The Pakistan Development Review
32:4 Part II (Winter 1993) pp. 687–698

Multinomial Logit Model of Occupational Choice: A Latent Variable Approach

TAYYEB SHABBIR

INTRODUCTION

Economists and other social scientists have had a long-standing interest in studying the different aspects of an individual's occupational choice. An important issue in this regard is an econometric analysis of the determinants of occupational choice. A rather well-known example of such a work is Schmidt and Strauss (1975) which uses a maximum likelihood procedure to estimate a multinomial logit model (MNL) where occupational choice is determined by an individual's education, experience, race and sex. Regarding the above genre of models (as, in fact, in the parallel and closely related literature on earnings functions), one is often interested in ascertaining the unbiased marginal effect of education on the dependent variable. Not only these estimates allow tests of the human capital theory against alternative hypotheses, they also have important public policy implications particularly for the developing countries which are typically contemplating expansion of their educational sectors. However, these estimates may become biased in the event that a relevant regressor is left out of the specification. Particularly difficult problems arise when such an excluded variable is a latent (or unobserved) one.

Using the Schmidt-Strauss multinomial logit model of occupational choice as its starting point, the present paper:

(a) Motivates inclusion of a latent variable as one of the relevant regressors.
(b) Shows how the omission of such a latent variable may bias the maximum likelihood estimates of the coefficients of included variables. Further, the conditions under which the direction of such a bias can be ascertained are noted and finally, this paper.
(c) Describes a methodology that would provide unbiased coefficient estimates given that certain conditions are met. The proposed procedure requires data on siblings.

In the last ten to fifteen years, a problem similar to the one described above has been considered in the context of the human capital type of (log-linear) earnings

Tayyeb Shabbir is Senior Research Economist at the Pakistan Institute of Development Economics, Islamabad.

Author's Note: An earlier version of this paper has benefited from very valuable comments by Paul Taubman, University of Pennsylvania and Fatah Ullah Bagheri, University of North Dakota.

functions. One aspect of this literature dealt with the question of how the OLS coefficient estimates of the various included variables would become biased if a relevant latent variable were omitted. See Taubman (1977); Griliches (1979); Behrman *et al.* (1980) and Shabbir (1987). However, compared to the treatment of the omitted variable problem for the (log-) linear regression model, the issue has not been analysed much for the class of descrete probability models.[1] However, Lee (1980) is one of the few studies dealing with the question of the omitted variable bias in MNL models.[2] Though Lee's results are not explicitly derived for the case of the omitted variable being a latent one, the present paper is able to extend his analysis to this case as well.

Rest of the paper is organised such that Section 1 outlines the essential features of the Schmidt-Strauss type of MNL model of occupational choice while Section 2 briefly presents the maximum likelihood estimation procedure for this model. Section 3 considers the implications of omitting a relevant latent or unobserved variable in the specification for the above model and Section 4 presents a procedure to handle this problem. This procedure requires data on siblings. Finally, Section 5 contains some concluding remarks and comments.

1. MULTINOMIAL LOGIT MODEL OF OCCUPATIONAL CHOICE

Consider a model of occupational choice where individuals choose one amongst the $L > 1$ alternatives facing them. Their behaviour can be represented in terms of a ordered polychotomous response variable $y = 0, 1, 2, ... L$ which has $(L + 1)$ mutually exclusive and exhaustive categories where the occupational alternative 0 has the lowest rank and L has the highest one. Let X be a $(j x 1)$ vector of an individual's characteristics (such as schooling, job experience, family background, etc.) which affect the occupational choice and let α_i be a $(1 x j)$ vector of coefficients attached to X. Then, a Multinomial Logit Model (MNL) of occupational choice can be specified as follows:

$$P(y = i \mid X) \quad = \quad \frac{exp(\alpha_i X)}{1 + \sum_{k=1}^{L} exp\,(\alpha_k X)} \qquad \ldots \qquad \ldots \qquad (1.1)$$

[1] This extension from continuous (earnings functions etc.) to discrete dependent variable (MNL etc.) is not merely a question of extending the scope of a particular methodology since, as *a* is shown in this paper, there are significant differences in the analytical results in the two cases e. g. in the case of discrete dependent variable, the direction of the bias cannot be determined without additional assumptions (of normality) regarding the distribution of the latent variable, *z*. Also, whereas the assumption of 'unconditional independence' was enough in the continuous dependent variable case to obtain unbiased estimates here we require 'conditional independence'.

[2] Some of the analysis presented in Lee (1980) is based on the results derived in Nerlove and Press (1973).

$$P(y=0\,|\,X) \;=\; \frac{1}{1 + \sum_{k=1}^{L} exp\,(\alpha_k X)} \qquad \ldots \qquad \ldots \qquad (1.2)$$

where $i = 1, 2, \ldots, L$

In the above context, we would further require that the probabilities add up to unity i. e.

$$\{P(y=0\,|\,X) + \sum_{i=1}^{L} P(y=i\,|\,X)\} = 1.$$

The MNL model given by (1.1) and (1.2) can be equivalently written in the so-called "odds" form as follows:

$$ln\, \frac{P(y=i\,|\,X)}{P(y=0\,|\,X)} \;=\; \alpha_i X \;=\; 1, \ldots, L \qquad \ldots \qquad \ldots \qquad (2)$$

where occupational category 0 is being used as the 'numeraire' category and *ln* represents natural logarithm.

2. INDIVIDUAL LEVEL ESTIMATES OF THE MNL MODEL OF OCCUPATIONAL CHOICE

The estimation of the parameters of the Multinomial Logit Model (2) [with selection probabilities given by (1.1) and (1.2)] can be carried out by using maximum likelihood procedures.[3] Such estimates will be consistent and efficient asymptotically provided that MNL model is correctly specified.[4] These estimates can then be used to calculate the appropriate selection probabilities.

3. THE OMITTED VARIABLE BIAS IN THE MULTINOMIAL LOGIT MODEL

The maximum likelihood estimates of the α coefficients from (2) may be biased if relevant variables have been left out of the specification. Such omitted variables can be measured or unmeasured, i.ė. latent. In this paper, however, we focus on the latent variable case.

Reconsider the MNL model described in (2) with the simplifying assumption that X consists only of a scalar, x. Then the ith equation from (2) is given as below:

[3] In this regard, the appendix of Schmidt and Strauss (1975) gives further details.

[4] For an interpretation of these coefficient estimates, see Pindyck and Rubinfeld (1981). However, there may be an inference problem in such models when, as is the case here, there are more than two occupational categories to choose from. For a discussion of the above and related problems as well as some suggested solutions, see Crawford and Pollak (1988).

$$In \frac{P(y=i\,|\,x)}{P(y=0\,|\,x)} = \alpha_{i0} + \alpha_{i1}\,x = i = 1, 2, ..., L \quad \ldots \quad \ldots \quad (3)$$

Now consider the possibility of (3) being misspecified because it excludes a variable z which represents family background that may be defined as 'everything that siblings born and raised in a given family share together'. The variable z is often treated as a latent variable since direct measures of it are not readily available. Incidentally, other studies of the determinants of earnings and other measures of the socioeconomic achievement of individuals have shown latent variables similar to our z to be important influences on the regressand [for instance, see Taubman (1977) and Behrman and Wolfe (1984)]. In any event, exclusion of z would entail that (3) would represent a misspecified model while the true MNL model would be given as follows:

$$In \frac{P^*(y=i\,|\,x,z)}{P^*(y=0\,|\,x,z)} = \alpha_{i0} + \alpha_{i1}\,x + \beta_i z \quad \ldots \quad \ldots \quad (4)$$

where $i = 1, 2, ..., L$

or in an equivalent form:

$$P^*(y=i\,|\,x,z) = \frac{exp(\alpha_{i0} + \alpha_{i1}\,x + \beta_i z)}{1 + \sum_{i=1}^{L} exp\,(\alpha_{i0} + \alpha_{i1}\,x + \beta_i z)} \quad \ldots \quad (4.1)$$

and

$$P^*(y=0\,|\,x,z) = \frac{1}{1 + \sum_{i=1}^{L} exp\,(\alpha_{i0} + \alpha_{i1}\,\mathrm{x} + \beta_i z)} \quad \ldots \quad (4.2)$$

where P^* represent the correctly specified logistic probability function. Also, note that since z is latent, we can define its (arbitrary) units such that its coefficient is unity. We also assume that z is continuous.

If the misspecified MNL model (3) is estimated rather than the true model (4), than $\hat{\alpha}_i$, the maximum likelihood estimates of $\hat{\alpha}_i$ may be biased; (i=1, ..., L). In particular, we will concentrate on α_{i1} i.e. the estimated 'slope' coefficient. In order to investigate the bias question more closely let us consider the following definitions:

(i) Definition: "Unconditional independence of z and x."

If $E(zx) = 0$, z and x are called unconditionally independent. Unconditional independence implies $r_1 = 0$ where r_1 is the coefficient estimate of x in the (auxiliary) regression of z on x, i.e., $z = r_0 + r_1\,x$.

(ii) Definition: "Conditional independence of z and x."

If $E(zx \mid y) = 0$, z and x are called conditionally independent. Conditional independence implies $\delta_1 = 0$, where δ_1 in the coefficient estimate of x in the auxiliary regression of z *and* y, i.e., $z = \delta_0 + \delta_1 x + \delta_2 y$.

Proposition 1: Sufficient Condition for Unbiasedness

Conditional independence of z and x as defined in (ii) above, is sufficient condition for $\hat{\alpha}_{i1}$ to be unbiased.

Proposition 2: Necessary and Sufficient Condition for Unbiasedness

When the omitted relevant variable, z, conditional on y and x is normally distributed, the sufficient condition, i.e., conditional independence of z and x is also a necessary one for $\hat{\alpha}_{i1}$ to be unbiased.

Proposition 3: Direction of the Bias in $\hat{\alpha}_{i1}$

The asymptotic bias in $\hat{\alpha}_{i1}$ is given by:

$$Plim(\hat{\alpha}_{i1} - \alpha_{i1}) = \delta_1 \beta_i$$

where β_i is the coefficient (in the ith equation) of the omitted (latent) variable z and δ_1 is the association of z and x, conditional on y (see definition (ii) above).

The direction of the bias in $\hat{\alpha}_{i1}$ can be determined if we make the assumption that conditional on y and x, z is normally distributed. Then, if the latent explanatory variable z is omitted from the true MNL model as given in (4), the maximum likelihood estimates, $\hat{\alpha}_{i1}$, of the included explanatory variable x will be:

(a) **Unbiased** if and only if either $\beta_i = 0$ or conditional on y, z is independent of x;

(b) **Biased upward** if either $\beta_i > 0$ and $\delta_1 > 0$ or $\delta_1 < 0$ and $\beta_i < 0$; and

(c) **Biased downward** if either $\beta_i > 0$ and $\delta_1 < 0$ or $\beta_i < 0$ and $\delta_1 > 0$.

4. 'WITHIN-FAMILY DEVIATION FORM' ESTIMATES OF THE MNL MODEL: SIBLING DATA TO THE RESCUE

As noted above, the maximum likelihood estimates, $\hat{\alpha}_{i1}$, will be biased if the (misspecified) MNL Model (3) is estimated.[5] This bias arises since a relevant

[5] In our particular model where z is a latent variable which is constrained to have unity coefficient, the nature and the direction of bias in $\hat{\alpha}_{i1}$ would then depend only on δ_1, which is the coefficient of x in the auxiliary regression of z on x, and y (i.e., y is also being controlled for).

Now, it is likely that $\delta_1 > 0$. With reference to the specific MNL model given as Equation (4) in the text, let us interpret x as *ED* or years of schooling and interpret z as a measure of ability such as *IQ* or a dimension of shared family invironment such as parental schooling levels. Since the regressand y is really different occupational categories, then δ_1 is simply a measure of within occupational category (linear) association between *ED* and *IQ* (or the latent variable z, to be more exact).

Thus, if we agree that $\delta_1 > 0$ and since $\beta_i = 1$ by virtue of our choice of the units of measurement for z, we will expect $\hat{\alpha}_{i1}$, the coefficient estimate of x to be upward biased.

variable, z, is omitted from the specification and conditional on y and z, it is associated with x. However, it is still possible to get unbiased $\hat{\alpha}_{i1}$ if we estimate the following 'within-family deviation' version of the misspecified model given in (3):

$$ln\ \left(P_i^w / P^w_m\right) = \alpha_{i0.m} + \alpha_{i1.m}\Delta x \qquad i, m = 0, 1, ..., L \quad ; i > m \qquad ...\ (5)$$

where the superscript w refers to 'within-family' (as against the individual level variables described earlier) and the subscript m refers to the values of the relevant variables for the 'numeraire' or 'reference' sibling defined here as the one whose occupation has the lowest rank amongst his/her siblings or 'within' that particular family. Then, for each individual, $\Delta x = (x - x_m)$ represents the deviation of his/her x value from the corresponding value, x_m, where the latter is the value of x for the 'numeraire' sibling. Note that whereas the 'numeraire' occupational category was fixed (and set at 0) for the individual level version of the MNL Model (3), in the above 'deviation-form' version (5), the numeraire category, P_m, could vary across families.

5. COMMENTS/CONCLUDING REMARKS

(a) Comments on the 'Deviation' Form vs. Individual Level

The following two comments are pertinent to the above discussion of the two estimation methods i.e. Deviation Form and Individual Level.

Estimation Method

Comment 1: The most significant point about the MNL Model (5) is that the maximum likelihood estimates of α_{i1} (i.e., $\hat{\alpha}_{i1.m}$) would now be (asymptotically) unbiased since z is assumed to be identical across all siblings in a given family which implies $\Delta z = (z - z_m) = 0$. Thus, there would be no omitted variable in (5) and hence no omitted variable bias to contaminate $\hat{\alpha}_{i1}$.

The above point can be further elaborated by considering the following three models together which have, in fact, been already introduced separately:

$$ln\ \left(P^*_i / P^*_0\right) = \alpha_{i0.0} + \alpha_{i1.0}\, x + \alpha_{i2.0}\, z \qquad i = 1,.., L \qquad ... \qquad (6)$$

$$ln\ \left(P_i / P_0\right) = \alpha_{i0.0} + \alpha_{i1.0}\, x \qquad i = 1,.., L \qquad ... \qquad (7)$$

$$ln\ \left(P_i^w / P_m^w\right) = \alpha_{i0.m} + \alpha_{i1.m}\Delta x \qquad i, m = 0, 1, ., L \quad i > m \qquad ... \qquad (8)$$

Note that (6) is nothing but the correctly specified MNL model already introduced as (4) while (7) is the (misspecified) model based on individual level data and was introduced as (2) and finally, (8) is the deviation form model, we just finished introducing as (5). Regarding, the subscript notation for the parameters in (6)–(8) note that for $\alpha_{is,t}$, i = occupational category chosen, s = sequence number of the regressand in the ith equation and t = the 'numeraire' occupational category used in the ith equation.

Further, note the following about the models (6)–(8).

P_i^* = True probability of an individual choosing the occupational category i— since it is estimated from the true Model (6).

P_i = (Incorrect) probability of an individual choosing the occupational category i as estimated from the Model (7) above.

P_i^w = Probability of an individual choosing the occupational category i when 'deviation-form' version of (8) is estimated. Note that under the assumption $\Delta z = 0$, $P_i^w \equiv P_i^*$.

Let us now compare individual level estimates from (7) to the 'deviation-form' equation set (8) when $m = 0$ and $i = 1, ..., L$. These latter estimates $\hat{\alpha}_{i1}$ would be unbiased and comparing them to the corresponding ones obtained from the individual level estimates using (7) would give us a (point) estimate of the extent of the bias due to the omission of z.

Thus, given our assumptions about z and its relationship to y and x, the methodology of 'within-family deviation form' maximum likelihood estimation of MNL model of occupational choice gives us asymptotically unbiased estimates of α_{i1} where $i = 1, ..., L$.

Comment 2: Existing Related Literature: The above methodology of employing 'within-family deviations' to flush out the latent omitted variable has also been used in some of the earlier studies. However, mostly such studies dealt with models where the regressand, y, was a continuous variable such as (log) earnings or years of schooling. [For instance, see Taubman (1977); Behrman and Wolfe (1984) and Shabbir (1989)]. However, there are relatively few studies where y is discrete. One notable exception is the discussion by Chamberlain (1984) of a binary logit model due to Rasch (1960) and a multinomial logit model based on McFadden (1974). The 'within-family' deviation methodology in the above noted studies is based on taking differences between pairs of siblings.

(b) Concluding Remarks/Empirical Research in Progress

We have outlined the possibility of a bias in the coefficient estimate for the included explanatory variable(s) if a latent variable that is equally shared by siblings in a family is omitted from the specification of MNL model of occupational choice.[6] However, under certain assumptions regarding the latent variable z, in particular, that it is purely familial,[7] we can estimate a 'within-family deviation' version of the MNL model where the maximum likelihood estimates would not be biased. The next step would be to conduct an empirical estimation of this model using data on siblings.

REFERENCES

Behrman, Jere R., and Barbara L. Wolfe (1984) The Socio-economic Impact of Schooling in a Developing Country. *Review of Economics and Statistics* 65:2.

Behrman, Jere R., Z. Hrubic, Paul Taubman and T. Wales (1980) *Socioeconomic Success*. New York: North Holland.

Chamberlain, Gary (1984) Panel Data. In Z. Griliches and M. D. Intrilligator (eds) *Handbook of Econometrics*. Volume 2. New York: North Holland.

Crawford, D. L., and R. Pollak (1988) Order and Inference in Qualitative Response Models. Discussion Paper, NBER.

Griliches, Zvi (1979) Sibling Models and Data in Economics: Beginning of a Survey. *Journal of Political Economy* 87:5.

Lee, L. (1980) Specification Error in Multinominal Logit Models: Analysis of the Omitted Variable Bias. Minneapolis, Minnesota: Center for Economic Research, Department of Economics, University of Minnesota. (Discussion Paper No. 80–131).

McFadden, D. (1974) Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (ed) *Frontiers in Econometrics*. New York: Academic Press.

Nerlove, M., and J. Press (1973) Multivariate and Log Linear Probability Models in Econometrics. Center for Statistics and Probability, Northwestern University. (Discussion Paper No. 1.)

Pindyck, R., and D. Rubinfeld (1981) *Econometric Models and Economic Forecasts*. New York: McGraw Hill.

Rasch, G. (1960) Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Denmark's Paedagogiske Institute.

Schmidt, P., and R. Strauss (1975) The Prediction of Occupation Using Multiple Logit Models. *International Economic Review* 16: 471–486.

Shabbir, Tayyeb (1987) Across and Intrahousehold Effects in a Model of Earnings and Schooling with Controls for Latent Factors. Unpublished Ph.D. Dissertation. Philadelphia: University of Pennsylvania.

[6] See the discussion in footnote 5 given earlier.

[7] In fact, the structure of z may be more complicated; it may also contain individual specific components which would require more complicated model specification and more complex estimation techniques than those suggested in this paper. In the context of the related literature on earnings functions, some of these issues have been discussed in Griliches (1979) or Behrman *et al.* (1980). However, I feel that such relatively more complicated models that are able to ask finer questions often can do so only after making correspondingly heuristic assumptions.

Shabbir, Tayyeb (1989) Latent Structure of Earnings Models. *The Pakistan Development Review* 28:4.

Taubman, Paul (ed) (1977) *Kinometrics: Determinants of Socioeconomic Success Within and Between Families*. New York: North-Holland Publishing Company.

Comments on "Multinomial Logit Model of Occupational Choice: A Latent Variable Approach"

We all know that in a standard linear regression model, the omission of relevant explanatory variables introduces bias in regression coefficient estimates. Furthermore, orthogonality conditions are readily available (between omitted and included regressors) under which biases disappear. In general, the directions of bias can also be related to the directions of correlations between omitted and included regressors. If there is no correlation (the orthogonality condition) between these two groups of variables, the bias is equal to zero.

Dr Shabbir's paper explores these issues in the context of a multinomial logit model–an interesting model with vast practical applications and containing substantial complications, vis-à-vis the standard linear regression model, which prevent a straightforward translation of the results I have enumerated in the preceding paragraph. Thus, Dr Shabbir's analytical results regarding bias and directions of bias in estimated "slope" coefficients on a multinomial logit model when relevant variables are inadvertently excluded indeed represent a significant contribution to the literature on econometric methodology.

There is also practical importance in Dr Shabbir's results. The use of sibling data, which he proposes in the paper, provides an operational procedure for modifying the estimator to erase bias. In a related study reported in Mariano, Reyes and Lim (1989) dealing with measurement errors in qualitative response models, corrections for bias in estimated slope coefficients can lead to substantial changes in the estimated relative effects of explanatory variables. For example, this is one major observation we arrived at in our analysis of farmers' decisions in the Philippines regarding the adoption of modern technology in agriculture. Differential effects of price subsidies, extension work, type of farm ownership and other factors change in character once corrections are introduced for biases due to measurement errors. Modern technology in our example concerns high yielding varieties of rice, fertilizers and pesticides.

Note that, through appropriate transformations to what we would call canonical form, we can show that the problem of omitted variables can be treated equivalently in terms of measurement errors. Thus, the practical lessons coming out of our study of the latter carry over, with appropriate transformations, to the study of the former. Incidentally, this same comment applies to the analytical phase of Dr Shabbir's study. The analytical results reported in Tayyeb's paper, can also be derived through the interpretation of the problem in terms of measurement errors. Incidentally, related analytical results are also reported in Yatchew and Griliches as well as in Kiefer.

Moving on to other technical issues regarding the paper, let me first point out that there is some ambiguity in the literature in the use of the phrase "multinomial logit model". The model that Tayyeb labels in the paper as multinomial logit is indeed the appropriate one–it allows for changes in factor effects across alternatives or choices. That is, his model described in his Equation (1.1) has α-coefficients which are subscripted by "*i*".

In models of this type, we can have *x*-variables which are constant, regardless of choices made. Examples of these are characteristics of individuals such as educational background, income, gender, and so on. On the other hand, there are other variables which vary across choices. These are attributes related to the choices themselves, such as, for example, compensation for the occupations covered in Tayyeb's study. If the α-parameters in the model do not change across choices, coefficients of variables in the first category will not be identifiable.

The next technical issue that I would like to raise concerns model estimation. There are various procedures that can be used. One is the approach discussed in this paper–least squares based on the linearity of the log odds ratio vis-à-vis the explanatory variables. A second procedure is a weighted least squares variation of the first. The log odds ratio itself is not observable and before relationships, like (3) in the paper, can be estimated, estimates of the log odds ratio must be calculated from the data (basically subsample proportions). These calculated values are then used as "observations" for the dependent variable in (3). The fact that these are estimates introduces heteroskedasticity in the version of (3) for the *estimated* log-odds ratio.

A third method of estimation is the maximisation of the "pseudo" likelihood function based on the model as specified. I call the likelihood a "pseudo" contstruct because this is not the correct likelihood that corresponds to the appropriate data generating process if the model has been misspecified (because of omission of variables).

All three procedures are distinct from each other. And they would be affected by omission of variables in different ways.

It would be interesting for Tayyeb to work out the analytical bias expressions for these three estimators and to show, in his empirical applications, how these three estimates differ from each other and from the bias-corrected least-squares procedure that he proposes.

Ladies and gentlemen, it is my pleasure to have served as discussant of Tayyeb's paper. Let me conclude my comments by stating once more that Tayyeb's paper presents new analytical results which are important not only on their own technical merits but also in their practical implications in the econometric study of occupational choice and other processes dealing with qualitative and other limited dependent variables.

Roberto S. Mariano

University of Pennsylvania,
Philadelphia, USA.

REFERENCE

Mariano, Roberto S., C. R. Reyes and P. C. Lim (1989) Measurement Errors in Limited-dependent Variable Models–Theory and Applications to Adoption of Technology in Philippine Agriculture. Philadelphia: Department of Economics, University of Pennsylvania. (Mimeographed.)