0165-1765/93/$06.00 © 1993 Elsevier Science Publishers B.V. All rights reserved

Missing measurements in econometric models with no auxiliary relations

Marno Verbeek *

Department of Econometrics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, Netherlands

Received 15 October 1992
Final revision received 20 July 1993
Accepted 12 August 1993

Abstract

In this paper it is argued that maximizing the complete data (log) likelihood function with respect to the missing data and the unknown parameters will not improve the efficiency of the estimators but may affect consistency instead. If no auxiliary relations are available or additional assumptions are made, the maximum likelihood estimator based on the observed data is (asymptotically) the most efficient estimator.

1. Introduction

Recently, several authors have discussed the problem of whether or not incomplete observations contain any information about the parameters of interest that can be used to improve the efficiency of estimators. Kmenta (1981) and Kmenta and Balestra (1986) considered the normal linear regression model and concluded that there is no gain in using the incomplete data, both in the case of missing measurements for the dependent variable as well as the case with missing values for the explanatory variable(s). Although this conclusion is obviously correct, the way in which these authors derive it is not, as will be shown below. Lien and Rearden (1988, 1990) mistakenly choose the same approach as Kmenta and Balestra to generalize the results to discrete response and tobit models. In this paper it is argued that the maximization of the complete data likelihood function with respect to the unknown parameters as well as the missing data will not improve the efficiency of estimators for the parameters of interest. Even worse, this approach can easily result in inconsistent estimators.

Suppose we have a data set with N observations stacked in a vector y and an $N \times k$ matrix X. Our interest lies in the parameter vector ϑ characterizing the distribution of y conditional on X, for example (in the case of a normal linear regression model) in the parameters β and σ^2 in

$$y = X\beta + \varepsilon\ , \quad \varepsilon \sim \mathrm{N}(0, \sigma^2 I)\ . \tag{1}$$

The maximum likelihood estimator $\hat{\vartheta}$ will be the solution of

* Helpful discussions with Bertrand Melenberg and Theo Nijman are gratefully acknowledged. Any remaining errors are my own.

$$\max_{\vartheta} \log f(y \mid X, \vartheta) . \tag{2}$$

Now suppose that only n of the total N observations are complete, while for the remaining $N-n$ observations the values for one of the variables is missing. We shall first, in section 2, pay attention to the case with missing measurements on y and, subsequently, in section 3 to missing measurements on X. Throughout it is assumed that the mechanism that generates the missing data is ignorable [Rubin (1976)], i.e. that the data are missing at random.

2. Missing measurements for the dependent variable

If we denote the observed part and the missing part of the y-vector as y^{obs} and y^{mis}, respectively, the maximum likelihood estimator, $\hat{\vartheta}^*$, based on the observed data is the solution of

$$\max_{\vartheta} \log f(y^{\text{obs}} \mid X, \vartheta) = \max_{\vartheta} \log \int f(y^{\text{obs}}, y^{\text{mis}} \mid X, \vartheta) \, \mathrm{d}\mu(y^{\text{mis}}) , \tag{3}$$

where $f(\cdot)$ is generic notation for any density/mass function and where μ is the measure with respect to which the density of y^{mis} is defined. Because the data on y are assumed to be missing at random, $\hat{\vartheta}^*$ is a consistent estimator for ϑ. Writing

$$\log f(y \mid X, \vartheta) = \log f(y^{\text{obs}} \mid X, \vartheta) + \log f(y^{\text{mis}} \mid y^{\text{obs}}, X, \vartheta) \tag{4}$$

it will be clear that $\hat{\vartheta}^*$ is less efficient than the ML estimator $\hat{\vartheta}$ that would have been obtained from a complete sample. Kmenta and Balestra (1986) and Lien and Rearden (1988, 1990) pose the question whether the efficiency can be improved by taking into account the incomplete observations, i.e. whether the second term in the RHS of (4) contains information on ϑ even if y^{mis} is not observed. Because (3) is the correct maximum likelihood procedure the answer is clearly no: $\hat{\vartheta}^*$ is the most efficient estimator for ϑ given the available data.

The approach adopted by the authors mentioned above is the maximization of the complete data log likelihood given in (4) with respect to the unknown parameters ϑ and the missing values y^{mis}, after which the resulting estimator, $\tilde{\vartheta}$, for ϑ is compared with $\hat{\vartheta}$. As argued by Hsiao (1980) and Little and Rubin (1983) this method is *not* maximum likelihood and, moreover, it can easily result in inconsistent parameter estimators. Noting that

$$\max_{\vartheta, y^{\text{mis}}} \log f(y \mid X, \vartheta) = \max_{\vartheta} \log f(y^{\text{obs}} \mid X, \vartheta) + \max_{\vartheta, y^{\text{mis}}} \log f(y^{\text{mis}} \mid y^{\text{obs}}, X, \vartheta) \tag{5}$$

there is obviously no reason why this approach should lead to consistent estimators if (3) does. Only in special cases does the second term in the RHS of (5) (asymptotically) not affect the estimator for (elements in) ϑ and $\hat{\vartheta}^* = \tilde{\vartheta}$ (for those elements). Unless the fraction of missing data tends to zero if the sample size increases, any different between $\hat{\vartheta}^*$ and $\tilde{\vartheta}$ will be caused by the inconsistency of the latter.

The statements above are valid for all types of models with one endogenous variable which has missing measurements, including, for example, limited dependent variable models. To illustrate our arguments we shall explicitly pay attention to the normal linear regression model. For this model the complete data log likelihood is given by

$$L(y\,|\,X,\vartheta) = k - \frac{N}{2}\log\sigma^2 - \tfrac{1}{2}(y^{\mathrm{obs}} - X^0\beta)'(y^{\mathrm{obs}} - X^0\beta)/\sigma^2$$
$$- \tfrac{1}{2}(y^{\mathrm{mis}} - X^1\beta)'(y^{\mathrm{mis}} - X^1\beta)/\sigma^2\,, \tag{6}$$

where k is a constant and X^0 and X^1 are the X's corresponding to y^{obs} and y^{mis}, respectively. If this function is maximized with respect to β, σ^2 and y^{mis} it is easily seen that the resulting estimators, $\tilde{\beta}$ and $\tilde{\sigma}^2$, satisfy $\tilde{\beta} = \hat{\beta}^*$ and $\tilde{\sigma}^2 = (n/N)\hat{\sigma}^{*2}$. So, as long as n/N does not tend to one as N increases (i.e. the fraction of missing observations does not tend to zero as the sample size increases), $\tilde{\sigma}^2$ is an inconsistent estimator for σ^2. For non-linear models (like the tobit model) this inconsistency is likely to be transmitted to the β parameters as well (cf. Lien and Rearden (1988)].

3. Missing measurements for the explanatory variable

Now we shall pay attention to the case where values for one of the explanatory variables may be missing and no auxiliary relations are postulated [as in Kmenta and Balestra (1986) and Lien and Rearden (1988, 1990)]. For ease of presentation we shall restrict attention to the case with one explanatory variable, x, only. This approach implies that the variable x is not endogenized. Consequently, the density of y conditional on x has no clear interpretation when the value for x is missing. The likelihood approach in this case has two alternatives. First, one can restrict attention to the complete observations only, i.e.

$$\max_{\vartheta} \log f(y^0\,|\,X^{\mathrm{obs}},\vartheta)\,, \tag{7}$$

where y^0 is the part of y corresponding to X^{obs}. Alternatively, one can use all available information by

$$\max_{\vartheta,\gamma} \{\log f(y^0\,|\,X^{\mathrm{obs}},\vartheta) + \log f(y^1\,|\,\gamma)\}\,, \tag{8}$$

in which case marginal information on y is used as well. However, unless clear relationships between ϑ and γ exist, this will not improve upon the efficiency of the estimator for ϑ, while the existence of such relationships will require that an assumption for the process of the exogenous variable is made (e.g. normality). If the explanatory variables are not endogenized one cannot improve upon the efficiency of the estimator $\hat{\vartheta}^*$ obtained from (7).

The approach considered by Kmenta (1981), Kmenta and Balestra (1986) and Lien and Rearden (1988, 1990) implies treating the missing x-values as unknown parameters and maximizing the complete data log likelihood given in (4) with respect to these missing values as well. This is a somewhat curious approach since one is choosing parameter values and x-values such that the conditional likelihood of observing y given these (and the observed) x-values is maximal. For the normal linear regression model characterized by (8) one can easily show that the resulting estimator for b equals the correct maximum likelihood estimator, $\hat{\beta}^*$, while – as before – the estimator for σ^2 is n/N times the ML estimator $\hat{\sigma}^{*2}$, which is clearly inconsistent. It is possible to consider this problem as a case of the well-known incidental parameters problem [Neyman and Scott (1948)] in a reparameterized model in which, when the x-value is missing, the RHS is replaced by some function of unknown parameters. In the case of only one explanatory variable, unique x-values are obtained from this procedure. however, there is definitely no reason to view these values as reasonable predictions for the true x-values.

Even if the explanatory variables are endogenized the approach of maximizing the complete data likelihood function with respect to the missing x-values is incorrect. Let us, as an illustration, consider the bivariate normal linear case in which both y and x are normal. Then, the complete data log likelihood function is given by

$$\log f(y, x \mid \vartheta, \phi) = \log f(y \mid x, \vartheta) + \log f(x \mid \phi) . \tag{9}$$

If (9) is maximized with respect to ϑ, ϕ and the missing x-values, the resulting estimator $\tilde{\beta}$ for β in $y = \beta' x + \varepsilon$ is inconsistent. In particular, one can show that

$$\operatorname*{plim}_{N\to\infty} \tilde{\beta} = \frac{\beta}{p + (1-p)\rho_{xy}^2} , \tag{10}$$

where it is assumed that $n/N \to p$ if $N \to \infty$, and where ρ_{xy} is the correlation coefficient between x and y. Only if $p = 1$, i.e. if the fraction of missing observations tends to zero if N increases, will $\tilde{\beta}$ be consistent for β.

4. Conclusions

To summarize our results, treating the missing observations as unknown parameters and maximizing the complete data likelihood function with respect to these parameters as well is not a correct maximum likelihood procedure and can easily result in inconsistent estimators. This result holds true for both the situation where the endogenous variable may be missing as well as the one where observations on exogenous variables are missing. Consequently, it is not correct to consider the question of whether or not incomplete observations contain any information about the parameters of interest in this way, as done by Kmenta and Balestra (1986) and Lien and Rearden (1988, 1990). Unless additional assumptions are made, one cannot improve upon the efficiency by maximizing the complete data log likelihood with respect to the unknown parameters and the missing values.

References

Hsiao, C., 1980, Missing data and maximum likelihood estimation, Economics Letters 6, 249–253.

Kmenta, J., 1981, On the problem of missing measurements in the estimation of economic relationships, in: E.G. Charatsis, ed., Proceedings of the Econometric Society European Meeting 1979 (North-Holland, Amsterdam).

Kmenta, J. and P. Balestra, 1986, Missing measurements in a regression problem with no auxiliary relations, in: D.J. Slottje, ed., Advances in econometrics, vol. 5 (JAI Press, Greenwich).

Lien, D. and D. Rearden, 1988, Missing measurements in limited dependent variable models, Economics Letters 26, 33–36.

Lien, D. and D. Rearden, 1990, Missing measurements in discrete response models, Economics Letters 32, 231–235.

Little, R.J.A. and D.B. Rubin, 1983, On jointly estimating parameters and missing data by maximizing the complete-data likelihood, The American Statistician 37, 218–220.

Neyman, J. and E.L. Scott, 1948, Consistent estimates based on partially consistent observations, Econometrica 16, 1–32.

Rubin, D.B., 1976, Inference and missing data, Biometrika 63, 581-592.