UvA Econometrics

Discussion Paper: 2006/07

# Robustness of Logit Analysis: Unobserved Heterogeneity and Misspecified Disturbances

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

# Robustness of Logit Analysis: Unobserved Heterogeneity and Misspecified Disturbances[1]

J. S. CRAMER

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

## Abstract

In probit and logit models, the $\beta$ coefficients vary inversely with the variance of the disturbances. The omission of a relevant orthogonal regressor leads to increased unobserved heterogeneity, and this depresses the $\beta$ coefficients of the remaining regressors towards zero. For the probit model, Wooldridge (2002) has shown that this bias does not carry over to the *effect* of these regressors on the outcome. We find by simulations that this also holds for the logit model, even when omitting a variable leads to severe misspecification of the disturbance. More simulations show that logit analysis is quite insensitive to pure misspecification of the disturbance as such.[2] [3]

## 1 Introduction

The observed variation of an outcome $Y$ may be related to any number of covariates $X_j$. As a rule, some but not all of these determinants are explicitly included in a statistical analysis, and the effect of the remainder is relegated to a disturbance term which is treated as a random variable. With cross-section data from a sample survey, the variation of the covariates is determined by the sample design, and their observation by the manner of data collection. Together, these determine the character of the disturbances.

In classic linear regression it is an important issue whether the missing covariates are correlated with the included explanatory variables or not. If they are, the effect of included variables is confounded with that of the omitted variables, and the estimates of the linear coefficients are biased as a result. But in the special case that the omitted variables are orthogonal to the in-

[1]I have benefited from the comments of Jan Kiviet, Bill Greene and one of the editors on an earlier draft of this paper.

[2]JEL Classification numbers: C13 C15 C25

[3]*Keywords: discrete choice, unobserved hetereogeneity, misspecified disturbances*

cluded variables, no bias ensues, and their absence leads only to an increase in the variance of the disturbance term which is of no particular concern.

No such comforting theorem exists for nonlinear models. In the study of durations, for example, it is fully recognized that the variance of the disturbances (here usually denoted as unobserved heterogeneity) does affect the model coefficients. For models of discrete choice, several authors like Lee (1982, p.208), Ruud (1983, p.228) and Gourieroux (2000, p.33) have tried to establish conditions that would render orthogonal omitted variables equally harmless as in linear regression. But these attempts have been to no avail. Indeed, Yatchew and Griliches (1985) (while mainly concerned with other matters) demonstrated for the binary discrete choice model that omitting orthogonal relevant variables does bias $\beta$ towards zero, and this is generally accepted by practitioners in the field. Recently, the argument has been carried a step further by Wooldridge (2002), who has shown for the probit model that while $\beta$ is affected by omitting orthogonal variables, the partial effect of the regressors on the outcome is not.

Below, we shall first retrace the arguments of Yatchew and Griliches and of Wooldridge. We then report simulations which extend Wooldridge's result to the logit model. This naturally raises the issue of misspecification of the disturbances. We find that this is of little importance in the present context, and also that logistic regression is insensitive to misspecification of the disturbances as such.

## 2 The latent variable regression equation

We derive the logit (or probit) model from a 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, $\varepsilon_i^*$ is a random disturbance that is uncorrelated with the regressors, and $\boldsymbol{\beta}^*$ is a vector of unknown parameters. In an ordinary regression equation, the $Y_i^*$ are observed, and $\boldsymbol{\beta}^*$ can be estimated by Ordinary Least Squares. In a discrete model, the $Y_i^*$ are not observed but constitute *latent* variables, and their 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 a symmetrical distribution function $F_\varepsilon$ of $\varepsilon$ this gives

$$P(Y_i = 1) = F_\varepsilon(\boldsymbol{x}_i^T \boldsymbol{\beta}^*). \tag{3}$$

Both in ordinary regression and in the discrete model identification of the parameters calls for further assumptions about the disturbances $\varepsilon_i^*$. In both models, identification of the constant $\beta_0^*$ requires that their *mean* is specified; it is invariably set at zero. In the discrete model, their variance $\sigma^{*2}$ must be specified, too, since the inequality (2) is invariant to scaling of $Y_i^*$, and hence to scaling of $\varepsilon_i^*$ and of $\boldsymbol{\beta}^*$, so that neither $\sigma^*$ nor $\boldsymbol{\beta}^*$ are identified. This indeterminacy is resolved by imposing a set value $C$ on $\sigma^*$. Both sides of (1) are then multiplied by $C/\sigma^*$, and it is replaced by

$$Y_i^+ = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{4}$$

with

$$Y_i^+ = Y_i^* \frac{C}{\sigma^*}, \quad \boldsymbol{\beta} = \boldsymbol{\beta}^* \frac{C}{\sigma^*}, \quad \varepsilon_i = \varepsilon_i^* \frac{C}{\sigma^*} \tag{5}$$

and

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

The observed $Y_i$ are now defined by the equivalent of (2)

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

In the probit model $\varepsilon_i$ has a standard normal distribution and $C$ equals 1; in the logit model $\varepsilon_i$ has a logistic distribution and $C$ equals $\lambda = \pi/\sqrt{(3)} \approx 1.8138$[4]. 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 by (5) they then vary inversely with $\sigma^*$.

These identifying restrictions are a matter of convenience, not of conviction. It is seldom argued that the zero mean of $\varepsilon$ is a 'natural' value[5], or that there are grounds for the standard normal distribution of $\varepsilon$ of the probit model. Nor do I know of a rational justification of the logistic distribution. Identifying restrictions of this sort are intrinsically arbitrary, and they should not materially affect the results of statistical analysis.

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

We trace the effect of omitting a relevant determinant from the analysis by examining the removal of $X_2$ from an equation with two independent regressors $X_1$ and $X_2$. The orthogonality of the regressors makes this a special

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

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

case, which brings out the difference between discrete models and classic linear regression. It is of course much more tractable than the general case of correlated regressors; if needs be, the latter can be treated along the same lines as in linear regression, but while this would complicate the argument it would not alter its drift. We may add that the case of uncorrelated regressors is not quite as unrealistic as it may seem: in cross-section microdata, covariates are often only weakly correlated. Besides, the assumption that disturbances are independent of the regressors is seldom contested, while they are generally understood to reflect *inter alia* the effect of neglected covariates.

We start from the *full* equation with two orthognal regressors

$$Y_i^* = \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2i} + \varepsilon_i^*. \tag{6}$$

This satisfies all the standard requirements listed before: in particular $\varepsilon_i^*$ has zero mean, variance $\sigma^{*2}$, 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

$$Y_i^* = (\beta_0^* + \beta_2^* \bar{X}_2) + \beta_1^* X_{1i} + \varepsilon_i^\bullet \tag{7}$$

with

$$\varepsilon_i^\bullet = \varepsilon_i^* + \beta_2^* (X_{2i} - \bar{X}_2). \tag{8}$$

Since $X_2$ is orthogonal to $X_1$ this has all the required properties too. Upon comparing (6) and (7), 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 $\sigma^{*2}$ to

$$\sigma^{\bullet 2} = \sigma^{*2} + \beta_2^{*2} var(X_2). \tag{9}$$

This depresses the slope coefficients of the discrete model towards zero, for instead of (5) we now have

$$\boldsymbol{\beta}^\bullet = \boldsymbol{\beta}^* \frac{C}{\sigma^\bullet}. \tag{10}$$

For the coefficient of the remaining regressor $X_1$ of (7), the equations (5), (10) and (9) give

$$\frac{\beta_1^\bullet}{\beta_1} = \frac{\sigma^*}{\sigma^\bullet} = \lambda = \frac{1}{\sqrt{1 + \beta_2^2 var(X_2)/C^2}} \tag{11}$$

or

$$\beta_1^\bullet = \lambda \beta_1. \tag{12}$$

with $\lambda < 1$.

This factor $\lambda$ 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 $\beta_2$ and hence, by (11), of $\lambda$. For an example see Cramer (2003, section 5.5).

The addition of $\beta_2^2 var(X_2)$ to $\varepsilon^*$ in (8) also affects the distribution of the disturbance. In the logit model, $\varepsilon^*$ of the full equation is assigned a logistic distribution; by (8) $X_2$ must have a very special sample distribution indeed for $\varepsilon^o$ 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 to 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.

Leaving the misspecification issue aside, we conclude that even with orthogonal regressors omitted variables depress the remaining coefficients towards zero, relatively to their value in the full equation. In other words, the $\hat{\boldsymbol{\beta}}$ of discrete models vary inversely with the variance of the disturbance, or with the extent of unobserved heterogeneity. This is somewhat disturbing, even though it does not affect all practical conclusions from these models: the relative sizes of the coefficients are preserved, and so is the ranking of cases by estimated probabilities used in the selection of likely propects in marketing and finance. But the $\hat{\boldsymbol{\beta}}$ are biased, and estimates from analyses that differ in the size of the disturbance variance are not directly comparable.

## 4 The effect on derivatives

In an ordinary regression equation, $\boldsymbol{\beta}$ represents the partial 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 $\boldsymbol{x}^\circ$, follows from (3) as

$$\frac{\delta P(Y_i = 1)}{\delta X_k} = f_\varepsilon(\boldsymbol{x}_i^T \boldsymbol{\beta}^*).$$

In a probit model, $F$ is the standard normal model, and this derivative is

$$\phi(\boldsymbol{x}^{\circ T} \boldsymbol{\beta}) \beta_k \tag{13}$$

with $\phi$ the normal density; for the logit model, it is

$$P^\circ (1 - P^\circ) \beta_k \tag{14}$$

with $P^\circ = P(\boldsymbol{x}^{\circ T}\boldsymbol{\beta})$ the logit probability. In these derivatives, the downward movement of $\beta_k$ may be compensated by inverse changes in the other terms. Wooldridge (2002, section 15.7.1) has shown for the probit model and a normal distribution of $X_2$ that 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 an equation with two regressors, the partial effect of $X_1$ is the derivative (13) at $(X_1^\circ, X_2^\circ)$. 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 the probit model, this gives (in present notation)

$$APE_p = \mathcal{E}_{X_2}\ [\beta_1\ \phi(X_1^\circ\beta_1 + X_2\beta_2)]\,.$$

If $X_2$ follows a normal distribution with zero mean and variance $\tau^2$, this expectation is

$$APE_p = \int_{X_2} \frac{1}{\sqrt{2\pi}}\, exp{-}1/2(\beta_1 X_1 + \beta_2 X_2)^2\, \frac{1}{\tau\sqrt{2\pi}}\, exp{-}1/2\left(\frac{\beta_2 X_2}{\tau}\right)^2 dX_2.$$

This is analytically tractable; the exponents can be rearranged, the square completed, and the expectation established. Making use of

$$\lambda = \frac{1}{\sqrt{1+\beta_2^2\tau^2}}$$

from (11), it is found that

$$APE_p = \lambda\ \beta_1\ \phi(X_1^\circ\lambda\beta_1)$$

or, by (12),

$$APE_p = \beta_1^\bullet\ \phi(X_i^\circ\beta_i^\bullet)$$

i.e. $APE$ is equal 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 intuitive argument applies to the derivative of the logit model of (14), for as $\boldsymbol{\beta}$ moves towards zero, $P$ goes towards .5 and $P(1-P)$ towards its maximum value of .25, and this will counteract the downward bias of $\boldsymbol{\beta}$. If we write the average partial effect at a given $X_1^\circ$ for the logit, we have

$$APE_l = \mathcal{E}_{X_2}\ \beta_1\ P^\circ(1-P^\circ),$$

where $P^\circ$ is the logit probability with argument $\beta_1 X_1^\circ + \beta_2 X_2$. If this is written in full as an integral, with a normal distribution of $X_2$, as before, it

does not seem to offer the same scope for an analytical solution as the logit did. It cannot be excluded that there is some other smart distribution of $X_2$ than the normal which does permit an elegant solution, but this is as yet unknown. Here a number of simulations have been performed in order to find out how the logit case works out in practice.

These simulations bear on a simple two-variable regression equation like (6) with $\beta_0^* = 0, \beta_1^* = \beta_2^* = 1$, or

$$Y_i^+ = X_{1i} + X_{2i} + \varepsilon_i^*. \tag{15}$$

The disturbance $\varepsilon_i^*$ is a logistic variate with mean zero and standard deviation 1.8138 or variance 3.29, so that no further normalization is necessary and $\boldsymbol{\beta} = \boldsymbol{\beta^*}$. The two regressors $X_i$ and $X_2$ are independent normal variates with mean zero and variance also equal to 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 (11) the rescaling factor is .70. Apart from the scaling of regressors and coeffients, which is matter of convenience, this design is an analogue of the probit case treated by Wooldridge, and it is also reasonably similar to the sample data of marketing research and finance.

We generate a sample of $3,000$ observations of the three right-hand variables of (15), and set

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

as in (12). By the values that have been adopted the sample frequency of $Y_i = 1$ will be close to .5. The $\beta$ of (15) - with true $\boldsymbol{\beta}$ $(0, 1, 1)$ - are estimated in the usual Maximum Likelihood manner, and this is repeated for the curtailed equation with $X_1$ alone. In addition to the estimates of $\beta_1$ we also calculate the means of the derivative (14) over all observations of the original sample, both for the full and for the curtailed equation. This is the *average sample effect* or $ASE$

$$ASE = \frac{1}{n} \sum \hat{P}_i (1 - \hat{P}_i) \hat{\beta}_1.$$

It is a sample mean, not an expectation, and it does not refer to a single fixed $X_1^\circ$; but otherwise it is quite similar to Wooldridge's $APE$. It is the partial derivative of the expected sample frequency with respect to a shift in all $X_{1i}$.

We illustrate the results of these calculations by the follIowing example.

| | full equation | curtailed equation | ratio |
|---|---|---|---|
| $\hat{\beta}_1$ | .96 | .67 | .69 |
| $s.d.$ | (.04) | (.03) | |
| $ASE_1$ | .12 | .13 | 1.05 |

Upon the removal of $X_2$, $\hat{\beta}_1$ declines in line with the rescaling factor of .70, but $ASE_1$ is hardly affected. This is the result of changes in $\hat{P}_i$. As $\hat{\beta}_2$ is set at zero and $\hat{\beta}_1$ is reduced, the sample $\hat{P}_i$ move towards .5. In the present case, this is their mean, so that their dispersion is reduced; their standard deviation declines from .35 for the full equation to .17 for the curtailed equation. As a result, the $\hat{P}_i(1-\hat{P}_i)$ increase, and this compensates $ASE$ for the reduction of $\hat{\beta}_1$. At the same time,the values of $\hat{P}_i$ for $Y_i = 1$ and of $1-\hat{P}_i$ for $Y_i = 0$ are reduced: the loglikelihood declines from $-1198.58$ for the full equation to $-1747.86$ for the curtailed equation. - The interesting point is however not so much how these mechanisms work, but that they compensate $ASE_1$ so well.

In Table 1 we report the mean and standard deviation of the estimates for 100 replications of this simulation. These confirm the main result: while $\beta_1$ is biased by the rescaling factor, in most cases the derivative $ASE_1$ is not. Subsidiary findings are that the estimates from the curtailed equation do not have a greater dispersion than those from the full equation, as one might expect. And then the reduction in $\hat{\beta}_1$ is somewhat larger than the rescaling factor of .70. This difference of $-.036$ is slight but significant. We attribute the further reduction of the $\hat{\beta}_1$ by a factor $.664/.700 = .95$ to the misspecification of the disturbance in the curtailed equation, where a normal variate has been added to the original logistic term.

TABLE 1
*Mean and standard deviation of estimates in* 100 *replications, normal distribution of* $X_2$

| | *Full equation* | *Curtailed equation* | *Ratio* |
|---|---|---|---|
| $\hat{\beta}_1$: | | | |
| mean | 1.000 | .664 | .665 |
| s. d. | .044 | .028 | .024 |
| $ASE_1$: | | | |
| mean | .129 | .129 | 1.001 |
| s. d. | .003 | .004 | .022 |

This issue is further explored in Table 2, which reports experiments with various other distributions of $X_2$. In all cases $X_2$ is scaled to have zero mean and variance 3.29, so that the rescaling factor is always .70; the .50/.50 binary dummy, for example, takes the values $-1.81, +1.81$, each for half of the observations. The distributions differ in their kurtosis, and sometimes in their skewness.

The original disturbances of the full equation have a logistic distribution, which is symmetrical, with rather fat tails: the kurtosis is 1.2. First we consider four symmetrical distributions. The normal distribution of Table 1 has kurtosis zero, and this gave a misspecification effect of .664/.70 = .95. The alternative of the first line of Table 2 is to give $X_2$ a logistic distribution with the higher kurtosis, too[6]. This gives a misspecification effect of .679/.700 = .97. The opposite case of slim tails arises if we make $X_2$ a binary .50/.50 dummy: this has kurtosis $-2$. This turns out to be a severe misspecification, with major effects. The downward misspecification bias increases to .606/.700 = .87; moreover $ASE_1$ of the full equation differs from the values for the other specifications. But even so it is not affected by omitting $X_2$.

[6]Note that the sum of two logistic variates does *not* have a logistic distribution.

TABLE 2
*Mean and standard deviation of estimates in* 100 *replications, 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 |
| $ASE_1$ | .131 | .131 | .999 |
| | .003 | .004 | .025 |
| Binary Dummy: | | | |
| $\hat{\beta}_1$ | 1.002 | .606 | .605 |
| | .044 | .026 | .023 |
| $ASE_1$ | .121 | .121 | 1.000 |
| | .003 | .003 | .028 |
| $t(6)$: | | | |
| $\hat{\beta}_1$ | 1.006 | .687 | .683 |
| | .038 | .036 | .022 |
| $ASE_1$ | .132 | .131 | .999 |
| | .003 | .004 | .022 |
| Lognormal: | | | |
| $\hat{\beta}_1$ | 1.005 | .695 | .691 |
| | .046 | .027 | .027 |
| $ASE_1$ | .132 | .132 | 1.000 |
| | .005 | .003 | .023 |
| Skew Dummy: | | | |
| $\hat{\beta}_1$ | .997 | .677 | .679 |
| | .040 | .031 | .022 |
| $ASE_1$ | .130 | .130 | .996 |
| | .003 | .004 | .023 |

If a (much) lower kurtosis means a larger reduction, one might expect an opposite effect for a *higher* kurtosis, like the $t$ distribution with 6 degrees of freedom, with kurtosis 3. But in fact it does not produce an upward bias: the misspecification effect remains at .687/.700 = .98. - The performance

of the last two distributions, which are skewed, is no different. For the lognormal, with skewness 1.32 and kurtosis 3.18, the misspecification effect is .692/.700 = .99, and for the skew dummy with proportions (.20, .80) with skewness −1.50 and kurtosis .25 it is .679/.700 = .97.

The overall conclusion is that the attenuation bias affects $\hat{\beta}_1$ but not $ASE_1$. As for the misspecification effects, with one exception (the binary dummy) they are slight, no more than a few percent, even though they may well be significant.

## 5 Pure misspecification of the disturbance

So far we have added various $X_2$ to the logistic disturbances of the initial full equation, and found only minor misspecification effects. This suggests that logit analysis may well be insensitive to *any* misspecification of the distribution of the disturbances.

TABLE 3
*Mean and standard deviation of $\hat{\beta}_1$ of the full equation in 100 replications, for various distributions of the disturbance.*

| *Distribution of disturbance* | *Skewness* | *Kurtosis* | $\hat{\beta}_1$ | $s.e.\hat{\beta}_1$ | $ASE_1$ |
|---|---|---|---|---|---|
| Logistic | 0 | 1.2 | 1.005 | .042 | .128 |
| | | | .046 | .002 | .004 |
| Normal | 0 | 0 | .956 | .040 | .126 |
| | | | .038 | .001 | .004 |
| Binary Dummy | 0 | -2 | .835 | .036 | .121 |
| | | | .032 | .001 | .004 |
| $t(6)$ | 0 | 3 | 1.028 | .042 | .130 |
| | | | .041 | .001 | .003 |
| Lognormal | 1.3 | 3.2 | 1.036 | .043 | .130 |
| | | | .036 | .001 | .003 |
| Skew Dummy | -1.5 | 0.3 | 1.005 | .042 | .129 |
| | | | .041 | .001 | .009 |

This robustness is confirmed by simulations of the full equation (15). This is a simple latent variable equation with two independent normal regressors with $\boldsymbol{\beta}$ $(1, 1)$. Both regressors and the disturbances have been scaled to

have zero mean and variance 3.29, so that the systematic part accounts for two-thirds of the variation of $Y_i^+$. Table 3 shows the results for $\hat{\beta}_1$, for its estimated standard error and for $ASE_1$; the results for $\hat{\beta}_2$ are of course almost identical.

The first line of Table 3 refers to the reference case or correct specification, and the second to a normal distribution that is not very much different. Yet the normal induces a significant reduction of $\hat{\beta}_1$, since the difference of the mean from the true value of 1 is .044 and the standard deviation of the mean is .0038. The binary dummy that follows is an extreme case, leading to a substantial decline of $\hat{\beta}_1$. But, as before, the other extreme of the $t(6)$ distribution gives only a slight bias (even though it is significant) and the same holds for the lognormal. The effect of the skew dummy is even not significant. The effects on the standard errors of the estimate (derived from a misspecified model) are quite modest; they are in line with the standard deviations of the estimates over the replications.

Apart from one case (the .50/.50 dummy), the misspecification effect on $\hat{\beta}_1$ is no more than a few percent. While this may well be significant, it is a quite acceptable defect for most empirical work. One expects that the stark misspecifications of the disturbances that we have considered will be brought to light by the proper statistical tests, but in the end they do not greatly matter for the substantive results of applied work.

With a downward misspecification bias, the same compensating mechanism as before mitigates the effect on $ASE_1$, though not as perfectly as earlier.

## 6 Concluding Remarks

In probit and logit analysis, omitting a variable will bias $\hat{\beta}$ of the remaining regressors towards zero. For the probit model, Wooldridge (2002) has proved that this bias does not carry over to the partial *effect* of the remaining regressors, or the derivatives of the outcome in their respect. For the logit model, simulations confirm that it shares this property of the probit, as is so often the case. And while omitting a variable always implies misspecification of the disturbance, the additional effect of this on the $\hat{\beta}$, while significant, is generally slight, of the order of a few percent or so. The same holds for pure misspecification of the disturbances, outside the context of omitted variables.

In field work, we never know how the disturbances are actually distributed; although the necessary assymptions may be put to the test, such statistical tests are rarely employed in empirical work. The present findings suggest that this is not an important issue. The blatant misspecifications of the disturbances that we have introduced in the last section should surely

be significant by any proper test, yet their effects are so slight that they are of little consequence for the substantive results. As an empirical tool, logistic regression is quite robust with respect to deviations of the disturbance distribution from the model.

## References


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

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

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

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

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

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

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