Discussion Papers
No. 1987.3

Missing Measurements in
Limited Dependent Variable Models

by

Donald Lien and

David Rearden

Institute for Public Policy
and Business Research

The University of Kansas

MISSING MEASUREMENTS IN LIMITED DEPENDENT VARIABLE MODELS

Da-Hsiang Donald Lien*
Department of Economics
and
David Rearden
Research Assistant
Institute for Public Policy and Business Research
University of Kansas
Lawrence, Kansas 66045

Discussion Paper
No. 1987.3

Anthony Redwood, Executive Director
Institute for Public Policy and Business Research
University of Kansas
Lawrence, Kansas 66045

*We wish to acknowledge De-Min Wu for helpful comments and IPPBR at the University of Kansas for financial support. Of course, we are responsible for any remaining errors.

ABSTRACT

In this paper, we investigate the problem, whether or not incomplete observations contain any information about the parameteres that could be used in their estimation, as it applies to limited dependent variable models. The answers differ between binary response models and Tobit models when serial correlation prevails given missing observations on independent variables. If, instead, missing observations pertain to the dependent variable, we have negative answers for both models.

MISSING MEASUREMENTS IN LIMITED DEPENDENT VARIABLE MODELS

Donald Lien and David Rearden

In a recent article, Kmenta and Balestra (1986) consider the problem of estimating the coefficients of a linear regression model with missing measurements when no auxiliary relations can be justified and when the omission of incomplete observations leaves the sample selection rule unaffected. They show that the incomplete observations contain no information about the regression parameters that could be used in their estimation. A remaining question of interest is whether or not their conclusion could be generalized to nonlinear econometric models. Here, we investigate the problem as it applies to binary response (BR) models and Tobit models.

For BR models, our results show that the answer is positive when we have missing observations on independent variables, but not in the case of missing observations on the dependent variable. More specifically, in the latter case, the only positive result appears when the missing values are estimated by their respective expected values using the estimated parameters derived from the complete observations. The procedure, although directly applicable to linear models, ignores the dichotonomous property of the underlying variable. For Tobit models, the result is different. First, if missing observations pertain to independent variables, Kmenta-Balestra's conclusion sustains in the absence of serial correlation assuming the variance-covariance matrix is known. A simple example with serial correlation is provided to invalidate the above result. In the case of missing observations on the dependent variable, there exist no reasonable estimates for the missing values that could retain Kmenta-Balestra's conclusion.

To present our results, we begin with the following BR model:

$$y_t = 1, \text{ if } x_t\beta + \epsilon_t \geq 0;\ 0, \text{ otherwise}, \ \forall \ t = 1, \ldots, T, \quad (1)$$

where the random vector $\epsilon = (\epsilon_1, \ldots, \epsilon_T)'$ has a joint density function h(.). We now decompose the observation matrix as follows: $Y = [Y_1' \ Y_2']'$, $X = [X_1' \ X_2']'$, where Y is a (T x 1) column vector, X is a (T x k) matrix; Y_i is a (T_i x 1) column vector, X_i is a (T_i x k) matrix, $\forall$ i = 1, 2 such that $T_1 + T_2 = T$. Consider the case that X_2 is unqbservable while X_1, Y_1 and Y_2 are all known. Let $\hat{\beta}$ be the maximum likelihood (ML) estimator using only complete observations (Y_1, X_1). To see whether or not X_2 contains any information about β, we adopt Kmenta-Balestra's approach. Specifically, we treat X_2 as unknown parameters to be estimated along with β in the framework of maximum likelihood estimation such that $\tilde{\beta}$ is the new ML estimator. The problem may be written as:

$$\max_{\beta, X_2} \int_{-\infty}^{-x_r\beta} \cdots \int_{-x_s\beta}^{\infty} h(\epsilon)\, d\epsilon_s \ldots d\epsilon_r. \quad (2)$$

Since h(.) is nonnegative, the likelihood increases as the integration region increases. To achieve the maximum, we choose $x_r\beta = -\infty$ if $-x_r\beta$ is the upper limit of the integration (i.e., $y_r = 0$), $\forall$ $r = T_1 + 1, \ldots, T$; $x_s\beta = \infty$ if $- x_s\beta$ is the lower limit (i.e., $y_s = 1$), $\forall\ s = T_1 + 1,\ldots,T$. As a consequence, equation (2) reduces to the maximum likelihood estimation problem associated with complete observations, which implies $\tilde{\beta} = \hat{\beta}$; and hence X_2 contains no information about β that could be used in the estimation procedures for the most general BR models.

We now turn to the Tobit models:

$$y_t = x_t\gamma + \epsilon_t \text{ if } x_t\gamma + \epsilon_t \geq 0 \text{ ; } 0, \text{ otherwise, } \forall\ t=1,\ldots,T. \quad (3)$$

where $\{\epsilon_t\}$ is a sequence of i.i.d. normal random variables with $E(\epsilon_t) = 0$, $var(\epsilon_t) = \sigma^2$. Since σ^2 is assumed to be known, the distribution [resp. density] function of ϵ_t is denoted by $G(\cdot)$ [resp. $g(\cdot)$]. Without implicating any confusion, let $T_1 = \{t{:}1 \leq t \leq T_1\}$, $T_2 = \{t{:}T_1{+}1 \leq t \leq T\}$, and define $R = \{t{:}\ 1 \leq t \leq T,\ y_t > 0\}$, $S = \{t{:}\ 1 \leq t \leq T,\ y_t = 0\}$. Upon maximizing the log-likelihood function associated with complete observations, the ML estimator $\hat{\gamma}$ satisfies

$$- \sum_{S\cap T_1} [G(-x_t\hat{\gamma})]^{-1} g(-x_t\hat{\gamma})x_t' + \sigma^{-2} \sum_{R\cap T_1} (y_t - x_t\hat{\gamma})x_t' = 0.$$

On the other hand, if we treat X_2 as unknown parameters, the corresponding maximum likelihood estimation problem is

$$\max_{\gamma, X_2} \sum_{t\in S} \log(G(-x_t\gamma)) + \sum_{t\in R} \log g(y_t - x_t\gamma).$$

The first order condition associated with X_2 leads to the following property:

$$[G(-\tilde{x}_t\tilde{\gamma})]^{-1} g(-\tilde{x}_t\tilde{\gamma}) = 0 \text{ when } y_t = 0,\ \forall\ t = T_1{+}1,\ldots,T \text{ ; } \quad (4a)$$

$$y_t = \tilde{x}_t\tilde{\gamma} \quad \text{when } y_t > 0, \ \forall \ t = T_1+1, \ldots, T, \qquad (4b)$$

where $(\tilde{x}_t, \tilde{\gamma})$ is the optimal solution. (Note that (4a) implies $\tilde{x}_t\tilde{\gamma} = -\infty$ when $y_t = 0$). Upon substituting (4a)-(4b) into the other first order condition associated with γ, we then establish $\tilde{\gamma} = \hat{\gamma}$. Following similar approach, the result is maintained in the presence of heteroscedasticity provided that all the variance terms are known. Thus, in the context of Tobit models with the underlying normal disturbances being independent, incomplete observations due to missing values on independent variables contain no information about the parameters.

In case that serial correlation prevails, the above result becomes invalid. For example, assume that ϵ_t is a standard normal variable such that $E(\epsilon_p\epsilon_q) = \rho \neq 0$, $q < T_1 < p$ while all the other covariances are zero. The corresponding log-likelihood function is

$$L = \sum_{t\in S'} \log(\Phi(-x_t\gamma)) + \sum_{t\in R'} \log(\phi(y_t - x_t\gamma))$$

$$+ \log\left[\int_{-\infty}^{-x_q\gamma} h(y_p - x_p\gamma, \epsilon_q)d\epsilon_q\right]$$

where $S' = S - \{p,q\}$, $R' = R - \{p,q\}$; $\Phi(.)$ [resp. $\phi(.)$] is the cumulative [resp. probability] density function of the standard normal variable; $h(.,.)$ is the bivariate normal density function for (ϵ_p, ϵ_q). Upon maximizing L, the first order condition associated with x_p implies

$$\int_{-\infty}^{-x_q\gamma} \frac{\partial h(y_p - x_p\gamma,\ \epsilon_q)}{\partial \epsilon_p}\, d\epsilon_q = 0, \tag{5}$$

if γ is nontrivial (i.e., $\gamma \neq 0$). Further, upon substitution equation (5) into the other first order condition associated with γ, we establish that $\hat{\gamma} = \tilde{\gamma}$ if and only if

$$h(y_p - x_p\hat{\gamma},\ -x_q\hat{\gamma})\Phi(-x_q\hat{\gamma}) = \phi(-x_q\hat{\gamma}) \int_{-\infty}^{-x_q\hat{\gamma}} h(y_p - x_p\hat{\gamma},\ \epsilon_q)\, d\epsilon_q, \tag{6}$$

whenever equation (5) holds. After algebraic manipulation, the two equations reduce to

$$\theta z\Phi((a - \rho z)/\theta) + \rho\phi((a - \rho z)/\theta) = 0; \tag{5'}$$

$$\phi(a)\Phi((a - \rho z)/\theta) = \phi((a - \rho z)/\theta)\Phi(a)/\theta, \tag{6'}$$

respectively where $\theta = (1 - \rho^2)^{1/2}$, $z = y_p - x_p\hat{\gamma}$, $a = -x_q\hat{\gamma}$. Solving the above two equations, we have $z = -\rho\phi(a)/\Phi(a)$. Therefore, $\tilde{\gamma} = \tilde{\gamma}$ if and only if

$$\begin{aligned} Q(a,\rho) = {} & \phi(a)\Phi((a\Phi(a) + \rho^2\phi(a))/\theta\Phi(a)) \\ & - \Phi(a)\phi((a\Phi(a) + \rho^2\phi(a))/\theta\Phi(a))/\theta \end{aligned}$$

is identically zero. It can be shown that $Q(a, \rho) \equiv 0$ when $\rho = 0$. Nonetheless, when $\rho \neq 0$, $Q(a, \rho)$ is not zero everywhere; hence in general $\tilde{\gamma} \neq \hat{\gamma}$.

The above results are derived under the assumption that the variance-covariance matrix is known. Now, suppose instead that the matrix is unknown, then the parameters of concern may not be identified. For example, in probit models with i.i.d. disturbance terms such that $var(\epsilon_t) = \sigma^2$, $\forall$ t, only (β/σ) can be identified. However, we can show that $\widehat{(\beta/\sigma)} = \widetilde{(\beta/\sigma)}$; hence X_2 contains no information about the identified parameters. Similar result does not carryover to Tobit models where both γ and σ^2 are identifiable.[1] The main reason is that the log-likelihood function depends upon the number of observations of the dependent variable that take their values at zero. As a result, $\hat{\beta} = \tilde{\beta}$ only when $y_t > 0$, $\forall$ $t = T_1+1,\ldots,T$.

Finally, assume that X_1, X_2 and Y_1 are observable but Y_2 is not. The approach is then finding predictions $\hat{Y}_2$ based on X_1, X_2 and Y_1 and apply ML estimation to the augmented samples thereafter. For BR models with i.i.d. disturbance terms, maximizing the log-likelihood function leads to

$$\sum_{t=1}^{T_1} \frac{y_t-(1-F(-x_t\bar{\beta}))}{F(-x_t\bar{\beta})\ [1-F(-x_t\bar{\beta})]} f(-x_t\bar{\beta})x_t'$$

$$+ \sum_{t=T_1+1}^{T} \frac{\hat{y}_t-(1-F(-x_t\bar{\beta}))}{F(-x_t\bar{\beta})\ [1-F(-x_t\bar{\beta})]} f(-x_t\bar{\beta})x_t' = 0, \qquad (7)$$

where $\bar{\beta}$ is the ML estimator for the augmented samples; f(.) [resp. F(.)] is the probability [resp. cumulative] density function of ϵ_t. It is readily seen[2] that $\bar{\beta} = \hat{\beta}$ if and only if $\hat{y}_t = 1-F(-x_t\hat{\beta})$, $\forall$ $t = T_1 + 1,\ldots,T$. The prediction rule, however, ignores the dichotonomous property of the dependent variable. Similarly, in Tobit models with i.i.d. disturbance terms, $\bar{\gamma} = \hat{\gamma}$ if and only if $\hat{y}_t = x_t\hat{\gamma}$, $\forall\ t = T_1+1,\dots,T$. Since $x_t\hat{\gamma}$ may be negative, the prediction rule deems unreasonable. At this point, we may conclude that Kmenta-Balestra's result does not apply to limited dependent models when the missing observations pertain to the dependent variable.

FOOTNOTES

1. As noted in Salkever (1976) for linear models, the problem considered in this paper when X_2 is unobservable is actually equivalent to the problem studied in the literature of unique event dummy variable models. Thus, Kmenta-Balestra's results correspond to Salkever's while our results correspond to Anderson (1987)'s. Nonetheless, both Salkever and Anderson only discuss i.i.d. cases.

2. In fact, $\hat{y}_t = 1-F(-x_t\hat{\beta})$ is the mean of y_t upon replacing unknown β by $\hat{\beta}$. The prediction rule is hence the same as that applied in linear models. For Tobit models, the rule will set $\hat{y}_t = x_t\hat{\gamma} + \sigma^2\phi(-x_t\hat{\gamma})/(1-\Phi(-x_t\hat{\gamma}))$.

REFERENCES

Anderson, G.J., "Prediction Tests in Limited Dependent Variable Models," Journal of Econometrics 34 (Jan.-Feb., 1987), 253-261.

Kmenta, J., and P. Balestra, "Missing Measurements in a Regression Problem with No Auxiliary Relations" in Slottje, D.J. (ed.), Advances in Econometrics (Greenwich: JAI press, 1986).

Salkever, "The Use of Dummy Variables to Compute Predictions, Prediction Errors, and Confidence Intervals," Journal of Econometrics 4 (Nov., 1976), 393-397.