UvA Econometrics

Discussion Paper: 2005/05

Omitted Variables and Misspecified Disturbances in the Logit Model

J.S. Cramer

www.fee.uva.nl/ke/UvA-Econometrics

Amsterdam School of Economics
Department of Quantitative Economics
Roetersstraat 11
1018 WB AMSTERDAM
The Netherlands

UvA Universiteit van Amsterdam

Omitted Variables and Misspecified Disturbances in the Logit Model

J.S. Cramer *

September 2005

Abstract

In binary discrete regression models like logit or probit the omission of a relevant regressor (even if it is orthogonal) depresses the remaining β coefficients towards zero. For the probit model, Wooldridge (2002) has shown that this bias does not carry over to the effect of the regressor on the outcome. We find by simulations that this also holds for logit models, even when the omitted variable leads to severe misspecification of the disturbance. More simulations show that estimates of these effects by logit analysis are also impervious to pure misspecification of the disturbance.

*University of Amsterdam and Tinbergen Institute, Amsterdam; e-mail address: cramer@tinbergen.nl

1 Introduction and summary

In a classic regression equation the estimated $\hat{\beta}$ is little affected by omitted variables: provided these are orthogonal to the remaining regressors, the estimates are still consistent and unbiased, and the only inconvenience is an increase of the residual variance and hence of the estimated standard deviations of $\hat{\beta}$. No such comforting theorem exists for the $\hat{\beta}$ of discrete models, though not for want of trying.

In the original field of probit and logit, the bio-assay of insecticides and other stimuli under controlled conditions, the issue of omitted variables hardly arose. But it is quite relevant for the analysis of survey data in epidemiology and in the social sciences, in marketing and in finance, for here the set of explanatory variables is seldom complete and unobserved heterogeneity is the rule. Yet the literature consists of a few isolated articles. In an early paper, Amemiya and Nold (1975) allow for omitted variables by an extra disturbance for the logit transform of grouped data, and then employ a variant of Berkson's minimum chi-squares estimation. Lee (1982), who considers a multinomial model in the framework of categorical or grouped data, argues that the other coefficients are not affected if the omitted and retained regressors are independent *conditional on the outcome*; but he remarks in an aside that in the binary logit the absence of an orthogonal variable will bias the remaining coefficients towards zero (p.208). Ruud (1983) also deals with multinomial models in an analysis of misspecification of the disturbance distribution, which is closely related to the omitted variable issue. He comes up with conditions that preserve the consistency of $\hat{\beta}$ *up to a scaling factor*, but admits that this result "rests on an assumption ... that is too restrictive to be generally applicable" (p.228). Yatchew and Griliches (1985) are the first to give a straightforward derivation of the downward omitted variable bias of $\hat{\beta}$ in the binary model, though they then go on to entirely different matters. Gourieroux (2000)[1] very briefly considers a condition whereby omitted variables would not affect $\hat{\beta}$ but adds that it is "of no practical use" (p.33).

The only lasting contribution of these studies is the demonstration by Yatchew and Griliches, which is probably at the bottom of statements by practitioners like Baltas and Doyle (2001) who write "An interesting property [of discrete models] is the effect of increasing unexplained stochastic variation on the identified coefficients... As unobserved variation decreases, the value of the identified price parameters increases and vice versa" (p.116, second column). But lately the argument has been carried a step further by Wooldridge (2002), who has shown that while $\hat{\beta}$ is affected by omitted variables, the partial effect of the remaining regressors on the outcome is not.

[1] Published earlier, in French, in 1991.

The effect of omitted variables has also been treated from an altogether different angle in the biomedical literature under the heading of adding further covariates in a simple case-control study. Robinson and Jewell (1991) argue that this will lead to *less precise* estimates of the main effect; Gail et al (1984) stress the inconsistency of the method of estimation, due to the change in the disturbance distribution, which is a particular feature of non-linear models where this distribution determines the form of the relation. The argument has been taken up again by Ford et al (1995).

Below, we shall first retrace the arguments of Yatchew and Griliches and of Wooldridge. We then report some simulations which extend Wooldridge's result to the logit model. This naturally raises the issue of misspecification of the disturbances. We find that this, too, is of little importance for the usual logit analyses.

2 The latent variable regression equation

We derive the logit (or probit) model from the familiar latent variable regression equation

$$Y_i^* = \boldsymbol{x}_i^T \boldsymbol{\beta}^* + \varepsilon_i^* \tag{1}$$

with the standard properties: the regressor vector $\boldsymbol{x}_i$ (which always includes a unit variable X_0) represents known constants, ε_i^* is a random disturbance that is uncorrelated with the regressors, and $\boldsymbol{\beta}^*$ is a vector of unknown parameters. If the Y_i^* are observed, this is an ordinary regression equation, and $\boldsymbol{\beta}^*$ can be estimated by Ordinary Least Squares. In discrete models, Y_i^* is not observed but constitutes a *latent* variable; its sign determines the $(0, 1)$ indicator variable Y_i that *is* observed, as in

$$\begin{array}{rcl} Y_i & = & 1 \text{ iff } Y_i^* > 0, \\ Y_i & = & 0 \text{ otherwise.} \end{array} \tag{2}$$

For some symmetrical distribution function F_ε of ε this gives

$$P(Y_i = 1) = F_\varepsilon(\boldsymbol{x}_i^T \boldsymbol{\beta}^*).$$

Both in ordinary regression and in the discrete model identification of the parameters requires further assumptions about the disturbances. In both models, their *mean* must be specified, or the constant β_0^* is not identified; it is invariably set at zero. In the discrete model, the variance of the disturbances σ^{*2} must be specified, too, since the inequality (2) is invariant to scaling of Y_i^*, and hence to scaling of ε_i^* and of $\boldsymbol{\beta}^*$, so that neither σ^* nor $\boldsymbol{\beta}^*$ are

identified:This indeterminacy is resolved by imposing a set value C on σ^*. Both sides of (1) are multiplied by C/σ^*, and it is replaced by

$$Y_i^+ = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{3}$$

with

$$Y_i^+ = Y_i^* \frac{C}{\sigma^*}, \quad \boldsymbol{\beta} = \boldsymbol{\beta}^* \frac{C}{\sigma^*}, \quad \varepsilon_i = \varepsilon_i^* \frac{C}{\sigma^*} \tag{4}$$

and

$$var(\varepsilon_i) = C^2.$$

The observed Y_i are now defined by

$$\begin{aligned} Y_i &= 1 \text{ iff } Y_i^+ > 0, \\ Y_i &= 0 \text{ otherwise.} \end{aligned}$$

In the probit model ε_i has a standard normal distribution and C equals 1; in the logit model ε_i has a logistic distribution and C equals $\lambda = \pi/\sqrt{(3)} \approx 1.8138$[2]. In either case the *normalized* parameters $\boldsymbol{\beta}$ that are estimated may be regarded as derived or reduced form coefficients with respect to the original $\boldsymbol{\beta}^*$, and they vary inversely with σ^*.

On the whole, these identifying restrictions are a matter of convenience, not of conviction. Thus, it is seldom argued that the zero mean of ε is a 'natural' value[3], or that there are grounds for the normal distribution of ε of the probit model. Nor do I know of a rational justification of the logistic distribution. But insofar as they are arbitrary, identifying restrictions of this sort should not materially affect the results of statistical analysis.

3 Omitting a variable: the effect on $\boldsymbol{\beta}$

The effect of omitting a relevant determinant from the analysis can be traced by examining the removal of X_2 from an equation with two independent regressors

$$Y_i^* = \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2i} + \varepsilon_i^*. \tag{5}$$

We shall call this the *full* equation. It satisfies all the standard requirements: in particular ε_i^* has zero mean, variance σ^{*2}, and is uncorrelated with both regressors. In the *curtailed* equation X_2 is omitted, and its contribution to Y_i^* relegated to the disturbance term, as in

$$Y_i^* = (\beta_0^* + \beta_2^* \bar{X}_2) + \beta_1^* X_{1i} + \varepsilon_i^\circ \tag{6}$$

[2]The difference between these values accounts for the difference of logit and probit coefficients from the same data.

[3]For a counterexample, see Greene (1990, p.147), not repeated in later editions of the book.

with

$$\varepsilon_i^\circ = \varepsilon_i^* + \beta_2^*(X_{2i} - \bar{X}_2). \tag{7}$$

This has all the required properties too. Upon comparing (5) and (6), and ignoring the intercept (by common usage), we find that the coefficient of X_1 is the same, but that the disturbance variance has increased from σ^{*2} to

$$\sigma^{\circ 2} = \sigma^{*2} + \beta_2^{*2} var(X_2). \tag{8}$$

This depresses the slope coefficients of a discrete model towards zero, for instead of (4) we now have

$$\beta^o = \beta^* \frac{C}{\sigma^o}. \tag{9}$$

In the present case, (4) and (8) give

$$\frac{\beta_1^o}{\beta_1} = \lambda = \frac{\sigma^*}{\sigma^o} = \frac{1}{\sqrt{1 + \beta_2^2 var(X_2)/C^2}} \tag{10}$$

or

$$\beta_1^\circ = \lambda \beta_1.$$

with $\lambda < 1$.

λ was called the *rescaling factor* by Yatchew and Griliches (1985), who first put forward the above argument, and the *attenuation bias* by Wooldridge (2002). The argument is easily generalized to more than two regressors. It can be empirically verified by deleting successive regressors from a large set (provided they are more or less orthogonal); at each stage the full equation provides an estimate of β_2 and hence of λ. I have done so elsewhere (see Cramer (2003), section 5.5).

The addition of $\beta_2^2 var(X_2)$ to ε^* will also affect the distribution of the disturbance. For logit models, ε^* of the full equation is assumed to have a logistic distribution, and X_2 must by (7) have a very special sample distribution indeed for ε^o of the curtailed equation to have a logistic distribution, too. In practice, at least one of the two models is misspecified, and this may lead to further systematic changes in the estimated coefficients. A similar argument applies for probit models. Here, X_2 must be normal for both equations to have normal disturbances. Most people feel more comfortable with this, but it is of course equally restrictive.

Even with orthogonal regressors, then, omitted variables depress $\hat{\boldsymbol{\beta}}$ towards zero, relatively to its value in the full equation. In other words, the $\hat{\boldsymbol{\beta}}$ of discrete models vary inversely with the extent of unobserved heterogeneity. The practical consequence is that estimates from samples that differ in this respect are not directly comparable.

4 The effect on derivatives

In an ordinary regression equation, $\boldsymbol{\beta}$ represents the derivatives of Y in respect of the regressors, and hence their *effect* on the outcome. But in a discrete model, this is not so. The derivative of $P(Y_i = 1)$ with respect to some X_k, evaluated at the regressor vector $\boldsymbol{x}^\circ$, is

$$\phi(\boldsymbol{x}^{\circ T}\boldsymbol{\beta})\beta_k$$

for the probit model, with ϕ the normal density, and

$$P^\circ(1 - P^\circ)\beta_k \tag{11}$$

with $P^\circ = P(\boldsymbol{x}^{\circ T}\boldsymbol{\beta}$ for the logit. In these derivatives, the downward movement of $\boldsymbol{\beta}$ may be compensated by inverse changes in the other terms, and Wooldridge (2002, section 15.7.1) has shown that for the logit model and a normal distribution of X_2 this is indeed the case.

Wooldridge considers the *average partial effect* or APE of X_k on P at a given point $\boldsymbol{x}^\circ$. In the present case of two regressors the partial effect of X_1 is the derivative $\delta P/\delta X_1$ at X_1°, X_2°. If the X_2 are unknown (as in the curtailed equation) we take the average or expected value of the derivative over the distribution of X_2, and this is the APE. For a probit model, this gives (in our notation)

$$APE = \mathcal{E}_{X_2}\, \beta_1\, \phi(X_1^\circ\beta_1 + X_2\beta_2)$$

and if X_2 follows a normal distribution it turns out that this is equal to

$$\lambda\, \beta_1\, \phi(X_1^\circ\lambda\beta_1)$$

i.e to the partial derivative given by the curtailed equation. Estimates of the partial derivative from this equation are therefore not subject to attenuation bias.

A similar argument applies to the derivative of the logit model of (11), for with $\boldsymbol{\beta}$ moving towards zero, P goes towards .5 and $P(1-P)$ towards its maximum value of .25. To find out how this works out in practice we have performed a number of simulations. These bear on a simple two-variable regression equation like (3) with $\beta_0 = 0, \beta_1 = \beta_2 = 1$, or

$$Y_i^+ = X_{1i} + X_{2i} + \varepsilon_i. \tag{12}$$

Both X_i and X_2 are independent normal variates with mean zero and variance equal to 3.29, and ε is a logistic variate, also with mean zero and variance 3.29. The three components thus contribute equally to the variation of Y_i^+;

in the full equation the systematic component is two-thirds of the total, and in the curtailed equation it is one third. By (10) the rescaling factor is .70.

We generate a sample of 3000 observations of the three right-hand variables, and set

$$\begin{aligned} Y_i &= 1 \text{ iff } Y_i^+ > 0, \\ Y_i &= 0 \text{ otherwise,} \end{aligned}$$

as before. By the values that have been adopted the sample frequency of $Y_i = 1$ will be close to .5. The β of (12) - with true β $(0, 1, 1)$ - are estimated in the usual Maximum Likelihood manner, and this is repeated for the curtailed equation with X_1 alone.

In addition to the estimate of β_1 we also calculate the mean of the derivatives (11) over all observations of the original sample. This is the *average sample effect* or ASE

$$ASE = \frac{1}{n} \sum \hat{P}_i(1 - \hat{P}_i)\beta_1.$$

It is a sample mean, not an expectation, and it does not refer to a single fixed X_1°, but otherwise it is quite similar to Wooldridge's APE. It is the partial derivative of the sample aggregate frequency with respect to X_1.

An example of the result of this exercise reads as follows.

	full equation	curtailed equation	ratio
$\hat{\beta}_1$	.96	.67	.69
$s.d.$	(.04)	(.03)	
ASE	.12	.13	1.05

Upon the removal of X_2, $\hat{\beta}_1$ declines a little more than the rescaling factor of .70, but ASE is not so affected. This can be traced to a general movement of the sample $\hat{P}_i$ towards .5. By the way the simulation has been set up, .5 is the mean of the $\hat{P}_i$; upon the removal of X_2, the $\hat{P}_i$ move towards this value. Their dispersion declines from a standard deviation of .35 for the full equation to only .17 for the curtailed equation.

Table 1. Mean and standard deviation in 100 replications, normal distribution of X_2.

distribution of X_2	full equation	curtailed equation	ratio
$\hat{\beta}_1$:			
mean	1.000	.664	.665
s. d.	.044	.028	.024
ASE			
mean	.129	.129	1.001
s. d.	.003	.004	.022

In Table 1 we report the results for 100 replications of this simulation with a normal distribution of X_2. The table gives the mean and standard deviation of the estimates over the 100 replications, not standard deviations of estimates as reported by the estimation programme. These replications confirm the findings of the above example. The principal result is that β_1 is biased but the derivative ASE is not. Note further that the estimates from the curtailed equation do not have a greater dispersion than those from the full equation, as one would expect. Finally, the reduction in $\hat{\beta}_1$ is definitely larger than the rescaling factor of .70. As the standard deviation among the 100 replications is .0244, the standard deviation of the mean of .664 is .0024, and the difference of $.700 - .664 = .036$ is significant. We attribute this further reduction by a factor $.664/.700 = .95$ to the misspecification of the disturbance in the curtailed equation.

In order to find out whether this is indeed so we have experimented with three distributions for X_2 other than the normal, viz. a logistic distribution, a binary $0, 1$ dummy and a t distribution. In all cases X_2 is scaled to have zero mean and variance 3.29, so that the rescaling factor is always .70. The binary dummy, for example, takes the values $-1.81, +1.81$, each for half of the observations. Since all four distributions are moreover symmetrical, they have the first three moments in common at zero, 3.29, and zero, and differences arise only in their fourth moment or *kurtosis*. Table 2 reports the outcome of these simulations.

Table 2. Mean and standard deviation in 100 replications
for various distributions of X_2.

distribution of X_2	full equation	curtailed equation	ratio
logistic:			
$\hat{\beta}_1$	1.001	.680	.679
	.041	.031	.021
ASE	.131	.131	.999
	.003	.004	.025
binary dummy:			
$\hat{\beta}_1$	1.002	.606	.605
	.044	.026	.023
ASE	.121	.121	1.000
	.003	.003	.028
$t(6)$:			
$\hat{\beta}_1$	1.006	.687	.683
	.038	.036	.022
ASE	.132	.131	0.999
	.003	.004	.022

The overall result of these exercises is that the $\hat{\beta}_1$ are reduced, both by the rescaling factor and by the additional misspecification effect, but that the ASE are not; moreover the estimates of ASE do not vary very much, whatever the distribution of X_2.

The misspecification effect varies with the distribution of X_2 and we have seen that these differ in their kurtosis. The original disturbances of the full equation have a logistic distribution, with rather fat tails and a kurtosis of 1.2, as opposed to 0 for the normal distribution of X_2 of Table 1. This gave a misspecification effect of .95. The first alternative (in the top panel of Table 2) is to give X_2 a logistic distribution, too, with the same kurtosis as the disturbance (but note that the sum of two logistic variates does *not* have a logistic distribution). This reduces the misspecification effect somewhat to $.679/.700 = .97$. A much more extreme case of slim tails arises if we make X_2 a binary dummy: this has kurtosis -2, and the downward misspecification bias increases to $.606/.700 = .87$. If a lower kurtosis thus means a larger

reduction, one might expect an effect in the opposite direction for a distribution with a higher kurtosis than the logistic. To this end we have employed a t distribution with a small number of degrees of freedom[4]. With 6 degrees of freedom, the kurtosis is 3, well in excess of the 1.2 of the logistic distribution. But this does not produce an upward bias bias: the misspecification effect remains at $.687/, 700 = .98$.

5 Pure misspecification of the disturbance

So far we have added various X_2 to the correctly specified logistic disturbances of the initial full equation, and found that this hardly affects the substantive results of the analysis. A similar conclusion arose in an altogether different context, when bank loans were screened for the possibility of default. From a statistical viewpoint it was quite clear that the data did not support the logit model, yet its performance in classifying new bank loans was hardly inferior to that of a more appropriate model (Cramer (2004)). Both results suggest that the substantive results of a logit analysis are insensitive to misspecification of the distribution of the disturbances.

Table 3. Mean and standard deviation in 100 replications
for various distributions of the disturbance of the full eauwtion.

distribution of disturbance	$\hat{\beta}_1$	s.e.$\hat{\beta}_1$	ASE_1
logistic	1.005	.042	.128
	.046	.002	.004
normal	.956	.040	.126
	.038	.001	.004
binary dummy	.835	.036	.121
	.032	.001	.004
$t(6)$	1.028	.042	.130
	.041	.001	.003

This robustness is confirmed by simple simulations of the full equation (12) for different disturbance distributions. We recall that this is a simple latent variable equation with two independent normal regressors with $\boldsymbol{\beta}$ given as $1, 1$. Both regressors and the disturbances have all been scaled to have

[4]I owe his suggestion to Casper de Vries.

zero mean and the same variance 3.29, so that the systematic part accounts for two-thirds of the variation of Y_i^+. Table 3 shows the results for $\hat{\beta}_1$, its standard error and the corresponding ASE; the results for X_2 are of course almost identical.

The first panel of Table 3 refers to the correct specification and the second to a normal distribution that is not so very much different; the normal induces a small (but significant) reduction of $\hat{\beta}_1$. The binary dummy is an extreme case and leads to a substantial decline of $\hat{\beta}_1$. As before, the other extreme of the $t(6)$ distribution gives no misspecification bias. The effects on the standard errors of the estimate (derived from a misspecified model) are quite modest. But the salient fact is that ASE is hardly affected; only in the extreme case of a binary dummy disturbance it is somewhat reduced. The overall conclusion is that the substantive results of logit analyses are quite robust.

6 References

Amemiya, Takeshi, and Frederick Nold (1975) A modified logit model. *The Review of Economics and statistics*, **57**, 255-257.

Baltas, George, and Peter Doyle (2001) Random utility models in market research: a survey. *Journal of Business Research*, **51**, 115-125.

Cramer, J.S. (2003) *Logit Models From Economics and Other Fields.* Cambridge: Cambridge University Press.

Cramer, J.S. (2004) Scoring bankl loans that may go wrong: a case study. *Statistica Neerlandica*, **58**, 365-380.

Ford, Ian, John Norris and Susan Ahmadi (1995) Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine*, **14**, 735-746.

Gail, M.H., S. Wienand and S. Piantadosi (1984) Diased estimates of treatment effects in randomized experiments with nonlinear regressions and omitted variables. *Biometrika* **71**, 431-444.

Gourieroux, Christian (2000) *Econometrics of Qualitative Dependent Variables.* Cambridge: Cambridge University Press.

Greene, William H. (1990) *Econometric Analysis.* Englewood Cliffs: Prentice Hall.

Lee, Lung-Fei (1982) Specification error in multinomial logit models. *Journal of Econometrics*, **20**, 197-209.

Manski, Charles F. (1977) The structure of random utility models. *Theory and Decision*, **8**, 229-254.

Robinson, Laurence D., and Nicholas P. Jewell (1991) Some surprising results about covariate adjustment in logistic regression. *International Statistical Review*, **58**, 227-240.

Ruud, Paul A. (1983) Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete models. *Econometrica*, **51**, 225-228.

Wooldridge, Jeffrey M. (2002) *Econometric Analysis of Cross Section and Panel Data.* Cambridge. Mass: MIT Press.

Yatchew, A., and Z. Griliches (1985) Specification error in probit models. *The Review of Economics and Statistics*, **67**, 134-139.