta^P) - L_\vartheta(\vartheta^P))$ goes to zero in probability as the sample size $T \to \infty$, has the same limiting distribution as the LML estimator. We will use this generalization of LML and denote it as the GML estimator.

Recently, an alternative algorithm for the ML estimation of models with incomplete data has been offered by Dempster *et al.* [1977]. This method, called the EM algorithm, is an iterative procedure with each iteration comprised of two steps: an Expectation (E) step which can be interpreted as a guideline for filling in the missing data; and a Maximization (M) step which proceeds with a

maximization task. Since this seminal paper, an extensive literature has emerged dealing with the convergence properties of the algorithm and its possible aplications (e.g., Wu [1983] and the reference cited therein).

The purpose of the research reported here is to investigate efficient estimation methods for some LDV models in the context of the EM algorithm.[1] The use of the term EM *approach* rather than algorithm is adopted because the M step can take on differing formats in the algorithm and we will modify this step substantially in several cases. Two procedures will be considered: an iterative procedure resulting from a straightforward application of the EM algorithm; and a single iteration procedure, corresponding to the LML procedure, resulting from the EM approach. Each method is evaluated under two scenarios: single equation models (Section 2) and two-equation models (Section 3). Each scenario involves two models corresponding to different patterns of missing data.

The conclusions emanating from the analysis include:

1. ML estimators can be derived via iterative Ordinary Least Squares (OLS) in single equations models, or Generalized Least Squares (GLS) in multiple equation models if the EM algorithm is employed.

2. The EM approach provides a unique method for deriving and interpreting the LML estimator which basically involves the application of an Instrumental Variable (IV) regression.

3. The approach lends itself to an integrative framework which unifies the analysis of the many possible LDV models and links it to classical linear model theory.

---

[1] The concept of efficiency used in this work together with the conditions that assure the efficiency of the ML estimators are given in Appendix C. Consequently, the terms "efficient estimator" and "ML estimator" are used interchangeably.

The basic notation is as follows: the parameter set is denoted by Θ with ϑ representing an arbitrary member. A specific member of Θ is denoted by a super-script on ϑ except for the cases of ML, LML, and GML estimators which are denoted by $\hat{\vartheta}$, ϑ_{LML}, and ϑ_{GML} respectively. Other exceptions are indicated in the text. The true parameter value is given by ϑ^o. Superscripts in general indicate evaluations of the quantity at the specific parameter value. A superscript "*" indexing a variable indicates that the variable is incompletely observed. The density and distribution functions of a standard normal variate are denoted by $\varphi(.)$ and $\Phi(.)$ respectively. The line segment joining two distinct points ϑ' and ϑ'' of Θ is denoted by $\vartheta'\,\vartheta''$. The log likelihood function is denoted by $L(\vartheta)$ with $L_\vartheta(.)$ and $L_{\vartheta\vartheta'}(.)$ indicating the gradient vector of first derivatives and the Hessian matrix of second derivatives respectively. An estimate ϑ^P is consistent of order $T^{-\frac{1}{2}}$, or $\sqrt{T}$-consistent, if $\sqrt{T}(\vartheta^P - \vartheta^o)$ is bounded in probability as $T \to \infty$.

2. SINGLE EQUATION MODELS

Two basic LDV single equation models, the Tobit and probit models, corresponding to two different patterns of missing data, are considered in this section.

The underlying structure is given by

$$(2.1)\quad y_t^* = x_t'\beta^o + u_t\ ,\ t=1,2,\ldots,T$$

$$(2.2)\quad u_t \overset{iid}{\sim} N(0,\sigma^{o2}),\ t = 1,2,\ldots,T$$

The Tobit model results from the following observation rule

$$(2.3)\quad y_t = \begin{cases} y_t^* \text{ if } y_t^* \geq 0 \\ 0 \ \ otherwise \end{cases} \qquad t=1,2,\ldots,T$$

and for the probit model

$$(2.3') \quad y_t = \begin{cases} 1 \text{ if } y_t^* \geq 0 \\ 0 \text{ otherwise} \end{cases} \qquad t = 1,2,\ldots,T$$

For both models it is assumed that x_t is observed for all t, where $X = (x_1,x_2,\ldots,x_T)'$ is a Txk matrix, $y^* = (y^*_1,y^*_2,\ldots,y^*_T)'$ is a Tx1 vector of quantities for which we observe $y=(y_1,y_2,\ldots,y_T)'$ according to (2.3) or (2.3'), β^o and σ^{o2} are respectively kx1 and 1x1 vectors of parameters to be estimated, and $u=(u_1,u_2,\ldots,u_T)'$ is a Tx1 vector of unobserved errors distributed according to (2.2). It is assumed that X is of full rank, that $\lim_{T\to\infty} \frac{1}{T}X'X$ exists and is positive definite and that each column of X is statistically independent of the error vector u.

The missing data pattern corresponding to the probit model creates identification problems in that any structure that agrees in sign with (2.1) results in the same observations. Specifically the structure $\frac{y_t^*}{a} = \frac{x_t'\beta^o}{a} + \frac{u_t}{a}$ for an arbitrary $0 < a < \infty$ is observationally equivalent to (2.1) and so a normalization rule is required to identify any of the parameters. The normalization usually employed and used in this study is $\sigma^o = 1$. No identification problems arise in the Tobit model.

Let q_t be defined as the indicator variable

$$(2.4) \quad q_t = \begin{cases} 1 \text{ if } y_t^* \geq 0 \\ 0 \text{ otherwise} \end{cases} \qquad t = 1,2,\ldots,T$$

The sample q_t; t = 1,2, . . ., T is a realization of T independent Bernoulli trials with the probability of success for the t^{th} trial equal to $1 - \Phi(-x_t'\beta^o / \sigma^o)$. This

captures the randomness inherent in the observation rule.

In the application of the EM algorithm, discussed below, the following definitions and results will be used

$$(2.5a)\quad \lambda_t^- = \frac{-\varphi(-\mu_t/\sigma)}{\Phi(-\mu_t/\sigma)}$$

$$(2.5b)\quad \lambda_t^+ = \frac{\varphi(-\mu_t/\sigma)}{1-\Phi(-\mu_t/\sigma)}$$

where $\mu_t = x_t'\beta$. It follows that

$$(2.6a)\quad \frac{\partial\lambda_t^-}{\partial\mu_t} \overset{def}{=} \Lambda_t^- = -\frac{1}{\sigma}\lambda_t^-(\mu_t/\sigma + \lambda_t^-)$$

$$(2.6b)\quad \frac{\partial\lambda_t^+}{\partial\mu_t} \overset{def}{=} \Lambda_t^+ = -\frac{1}{\sigma}\lambda_t^+(\mu_t/\sigma + \lambda_t^+)$$

Defining $\tilde{\tau}_t^- = var(u_t \mid u_t < -\mu_t)$ and $\tilde{\tau}_t^+ = var(u_t \mid u_t \geq -\mu_t)$ it can be demonstrated that (See Appendix A)

$$(2.7a)\quad \tilde{\tau}_t^- = \sigma^2(1 + \sigma\Lambda_t^-)$$

$$(2.7b)\quad \tilde{\tau}_t^+ = \sigma^2(1 + \sigma\Lambda_t^+)\,.$$

For any given parameter value, ϑ^κ, we define $y_t^\kappa = E_{\vartheta^\kappa}\{y_t^* \mid y_t\}$ where $E_{\vartheta^\kappa}\{.\}$ denotes the expectation when ϑ^κ holds. So

$$(2.8)\quad y_t^\kappa = q_t y_t + (1-q_t)(\mu_t^\kappa + \sigma^\kappa\lambda_t^{-\kappa}) \quad \textit{under Tobit}$$

and

$$(2.8')\quad y_t^\kappa = q_t(\mu_t^\kappa + \lambda_t^{+\kappa}) + (1-q_t)(\mu_t^\kappa + \lambda_t^{-\kappa}) \quad \textit{under probit.}$$

It can be verified that

(2.9) $E_{\vartheta^\kappa}\{(y_t^* - \mu_t)^2 | y_t\} = (y_t^\kappa - \mu_t)^2 + \tilde{\tau}_t^\kappa$

where

(2.10) $\tilde{\tau}_t = (1 - q_t)\tilde{\tau}_t^-$ *under Tobit*

and

(2.10') $\tilde{\tau}_t = (1 - q_t)\tilde{\tau}_t^- + q_t\tilde{\tau}_t^+$ *under probit* .

The log of the joint density function of y_t^*; t = 1,2, . . ., T, is given by (disregarding the constant term)

(2.11) $f(y^*;\vartheta) = -\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t^* - \mu_t)^2 + \frac{T}{2}\log\frac{1}{\sigma^2}$

From (2.9) it follows that

(2.12) $E_{\vartheta^\kappa}\{f(y^*;\vartheta)|y\} = -\frac{1}{2\sigma^2}\{\sum_{t=1}^{T}[(y_t^\kappa - \mu_t)^2 + \tilde{\tau}_t^\kappa]\} - \frac{T}{2}\log\sigma^2$.

Maximizing (2.12) over β and σ^2 yields

(2.13a) $\beta^{\kappa+1} = (X'X)^{-1}X'y^\kappa$

(2.13b) $\sigma^{\kappa+1^2} = \frac{1}{T}\sum_{t=1}^{T}(y_t^\kappa - x_t'\beta^{\kappa+1})^2 + \frac{1}{T}\sum_{t=1}^{T}\tilde{\tau}_t^\kappa$

Equations (2.12) and (2.13) form, respectively, the E step and M step of one iteration of the EM algorithm which was shown (Dempster *et al.* [1977]) to yield a root of the likelihood function. Therefore ML estimate can be derived via iterative OLS procedures.[2]

[2] Note that σ^{o^2} is not an unknown parameter under the Probit model so (2.13b) is irrelevant in this case.

The same results were obtained by Hartley [1976] Fair [1977] (for the Tobit model), and Green [1982b]. The first two by considering the first order conditions of the log likelihood function and the later by using the EM method. Hence, in single equation models the standard approach of maximizing the log likelihood function can yield the same procedure as the EM algorithm of (2.11) to (2.13). However, as is demonstrated in Section 3, in multiple equation models the EM algorithm provides a unique procedure with advantages that have not previously been recognized.

The other estimation procedure considered in this study is the single step procedure. In this procedure an initial "good" estimate (ϑ^P) is first derived and used to construct $R(\vartheta^P)$ and $\omega_\vartheta(\vartheta^P)$. The final estimate is given by $\vartheta_{GML} = \vartheta^P - R(\vartheta^P)^{-1}\omega_\vartheta(\vartheta^P)$. Following Le Cam [1960, 1969], and Rothenberg and Leenders [1964], ϑ_{GML} is asymptotically efficient provided that ϑ^P is consistent of order $T^{-\frac{1}{2}}$, $R(\vartheta^P) \xrightarrow{P} \Psi(\vartheta^o)$ the information matrix, and $\sqrt{T}\,(\omega_\vartheta(\vartheta^P) - L_\vartheta(\vartheta^P)) \xrightarrow{P} 0$. ϑ_{LML} is accepted from ϑ_{GML} by putting $\omega_\vartheta(\vartheta^P) \equiv L_\vartheta(\vartheta^P)$. ϑ_{GML} is not unique since there are different $R(.)$'s and $\omega_\vartheta(.)$'s satisfying the conditions above (e.g., R(.) can be the Hessian matrix of the second derivatives of the log likelihood function or the one suggested by Berndt *et al.* [1974]). Our approach provides another method to construct ϑ_{GML} which produces an outcome that is identical to the standard method of putting $R(.) = L_{\vartheta\vartheta'}(\vartheta^P)$ and $\omega_\vartheta(.) = L_\vartheta(\vartheta^P)$, in single equation models but, as will be shown in Section 3, is unique in multiple equation models.[3] This is the subject of the following discussion which begins by considering the case of known σ^{o2} (which includes the Probit model automatically).

[3] It is worth noting that, starting from a $\sqrt{T}$-consistent estimate, one EM iteration (of the form given in (2.12)-(2.13)) will not yield an efficient estimator. This is true also in multiple equations models.

2.A *The Case of Known* σ^{o2}

The first step is to characterize the asymptotic properties of an efficient estimator. From (2.13a) it follows that $\beta^{\kappa+1} = \beta^{\kappa} + (X'X)^{-1}X'u^{\kappa}$ where u^{κ} is a T×1 vector with the t^{th} component defined as

(2.14) $u_t^{\kappa} = y_t^{\kappa} - x_t'\beta^{\kappa}$

Since the ML estimator $\hat{\beta}$ is a fixed point of (2.13a) the following orthogonality relation holds at $\hat{\beta}$

(2.15) $Xu^{\hat{\beta}} = \sum_{t=1}^{T} x_t u_t^{\hat{\beta}} = 0$

Expanding $\lambda_t^{-\beta}$, $\lambda_t^{+\beta}$ around λ_t^{-o}, λ_t^{+o} respectively allows one to express u_t^{β} as

$$u_t^{\beta} = [1 - (1 - q_t)(1 + \sigma\Lambda_t^{-\tilde{\beta}})]x_t'(\beta^o - \hat{\beta}) + \xi_t$$

under Tobit, where the super-script "o" is dropped from σ until otherwise indicated, and under the probit model as

$$u_t^{\beta} = [1 - (1 - q_t)(1 + \Lambda_t^{-\tilde{\beta}}) - q_t(1 + \Lambda_t^{+\tilde{\beta}})]x_t(\beta^o - \hat{\beta}) + \xi_t$$

where $\tilde{\beta} \in \beta^o\hat{\beta}$ and ξ_t is an error term defined as

(2.16) $\xi_t = u_t - (1-q_t)(u_t - \sigma\lambda_t^{-o})$ *under Tobit*

and

(2.16') $\xi_t = u_t - (1-q_t)(u_t - \lambda_t^{-o}) - q_t(u_t - \lambda_t^{+o})$ *under probit.*

Using (2.7) u_t^{β} can be expressed as

(2.17) $u_t^{\beta} = \frac{1}{\sigma^2}[\sigma^2 - \tilde{\tau}_t^{\tilde{\beta}}]x_t'(\beta^o - \hat{\beta}) + \xi_t$

where $\tilde{\tau}_t$ is defined in (2.10) and (2.10') and is rewritten below for convenience

(2.10) $\tilde{\tau}^t = (1-q_t)\tilde{\tau}_t^-$ *under Tobit*

(2.10') $\tilde{\tau}_t = (1-q_t)\tilde{\tau}_t^- + q_t\tilde{\tau}_t^+$ *under probit.*

It follows from (2.15) and (2.17) that

(2.18) $\sqrt{T}(\hat{\beta}-\beta^o) = \{\frac{1}{T}\sum_{t=1}^{T} x_t \frac{1}{\sigma^2}(\sigma^2 - \tilde{\tau}_t^{\beta})x_t'\}^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\xi_t$

Accounting for the randomness underlying q_t the ξ_t's are independent random variables with zero mean and variance equal to:

(2.19) $var\ \{\xi_t\} = \sigma^0 - \tau_t^o\ ;\ t=1,2,\ldots,T$

where

(2.20) $\tau_t = \Phi(-\mu_t/\sigma)\tilde{\tau}_t^-$ *under Tobit*

and

(2.20') $\tau_t = \Phi(-\mu_t)\tilde{\tau}_t^- + [1-\Phi(-\mu_t)]\tilde{\tau}_t^+$ *under probit.*

Applying Liapounoff's Central Limit Theorem to $v'\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\xi_t$ for an arbitrary K-vector v yields

(2.21) $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\xi_t \xrightarrow{d} N(0, A^o - D^o)$

where

(2.22a) $A = \lim_{T\to\infty}\frac{1}{T}\sigma^2\sum_{t=1}^{T} x_t x_t'$

(2.22b) $D = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} x_t\tau_t x_t'$

Note that A-D $\geq$ 0 (i.e., is positive semi-definite) since $\sigma^2 - \tau_t \geq 0$ for all t.

Let v be a member of an arbitrary vector space (finite dimension) such that the $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} v_t v_t'$ exists and let d_t be a Bernoulli variate with success probability equal to P_t. Then the following result holds. (See Amemiya [1973]).

$$(2.24)\quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} v_t d_t v_t' = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} v_t p_t v_t'$$

This result together with the fact that $\tilde{\beta}$ is consistent implies that

$$(2.25)\quad \frac{1}{T}\sum_{t=1}^{T} x_t \frac{1}{\sigma^2}[\sigma^2 - \tilde{\tau}_t^{\tilde{\beta}}]x_t' \xrightarrow{P} \frac{1}{\sigma^2}[A^\circ - D^\circ]$$

It follows from (2.18), (2.22a and b), and (2.25) that

THEOREM:

$$(2.26)\quad \sqrt{T}\,(\hat{\beta}-\beta^\circ) \xrightarrow{d} N\{O,\ \sigma^4(A^\circ - D^\circ)^{-1}\}.$$

The inverse of the covariance matrix in the RHS of (2.26), denoted as the asymptotic precision (also the information matrix in this case) is equal to $\frac{1}{\sigma^4}A^\circ - \frac{1}{\sigma^4}D^\circ$. The first term is the asymptotic precision in the absence of missing observations. The second term is the asymptotical loss in precision (information) due to the missing data pattern.[4]

Having characterized the asymptotic distribution of an efficient estimator, we now assume that a "good" initial estimate, denoted as β^P, is available. Define

$$(2.27)\quad \tilde{y}_t^P = y_t^P - \frac{1}{\sigma^2}\tilde{\tau}_t^P x_t'\beta^P\ ;\quad \tilde{y}^P = (\tilde{y}_1^P, \tilde{y}_2^P, \ldots, \tilde{y}_T^P)'\ \ a\ T\times 1\ vector$$

[4] The theorem above is, of course, not new (e.g., see Amemiya [1973]), but the process of deriving it, motivated by the fact that the ML estimator, β, is a fixed point of (2.13a) thereby leading to (2.15) - (2.26), is a natural outcome of using the EM algorithm. This generalizes easily to more complicated LDV models and provides another way of deriving asymptotic distributions of ML estimators.

(2.28) $\tilde{x}_t^P = x_t' - \frac{1}{\sigma^2}\tilde{\tau}_t^P x_t' \; ; \quad \tilde{X}^P = (\tilde{x}_1^P, \tilde{x}_2^P, \ldots, \tilde{x}_T^P)'$ *a* $T \times k$ *matrix*

where y_t^P defined in (2.8) or (2.8') by substituting ϑ^P for ϑ^κ and $\tilde{\tau}_t^P$ is $\tilde{\tau}_t$ of (2.10) - (2.10') evaluated at ϑ^P. With the above definitions, the following relation is identified

(2.29) $\tilde{y}_t^P = \tilde{x}_t^P \beta^o + \xi_t \; ; \; t = 1,2,\ldots,T$

where ξ_t is the same error term defined in (2.16) or (2.16'). Our estimator, denoted by $\bar{\beta}$, is the result of applying IV regression to (2.29) using x' as the instrument. Formally

(2.30) $\bar{\beta} = (X'\tilde{X}^P)^{-1} X'\tilde{y}^P$

It can be verified that $\bar{\beta}$ is identical to $\beta_{LML} = \beta^P - L_{\beta\beta'}^{P-1} L_\beta^P$ by writing (2.30) as

$$\bar{\beta} = \beta^P + (X'\tilde{X}^P)^{-1} X' u^P$$

and noting that

$$\frac{\partial L(\beta^P)}{\partial \beta} = \frac{1}{\sigma^2} X' u^P$$

and

$$\frac{\partial^2 L(\beta^P)}{\partial \beta \partial \beta'} = -\frac{1}{\sigma^2}(X'\tilde{X}^P)$$

(see Appendix B). For the sake of completeness the asymptotic efficiency of $\bar{\beta}$ when β^P is consistent of order $T^{-1/2}$, is demonstrated below:

$$\bar{\beta} = (X'\tilde{X}^P)^{-1} X'\tilde{y}^P = \beta^P + (X'\tilde{X}^P)^{-1} X' u^P$$

$$= \beta^P + (X'\tilde{X}^P)^{-1} \sum_{t=1}^{T} \{x_t \frac{1}{\sigma^2}[\sigma^2 - \tilde{\tau}_t^P] x_t'(\beta^o - \beta^P) + x_t \xi_t\}$$

(*by replacing* β^P for $\hat{\beta}$ *in* (2.16) *where* $\tilde{\beta} \in \beta^P \beta^o$)

$$= \beta^P + (X'\tilde{X}^P)^{-1} \sum_{t=1}^{T} \{x_t \frac{1}{\sigma^2}[\sigma^2 - \tilde{\tau}_t^P - o_P(1)]x_t' (\beta^o - \beta^P) + x_t \xi_t\}$$

(where $h_T = o_P(1)$ if $h_T \xrightarrow{P} 0$ *as* $T \to \infty$)

$$= \beta^P + (X'\tilde{X}^P)^{-1}X'\{\tilde{X}^P(\beta^o - \beta^P) + X[o_P(1)(\beta^P - \beta^o)] + \xi\}$$

$$= \beta^o + (X'\tilde{X}^P)^{-1}X'\xi + (X'\tilde{X}^P)^{-1}(X'X)[o_P(1)(\beta^P - \beta^o)]$$

Hence

$$(2.31) \quad \sqrt{T}\,(\bar{\beta} - \beta^o) = (\frac{1}{T}X'\tilde{X}^P)^{-1} \frac{1}{\sqrt{T}}X'\xi +$$

$$[\frac{1}{T}X'\tilde{X}^P]^{-1}(\frac{1}{T}X'X)[o_P(1)\sqrt{T}\,(\beta^P - \beta^o)]$$

The convergence of $\frac{1}{T}X'\tilde{X}^P$ and $\frac{1}{T}X'X$ to a proper limit and the assumption that $\sqrt{T}\,(\beta^P - \beta^o)$ is bounded in probability assures that the second term in the RHS of (2.31) goes to zero in probability. By comparing (2.18) to (2.32) it follows that

$$\sqrt{T}\,(\hat{\beta} - \beta^o) - \sqrt{T}\,(\bar{\beta} - \beta^o) = \sqrt{T}\,(\hat{\beta} - \bar{\beta}) \xrightarrow{P} 0$$

This completes the demonstration. A straightforward implication of the above result is that $var(\bar{\beta}) = \sigma^2(X'\tilde{X}^P)^{-1}$ which is the usual formula for the variance of an IV regression coefficient.

The major result is that an efficient estimator can be achieved via instrumental variable regression on a suitably transformed data set. In general, however, σ^{o^2} is unknown and must be estimated along with β^o.

2.B *The Case of Unknown* σ^{o^2}

The normalization $\sigma^o = 1$ excludes the probit model from this case and so only the Tobit model is considered. It is assumed that a "good" initial estimates β^P and σ^{P2} are available. Let us redefine $\tilde{y}_t^P$ and $\tilde{x}_t^P$ in (2.27) and (2.28) by replacing σ with σ^P so $\bar{\beta}$ in (2.30) is now evaluated at σ^P. Denote Ψ as the information matrix and partition it according to β and σ^2; i.e.,

$$(2.32)\quad \Psi = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix}$$

where Ψ_{11}, $\Psi_{12} = \Psi_{21}'$, and Ψ_{22} are, respectively, $k \times k$, $k \times 1$, and 1×1 matrices. Finally, let

$$(2.33a)\quad R^P = -\frac{1}{T}(X'\tilde{X}^P)\frac{1}{\sigma^{P2}}$$

and

$$(2.33b)\quad \Delta^P = \beta^P - \bar{\beta}$$

It is verified in Appendix B that

$$(2.34a)\quad R^P = \frac{\partial^2 L(\vartheta^P)}{\partial\beta\partial\beta'}\frac{1}{T}$$

$$(2.34b)\quad \Delta^P = \left[\frac{\partial^2 L(\vartheta^P)}{\partial\beta\partial\beta'}\right]^{-1}\frac{\partial L(\vartheta^P)}{\partial\beta}$$

Consider now the LML estimator

$$\vartheta_{LML} = \begin{bmatrix} \beta_{LML} \\ \sigma^2_{LML} \end{bmatrix} = \begin{bmatrix} \beta^P \\ \sigma^{P2} \end{bmatrix} - \begin{bmatrix} \Psi_{11}^P & \Psi_{12}^P \\ \Psi_{21}^P & \Psi_{22}^P \end{bmatrix}^{-1} \begin{bmatrix} L_\beta(\vartheta^P) \\ L_{\sigma^2}(\vartheta^P) \end{bmatrix}$$

where $\Psi_{ij}^P = \Psi_{ij}(\vartheta^P)$ is any consistent estimate of $\Psi_{ij}(\vartheta^o)$ for i,j=1,2. From an expression for a partitioned inverse, ϑ_{LML} can be written as

$$\sigma^2_{LML} = \sigma^{P2} - [\ ]^{P-1}\,\Psi^P_{21}\,\Psi^{P-1}_{11}\,L^P_{\beta} - [\ ]^{P-1}\,L^P_{\sigma^2}$$

$$\beta_{LML} = \beta^P - \Psi^{P-1}_{11}\,L^P_{\beta} - \Psi^{P-1}_{11}\Psi^P_{12}[\ ]^{P-1}\Psi^P_{21}\Psi^{P-1}_{11}\,L^P_{\beta} + \quad \Psi^{P-1}_{11}\,\Psi^P_{12}[\ ]^{P-1}L^P_{\sigma^2}$$

where

$$[\ ]^P = \Psi^P_{22} - \Psi^P_{21}\Psi^{P-1}_{11}\Psi^P_{12}$$

Choosing $\Psi^P_{11} = R^P$ and using (2.34b) gives

(2.36a) $\sigma^2_{LML} = \sigma^{P2} + [\ ]^{P-1}(\Psi^P_{21}\,\Delta^P - L^P_{\sigma^2})$

(2.36b) $\beta_{LML} = \bar{\beta} + R^{P-1}\Psi^P_{12}\,(\sigma^{P2} - \sigma^2_{LML})$

$\bar{\beta}$, Δ^P and R^P are constructed using the method outlined above. The other terms needed in the derivation of ϑ_{LML} are Ψ^P_{12}, Ψ^P_{22}, and $L^P_{\sigma^2}$. They can be recovered from the first derivatives of the log likelihood function (Berndt, *at.al.* [1974]) which is given in Appendix B. This completes the discussion of our approach in analyzing single equation LDV models. The results generated by this approach are similar to those methods developed in the literature.

3. TWO EQUATION MODELS

The number of possible models corresponding to different patterns of missing data increases dramatically with the number of equations. This highlights the importance of a unified framework of analysis. Two models, resulting from two different patterns of missing data, are considered. Extension to other possible models is straightforward.

The underlying structure is given by

(3.1) $y^*_{1t} = x'_{1t}\beta^o + u_{1t}\ ;\ t = 1,2,\ldots,T$

$$(3.2)\quad y_{2t}^{*} = x_{2t}'\beta^{o} + u_{2t}\ ;\ t=1,2,\ldots,T$$

$$(3.3)\quad \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \overset{iid}{\sim} N(0\ ,\ \Sigma^{o});\ \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix};\ t=1,2,\ldots,T.$$

Model (I) results from the following observation rule

$$(3.4)\quad y_{1t} = \begin{cases} y_{1t}^{*} \ \text{ if } y_{2t}^{*} \geq 0 \\ \\ NA \ \ otherwise \end{cases} \qquad t\ =\ 1,2,\ldots,T$$

where NA denotes unobserved data

$$(3.5)\quad y_{2t} = \begin{cases} 1 \ \text{ if } y_{2t}^{*} \geq 0 \\ \\ 0 \ \ otherwise \end{cases} \qquad t\ =\ 1,2,\ldots,T$$

Model (II) results from

$$(3.4')\quad y_{1t} = \begin{cases} y_{1t}^{*} \ \text{ if } y_{2t}^{*} \geq 0 \\ \\ NA \ \ otherwise \end{cases} \qquad t\ =\ 1,2,\ldots,T$$

$$(3.5')\quad y_{2t} = \begin{cases} y_{2t}^{*} \ \text{ if } y_{2t}^{*} \geq 0 \\ \\ NA \ \ otherwise \end{cases} \qquad t\ =\ 1,2,\ldots,T$$

For both models it is assumed that x_{jt}, j=1,2 is observed for all t. For examples of situations that give rise to such models see Heckman [1976], Lee and Trost [1978], Lee [1978], and Hanemann and Tsur [1982]. The specifications of the quantities involved are as follows: $X_j = (x_{j1}, x_{j2}, \ldots, x_{jt})'$, $j=1,2$ is a $T \times k$ matrix for which we assume:

i) each column of X_j is statistically independent of $U_j = (u_{j1}, u_{j2}, \ldots, u_{jt})'$, $j=1,2$

ii) $\lim_{T\to\infty} \frac{1}{T} X_j' X_j$ exists and is positive definite, j = 1,2

iii) X_j is of full rank, j = 1,2.

$Y_j^* = (y_{j1}^*, y_{j2}^*, \cdots, y_{jT}^*)'$ is a $T \times 1$ vector for which we observe $Y_j = (y_{j1}, y_{j2}, \ldots, y_{jT})$ according to the observation rule (3.4), (3.5) or (3.4'), (3.5'), j=1,2. β^o and Σ^o are, respectively, $k \times 1$ vector and 2×2 positive definite symetric matrix of parameters to be estimated and $U_j = (u_{j1}, u_{j2}, \cdots, u_{jT})'$; j=1,2 is an unobserved error vector with $(u_{1t}, u_{2t})'$ distributed according to (3.3) independently for t = 1,2, . . ., T. The coefficient vectors in (3.1) and (3.2) are not necessarily identical. They can be distinct or have (some or all) common elements. However, it is always possible to merge them into one coefficient vector by a suitable redefinition of x_{1t} and x_{2t}. Hence there is no loss in generality in the way the model is presented. The missing data pattern can create identification problems. In model (I), without exogenous restrictions on the parameters, σ_2^o, σ_{12}^o, and the part of β^o corresponding only to (3.2) are not identifiable. Therefore, a normalization rule is required and is given by $\sigma_2^o = 1$. No identification problems arise in model (II).

Let us define

$$(3.6a)\quad y_t^* = \begin{bmatrix} y_{1t}^* \\ y_{2t}^* \end{bmatrix};\ Y^* = (y_{1t}^{*'}, y_{2t}^{*'}, \ldots, y_T^{*'})'\ \ a\ 2T \times 1\ vector$$

$$(3.6b)\quad x_t' = \begin{bmatrix} x_{1t}' \\ x_{2t}' \end{bmatrix};\ X = (x_1, x_2, \cdots, x_T)'\ \ a\ 2T \times k\ matrix$$

and

$$(3.6c)\quad u_t = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix};\ U = (u_1', u_2', \ldots, u_T')'\ \ a\ 2T \times 1\ vector$$

The structural model can be written now as

$$y_t^* = x_t'\beta^0 + u_t \; ; \; u_t \underset{\sim}{iid} N(0,\Sigma^0),\; t=1,2,\ldots,T.$$

With the observed dependent variable given by

$$(3.6d)\quad y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} \; ; \; Y = (y_1', y_2', \ldots, y_T')' \; a\; 2T \times 1 vector$$

where y_{jt}, j = 1,2 follows from (3.4) and (3.5) in model (I) and from (3.4') and (3.5') in model (II).

In the application of the EM approach, the following definitions and results will be used:

$$(3.7)\quad \mu_{jt} = x_{jt}'\beta \; , \; j=1,2$$

$$(3.8a)\quad \lambda_t^- = \frac{-\varphi(-\mu_{2t}/\sigma_2)}{\Phi(-\mu_{2t}/\sigma_2)}$$

$$(3.8b)\quad \lambda_t^+ = \frac{\varphi(-\mu_{2t}/\sigma_2)}{1-\Phi(-\mu_{2t}/\sigma_2)}.$$

From (3.8a and b)

$$(3.9a)\quad \Lambda_t^- = \frac{\partial \lambda_t^-}{\partial \mu_{2t}} = -\frac{1}{\sigma_2}\lambda_t^-\left(\frac{\mu_{2t}}{\sigma_2} + \lambda_t^-\right)$$

$$(3.9b)\quad \Lambda_t^+ = \frac{\partial \lambda_t^+}{\partial \mu_{2t}} = -\frac{1}{\sigma_2}\lambda_t^+\left(\frac{\mu_{2t}}{\sigma_2} + \lambda_t^+\right).$$

Let $\tilde{\tau}_{ijt}^- = cov(u_{it}, u_{jt} \,|\, u_{2t} \le -\mu_{2t})$ and $\tilde{\tau}_{ijt}^+ = cov(u_{it}, u_{jt} \,|\, u_{2t} \ge -\mu_{2t})$ for i,j = 1,2. Then using properties of truncated bivariate normal (Johnson and Kotz [1972]. See also Appendix A.), it also follows that

$$(3.10a)\quad \tilde{\tau}_{11t}^- = \sigma_1^2(1 + \rho^2\sigma_2\Lambda_t^-)$$

$$(3.10b)\quad \tilde{\tau}_{22t}^- = \sigma_2^2(1 + \sigma_2\Lambda_t^-)$$

(3.10c) $\tilde{\tau}^{-}_{12t} = \sigma_{12}(1 + \sigma_2\Lambda_t^{-})$

(3.11a) $\tilde{\tau}^{+}_{11t} = \sigma_1^2(1+\rho^2\sigma_2\Lambda_t^{+})$

(3.11b) $\tilde{\tau}^{+}_{22t} = \sigma_2^2(1 + \sigma_2\Lambda_t^{+})$

(3.11c) $\tilde{\tau}^{+}_{12t} = \sigma_{12}(1+\sigma_2\Lambda_t^{+})$

where $\rho = \dfrac{\sigma_{12}}{\sigma_1\sigma_2}$. Let $\tilde{\tau}_t^{-}$, $\tilde{\tau}_t^{+}$ be a 2x2 matrix with ij element equal to $\tilde{\tau}^{-}_{ijt}$, $\tilde{\tau}^{+}_{ijt}$ respectively and define y_{jt}^{κ} to be $E_{\vartheta^{\kappa}}(y_{jt}^{*} \mid y_{jt})$ for j=1,2. That is, under model (I):

(3.12a) $y_{1t}^{\kappa} = q_t y_{1t} + (1-q_t)(\mu_{1t}^{\kappa} + \sigma_{12}^{\kappa}\,\lambda_t^{-\kappa})$

(3.12b) $y_{2t}^{\kappa} = q_t(\mu_{2t}^{\kappa} + \lambda_t^{+\kappa}) + (1-q_t)\,(\mu_{2t}^{\kappa} + \lambda_t^{-\kappa})$

and under model (II):

(3.12a') $y_{1t}^{\kappa} = q_t y_{1t} + (1-q_t)(\mu_{1t}^{\kappa} + \dfrac{\sigma_{12}^{\kappa}}{\sigma_2^{\kappa}}\lambda_t^{-\kappa})$

(3.12b') $y_{2t}^{\kappa} = q_t y_{2t} + (1-q_t)(\mu_{2t}^{\kappa} + \sigma_2^{\kappa}\lambda_t^{-\kappa})$

where q_t is the indicator variable defined as

(3.13) $$q_t = \begin{cases} 1 \text{ if } y_{2t}^{*} \geq 0 \\ 0 \text{ otherwise} \end{cases} \qquad t = 1,2,\ldots,T.$$

With the definitions in (3.12) it is straightforward to verify that

(3.14) $E_{\vartheta^{\kappa}}\{(y_t^{*}-\mu_t)(y_t^{*}-\mu_t)' \mid y_t\} = (y_t^{\kappa}-\mu_t)(y_t^{\kappa}-\mu_t)' + \tilde{\tau}_t^{\kappa}$

where

$$(3.15)\quad \tilde{\tau}_t = (1-q_t)\tilde{\tau}_t^- + q_t \begin{bmatrix} 0 & 0 \\ 0 & \tilde{\tau}_{22t}^+ \end{bmatrix} \quad under\ model\ (I)$$

$$(3.15')\quad \tilde{\tau}_t = (1-q_t)\tilde{\tau}_t^- \qquad under\ model\ (II)$$

and

$$\mu_t = (\mu_{1t}, \mu_{2t})' .$$

The log of the joint density of y_t^*, t = 1,2, . . . , T (disregarding the constant term) is

$$(3.16)\quad f(y^*;\vartheta) = \frac{T}{2}\log |\Sigma^{-1}| - \frac{1}{2}\sum_{t=1}^{T} tr\{\Sigma^{-1}(y_t^*-\mu_t)(y_t^*-\mu_t)'\}$$

where "tr" indicates trace. Using (3.14) gives

$$(3.17)\quad E_{\vartheta^\kappa}\{f(y^*;\vartheta)|y\} = -\frac{1}{2}\sum_{t=1}^{T} tr\{\Sigma^{-1}[(y_t^\kappa-\mu_t)(y_t^\kappa-\mu_t)' + \tilde{\tau}_t^\kappa]\} + \frac{T}{2}\log |\Sigma^{-1}|$$

Maximizing (3.17) over β and Σ yields

$$(3.18a)\quad \beta^{\kappa+1} = X'(I \otimes \Sigma^{\kappa+1^{-1}})X)^{-1}X'(I \otimes \Sigma^{\kappa+1^{-1}})Y^\kappa$$

$$(3.18b)\quad \Sigma^{\kappa+1} = \frac{1}{T}\sum_{t=1}^{T}(y_t^\kappa-\mu_t^{\kappa+1})(y_t^\kappa-\mu_t^{\kappa+1})' + \frac{1}{T}\sum_{t=1}^{T}\tilde{\tau}_t^\kappa$$

where "I" is the identity matrix (of order T) and "$\otimes$" is the Kronecker product.

Equations (3.17) and (3.18) form, respectively, the E step and the M step of an iteration of the EM algorithm which was shown (Dempster *et al.*) to yield an estimate corresponding to a (local) maximum of the likelihood function. Solving for $\beta^{\kappa+1}$ and $\Sigma^{\kappa+1}$ that satisfy (3.18) is not trivial and in fact requires an iterative procedure itself. This can be simplified by first solving (3.18a) using Σ^κ and then solving (3.18b) using $\beta^{\kappa+1}$.

[4] this simplification coresponds to using a GEM procedure (see Dempster *at al.* p. 7).

Equation (3.18a) clearly reveals the similarity of our approach to the seemingly unrelated GLS estimation technique of Zellner [1962]. In fact, it implies that the ML estimator can be achieved via iterative GLS procedure where the missing data are being "filled-in" in each iteration according to (3.12a)-(3.12b) or (3.12a')-(3.12b'), depending on the observation rule (I) or (II) respectively. From there the analysis is carried out as if no data are missing with some modifications that account for the "filled-in" values.[5]

The advantage of this procedure is its simplicity of implementation which involves data transformations (that require the evaluation of the standard normal density and distribution functions) and the readily available GLS option. Its convergence properties and computational efficiency is yet to be compared to other iterative methods such as that of Berndt *et al.*

We turn now to the single iteration procedure. First, consider

3.A *The Case of Known* Σ^o

The first step is to characterize the limiting distribution of the ML estimator. Let us define

$$(3.19)\quad u_t^\kappa = y_t^\kappa - \mu_t^\kappa = \begin{bmatrix} y_{1t}^\kappa \\ y_{2t}^\kappa \end{bmatrix} - \begin{bmatrix} x_{1t}' \\ x_{2t}' \end{bmatrix} \beta^\kappa \; ; \; U^\kappa = \begin{bmatrix} u_1^\kappa \\ \cdots \\ u_T^\kappa \end{bmatrix}, \; a\ 2T\times 1\ vector\ .$$

From (3.18a) $\beta^{\kappa+1} = \beta^\kappa + (X'(I \otimes \Sigma^{-1})U^\kappa$. Since $\hat{\beta}$ is a fixed point of (3.18a) the following orthogonality relation holds

$$(3.20)\quad X'(I \otimes \Sigma^{-1})U^{\hat{\beta}} = 0$$

where the superscript "o" is dropped from Σ until otherwise indicated. By

[5] The identification problem of Model (I) requires the normalization $\sigma_2 = 1$. This constraint should be incorporated into the maximization of (3.18) which is done over β and the identifiable elements of Σ. No such problem arises under model (II).

replacing $\lambda_t^{-\hat{\beta}}$, $\lambda_t^{+\hat{\beta}}$ with their first order expansions around λ_t^{-o}, λ_t^{+o}, respectively, and rearranging we can express $u_t^{\hat{\beta}}$ under model (I) as

$$u_t^{\hat{\beta}} = \{\Sigma-(1-q_t)\begin{bmatrix}1 & \frac{\sigma_{12}}{\sigma_2}\Lambda_t^{-\hat{\beta}}\\ 0 & 1+\sigma_2\Lambda_t^{-\hat{\beta}}\end{bmatrix}\Sigma - q_t\begin{bmatrix}0 & 0\\ 0 & 1+\sigma_2\Lambda_t^{+\hat{\beta}}\end{bmatrix}\Sigma\}\Sigma^{-1}x_t'(\beta^o-\hat{\beta}) + \xi_t$$

and under model (II) as

$$u_t^{\hat{\beta}} = \{\Sigma-(1-q_t)\begin{bmatrix}1 & \frac{\sigma_{12}}{\sigma_2}\Lambda_t^{-\tilde{\beta}}\\ 0 & 1+\sigma_2\Lambda_t^{-\tilde{\beta}}\end{bmatrix}\Sigma\}\Sigma^{-1}x_t'(\beta^o-\hat{\beta})+\xi_t\ .$$

or by using (3.10) and (3.11) $u_t^{\hat{\beta}}$ can be expressed as

$$(3.21)\quad u_t^{\hat{\beta}} = \{\Sigma - \tau_t^{*\tilde{\beta}}\}\Sigma^{-1}x_t'(\beta^o - \hat{\beta}) + \xi_t$$

where

$$(3.22)\quad \tau_t^* = (1-q_t)\tilde{\tau}_t + q_t\begin{bmatrix}0 & 0\\ \tilde{\tau}_{21t}^{+} & \tilde{\tau}_{22t}^{+}\end{bmatrix}\ \textit{under model (I)}$$

$$(3.22')\quad \tau_t^* = (1-q_t)\tilde{\tau}_t^{-}\qquad\qquad \textit{under model (II)}$$

$$(3.23)\quad \xi_t = \begin{bmatrix}u_{1t}\\ u_{2t}\end{bmatrix} - (1-q_t)\begin{bmatrix}u_{1t} - \frac{\sigma_{12}}{\sigma_2}\lambda_t^{-o}\\ u_{2t} - \sigma_2\lambda_t^{-o}\end{bmatrix} - q_t\begin{bmatrix}0\\ u_{2t}-\sigma_2\lambda_t^{+0}\end{bmatrix}\ \textit{under model (I)}$$

$$(3.23')\quad \xi_t = \begin{bmatrix}u_{1t}\\ u2t\end{bmatrix} - (1-q_t)\begin{bmatrix}u_{1t} - \frac{\sigma_{12}}{\sigma_2}\lambda_t^{-o}\\ u_{2t} - \sigma_2\lambda_t^{-o}\end{bmatrix}\ \textit{under model (II)}$$

and $\tilde{\beta} \in \hat{\beta}\beta^o$.

It follows from (3.20) and (3.21) that

$$(3.24)\quad \sqrt{2T}\ (\hat{\beta}-\beta^o) = \{\frac{1}{2T}X'(I\otimes\Sigma^{-1})[(I\otimes\Sigma) - \Gamma^{*\tilde{\beta}}](I\otimes\Sigma^{-1})X\}^{-1}$$

$$\frac{1}{\sqrt{2T}} X'(I \otimes \Sigma^{-1})\,\xi$$

where Γ^* is a block diagonal matrix with the t^{th} block equal to τ_t^* and $\xi = (\xi_1',\ldots,\xi_T')'$. When taking the q_t in (3.23) to be a Bernoulli variate with success probability of $1-\Phi(-\mu_{2t}^0/\sigma_2^0)$ the ξ_t are found to be 2×1 independent random vectors with zero mean and

(3.25) $var(\xi_t) = \Sigma^0 - \tau_t^0 \;;\; t=1,2,\ldots T$

where

(3.30) $\tau_t = (1-\Phi_t)\begin{bmatrix} 0 & \tilde{\tau}_{12t}^{+} \\ \tilde{\tau}_{12t}^{+} & \tilde{\tau}_{22t}^{+} \end{bmatrix} + \Phi_t\,\tilde{\tau}_t^{-}$ *under model (I)*

(3.30') $\tau_t = \Phi_t\,\tilde{\tau}_t^{-}$ *under model (II)*

and $\Phi_t = \Phi(-\mu_{2t}/\sigma_2)$. Let Γ be a $2T\times 2T$ block diagonal matrix with the t^{th} block equal to τ_t and define

(3.31a) $A = \lim_{T\to\infty} \frac{1}{2T} X'(I \otimes \Sigma^{-1})X$

(3.31b) $D = \lim_{T\to\infty} \frac{1}{2T} X'(I \otimes \Sigma^{-1})\Gamma(I \otimes \Sigma^{-1})X$

(3.31c) $D^* = \lim_{T\to\infty} \frac{1}{2T} X'(I \otimes \Sigma^{-1})\Gamma^*(I \otimes \Sigma^{-1})X$

The assumed convergence of $\frac{1}{2T} X_j' X_j$, j = 1,2 assures the existence of A, D and D*. Note that under model (II) D = D*. This can be seen from the definitions of τ_t and τ_t^* in (3.30') and (3.22'), respectively, and from result (2.25). Formally,

(3.31d) $D = D^*$ *under model (II)*

Since $\tilde{\beta}$ is consistent, it follows that

$$(3.32)\quad \frac{1}{2T}X'(I\otimes\Sigma^{-1})[(I\otimes\Sigma)-\Gamma^{*\tilde{\beta}}](I\otimes\Sigma^{-1})X \xrightarrow{P} A^{o}-D^{*o}$$

and from (3.25)

$$(3.33)\quad \frac{1}{2T}X'(I\otimes\Sigma^{-1})\xi \xrightarrow{d} N(0,\ A^{o}-D^{o})$$

These two results together with (3.24) yield

THEOREM:

$$(3.34)\quad \sqrt{2T}\,(\hat{\beta}-\beta^{o}) \xrightarrow{d} N\{0,(A^{o}-D^{*o})^{-1}(A^{o}-D^{o})(A^{o}-D^{*o})^{-1}\}\,.$$

Note that in model (II) D = D*. Hence the asymptotic covariance matrix reduces to $(A^{o}-D^{o})^{-1}$.

The process of obtaining result (3.34) relies heavily on the EM approach (starting from the fact that $\hat{\beta}$ is a fixed point of (3.18a) hence (3.20), etc.) and is a straightforward extension of the single equation case that can be generalized easily to other multiple equation LDV models. The asymptotical precision (or information in these cases) is given by the inverse of the covariance matrix and is equal to $A-[D-(D-D^{*})(A-D)^{-1}(D-D^{*})]$ for model (I) and $A-D$ for model (II). In both cases A is the asymptotic precision in the absence of unobservables. Hence the second term can be interpreted as the asymptotic loss of information due to the unobservable pattern.

Suppose now that a "good" initial estimate β^{P} is available. Let us define

$$(3.35)\quad \tilde{y}_t^{P} = y_t^{P} - \tau_t^{*P}\Sigma^{-1}x_t'\beta^{P}\ ;\ \tilde{Y}^{P} = (\tilde{y}_1^{P},\tilde{y}_2^{P},\ldots,\tilde{y}_T^{P})'\ \ a\ 2T\times 1\ vector$$

$$(3.36a)\quad \tilde{x}_t^{*P} = x_t' - \tau_t^{*P}\Sigma^{-1}x_t'\ ;\ \tilde{X}^{*P} = (\tilde{x}_1^{*P},\tilde{x}_2^{*P},\ldots,\tilde{x}_T^{*P})'\ \ a\ 2T\times k\ matrix$$

$$(3.36b)\quad \tilde{x}_t^{P} = x_t' - \tau_t^{P}\Sigma^{-1}x_t'\ ;\ \tilde{X}^{P} = (\tilde{x}_1^{P},\tilde{x}_2^{P},\ldots,\tilde{x}_T^{P})'\ \ a\ 2T\times k\ matrix$$

where τ_t^{*} and τ_t defined in (3.22)-(3.22') and (3.30)-(3.30') respectively.[6] With the

[6] Note that under model (II) $\tau_t^{*} = \tau_t$ which implies $\tilde{x}_t^{*P}=\tilde{x}_t^{P}$.

above definitions the following relation can be identified

(3.37) $\tilde{y}_t^P = \tilde{x}_t^{*P\prime}\beta^o + \xi_t;\ t=1,2,\ldots,T$

where ξ_t is the same error defined in (3.23) or (3.23'). Our estimator $\bar{\beta}$ is the outcome of applying an IV regression on (3.37) using $X'(I \otimes \Sigma^{-1})$ as the instrument. That is

(3.38) $\bar{\beta} = (X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}X'(I \otimes \Sigma^{-1})\tilde{Y}^P$

Unlike in one equation models, in this case it is impossible to show the equivalence of $\bar{\beta}$ and $\beta_{LML} = \beta^P - L_{\beta\beta'}^{P-1}L_\beta^P$ by identifying the components at the RHS of (3.38) as the appropriate derivatives of the log likelihood function. However, it is proven below that

THEOREM:

$\bar{\beta}$ is asymptotically efficient, provided that β^P is consistent of order $T^{-1/2}$.

PROOF:

$$\bar{\beta} = (X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}X'(I \otimes \Sigma^{-1})\tilde{Y}^P$$

$$= \beta^P + (X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}X'(I \otimes \Sigma^{-1})U^P$$

(where $U^P = \tilde{Y}^P - \tilde{X}^{*P}\beta^P = Y^P - X\beta^P$*)*

$$= \beta^P + (X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}\sum_{t=1}^{T} x_t\Sigma^{-1}u_t^P$$

$$= \beta^P + (X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}\sum_{t=1}^{T} x_t\Sigma^{-1}\{(\Sigma - \tau_t^{*\tilde{\beta}})\Sigma^{-1}x_t'(\beta^o - \beta^P) + \xi_t\}$$

(from (3.21) *by replacing* $\hat{\beta}$ *with* β^P and *evaluating at* $\tilde{\beta} \in \beta^P\beta^o$ *)*

$$= \beta^P + (X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}\sum_{t=1}^{T} x_t\Sigma^{-1}\{(\Sigma - \tau_t^{*P} - o_P(1)I)\Sigma^{-1}x_t'(\beta^o - \beta^P) + \xi_t\}$$

(where the identify matrix multiplying $o_P(1)$ is of order 2)

$$= \beta^P + (X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}X'(I \otimes \Sigma^{-1})\{\tilde{X}^{*P}(\beta^o - \beta^P) + \xi + X[o_P(1)(\beta^P - \beta^o)]\}$$

$$= \beta^o + (X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}X'(I \otimes \Sigma^{-1})\xi +$$

$$(X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}X'(I \otimes \Sigma^{-1})X[o_P(1)(\beta^P - \beta^o)]$$

Hence

$$\sqrt{2T}\,(\bar{\beta} - \beta^o) = (\frac{1}{2T}X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}\frac{1}{\sqrt{2T}}X'(I \otimes \Sigma^{-1})\xi +$$

$$(\frac{1}{2T}X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}\ \frac{1}{2T}X'(I \otimes \Sigma^{-1})X[o_P(1)\ \sqrt{2T}\ (\beta^P - \beta^o)]$$

The $\sqrt{T}$-consistency of β^P and the convergence of $\frac{1}{2T}X'(I \otimes \Sigma^{-1})\tilde{X}^{*P}$ and $\frac{1}{2T}X'(I \otimes \Sigma^{-1})X$ to proper limits assures that the second term in the RHS above approaches zero in probability as T → ∞, that is

$$\sqrt{2T}\ (\bar{\beta} - \beta^o) = (\frac{1}{2T}X'(I \otimes \Sigma^{-1})\tilde{X}^{*P})^{-1}\frac{1}{2T}X'(I \otimes \Sigma^{-1})\xi + o_P(1)$$

From (3.24) then

$$\sqrt{2T}\ (\hat{\beta} - \beta^o) - \sqrt{2T}\ (\bar{\beta} - \beta^o) = \sqrt{2T}\ (\hat{\beta} - \bar{\beta}) \xrightarrow{P} 0 \quad Q.E.D.$$

A direct implication of the above theorem is that $\bar{\beta}$ defined in (3.38) above is a GML estimator. To see this from another angle define

(3.39a) $Q^{*P} = X'(I \otimes \Sigma^{-1})\tilde{X}^{*P}$

(3.39b) $Q^P = X'(I \otimes \Sigma^{-1})\tilde{X}^P$.

Then from (3.34) and (3.31a) to (3.31c) it follows that

(3.40) $R^P = \frac{1}{2T} Q^{*P} Q^{P-1} Q^{*P} \xrightarrow{P} \Psi_{11}(\vartheta^\circ)$ *the information matrix.*

Now $\bar{\beta}$ can be expressed as

(3.41) $\bar{\beta} = \beta^P - R^{P-1} V^P$

where

(3.42) $V^P = \frac{1}{2T} Q^{*P} Q^{P-1} X'(I \otimes \Sigma^{-1}) U^P$.

The fact that $\bar{\beta}$ and β_{LML} are both efficient implies:

LEMMA:

Let α be the vector comprised of the distinct elements of Σ, V(.) as in (3.42), and β^P a $\sqrt{T}$-consistent estimate of β°. Then

(3.43) $\sqrt{T}\ [V(\beta^P, \alpha^\circ) - L_\beta(\beta^P, \alpha^\circ)] \xrightarrow{P} 0$.

Proof is given in Appendix C. The above Lemma together with (3.40) imply that $\bar{\beta}$ is a GML estimator.

The critical point here is that, in the presence of a "good" initial estimate, efficiency is attainable via (single) IV regression on a suitably transformed data set, using filled-in values for the "missing" dependent variables.

Achieving "good" initial estimators is quite feasible (Amemiya [1973], Heckman [1976], Quandt and Ramsey [1978], Green [1981,1983]) however, it is rarely the case that the covariance matrix is known *a priori.*

3.B *The Case of Unknown* Σ°

It is assumed that "good" initial estimators β^P, α^P, of β° , α° are available where α is the vector of the identifiable parameters of Σ. That is

(3.44) $\alpha = (\sigma_1^2, \sigma_{01}/\sigma_2)$ *a* 2×1 *vector, under model* (I)

(3.44′) $\alpha = (\sigma_1^2, \sigma_{01}, \sigma_2^2)'$, *a* 3×1 *vector, under model* (II)

The information matrix, partitioned according to β and α, is given by

$$(3.45) \quad \Psi = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix}$$

where Ψ_{11}, $\Psi_{12} = \Psi_{21}'$, and Ψ_{22} are, respectively, $k \times k$, $k \times 2$ ($k \times 3$), and 2×2 (3×3) matrices in model (I) (model (II)). The GML estimator is defined as

$$(3.46) \quad \vartheta_{GML} = \begin{bmatrix} \beta_{GML} \\ \alpha_{GML} \end{bmatrix} = \begin{bmatrix} \beta^P \\ \alpha^P \end{bmatrix} - \begin{bmatrix} \Psi_{11}^P & \Psi_{12}^P \\ \Psi_{21}^P & \Psi_{22}^P \end{bmatrix}^{-1} \begin{bmatrix} \omega_\beta^P \\ \omega_\alpha^P \end{bmatrix}$$

where Ψ_{ij}^P is any consistent estimate of Ψ_{ij}^0, i, j = 1,2; and $(\omega_\beta^{P\prime}, \omega_\alpha^{P\prime})'$ satisfy

$$\sqrt{T}\,[(\omega_\beta^{P\prime}, \omega_\alpha^{P\prime}) - (L_\beta^{P\prime}, L_\alpha^{P\prime})] \xrightarrow{P} 0 .$$

When using expressions for the inverse of a partitioned matrix, β_{GML} and α_{GML} can be written as

$$(3.47a) \quad \alpha_{GML} = \alpha^P - [\ ]^{P-1}\Psi_{21}^P\Psi_{11}^{P-1}\omega_\beta - [\ ]^{P-1}\omega_\alpha^P$$

$$(3.47b) \quad \beta_{GML} = \beta^P - \Psi_{11}^{P-1}\omega_\beta^P - \Psi_{11}^{P-1}\Psi_{12}[\ ]^{P-1}\Psi_{21}^P\Psi_{11}^{P-1}\omega_\beta^P + \Psi_{11}^{P-1}\Psi_{12}[\ ]^{P-1}\omega_\alpha^P$$

$$\textit{where } [\ ]^P = \Psi_{22}^P - \Psi_{21}^P\Psi_{11}^{P-1}\Psi_{12}^P$$

Let us redefine $\tilde{y}_t^P$, $\tilde{x}_t^{*P}$, and $\tilde{x}_t^P$ in (3.35), (3.36a), and (3.36b) by substituting Σ^P for Σ. Also redefine τ_t of (3.30) or (3.30′) by replacing $\Phi(-\mu_{2t}/\sigma_2)$ with $1 - q_t$. Finally define

$$(3.48) \quad \Delta^P = \beta^P - \bar{\beta} = R^{P-1}V^P$$

That R^P, defined in (3.40) with Σ^P replacing Σ, remains a consistent estimator of

Ψ_{11}^{o} is not surprising. What is less obvious is that, under some weak conditions involving the existence of $\Psi(.)$ and $\partial V(.)/\partial\alpha$ at an appropriate vicinity of ϑ^o, (3.43) implies

LEMMA:

$$(3.48)\quad \sqrt{T}\ (V(\beta^P, \alpha^P) - L_\beta(\beta^P, \alpha^P)) \xrightarrow{P} 0$$

Provided that β^P, α^P are $\sqrt{T}$-consistent estimates of β^o, α^o. Proof is given in Appendix C.

The fact that $R^P \xrightarrow{P} \Psi_{11}^{o}$ together with (3.48) allows one to substitute $\Delta^P = R^{P^{-1}} V^P$ for $\Psi_{11}^{P^{-1}}\omega_\beta^P$ and R^P for Ψ_{11}^P in (3.47a) and (3.47b) and to arrive with the following expressions for ϑ_{GML}:

$$(3.49a)\quad \alpha_{GML} = \alpha^P + [\ ]^{P^{-1}} \{\ \Psi_{21}^P \Delta^P - \omega_\alpha^P\ \}$$

$$(3.49b)\quad \beta_{GML} = \bar{\beta} + R^{P^{-1}} \Psi_{12}^P (\alpha^P - \alpha_{GML})$$

$\bar{\beta}$, Δ^P and R^P are constructed using the method outlined above. The other terms needed in the derivation of ϑ_{GML} are Ψ_{12}^P, Ψ_{22}^P, and ω_α^P. They can be recovered from the first derivatives of the log likelihood function (berndt *at al* [1974] which are given in Appendix B.

4. SUMMARY

Methods of Maximum Likehood estimation of Limited Dependent Variables models are investigated in the context of the EM algorithm developed by Dempster *et al* [1977]. It is shown that this approach lends itself to an integrative framework of analysis that unifies various LDV models and link them to classical linear model theory. The computational consequences of this approach are: i) ML estimators can be obtained via iterative OLS or GLS procedure when the EM

algorithm is applied directly; ii) A unique method for deriving an asymptotically efficient single iteration estimator is achieved which produces an outcome identical to that of the standard (Hessian) method under single equation models but differs from those of other methods in multiple equation models.

The extension of this technique to models of more then two equations is straightforward but the use of this approach in simultaneous equatons LDV models (eg., Amemiya [1974,1978], Heckman [1978], Lee [1981]) requires further research. Another topic for investigation is the practical question of convergence and computational efficiency in empirical examples. In Tsur [1983] I have estimated a 3-equation LDV model using the EM algorithm outlined above as well as two nonlinear ML algorithms. The results showed that EM wase the only method which converged over the whole parameter space, including the identifiable covariance terms, where as the ML algorithms broke down. However, further computatonal experience is required before a comprehensive evaluation of the role of the EM method in LDV models can be made.

APPENDIX A

Let $(u_1,u_2)'$ be $N\{0, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ & \sigma_2^2 \end{bmatrix}\}$ and define $\tilde{u}_1 = u_1/\sigma_1$, $\tilde{u}_2 = u_2/\sigma_2$, and $\rho = \sigma_{12}/\sigma_1\sigma_2$. The density function of $\tilde{u}_2 | \tilde{u}_2 > a$ and $\tilde{u}_2 \mid \tilde{u}_2 \leq a$ are respectively $[1-\Phi(a)]^{-1}\varphi(x)$ for x > a and $[\Phi(a)]^{-1}\varphi(y)$ for y ≤ a.

$$(A.1)\quad E(\tilde{u}_2 \mid \tilde{u}_2 > a) = [1-\Phi(a)]^{-1} \int_a^{\infty} x\varphi(x)dx$$

$$= [1-\Phi(a)]^{-1} \int_a^{\infty} -\frac{d\varphi(x)}{dx}dx$$

$$= [1-\Phi(a)]^{-1}\varphi(a) \stackrel{def}{=} \lambda^+(a)$$

In the same manner

$$(A.2)\quad E(\tilde{u}_2 \mid \tilde{u}_2 \leq a) = [\Phi(a)]^{-1} \int_{-\infty}^{a} x\varphi(x)dx$$

$$= -[\Phi(a)]^{-1}\varphi(a) \stackrel{def}{=} \lambda^-(a)$$

$$E(\tilde{u}_2^2 \mid \tilde{u}_2 > a) = [1-\Phi(a)]^{-1} \int_a^{\infty} x^2\varphi(x)dx$$

(integrating by parts)

$$= 1 + a\lambda^+(a)$$

Hence

$$(A.3)\quad var(\tilde{u}_2 \mid \tilde{u}_2 > a) = 1+a\lambda^+(a) - [\lambda^+(a)]^2 = 1+\Lambda^+(a)$$

where

$$\Lambda^+(a) \stackrel{def}{=} \frac{-\partial\lambda^+(a)}{\partial a} = -\lambda^+(a)[-a + \lambda^+(a)]$$

Note that in the text $a = -\mu_t/\sigma$ and $\Lambda_t^s = \partial\lambda_t^s/\partial\mu = \frac{1}{\sigma}\frac{\partial\lambda_t^s}{\partial(\mu_t/\sigma)}$ for s = - and +.

In the same manner

$$(A.4)\quad var(\tilde{u}_2 \mid \tilde{u}_2 \le a) = 1 + a\lambda^-(a) - [\lambda^-(a)]^2$$

$$= 1 + \Lambda^-(a)$$

where

$$\Lambda^-(a) \overset{def}{=} -\frac{\partial\lambda^-(a)}{\partial a} = a\lambda^-(a) - [\lambda^-(a)]^2.$$

The variate $\tilde{u}_1$ can be expressed as

$$(A.5)\quad \tilde{u}_1 = \rho\tilde{u}_2 + (1-\rho^2)^{1/2}z$$

where $z \sim N(0, 1)$ independent of $\tilde{u}_2$. Therefore

$$(A.6)\quad E(\tilde{u}_1 \mid \tilde{u}_2 > a) = \rho E(\tilde{u}_2 \mid \tilde{u}_2 > a) = \rho\lambda^+(a)$$

$$(A.7)\quad E(\tilde{u}_1 \mid \tilde{u}_2 \le a) = \rho E(\tilde{u}_2 \mid \tilde{u}_2 \le a) = \rho\lambda^-(a)$$

Using (A.5)

$$E(\tilde{u}_1^2 \mid \tilde{u}_2 > a) = E\{\rho^2\tilde{u}_2^2 + (1-\rho^2)z^2 + \rho(1-\rho^2)^{1/2}\tilde{u}_2 z \mid \tilde{u}_2 > a\}$$

$$= \rho^2 E(\tilde{u}_2^2 \mid \tilde{u}_2 > a) + 1 - \rho^2$$

$$(A.8)\quad var(\tilde{u}_1 \mid \tilde{u}_2 > a) = \rho^2 E(\tilde{u}_2^2 \mid \tilde{u}_2 > a) + 1 - \rho^2 - \rho^2[\lambda^+(a)]^2$$

$$= \rho^2\, var(\tilde{u}_2 \mid \tilde{u}_2 > a) + 1 - \rho^2$$

$$= \rho^2(1 + \Lambda^+(a)) + 1 - \rho^2$$

$$= 1 + \rho^2\Lambda^+(a)$$

In the same manner

$$(A.9)\quad var(\tilde{u}_1 \mid \tilde{u}_2 \leq a) = 1+\rho^2\Lambda^-(a)$$

Using (A.5) again

$$E(\tilde{u}_1 \tilde{u}_2 \mid \tilde{u}_2 > a) = E\{\rho\tilde{u}_2^2+(1-\rho^2)^{\frac{1}{2}}z\tilde{u}_2 \mid \tilde{u}_2 > a\}$$

$$= \rho[var(\tilde{u}_2 \mid \tilde{u}_2 > a) + [\lambda^+(a)]^2]$$

$$= \rho(1+\Lambda^+(a)) + \rho[\lambda^+(a)]^2$$

Hence

$$(A.10)\quad cov(\tilde{u}_1, \tilde{u}_2 \mid \tilde{u}_2 > a) = \rho(1+\Lambda^+(a)) + \rho[\lambda^+(a)]^2 -$$

$$\rho[\lambda^+(a)]^2$$

$$= \rho(1+\Lambda^+(a))$$

In the same manner

$$(A.11)\quad cov(\tilde{u}_1,\tilde{u}_2 \mid \tilde{u}_2 \leq a) = \rho(1+\Lambda^-(a))$$

The extension to the corresponding moments of $u_1 = \sigma_1\tilde{u}_1$ and $u_2 = \sigma_2\tilde{u}_2$ is straightforward.

APPENDIX B

This appendix provides likelihood functions and their first derivatives for the Tobit model and models (I) and (II) of the two-equation case.

The Tobit Model

The log of the likelihood of the t^{th} observation is denoted by $L_t(\vartheta)$ and is given by

$$(B.1)\quad L_t(\vartheta) = q_t\{-\tfrac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_t - x_t\beta)^2\} + (1-q_t)\log\Phi(-x_t'\beta/\sigma) + constant$$

$$(B.2)\quad \frac{\partial L_t(\vartheta)}{\partial\beta} = x_t\,\frac{1}{\sigma^2}\{q_t(y_t - x_t'\beta) + (1-q_t)\sigma\lambda_t^-\}$$

$$(B.3)\quad \frac{\partial L_t(\vartheta)}{\partial\sigma^2} = \frac{1}{2\sigma^4}\{q_t[(y_t - x_t'\beta)^2 - \sigma^2] + (1-q_t)\sigma\lambda_t^- x_t'\beta\}$$

Note from (B.2) that

$$\frac{\partial L(\vartheta^P)}{\partial\beta} = \frac{1}{\sigma^{P2}}\sum_{t=1}^{T} x_t\{q_t(y_t - x_t'\beta^P) + (1-q_t)\sigma^P\lambda_t^{-P}\}$$

$$= \frac{1}{\sigma^{P2}}\sum_{t=1}^{T} x_t u_t^P = \frac{1}{\sigma^{P2}}X'u^P$$

where

$$u_t^P = y^P - x_t'\beta^P\ ;\ u^P = (u_1^P, u_2^P, \cdots, u_t^P)'\ .$$

Likewise

$$\frac{\partial^2 L(\vartheta^P)}{\partial\beta\partial\beta'} = -\frac{1}{\sigma^{P2}}\sum_{t=1}^{T} x_t\{q_t - (1-q_t)\Lambda_t^{-P}\}x_t'$$

$$= -\frac{1}{\sigma^{P2}}\sum_{t=1}^{T} x_t\{1 - \frac{1}{\sigma^{P2}}\tilde{\tau}_t^P\}x_t'$$

$$= -\frac{1}{\sigma^{P2}}X'\tilde{X}^P$$

where $\tilde{\tau}_t$ and $\tilde{X}^P$ are defined in (2.18) and (2.29) respectively.

Two-Equation Model (I)

$$(B.4)\quad L_t(\vartheta) = q_t\{-\tfrac{1}{2}\log\sigma_1^2 - \frac{1}{2\sigma_1^2}(y_{1t}-x_{1t}'\beta)^2 + \log\Phi(g_t)\} + (1-q_t)\log\,\Phi(-x_{2t}'\beta) + constant$$

where

$$(B.5)\quad g_t = \{x_{2t}'\beta + \frac{\sigma_{12}}{\sigma_1^2}(y_{1t}-x_{1t}'\beta)\} \,/\, (1-\rho^2)^{\frac{1}{2}}$$

and recall that σ_2 is normalized to unity.

$$(B.6)\quad \frac{\partial L_t(\vartheta)}{\partial\beta} = q_t\{x_{1t}\,\frac{1}{\sigma_1^2}(y_{1t}-x_{1t}'\beta) + \frac{h_t}{(1-\rho^2)^{\frac{1}{2}}}[x_{2t}-\frac{\sigma_{12}}{\sigma_1^2}x_{1t}]\} + (1-q_t)\lambda_t^{-}x_{2t}$$

where

$$(B.7)\quad h_t = \frac{\varphi(g_t)}{\Phi(g_t)}$$

$$(B.8)\quad \frac{\partial L_t(\vartheta)}{\partial\sigma_1^2} = q_t\,\frac{1}{2\sigma_1^4}\{-\sigma_1^2+(y_{1t}-x_{1t}'\beta)^2 - h_t\,\frac{2\sigma_{12}}{1-\rho^2}[(y_{1t}-x_{1t}'\beta)+\frac{g_t}{2}(1-\rho^2)^{-\frac{1}{2}}]\}$$

$$(B.9)\quad \frac{\partial L_t(\vartheta)}{\partial\sigma_{12}} = q_t h_t\,\frac{1}{\sigma_1^2(1-\rho^2)^{\frac{1}{2}}}[(y_{1t}-x_{1t}'\beta) + g_t\sigma_{12}(1-\rho^2)^{-\frac{1}{2}}]$$

Two-Equation Model (II)

$$(B.10)\quad L_t(\vartheta) = q_t\{\tfrac{1}{2}\log|\Sigma^{-1}| - \tfrac{1}{2}(y_t-x_t'\beta)'\Sigma^{-1}(y_t-x_t'\beta)\} + (1-q_t)\log\,\Phi(-x_{2t}'\beta/\sigma_2) + constant$$

$$(B.11)\quad \frac{\partial L_t(\vartheta)}{\partial\beta} = q_t\{-2x_t'\Sigma^{-1}y_t + 2x_t\Sigma^{-1}x_t'\beta\} - (1-q_t)\lambda_t^{-}\,\frac{x_{2t}}{\sigma_2}$$

$$(B.12)\quad \frac{\partial L_t(\vartheta)}{\partial \sigma_1^2} = tr\left\{ \frac{\partial L_t(\vartheta)}{\partial \Sigma^{-1}} \frac{\partial \Sigma^{-1}}{\partial \sigma_1^2} \right\}$$

$$= tr\left\{ \frac{q_t}{2}[\Sigma - (y_t - x_t'\beta)(y_t - x_t'\eta)'] \frac{-1}{(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)^2} \begin{bmatrix} \sigma_2^{11} & -\sigma_{12}\sigma_2^2 \\ -\sigma_{12}\sigma_2^2 & -\sigma_{12}^2 \end{bmatrix} \right\}$$

$$(B.13)\quad \frac{\partial L_t(\vartheta)}{\partial \sigma_{12}} = tr\left\{ \frac{\partial L_t(\vartheta)}{\partial \Sigma^{-1}} \frac{\partial \Sigma^{-1}}{\partial \sigma_{12}} \right\}$$

$$= tr\left\{ \frac{q_t}{2}[\Sigma - (y_t - x_t'\beta)(y_t - x_t'\beta)'] \frac{1}{(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)^2} \begin{bmatrix} -2\sigma_{12}\sigma_2^2 & -\sigma_1^2\sigma_2^2 - \sigma_{12}^2 \\ -\sigma_{1_2}\sigma_2^2 - \sigma_{12}^2 & -2\sigma_{12}\sigma_1^2 \end{bmatrix} \right\}$$

$$(B.14)\quad \frac{\partial L_t(\vartheta)}{\partial \sigma_2^2} = tr\left\{ q_t \tfrac{1}{2}[\Sigma - (y_t - x_t'\beta)(y_t - x_t'\beta)'] \frac{-1}{(\sigma_1^2\sigma_2^2 - \sigma^{12^2})^2} \begin{bmatrix} -\sigma_{12}^2 & -\sigma_{12}\sigma_1^2 \\ -\sigma_{12}\sigma_1^2 & \sigma_1^4 \end{bmatrix} \right\} +$$

$$\frac{(1-q_t)}{2}\sigma_2\lambda_t^- \frac{x_{2t}'\beta}{\sigma_2^4}$$

APPENDIX C

This appendix clarifies what is meant by *efficiency* in this paper and provides proofs for (3.43) and (3.48).

The concept of *efficiency of an estimator* takes on several forms in statistical literature. (See Rao [1973, pp. 346-351] and Hájek [1972] for a discussion of the different approaches.) One such form entails consistency and attainment of the Cramer-Rao lower bound. That is, an estimator $\hat{\vartheta}_T$ is (asymptotically) efficient if

$$(C.1)\quad P_{\vartheta^0}\{\sqrt{T}\ (\hat{\vartheta}_T - \vartheta^0)\} \rightarrow N(0, \Psi^{-1}(\vartheta^0))$$

where $P_\vartheta\{x\}$ indicates the distribution law of x when ϑ holds and $\Psi(.)$ is the

information matrix. Definition (C.1) suffers from some deficiencies as is demonstrated in Rao [1973, p. 347]. Hence, a slightly modified definition is often employed (e.g., Bickel [1982]) and will be used here:

Definition: *An estimator* $\hat{\vartheta}_T$ *is* (*asymptotically*) *efficient* if

(C.2) $P_{\vartheta^T}\{\sqrt{T}\ (\hat{\vartheta}_T - \vartheta^T)\} \rightarrow N(0, \Psi^{-1}(\vartheta^0))$ for *any sequence* ϑ^T

such that $\sqrt{T}\ (\vartheta^T - \vartheta^0)$ *stay bounded.*

Since in the text the ML estimators are shown to satisfy (C.1), the question arises as to whether they also satisfy (C.2), and are therefore efficient.

Le Cam [1960] provided a general answer to this question. He proved that, in general, (C.1) implies (C.2) if:

(i) $\hat{\vartheta}_T$ is a "right" root of the likelihood function in the sense that it maximizes the likelihood at a "small" neighborhood of ϑ^0. That is, let $N_{\vartheta^0}(T) = \{\vartheta^T \mid \vartheta^T = \vartheta^0 + \frac{h_T}{\sqrt{T}}$, h_T *is a bounded sequence* $\}$ then $\hat{\vartheta}_T = \underset{\vartheta \in N_{\vartheta^0}}{MAX^{-1}}\{ L(\vartheta) \}$.

(ii) The Local Asimptotic Normality (LAN) assumption holds (Le Cam *op cit.*).

the LAN assumption implies:

$$\Lambda(\vartheta^0, h_T, T) \overset{def}{=} L(\vartheta^0 + \frac{h_T}{\sqrt{T}}) - L(\vartheta^0) = h'_T \Delta_T(\vartheta^0) - \tfrac{1}{2} h'_T \Psi(\vartheta^0) h_T + o_p(1)$$

where

$$P_{\vartheta^0}\{ \Delta_T(\vartheta^0) \} \rightarrow N\{ 0 , \Psi(\vartheta^0) \} \quad \text{and } h_T \textit{ stay bounded.}$$

A sketch of the proof is as follow:

maximizing $\Lambda(.)$ over h_T (disregarding the residual term) to get

$$\hat{h}_T = \Psi^{-1}(\vartheta^0)\Delta_T(\vartheta^0)$$

from (i)

$$\hat{\vartheta}_T = \vartheta^0 + \frac{\hat{h}_T}{\sqrt{T}}$$

and from (ii)

$$COV_{\vartheta^0}\{\sqrt{T}(\hat{\vartheta}_T - \vartheta^0), h'_T\Delta_T(\vartheta^0)\} \to h_T .$$

From (C.1) and LAN

$$P_{\vartheta^0}\{\sqrt{T}(\hat{\vartheta}_T - \vartheta^0), h'_T\Delta_T(\vartheta^0)\} \to N\{0, \begin{bmatrix}\Psi^{-1}(\vartheta^0) & h_T \\ h'_T & \Psi(\vartheta^0)\end{bmatrix}\}$$

which implies, using theorem of Le Cam [1960], that

$$P_{\vartheta^T}\{\sqrt{T}(\hat{\vartheta}_T - \vartheta^0)\} \to N\{h_T, \Psi^{-1}(\vartheta^0)\}$$

where $\vartheta^T = \vartheta^0 + \frac{h_T}{\sqrt{T}}$, hence

$$P_{\vartheta^T}\{\sqrt{T}(\hat{\vartheta}_T - \vartheta^T)\} \to N\{0, \Psi^{-1}(\vartheta^0)\}$$

For the models considered in the text the ML estimators satisfy (i) by cnstruction, and the asumption that $\frac{1}{T}X'X$ approaches a proper limit assures that the LAN condition holds.

A proof of (3.43) and (3.48):

Let B, A be a partition of Θ so $\vartheta = (\beta, \alpha)$; $\beta \in B$, $\alpha \in A$, and partition $\Psi(.)$

according to β and α;

$$(C.3)\quad \Psi(\vartheta) = \begin{bmatrix} \Psi_{11}(\vartheta) & \Psi_{12}(\vartheta) \\ \Psi_{21}(\vartheta) & \Psi_{22}(\vartheta) \end{bmatrix}.$$

It is assumed that:

(*AC*) $\Psi(.)$ *exists* and *is positive definite in a small vicinity* N_{ϑ^o} *of* ϑ^o.

Let us consider first the case of known α, so α is fixed at α^o, and take two asymptotically efficient estimates β_T^1 and β_T^2, that is

$$(C.4)\quad P_{(\beta^T,\alpha^0)}\{\sqrt{T}\ (\beta_T^j - \beta^T)\} \to N(0, \Psi_{11}^{-1}(\vartheta^0));\ j=1,2$$

for any sequence β^T such that $\sqrt{T}\,(\beta^T-\beta^0)$ stays bounded. Then, under (C.4), the following holds:

LEMMA:

$$(C.5)\quad \sqrt{T}\,(\beta_T^1-\beta_T^2) \xrightarrow{P} 0$$

PROOF:

See Theorem (4.1) *of Hajek* [1972].

Let $\beta_T^1 = \beta^P - \Psi_{11}^{P-1} L_\beta(\beta^P,\alpha^o)$, where β^P is $\sqrt{T}$-consistent estimate of β^o, Ψ_{11}^P is a consistent estimate of $\Psi_{11}(\vartheta^o)$ and $L_\beta(\beta^P,\alpha^o) = \partial L(\beta^P,\alpha^o)/\partial\beta$ and let $\beta_T^2 = \beta^P - \Psi_{11}^{P-1} V(\beta^P, \alpha^o)$ where V(.) is a $k\times 1$ vector. If β_T^2 is asymptotically efficient then (C.5) implies

$$(C.6)\quad \sqrt{T}\,(L_\beta(\beta^P,\alpha^o) - V(\beta^P,\alpha^o)) \xrightarrow{P} 0.$$

We now extend (C.6) to the case of unknown α and prove:

THEOREM:

Let β^P, α^P be $\sqrt{T}$-consistent estimates of β^o, α^o, and let $V(\beta,\alpha)$ be such that (C.6) holds at (β^P,α^o) and $V_\alpha(.) = \frac{\partial V(.)}{\partial \alpha}$ exists at $N_{\alpha^o} = N_{\gamma^o} \cap A$, finally assume that (Ac) holds. Then

$$(C.7)\quad \sqrt{T}\,[L_\beta(\beta^P,\alpha^P) - V(\beta^P,\alpha^P)] \xrightarrow{P} 0$$

PROOF:

Let us define

$$(C.8)\quad \Delta(\beta,\alpha) = L_\beta(\beta,\alpha) - V(\beta,\alpha)$$

and note that, from the $\sqrt{T}$-consistency, α^P can be expressed as

$$(C.9)\quad \alpha^P = \alpha^o + \frac{1}{\sqrt{T}} S_T \quad \textit{where}\ 0 < |S_T| < \infty$$

Hence

$$(C.10)\quad \sqrt{T}\ \Delta(\beta^P,\alpha^P) = \sqrt{T}\ \Delta(\beta^P,\alpha^o) + \sqrt{T}\,[\Delta(\beta^P,\alpha^o + \frac{S_T}{\sqrt{T}}) - \Delta(\beta^P,\alpha^o)]$$

$$= \sqrt{T}\ \Delta(\beta^P,\alpha^o) +$$

$$\frac{\Delta[\beta^P,\alpha^o + (1/\sqrt{T}\ S_T)] - \Delta(\beta^P,\alpha^o)}{S_T/\sqrt{T}} \sqrt{T}\ (\alpha^P - \alpha^o)$$

As $T \to \infty$, $\sqrt{T}\ \Delta(\beta^P,\alpha^o) \xrightarrow{P} 0$ by (C.6) and the second term on the R.H.S. of (C.10) tends to $\Delta_\alpha(\beta^P,\alpha^o)\sqrt{T}\ (\alpha^P - \alpha^o)$. Since $\sqrt{T}\ (\alpha^P - \alpha^o)$ is bounded in probability, to prove (C.7) it is sufficient to show

$$(C.11)\quad \Delta_\alpha(\beta^P,\alpha^o) \xrightarrow{P} 0\,.$$

Let us return to the case of known α and fix α at $\alpha = \alpha' \in N_{\gamma^o} \cap A$. All the

results attained under $\alpha = \alpha^o$ are now followed for $\alpha = \alpha'$ Therefore the following can be concluded:

$$\left.\begin{array}{l} \sqrt{T}\,\Delta(\beta^P,\alpha^o) \xrightarrow{P} 0 \\ \sqrt{T}\,\Delta(\beta^P,\alpha') \xrightarrow{P} 0 \end{array}\right\} \Longrightarrow \frac{\Delta(\beta^P,\alpha')-\Delta(\beta^P,\alpha^o)}{1/\sqrt{T}} \xrightarrow{P} 0 \quad as\ T \to \infty.$$

Now let $\alpha' \to \alpha^o$ to get (C.11). Q.E.D.

References

Afifi, A.A. and R.M.E. Elashoff; "Missing Observations in Multivariate Statistics--I. Review of the Literature," *Journal of the American Statistical Association*, 61(1966), 595-605.

Amemiya, T.; "Regression Analysis when the Dependent Variable is Truncated Normal," *Econometrica*, 41(1973), 997-1017.

_____; "Multivatiate Regression and Simultaneous Equation Models when the Dependent Variables are Truncated Normal," *Econometrica*, 42(1974), 999-1012.

_____; "The Estimation of Simultaneous Equation Generalized Probit Model," *Econometrica*, 46(1978), 1193-1206.

Berndt, E.B., B. Hall, R. Hall, and J.A. Hausman; "Estimation and Inference in Nonlinear Structural Models," *Annals of Economics and Social Measurement*, 3(1974), 653-665.

Bickel, P.J.; "On Adaptive Estimation," *Ann. Statist.*, 10(1982), 647-671.

Cosslett, S.; "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*, 49(1981), 1289-1316.

Dempster, A.P., N.M. Laird, and D.B. Rubin; "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39(1977), 1-22.

Duncan, Gregory M.; "Formulation and Statistical Analysis of the Mixed, Continuous/Discrete Dependent Variable Model in Classical Production Theory," *Econometrica*, 48(1980), 839-852.

Fair, R.C.; "A Note on the Computation of the Tobit Estimator," *Econometrica*, 45(1977), 1723-1727.

Green, W.H.; "On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model," *Econometrica*, 49(1981), 505-513.

_____; "Estimation of Limited Dependent Variable Models By Ordinary Least Squares and the Method of Moments," *Journal of Econometrics*, 21(1983), 195-212.

Griliches, Zvi, B.H. Hall, J.A. Hausman; "Missing Data and Self-Selection in Large Panels," *Annales de INSEE*, 30-31(1978), 137-176.

Hájek, J.; "Local Asymptotic Minimax and Admissibility in Esimation," *Sixth Berkeley Symp. Math. Statist. Prob.*, 1(1972), 175-194. Univ. of California Press, Berkeley.

Hanemann, W.M., and Yacov Tsur; "Econometric Models of Discrete/Continuous Supply Decisions under Uncertainty," *Dept. of Agricultural and Resource Economics, Univ. of Califonia, Berkeley, Giannini Foundation Working Paper, No. 195*, (1982).

Hartley, M.J.; "The Tobit and Probit Models: Maximum Likelihood by Ordinary Least Squares," *Dept. of Economics, State Univ. of New York at Buffalo, Discussion Paper, Number 374*, (1976).

Hausman, J.A., and D.A. Wise; "A Conditional Probit Model for Qualitative Choice: Discrete Data Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46(1978), 403-420.

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

_____; "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46(1978), 931-961.

Johnson, N., and S. Kotz; *Distribution in Statistics: Continuous Multivariate Distributions*, New York: John Wiley & Sons, 1972.

Le Cam, L.; "Locally Asymptotically Normal Families of Distributions," *Univ. California Publ. Statist.* 3(1960), 27-98.

_____; *Theorie Asymptotique de la Decision Statistique*. Les Presses de l'Universite de Montreal, 1969.

Lee, Lung-Fei; "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables," *International Economic Review*, 19(1978), 415-433.

_____; "Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables," *Econometrica*, 47(1979), 977-996.

_____; "Simultaneous Equations Models with Discrete and Censored Variables," *in* C.F. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge: MIT Press, 1981.

_____, and R.P. Trost; "Estimation of Some Limited Dependent Variable Models with Application to Housing Demand," *Journal of Econometrics*, 8(1978), 357-382.

Quandt, R.E., and J.B. Ramsey; "Estimating Mixtures of Normal Distributions and Switching Regressions," *Journal of the American Statistical Association*, 73(1978), 730-738.

Rao, C.R.; *Linear Statistical Inference and its Applications*, New York: Wiley, 1973.

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

Tsur, Y.; "The Formulation and Estimation of Discrete/Continous Supply Models Under Uncertainty," Ph.D. dissertation, Department of Agricultural and

Resource Economics, University of California, Berkeley, (1983).

Wu, C.F.J.; "On the Convergence Properties of the EM Algorithm," *Ann. Statist.*, 11(1983), 95-103.

Zellner, A.; "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for aggregation Bias," *Journal of the American Statistical Association*, 57(1962), 348-368.