SPECIFICATION ERROR IN MULTINOMIAL LOGIT MODELS:

ANALYSIS OF THE OMITTED VARIABLE BIAS

by

Lung-Fei Lee

Discussion Paper No. 80 - 131, May 1980

Center for Economic Research
Department of Economics
University of Minnesota
Minneapolis, Minnesota 55455

SPECIFICATION ERROR IN MULTINOMIAL LOGIT MODELS:
ANALYSIS OF THE OMITTED VARIABLE BIAS

by

Lung-Fei Lee

University of Minnesota, Minneapolis

Abstract

In this article, we analyze the omitted variable bias problem in the multinomial logistic probability model. Sufficient, as well as necessary, conditions under which the omitted variable will not create biased coefficient estimates for the included variables are derived. Conditional on the response variable, if the omitted explanatory and the included explanatory variable are independent, the bias will not occur. Bias will occur if the omitted relevant variable is independent with the included explanatory variable. The coefficient of the included variable plays an important role in the direction of the bias.

Lung-Fei Lee, Department of Economics, 1035 Business Administration,
271 19th Avenue South, Minneapolis, Minnesota 55455

February 1980

SPECIFICATION ERROR IN MULTINOMIAL LOGIT MODELS: ANALYSIS OF THE OMITTED VARIABLE BIAS

by

Lung-Fei Lee(*)

1. Introduction

The omitted variables problem has been thoroughly analyzed in the econometrics literature for the linear regression model; however, such problems have not been analyzed for discrete probability models. In this article, we concentrate on analyzing this specification error problem for the multinomial logistic probability model. We compare the problem in the logit model to that in the general linear regression model. Sufficient, as well as necessary, conditions under which the omitted variable will not create biased coefficient estimates for the included variables are derived. The direction of the omitted variable bias is also investigated.

The article is organized as follows. Section 2 compares the omitted variables problem in the multinomial logit model with that in the linear regression model. The asymptotic bias or inconsistency of the maximum likelihood estimates and the minimum chi-square estimates are examined. In section 3, we analyze the omitted variables bias in more detail for the models where the omitted variables are discrete. Sufficient as well as necessary conditions under which the omitted variable will not affect the coefficients of the included variables are derived. The

(*) I appreciate having financial support from the National Science Foundation under Grant SOC-78-07304.

exact asymptotic bias and the direction of the bias are analyzed for the case when all of the explanatory variables are discrete. Section 4 extends the analysis to the model of a continuous omitted variable. Finally we summarize our conclusions.

2. The Omitted Variable Bias in the Logit Model

Let x and z be two stochastic explanatory variables and let $y = 0, 1, \ldots, L$ be a polychotomous response variable with $L + 1$ mutually exclusive and exhaustive categories. The multinomial logistic probability model is specified as

$$P(y = i|x, z) = \frac{e^{\alpha_{i0} + x\alpha_{i1} + z\beta_i}}{1 + \Sigma_{\ell=1}^{L} e^{\alpha_{\ell 0} + x\alpha_{\ell 1} + z\beta_\ell}} \qquad i=1,2,\ldots,L \qquad (2.1)$$

and

$$P(y = 0|x, z) = \frac{1}{1 + \Sigma_{\ell=1}^{L} e^{\alpha_{\ell 0} + x\alpha_{\ell 1} + z\beta_\ell}} \qquad (2.2)$$

where $P(y = i|x, z)$ denotes the probability of a response in category i conditional on x and z. Without loss of generality, we assume in this section that x has zero mean. Suppose the above model is the correct probability model. The specification error occurs by omitting a relevant explanatory variable z from the fitted model. Specifically,

$$P^*(y = 1|x) = \frac{e^{\alpha_{i0} + x\alpha_{i1}}}{1 + \Sigma_{\ell=1}^{L} e^{\alpha_{\ell 0} + x\alpha_{\ell 1}}} \qquad i=1,\ldots,L \qquad (2.3)$$

and

$$P^*(y = 0|x) = \frac{1}{1 + \Sigma_{\ell=1}^{L} e^{\alpha_{\ell 0} + x\alpha_{\ell 1}}} \qquad (2.4)$$

where P^* denotes the misspecified logistic probability function. To facilitate a comparison of this probability model with the linear regression model, it is instructive to write the model in an equivalent logarithmic probability odds representation. The correctly specified logistic probability model is

$$\ell n \frac{P(y = i|x, z)}{P(y = 0|x, z)} = \alpha_{i0} + x\alpha_{i1} + z\beta_i \qquad i=1,\ldots,L \quad (2.5)$$

and the misspecified model is

$$\ell n \frac{P^*(y = i|x)}{P^*(y = 0|x)} = \alpha_{i0} + x\alpha_{i1} \qquad i=1,\ldots,L \quad (2.6)$$

With variable z omitted from (2.5), we would like to investigate how the coefficient α_{i1} estimated from the misspecified model is different from the coefficient α_{i1} in the true model.

To investigate the omitted variable bias, we need to examine the relationship between equations (2.5) and (2.6). Let $\delta_0 + x\delta_1$ be the best linear predictor (wide sense conditional expectation) of z. It follows

$$z = \delta_0 + x\delta_1 + \xi \qquad (2.7)$$

where $E(x\xi) = 0$, $E(\xi) = 0$ and $E(zx) = E(x^2)\delta_1$. Let $P(y = i|x)$ be the probability of $y = i$ conditional only on x and $\phi(\xi|x)$ be the conditional probability density function of ξ conditional on x. It follows from (2.1) and (2.2),

$$P(y = i|x) = e^{\alpha_{i0}+x\alpha_{i1} + (\delta_0+x\delta_1)\beta_i} \int \frac{e^{\xi\beta_i}}{1+\sum_{\ell=1}^{L} e^{\alpha_{\ell 0}+x\alpha_{\ell 1}+(\delta_0+x\delta_1)\beta_\ell+\beta_\ell\xi}} \cdot$$

$$\cdot\ \phi(\xi|x)d\xi \qquad (2.8)$$

and

$$P(y = 0|x) = \int \frac{1}{1+\sum_{\ell=1}^{L} e^{\alpha_{\ell 0}+x\alpha_{\ell 1}+(\delta_0+x\delta_1)\beta_\ell+\beta_\ell\xi}} \phi(\xi|x)d\xi \qquad (2.9)$$

$$\text{Let } G_i(x) = \ell n \int \frac{e^{\xi\beta_i}}{1+\sum_{\ell=1}^{L} e^{\alpha_{\ell 0}+x\alpha_{\ell 1}+(\delta_0+x\delta_1)\beta_\ell+\beta_\ell\xi}}\phi(\xi|x)d\xi$$

$$- \ell n \int \frac{1}{1+\sum_{\ell=1}^{L} e^{\alpha_{\ell 0}+x\alpha_{\ell 1}+(\delta_0+x\delta_1)\beta_\ell+\beta_\ell\xi}} \phi(\xi|x)d\xi\ .$$

We have

$$\ell n \frac{P(y = i|x)}{P(y = 0|x)} = \alpha_{i0}+\delta_0\beta_i + x(\alpha_{i1}+\delta_1\beta_i) + G_i(x) \quad i=1,\dots,L \qquad (2.10)$$

The probability function $P(y|x)$ is the correct probability for the occurrence of y conditional on x. Let $\hat{\alpha}_{i1}$ be the maximum likelihood estimate (or the minimum chi-square estimate) of α_{i1} derived by fitting the misspecified model (2.6). If the $G_i(x)$ are constant functions, i.e., $G_i(x) = c_i$ for all i, we have

$$\text{plim}\ \hat{\alpha}_{i1} = \alpha_{i1} + \delta_1\beta_i \qquad i=1,\cdots,L.$$

$\hat{\alpha}_{i1}$ is an asymptotically biased (inconsistent) estimate of α_{i1} if $\delta_1 \neq 0$ and $\beta_i \neq 0$. The direction of the asymptotic bias depends on the sign of the coefficient β_i of the omitted variable z and its covariance $E(zx)$ with the included variable x. For this case, the omitted variable bias is exactly analogous to the omitted variables bias in the standard linear regression model.

However, the above result holds only if the $G_i(x)$ are constant functions. Given the complicated functional form of the $G_i(x)$ in (2.10), it is very unlikely in general that the $G_i(x)$ will be constant functions. The estimate $\hat{\alpha}_{i1}$ may be asymptotically bias even if $\delta_1 = 0$ and $\phi(\xi|x) = \phi(\xi)$, i.e., z and x are independent, as the functions $G_i(x)$ are not constants. $\hat{\alpha}_{i1}$ is biased due to omission of the functions $C_i(x) \equiv (\delta_0 + x\delta_1)\beta_i + G_i(x)$ from the fitted model; consequently, the effects of omitting relevant variables in logit model are quite different from that in the linear regression model.

When the logit model is misspecified as in (2.3) and (2.4), the maximum likelihood estimates (MLE) of α_{i1}, $i=1,\ldots,L$ are biased whenever the functions $C_i(x) = (\delta_0 + x\delta_1)\beta_i + G_i(x)$ are not constants.[1/]

This can be shown as follows. The misspecified model in (2.3) and (2.4) implies the following logarithmic likelihood function,

$$\ell nL = \Sigma_{t=1}^{N}\Sigma_{i=0}^{L}\ I_{it}\,\ell nP^*(y_t = i|x_t) \tag{2.11}$$

where t refers to the t^{th} sample and I_i are dichotomous indicators which are defined as $I_{it} = 1$ if $y_t = i$ and 0 otherwise. The MLE

[1/] When $C_i(x)$ are constants, only the constant terms α_{i0} will be biased.

$\hat{\alpha}_{i0}$, $\hat{\alpha}_{i1}$, $i=1,\ldots,L$, are solved from the first order conditions which are

$$\Sigma_{t=1}^{T}(I_{it} - P^{*}(y_t = i|x_t)) = 0 \qquad i=1,\ldots,L \qquad (2.12)$$

$$\Sigma_{t=1}^{T}(I_{it} - P^{*}(y_t = i|x_t))x_t = 0 \qquad i=1,\ldots,L \qquad (2.13)$$

The equations (2.10) can be rewritten as

$$I_{it} = P(y_t = i|x_t) + \xi_{it} \qquad (2.10)'$$

where $E(\xi_{it}|x_t) = 0$. As T tends to infinity, we have in the probability limit,

$$\text{plim}\,\frac{1}{T}\,\Sigma_{t=1}^{T}\left(\frac{e^{\alpha_{i0}+x_t\alpha_{i1}+C_i(x_t)}}{1+\sum_{i=1}^{L} e^{\alpha_{i0}+x_t\alpha_{i1}+C_i(x_t)}} - \frac{e^{\tilde{\alpha}_{i0}+x_t\tilde{\alpha}_{i1}}}{1+\Sigma_{i=1}^{L}e^{\tilde{\alpha}_{i0}+x_t\tilde{\alpha}_{i1}}}\right) = 0 \qquad (2.14)$$

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

$$\text{plim}\,\frac{1}{T}\,\Sigma_{t=1}^{T}\left(\frac{e^{\alpha_{i0}+x_t\alpha_{i1}+C_i(x_t)}}{1+\Sigma_{i=1}^{L}e^{\alpha_{i0}+x_t\alpha_{i1}+C_i(x_t)}} - \frac{e^{\tilde{\alpha}_{i0}+x_t\tilde{\alpha}_{i1}}}{1+\Sigma_{i=1}^{L}e^{\tilde{\alpha}_{i0}+x_t\tilde{\alpha}_{i1}}}\right) x_t=0, \qquad (2.15)$$

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

where $\tilde{\alpha}_{i0} = \text{plim}\ \hat{\alpha}_{i0}$ and $\tilde{\alpha}_{i1} = \text{plim}\ \hat{\alpha}_{i1}$ denote the probability limits of the MLE estimates. From the above equations, evidently, the $\hat{\alpha}_{i1}$ are not consistent estimates of the α_{i1} as the functions are not constants.

When the variables x are discrete or x_t, $t=1,\ldots,T$ are T fixed constants, an alternative estimation procedure is the minimum chi-square procedure. For the minimum chi-square estimates as contrary to the MLE, the exact asymptotic bias can be easily derived. Assume that for each t, there are N_{it} observations on $y = i$. Let $n_{it} = \frac{N_{it}}{N_t}$, where $N_t = \Sigma_{i=1}^{L} N_{it}$, which is the MLE of $P_{it} \equiv P(y = i|x_t)$. The misspecified equations in (2.6) are

$$\ell n \frac{N_{it}}{N_{0t}} = \alpha_{i0} + x_t \alpha_{i1} + \varepsilon_{it} \qquad i=1,\ldots,L \qquad (2.16)$$

where $\varepsilon_{it} = \ell n \frac{N_{it}}{N_{0t}} - \ell n \frac{P_{it}}{P_{0t}}$. The asymptotic variance matrix of $\varepsilon_t' = (\varepsilon_{1t},\ldots,\varepsilon_{Lt})$ is

$$\Lambda_t = \frac{1}{N_t}\left\{\begin{pmatrix} P_{1t}^{-1} & & \\ & \ddots & \\ & & P_{Lt}^{-1} \end{pmatrix} + P_{0t}^{-1}\ell\ell'\right\}$$

$$= \frac{1}{N_t}\left\{\begin{pmatrix} P_{1t} & & \\ & \ddots & \\ & & P_{1t} \end{pmatrix} - \begin{pmatrix} P_{1t} \\ \vdots \\ P_{Lt} \end{pmatrix} [P_{1t},\ldots,P_{Lt}]\right\}^{-1}$$

where $\ell' = (1,\ldots,1)$ is a L vector of ones. Let $X_t = I_L \otimes (1\ x_t)$, $W_t' = (\ell n \frac{N_{1t}}{N_{0t}},\ldots,\ \ell n \frac{N_{Lt}}{N_{0t}})$ and $\alpha' = (\alpha_{10},\ \alpha_{11},\ \ldots,\ \alpha_{L0},\ \alpha_{L1})$. The minimum chi-square estimate of α for the misspecified model is given by

$$\hat{\alpha} = (\Sigma_{t=1}^{T} X_T' \tilde{\Lambda}_t^{-1} X_t)^{-1} \Sigma_{t=1}^{T} X_t' \tilde{\Lambda}_t^{-1} W_t \qquad (2.17)$$

where $\tilde{\Lambda}_t^{-1} = N_t\left\{\begin{pmatrix} n_{1t} & & \\ & \ddots & \\ & & n_{Lt} \end{pmatrix} - \begin{pmatrix} n_{1t} \\ \vdots \\ n_{1t} \end{pmatrix} [n_{1t}\ldots\ldots n_{Lt}]\right\}$ are used as weights.

As N_{it} goes to infinity for all i and t, the exact asymptotic bias of $\hat{\alpha}$ is

$$\text{plim}\ \hat{\alpha} - \alpha = \text{plim}\ (\Sigma_{t=1}^{T} X_t' \Lambda_t^{-1} X_t)^{-1} \Sigma_{t=1}^{T} X_t' \Lambda_t^{-1} C_t \qquad (2.18)$$

where $C_t' = [C_1(x_t),\ldots,C_L(x_t)]$. [2/]

[2/] If z_t is a function of x_t, $C_i(x_t) = z_t\beta_i$ and the bias formula (2.18) will be similar to the omitted variables bias in heteroscedastic linear regression model.

3. Discrete Omitted Variables

The above approach, based on the best linear prediction in (2.7), is illustrative but does not lead us very far. To analyze the problem in more detail, it is desirable to distinguish between cases of a discrete omitted variable z and a continuous omitted variable. In this section, we analyze the discrete omitted variable case.

Let $z = 0, 1, \ldots, M$ be the omitted discrete variable with $M + 1$ categories. The discrete variable z may be an ordered polychotomous or an unordered polychotomous variable. When z is unordered, different categories of z may have different effects. In any case, we can introduce M auxiliary dichotomous variables z_i as follows,

$$\begin{aligned} z_i &= 1 \text{ if and only if } z = i, \\ &= 0 \text{ otherwise.} \end{aligned}$$

The correctly specified logistic probability model is

$$\ln \frac{P(y = i|x, z)}{P(y = 0|x, z)} = \alpha_{i0} + x\alpha_{i1} + z_1\beta_{i1} + \ldots + z_M\beta_{iM} \qquad i=1,\ldots,L \tag{3.1}$$

where the explanatory variable x can be either discrete or continuous. When z is ordered, $\beta_{ij} = j\beta_i$ and model (3.1) is the same model as specified in (2.5).

The first problem is to derive conditions for omitting variable z without affecting the coefficient of the included variable x. The following proposition which generalizes the concept of collapsibility of contingency tables, Bishop et al [1975], provides the answer.

Proposition 1. The coefficient of x in the misspecified model (2.6) will not be biased by omitting the relevant discrete variable z [3] from the correct model (3.1) if, conditional on the response variable y, x and z are independent.

The above proposition can be proved with the following lemma.

Lemma 1. Equations (3.1) and (3.2),

$$\ell n \frac{P(z = j|x, y)}{P(z = 0|x, y)} = \delta_{j0} + x\delta_{j1} + y_1\beta_{1j} + \dots + y_L\beta_{Lj}, \quad j=1,\dots,M \tag{3.2}$$

are equivalent to equations (3.2) and (3.3),

$$\ell n \frac{P(y = i|x)}{P(y = 0|x)} = \alpha_{i0} + x\alpha_{i1} - \ell n \frac{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x\delta_{j1}}}{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x\delta_{j1}+\beta_{ij}}} \tag{3.3}$$

Proof of Lemma 1: By Bayes theorem, the following identity holds

$$\frac{P(y = i|x, z)}{P(y = 0|x, z)} = \frac{P(z|x, y = i)}{P(z|x, y = 0)} \frac{P(y = i|x)}{P(y = 0|x)} \tag{3.4}$$

(⇒ :) From equation (3.2),

$$P(z|x, y) = \frac{e^{\Sigma_{j=1}^{M}(\delta_{j0}+x\delta_{j1}+y_1\beta_{1j}+\dots+y_L\beta_{Lj})z_j}}{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x\delta_{j1}+y_1\beta_{1j}+\dots+y_2\beta_{Lj}}}$$

It follows that

[3] This means that all the auxiliary dichotomous variables in equation (3.1) are excluded.

$$\frac{P(z|x,\ y=i)}{P(z|x,\ y=0)} = \frac{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x\delta_{j1}}}{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x\delta_{j1}+\beta_{ij}}}\ e^{\Sigma_{j=1}^{M}\beta_{ij}z_j} \qquad (3.5)$$

and hence

$$\ell n\,\frac{P(y=i|x)}{P(y=0|x)} = \ell n\,\frac{P(y=i|x,\ z)}{P(y=0|x,\ z)} - \ell n\,\frac{P(z|x,\ y=i)}{P(z|x,\ y=0)}$$

$$= \alpha_{i0} + x\alpha_{i1} - \ell n\,\left(\frac{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x\delta_{j1}}}{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x\delta_{j1}+\beta_{ij}}}\right)\ .$$

(⇐ :) From equation (3.2), equation (3.5) holds. Hence, it follows from equations (3.4), (3.5) and (3.3),

$$\ell n\,\frac{P(y=i|x,\ z)}{P(y=0|x,\ z)} = \ell n\,\frac{P(z|x,\ y=i)}{P(z|x,\ y=0)} + \ell n\,\frac{P(y=i|x)}{P(y=0|x)}$$

$$= \alpha_{i0} + x\alpha_{i1} + \Sigma_{j=1}^{M}\beta_{ij}z_j \qquad \text{Q.E.D.}$$

Proof of Proposition 1.

(⇐ :) Since z and x are independent conditional on y, $P(z|x,\ y) = P(z|y)$. Let $P(z=j|y) = \Phi_j(y,\ \theta)$ be the (unspecified) conditional probability of $z = j$, with unknown parameter θ. As y is discrete, the probability function $P(z=j|y)$ can always be rewritten in a logistic functional form,

$$\ell n\,\frac{P(z=j|x,\ y)}{P(z=0|x,\ y)} = \ell n\,\frac{\Phi_j(y,\ \theta)}{\Phi_0(y,\ \theta)}$$

$$= \delta_{j0}+y_1\lambda_{1j}+\ldots+y_L\lambda_{Lj} \qquad j=1,\ldots,M$$

where δ_{j0}, $\lambda_{\ell j}$ are implicitly defined from the above equation. From this equation and equation (3.1), it can be easily shown that $\lambda_{\ell j}$ must satisfy the relations $\lambda_{\ell j} = \beta_{\ell j}$ in (3.1) to have a well defined joint probability function $P(y, z|x)$. (See, e.g., Nerlove and Press [1976] for the analysis of the general log linear model). It follows from lemma 1 that

$$\ell n \frac{P(y = i|x)}{P(y = 0|x)} = \alpha_{i0} + x\alpha_{i1} - \ell n \left(\frac{1+\Sigma_{j=1}^{M} e^{\delta_{j0}}}{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+\beta_{ij}}}\right)$$

$$= \overset{*}{\alpha}_{i0} + x\alpha_{i1} \qquad i=1,\ldots,L \qquad (3.6)$$

where $\overset{*}{\alpha}_{i0} = \alpha_{i0} - \ell n \left(\frac{1+\Sigma_{j=1}^{M} e^{\delta_{j0}}}{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+\beta_{ij}}}\right)$. Hence if the misspecified model (2.6) is estimated, we have

$$\text{plim } \hat{\alpha}_{i1} = \alpha_{i1}$$

since fitting equation (2.6) by the maximum likelihood procedure (or minimum chi-square if applicable) is equivalent to fitting equations in (3.6). Hence the coefficient of x will not be biased when z is omitted.[4/]

Q.E.D.

This condition, however, is not a necessary condition. Suppose that conditional on y, x and z are not independent. Let $P(z = j|x) = F_j(x)$, i=0,....,M be the (unspecified) probability function of z conditional on x. It follows that

[4/] The constant term will, however, be biased. But this term is of no interest to us.

$$\ell n \frac{P(z = j|x)}{P(z = 0|x)} = \ell n \frac{F_j(x)}{F_0(x)} \qquad j=1,\ldots,M \qquad (3.7)$$

From lemma 1, equations (3.1) and (3.7) imply

$$\ell n \frac{P(z = j|x, y)}{P(z = 0|x, y)} = J_j(x) + y_1\beta_{1j} +\ldots+ y_L\beta_{Lj} \quad j-1,\ldots,M \qquad (3.8)$$

where $J_j(x) = \ell n \frac{F_j(x)}{F_0(x)} + \ell n \left(\frac{1+\Sigma_{i=1}^{L} e^{\alpha_{i0}+x\alpha_{i1}}}{1+\Sigma_{i=1}^{L} e^{\alpha_{i0}+x\alpha_{i1}+\beta_{ij}}}\right)$. Since conditional on y, x and z are not independent, $J_j(x)$ are not constant functions for at least for some j. It follows from the lemma 1 and from equations (3.8) and (3.1),

$$\ell n \frac{P(y = i|x)}{P(y = 0|x)} = \alpha_{i0} + x\alpha_{i1} - G_i(x) \qquad (3.9)$$

where $G_i(x) = \ell n\left(\frac{1+\Sigma_{j=1}^{M} e^{J_j(x)}}{1+\Sigma_{j=1}^{M} e^{J_j(x)+\beta_{ij}}}\right)$. If the functions $G_i(x)$ are not constants, and the misspecified equations (2.6) omit the functions $G_i(x)$ the coefficient of x will be affected. Therefore, if the functions $G_i(x)$ are not constants, the sufficient condition in the proposition is also necessary. However, for some special probability function $F_j(x)$ of x, the functions $G_i(x)$ can be constants even if $J_j(x)$ are not constant functions. For example, the following probability functions $F_j(x)$ which satisfy

$$\frac{F_j(x)}{F_0(x)} = \frac{e^{c_i}-1}{1-e^{\beta_{ij}+c_i}} \; \frac{1+\Sigma_{i=1}^{L} e^{\alpha_{i0}+x\alpha_{i1}+\beta_{ij}}}{1+\Sigma_{i=1}^{L} e^{\alpha_{io}+x\alpha_{i1}}} A_j(x)$$

where $A_j(x)$ are not constant functions $\Sigma_{j=1}^{M} A_j(x) = 1$ and the constants $c_i \neq 1$, will imply $G_i(x) = c_i$ for all i while $J_j(x) = \ln \frac{e^{c_i}-1}{1-e^{\beta_{ij}+c_i}} A_j(x)$ are not constants.

When the omitted variable z is dichotomous, i.e., $M = 1$, the sufficient condition in the proposition will be necessary. Suppose $J_1(x)$ in (3.8) is not a constant function but $G_1(x) = c$ is a constant function. It follows

$$e^c = \frac{1 + e^{J_1(x)}}{1 + e^{J_1(x) + \beta_{11}}}$$

and

$$J_1(x) = \ln \frac{e^c - 1}{1 - e^{\beta_{11+c}}}$$

which is a constant function, a constradiction. Hence we have

<u>Corollary 1</u>. A necessary and sufficient condition for the coefficient of x in the misspecified model (2.6) to be unbiased when the omitted relevant variable z is dichotomous, is that conditional on the response variable y, x, and z are independent.

When the included explanatory variable x is also discrete, more exact analytical results on the omitted variables bias can be derived. Suppose x is a polychotomous variable with $K + 1$ categories. Denote for each i, $i = 1,\ldots,K$, $x_i = 1$ if x is in its i^{th} category, $x_i = 0$ otherwise. The correctly specified logistic probability model is

$$\ell n \frac{P(y = i|x, z)}{P(y = 0|x, z)} = \alpha_{i0}+x_1\alpha_{i1}+\ldots+x_k\alpha_{ik}+z_1\beta_{i1}+\ldots+z_M\beta_{iM} \qquad (3.10)$$

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

and the misspecified model which omits the variable z is

$$\ell n \frac{P^*(y = i|x)}{P^*(y = 0|x)} = \alpha_{i0}+x_1\alpha_{i1}+\ldots+x_K\alpha_{iK} \qquad i=1,\ldots,L \qquad (3.11)$$

The omitted variable bias of the coefficients for x_i can be derived as follows. As both z and x are discrete variables, we can rewrite $P(z|x, y)$ in a log-linear form,

$$\ell n \frac{P(z = j|x, y)}{P(z = 0|x, y)} = \delta_{j0}+x_1\delta_{j1}+\ldots+x_K\delta_{jK}+y_1\beta_{1j}+\ldots+y_L\beta_{Lj} \qquad (3.12)$$

$$j=1,\ldots,M$$

Using lemma 1, it follows from equations (3.10) and (3.12),

$$\ell n \frac{P(y = i|x)}{P(y = 0|x)} = \alpha_{i0}+x_1\alpha_{i1}+\ldots+x_K\alpha_{iK}-\ell n\left(\frac{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x_1\delta_{j1}+\ldots+x_K\delta_{jK}}}{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x_1\delta_{j1}+\ldots+x_K\delta_{jK}+\beta_{ij}}}\right)$$

$$= \alpha^*_{i0}+x_1\alpha^*_{i1}+\ldots+x_K\alpha^*_{iK}, \qquad i=1,\ldots,L \qquad (3.13)$$

where

$$\alpha^*_{i0} = \alpha_{i0} - G_i(0) \quad ,$$

$$\alpha^*_{ik} = \alpha_{ik} - G_i(k) + G_i(0), \qquad k=1,\ldots,K \qquad (3.14)$$

with $$G_i(x) = \ell n \left(\frac{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x_1\delta_{j1}+\ldots+x_K\delta_{jK}}}{1+\Sigma_{j=1}^{M} e^{\delta_{j0}+x_1\delta_{j1}+\ldots+x_K\delta_{jK}+\beta_{ij}}}\right) .$$

Fitting the misspecified model (3.11) is equivalent to fitting the equations (3.13). Hence

$$\begin{aligned} \text{plim } \hat{\alpha}_{ik} &= \alpha^{*}_{ik} \\ &= \alpha_{ik} + G_i(0) - G_i(k) \end{aligned} \tag{3.15}$$

and $G_i(0) - G_i(k)$ is the exact asymptotic bias for the MLE or minimum chi-square estimates $\hat{\alpha}_{ik}$.

When the omitted variable z is dichotomous, i.e., $M = 1$, one can look more closely at the direction of the bias.

Proposition 2. For the multinomial logistic probability model (3.10) where the omitted variable z is dichotomous and the included variable x is discrete, the coefficient α_{ik} of x_k in the misspecified model will be

(i) biased upward if either $\beta_i > 0$ and $\delta_k > 0$, or $\beta_i < 0$ and $\delta_k < 0$,

(ii) biased downward if either $\beta_i > 0$ and $\delta_k < 0$, or $\beta_i < 0$ and $\delta_k > 0$,

and

(iii) unbiased if either $\beta_i = 0$ or $\delta_k = 0$.[5/]

Proof: The asymptotic bias of $\hat{\alpha}_{ik}$ is

[5/] As $M = 1$, the subscripts j in the parameters δ_{jk}, β_{ij} are redundant and are deleted.

$$b_{ik} = G_i(0) - G_i(k)$$

$$= \ell n \left[\frac{1+e^{\delta_0}}{1+e^{\delta_0+\beta_i}}\right] - \ell n \left[\frac{1+e^{\delta_0+\delta_k}}{1+e^{\delta_0+\delta_k+\beta_i}}\right]$$

By the mean value theorem,

$$\ell n \left[\frac{1+e^{\delta_0+\delta_k}}{1+e^{\delta_0+\delta_k+\beta_i}}\right] = \ell n \left[\frac{1+e^{\delta_0}}{1+e^{\delta_0+\beta_i}}\right] + \frac{\partial}{\partial\delta_k} \ell n \left[\frac{1+e^{\delta_0+\delta_k}}{1+e^{\delta_0+\delta_k+\beta_i}}\right]\Bigg|_{\delta_k^*} \cdot \delta_k$$

where δ_k^* lies between 0 and δ_k. As

$$\frac{\partial}{\partial\delta_k} \ell n \left[\frac{1+e^{\delta_0+\delta_k}}{1+e^{\delta_0+\delta_k+\beta_i}}\right] = \frac{e^{\delta_0+\delta_k}}{(1+e^{\delta_0+\delta_k})(1+e^{\delta_0+\delta_k+\beta_i})} (1 - e^{\beta_i}),$$

we have

$$b_{ik} = \frac{e^{\delta_0+\delta_k^*}}{(1+e^{\delta_0+\delta_k^*})(1+e^{\delta_0+\delta_k^*+\beta_i})} (e^{\beta_i} - 1)\delta_k \qquad (3.15)' .$$

The conclusion follows immediately from the above equation. Q.E.D.

The above conclusion is still valid if the logit model (3.10) and the misspecified model (3.11) including other explanatory variables w in addition to x as long as conditional on x and y, z and w are independent.

The above results are of interest as compared to the omitted variable bias in the standard linear regression model. In the standard linear model, the direction of the bias depends on the signs of the coefficient of the omitted variable z and the unconditional correlation of the included and excluded explanatory variables. In the logistic model, it depends on the association i.e., δ_k, of the included and excluded variables conditional on the dependent variable y.

To emphasize the implications of conditional and unconditional arguments, let us investigate the marginal association of z and x and its effects on the omitted variables bias for the dichotomous response logit model. The unspecified probability function $P(z|x)$ can be represented as

$$\ell n \frac{P(z = 1|x)}{P(z = 0|x)} = \lambda_0 + x_1\lambda_1 + \ldots + x_K\lambda_K \quad . \tag{3.16}$$

By lemma 1, equations (3.16) and (3.10) with $L = 1$ and $M = 1$ imply [6/]

$$\begin{aligned}\ell n \frac{P(z = 1|x, y)}{P(z = 0|x, y)} &= \lambda_0 + x_1\lambda_1 + \ldots + x_K\lambda_K + y\beta \\ &\quad + \ell n \left\{\frac{1+e^{\alpha_0+x_1\alpha_1+\ldots+x_K\alpha_K}}{1+e^{\alpha_0+x_1\alpha_1+\ldots+x_K\alpha_K+\beta}}\right\} \\ &= \delta_0 + x_1\delta_1 +\ldots+ x_K\delta_K + y\beta \end{aligned} \tag{3.17}$$

where

$$\begin{aligned}\delta_0 &= \lambda_0 + G(0), \\ \delta_k &= \lambda_k + G(k) - G(0), \qquad k=1,\ldots,K\end{aligned} \tag{3.18}$$

with $\quad G(x) = \ell n \left\{\frac{1+e^{\alpha_0+x_1\alpha_1+\ldots+x_K\alpha_K}}{1+e^{\alpha_0+x_1\alpha_1+\ldots+x_K\alpha_K+\beta}}\right\}$

[6/] As y is dochotomous, the subscript i in the parameters $\alpha_{i\ell}$ is redundant and is deleted in the following expressions.

The direction of the omitted variables bias of coefficient $\hat{\alpha}_k$ of x_k depends on the sign of δ_k and β as shown in proposition 2. The coefficients δ_k are related to the coefficients λ_k. Explicitly,

$$\delta_k = \lambda_k + \ln \frac{1+e^{\alpha_0+\alpha_k}}{1+e^{\alpha_0+\alpha_k+\beta}} - \ln \frac{1+e^{\alpha_0}}{1+e^{\alpha_0+\beta}}$$

$$= \lambda_k + \frac{e^{\alpha_0+\alpha_k^*}}{(1+e^{\alpha_o+\alpha_k^*})(1+e^{\alpha_0+\alpha_k^*+\beta})} (1-e^{\beta})\alpha_k \tag{3.19}$$

where the second equality is derived by the mean value theorem and α_k^* is evaluated between α_k and 0. From (3.19), δ_k may be negative when $\lambda_k > 0$, $\beta > 0$ if α_k is positive. Therefore, there are cases that result in a downward bias of α_k when the coefficient β of the omitted variable z and the association λ_k of x and z unconditional on the response variable have the same sign. In this set up, the sign of the true coefficient α_k of the included variable x_k plays an important role in determining the direction as well as the magnitude of the omitted variable bias. For the case when x and z are independent, α_k uniquely determines the direction of the bias of $\hat{\alpha}_k$. This is shown as follows. When x and z are independent, $\lambda_k = 0$, $k = 1,\ldots,K$ Hence equation (3.19) becomes

$$\delta_k = \frac{e^{\alpha_0 + \alpha_k^*}}{(1+e^{\alpha_0+\alpha_k^*})(1+e^{\alpha_0+\alpha_k^*+\beta})} (1 - e^{\beta})\alpha_k \tag{3.20}$$

As shown in equation (3.15)',

$$\operatorname{plim}\hat{\alpha}_k - \alpha_k = \frac{e^{\delta_0+\delta_k^*}}{(1+e^{\delta_0+\delta_k^*})(1+e^{\delta_0+\delta_k^*+\beta})}(e^{\beta}-1)\delta_k$$

and it follows with δ_k in (3.20),

$$\operatorname{plim}\hat{\alpha}_k - \alpha_k = -\frac{e^{\alpha_0+\alpha_k^*+\delta_0+\delta_k^*}\quad(1-e^{\beta})^2}{(1+e^{\alpha_0+\alpha_k^*})(1+e^{\delta_0+\delta_k^*})(1+e^{\alpha_0+\alpha_k^*+\beta})(1+e^{\delta_0+\delta_k^*+\beta})}\alpha_k$$

Hence when x and z are independent, omitting the relevant explanatory dichotomous variable z results in an upward bias if the true α_k is negative and a downward bias if α_k is positive.

4. Continuous Omitted Variable

When the omitted variable z is continuous, the analysis becomes more complicated; however, some of the conclusions in the last section still hold. Recall that the correctly specified model is (2.5),

$$\ln \frac{P(y = i \mid x,\ z)}{P(y = 0 \mid x,\ z)} = \alpha_{i0} + x\alpha_{i1} + z\,\beta_i \qquad i=1,\ldots,L \qquad (4.1)$$

where x can be continuous or discrete variables, and the misspecified model is

$$\ln \frac{P^*(y = i \mid x)}{P^*(y = 0 \mid x)} = \alpha_{i0} + x\alpha_{i1} \qquad i=1,\ldots,L \qquad (4.2)$$

Proposition 3. If conditional on the response variable y, the omitted continuous variable z is independent with the included explanatory variable x, the coefficients of x in the misspecified multinomial logit model will not be affected.

Proof: Suppose conditional on y, z and x are independent. We have $P(z \mid x,\ y) = P(z \mid x)$ where P denotes the conditional density function of z. The identity

$$\frac{P(y = i \mid x)}{P(y = 0 \mid x)} = \frac{P(y = i \mid x,\ z)}{P(y = 0 \mid x,\ z)}\,\frac{P(z \mid y = 0,\ x)}{P(z \mid y = i,\ x)} \qquad (4.3)$$

together with equation (4.1) implies that

$$\ln \frac{P(y = i \mid x)}{P(y = 0 \mid x)} = \alpha_{i0} + x\alpha_{i1} + z\beta_i + \ln \frac{P(z \mid y = 0,\ x)}{P(z \mid y = i,\ x)}$$

$$= \alpha_{i0} + x\alpha_{i1} + c_i(z,\ x) \qquad (4.4)$$

where $c_i(z, x) \equiv z\beta_i + \ell n \frac{P(z|y = 0, x)}{P(z|y = i, x)}$. Under the conditional independence assumption, $c_i(z, x) = z\beta_i + \ell n \frac{P(z|y = 0)}{P(z|y = i)}$ does not depend on x. As the probability $P(y|x)$ depends solely on x, the functions $c_i(z, x)$ are constants; hence, the coefficient of x is not affected when z is omitted. Q.E.D.

For some specific conditional distributions of z, it can be shown that the sufficient condition is also necessary. The following lemma generalizes a result in Olsen [1978] to the multinational logistic model. It applies to the case when z_1 conditional on y and x, is normally distributed.

Lemma 2. The multinomial logistic model,

$$\ell n \frac{P(y = i|x, z)}{P(y = 0|x, z)} = \alpha_{i0} + x\alpha_{i1} + z\beta_i \qquad i=1,\ldots,L$$

and the conditional normality distribution

$$P(z|x, y) = (2\pi)^{-\frac{1}{2}}\sigma^{-1}\exp(-\frac{1}{2\sigma^2}(z-\delta_0-x\delta_1-y_1\theta_1 - \ldots - y_L\theta_L)^2) \qquad (4.5)$$

imply

(i) $\theta_i = \sigma^2\beta_i$

and (ii) $$\ell n \frac{P(y = i|x)}{P(y = 0|x)} = \alpha_{i0} + x\alpha_{i1} + \beta_i(\delta_0 + x\delta_1) + \frac{1}{2}\sigma^2\beta_i^2 \qquad (4.6)$$

for $i = 1,\ldots,L$.

Proof: From (4.5), it follows

$$\frac{P(z|x, y = 0)}{P(z|x, y = i)} = \exp(-\frac{\theta_i}{\sigma^2}(z - \delta_0 - x\delta_1) + \frac{\theta_i^2}{2\sigma^2})$$

Hence equation (4.4) becomes

$$\ell n \frac{P(y = i|x)}{P(y = 0|x)} = \alpha_{i0} + x\alpha_{i1} + z\beta_i - \frac{\theta_i}{\sigma^2} z + \frac{\theta_i}{\sigma^2} (\delta_0 + x\delta_1) + \frac{\theta_i^2}{2\sigma^2}$$

Since the above probability odds should not depend on z, $\sigma^2\beta_i = \theta_i$ must hold for logical consistency; consequently, (i) and (ii) follow immediately. Q.E.D.

The following proposition follows directly from the lemma.

Proposition 4. Suppose conditional on y and x, z is normally distributed as in (4.5). When the explanatory variable z is omitted from the multinomial logistic probability model (4.1), the coefficient α_{i1} of the included explanatory variable x will be

(i) unbiased if and only if either $\beta_i = 0$ or conditional on y, z is independent with x,

(ii) biased upward if either $\beta_i > 0$ and $\delta_1 > 0$ or $\beta_i < 0$ and $\delta_1 < 0$,

(iii) biased downward if either $\beta_i > 0$ and $\delta_1 < 0$ or $\beta_i < 0$ and $\delta_1 > 0$.

The proposition implies that the sufficient condition in the proposition 3 is also necessary for this case. When the coefficient α_{i1} is biased, the asymptotic bias is simply

$$\text{plim}\ \hat{\alpha}_{i1} = \alpha_{i1} + \beta_i\delta_1 .$$

This case is unusual. It implies that the marginal distribution of z conditional on x is a mixture of L + 1 normal distributions, where the density function

$$P(z|x) = \Sigma_{i=0}^{L}(2\pi\sigma^2)^{-\frac{1}{2}}\exp(-\frac{1}{2\sigma^2}(z-\delta_0-x\delta_1-\sigma^2\beta_i)^2)$$

$$\frac{\exp(\alpha_{i0}+\frac{1}{2}\sigma^2\beta_i^2+\beta_i\delta_0+x(\alpha_{i1}+\beta_i\delta_1))}{1+\Sigma_{j=1}^{L}\exp(\alpha_{j0}+\frac{1}{2}\sigma^2\beta_j^2+\beta_j\delta_0+x(\alpha_{j1}+\beta_j\delta_1))}$$

with $\alpha_{00} = \beta_0 = 0$ and $\alpha_{01} = 0$, involves parameters in the multinormal logistic probability model. For the case, z is normal conditional on x, the estimates $\hat{\alpha}_{i1}$ will be biased by arguments similar to those in section 2. The exact asymptotic bias for the minimum chi-square estimate can be derived as in (2.18).

5. Conclusions

In this article, we have analyzed the omitted variable bias problem in the multinomial logistic probability model. We have investigated the omitted variable bias in the coefficient of the included variable as compared to the omitted variable bias in the linear regression model. In the logit model, bias will occur even if the omitted variable is independent with the included explanatory variable. The coefficient of the included variable plays an important role in the direction of the bias when the relevant variable is omitted. The bias may be downward even if the omitted variable is positively associated with the included variable and it has positive coefficient in the model. Sufficient as well as necessary conditions under which relevant variables can be omitted without affecting the coefficients of the included variables are provided. Conditional on the response variable, if the omitted explanatory variable and the included explanatory variable are independent, the bias will not occur. For some cases, the omitted variable bias can be assessed.

Our analysis is concentrated on the logit model. The conclusions in the logit model may not be valid for other probability models. A simple example is given in the appendix to show that the conditional independence condition is not sufficient for the probit model. Much effort is needed to analyze this problem in other probability models which are beyond the scope of this paper.

References

Bishop, Y. M. M, S. E. Fienberg and Paul W. Holland (1975), Discrete Multivariate Analysis: Theory and Practice, Cambridge: The MIT Press.

Nerlove, M. and S. J. Press (1976), "Multivariate Log-Linear Probability Models for the Analysis of Qualitative Data", Discussion Paper No. 1, Center for Statistics and Probability, Northwestern University.

Olsen, R. J. (1978), "Comment on 'The Effect of Unions on Earnings and Earnings on Unions: A Mixed Logit Approach'", International Economic Review 19, 259-261.

Appendix

Proposition 1 does not hold for the probit model.

Consider the following simple probit model,

$$P(y = 1|x, z) = \Phi(\alpha_0 + x\alpha_1 + z\beta) \tag{A.1}$$

and $$P(y = 0|x, z) = 1 - \Phi(\alpha_0 + x\alpha_1 + z\beta)$$

where Φ is the standard normal distribution function and the explanatory variables x and z are both dichotomous. The misspecified probit model which omits the explanatory variable z is

$$P^*(y = 1|x) = \Phi(\alpha_0 + x\alpha_1) \tag{A.2}$$

and $$P^*(y = 0|x) = 1 - \Phi(\alpha_0 + x\alpha_1).$$

Suppose that conditional on y, x and z are independent. It follows

$$\ell n \frac{P(z = 1|x, y)}{P(z = 0|x, y)} = \delta_0 + y\delta_1 \tag{A.3}$$

We will show that this conditional independence assumption is not a sufficient condition for the coefficient of x to be unaffected when the variable z is omitted.

Equation (A.1) can be rewritten in a logistic functional form,

$$\ell n \frac{P(y = 1|x, z)}{P(y = 0|x, z)} = \ell n \frac{\Phi(\alpha_0 + x\alpha_1 + z\beta)}{1 - \Phi(\alpha_0 + x\alpha_1 + z\beta)}$$

$$= \theta_0 + x\theta_1 + z\theta_2 + xz\theta_3 \tag{A.4}$$

where

$$\theta_0 = \ell n \frac{\Phi(\alpha_0)}{1 - \Phi(\alpha_0)} \tag{A.5}$$

$$\theta_1 = \ell n \frac{\Phi(\alpha_0+\alpha_1)}{1 - \Phi(\alpha_0+\alpha_1)} - \ell n \frac{\Phi(\alpha_0)}{1 - \Phi(\alpha_0)} \tag{A.6}$$

$$\theta_2 = \ell n \frac{\Phi(\alpha_0+\beta)}{1 - \Phi(\alpha_0+\beta)} - \ell n \frac{\Phi(\alpha_0)}{1 - \Phi(\alpha_0)} \tag{A.7}$$

$$\theta_3 = \ell n \frac{\Phi(\alpha_0+\alpha_1+\beta)}{1 - \Phi(\alpha_0+\alpha_1+\beta)} - \ell n \frac{\Phi(\alpha_0+\alpha_1)}{1 - \Phi(\alpha_0+\alpha_1)} - \ell n \frac{\Phi(\alpha_0+\beta)}{1 - \Phi(\alpha_0+\beta)} + \ell n \frac{\Phi(\alpha_0)}{1 - \Phi(\alpha_0)} \; . \tag{A.8}$$

Since $\frac{P(y=1|x,\ z=1)}{P(y=0|x,\ z=1)} \ \frac{P(z=1|x,\ y=0)}{P(z=0|x,\ y=0)} = \frac{P(y=1|x,\ z=0)}{P(y=0|x,\ z=0)} \ \frac{P(z=1|x,\ y=1)}{P(z=0|x,\ y=1)}$,

it implies that $\theta_2 = \delta_1$ and $\theta_3 = 0$. The relation $\theta_3 = 0$ derived under the conditional independence assumption in (A.3) imposes restrictions on the coefficients α_0, α_1 and β in the probit equation. Evidently, if $\alpha_0 = 0$ and $\alpha_1 = -\beta$, θ_3 is zero. Hence, equivalently, equation (A.4) is

$$\ell n \frac{P(y = 1|x,\ z)}{P(y = 0|x,\ z)} = \theta_0 + x\theta_1 + z\delta_1 \tag*{(A.4)'}$$

As shown in the text, equations (A.3) and (A.4)' imply

$$\ell n \frac{P(y = 1|x)}{P(y = 0|x)} = \theta_0 + x\theta_1 - \ell n \left(\frac{1+e^{\delta_0}}{1+e^{\delta_0+\delta_1}} \right)$$

$$= \theta_0^* + x\theta_1 \tag{A.9}$$

where $\theta_0^* = \theta_0 - \ell n \left(\frac{1+e^{\delta_0}}{1+e^{\delta_0+\delta_1}} \right)$. Equation (A.9) can be rewritten as

$$P(y = 1|x) = \Phi(\alpha_0^* + x\alpha_1^*) \tag{A.10}$$

where

$$\alpha_0^* = \Phi^{-1} \left(\frac{e^{\theta_0^*}}{1+e^{\theta_0^*}} \right) \tag{A.11}$$

$$\alpha_1^* = \Phi^{-1} \left(\frac{e^{\theta_0^*+\theta_1}}{1+e^{\theta_0^*+\theta_1}} \right) - \Phi^{-1} \left(\frac{e^{\theta_0^*}}{1+e^{\theta_0^*}} \right) \tag{A.12}$$

Estimating the misspecified probit model is equivalent to estimating the probit equation in (A.10). Let $\hat{\alpha}_1$ be the MLE or minimum chi-square estimate of α_1. It follows

$$\begin{aligned} \text{plim } \hat{\alpha}_1 &= \alpha_1^* \\ &= \Phi^{-1} \left(\frac{e^{\theta_0^*+\theta_1}}{1+e^{\theta_0^*+\theta_1}} \right) - \Phi^{-1} \left(\frac{e^{\theta_0^*}}{1+e^{\theta_0^*}} \right) \end{aligned}$$

From (A.5) and (A.6), it is known that

$$\alpha_1 = \Phi^{-1} \left(\frac{e^{\theta_0+\theta_1}}{1+e^{\theta_0+\theta_1}} \right) - \Phi^{-1} \left(\frac{e^{\theta_0}}{1+e^{\theta_0}} \right)$$

Hence plim $\hat{\alpha}_1 \neq \alpha_1$ as $\theta_0^* \neq \theta_0$.