NBER WORKING PAPER SERIES

SPECIFICATION ERRORS IN LIMITED DEPENDENT VARIABLE MODELS

G.S. Maddala*
Forrest D. Nelson**

Working Paper No. 96

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

July 1975

Preliminary: not for quotation

*Research supported by National Science Foundation Grant SOC-74-13158 to the University of Rochester.

**NBER Computer Research Center. Research supported in part by National Science Foundation Grant GJ-1154X3 to the National Bureau of Economic Research, Inc.

ABSTRACT

A preliminary investigation of two specification error problems in truncated dependent variable models is reported. It is shown that heteroscedasticity in a tobit model results in biased estimates when the model is misspecified. This differs from the OLS model where estimates are still consistent though inefficient. The second problem examined is aggregation. An appropriate nonlinear least squares regression model is derived for situations when the micro-level model fits a tobit framework but only aggregate data are available.

1. INTRODUCTION

In the case of ordinary regression models there exists substantial literature on the problems of heteroscedasticity, autocorrelation, omitted variable biases, aggregation biases etc. There is no corresponding analysis in the case of limited dependent variable models though these models are increasingly being used nowadays, particularly in the analysis of microdata. The present paper presents some results on two of these problems: heteroscedasticity and aggregation.

2. HETEROSCEDASTICITY

In the general linear model OLS estimates are consistent but not efficient when the disturbances are heteroscedastic. It will be shown below, in the limited dependent variable model the estimates are not even consistent when the model is mis-specified to be homoscedastic.

Consider the model:

$$\begin{aligned} y_t &= \beta' X_t + u_t \qquad \text{if RHS} > 0 \\ &= 0 \qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{2.1}$$

Assume that the residuals are normally distributed and denote by f(t) the frequency function and F(t) the cumulative distribution function corresponding to a standard normal. The locus of expected values for y_t is given by

$$E(y_t) = \beta' X_t \cdot F\left(\frac{\beta' X_t}{\sigma}\right) + \sigma\, f\left(\frac{\beta' X_t}{\sigma}\right) \tag{2.2}$$

Now suppose that the "true" model is heteroscedastic with parameters σ_{ot} and β_o and designate the "true" locus as $E_o(y_t)$. Then

$$E_o(y_t) = \beta_o' X_t \cdot F\left(\frac{\beta_o' X_t}{\sigma_{ot}}\right) + \sigma_{ot}\, f\left(\frac{\beta_o' X_t}{\sigma_{ot}}\right) \tag{2.3}$$

while the "misspecified" locus is given by $E(y_t)$ in equation (2.2). It is immediately obvious that the presence of the variance term σ_{ot} in the expected value locus is the source of the difference in the two expressions and, in turn, the estimation bias.

To make the discussion more concrete we will assume that the heteroscedasticity takes the form

$$\sigma_{ot} = \lambda^2 x_{tk}^2$$

i.e., the variance is proportional to the square of the last of the k independent variables. Furthermore we restrict attention to models in which x_{tk} takes on only positive values so that the standard deviation may be written as λx_{tk}. It will be useful to write $\beta' X_t$ as

$$\beta' X_t = A_t + \alpha x_{tk} \tag{2.4}$$

where $A_t = \sum_{i=1}^{k-1} \beta_i x_{ti}$ and $\alpha = \beta_k$

On substitution (2.2) and (2.3) become (in what follows the subscript t is omitted for ease of exposition)

$$E(y) = (A + \alpha x_k) F\left(\frac{A + \alpha x_k}{\sigma}\right) + \sigma f\left(\frac{A + \alpha x_k}{\sigma}\right) \tag{2.5}$$

$$E_o(y) = (A_o + \alpha_o x_k) F\left(\frac{A_o + \alpha_o x_k}{\lambda_o x_k}\right) + \lambda_o x_k f\left(\frac{A_o + \alpha_o x_k}{\lambda_o x_k}\right) \tag{2.6}$$

Consider the behavior of the two expressions as $x_k \to \infty$. We have

$$E(y) \to A + \alpha x_k \tag{2.7}$$

and

$$E_o(y) \to A_o F\left(\frac{\alpha_o}{\lambda_o}\right) + \left[\alpha_o F\left(\frac{\alpha_o}{\lambda_o}\right) + \lambda_o f\left(\frac{\alpha_o}{\lambda_o}\right)\right] x_k \tag{2.8}$$

Expressions (2.7) and (2.8) suggest a first guess at the potential bias in the estimate of α obtained from a mis-specified model. It seems reasonable to assume that the estimate α will approach, on the average, the expression multiplying x_K in (3.8) rather than α_o as we would like. Thus, we can write

$$E(\hat{\alpha}) \simeq \alpha_o F\left(\frac{\alpha_o}{\lambda_o}\right) + \lambda_o f\left(\frac{\alpha_o}{\lambda_o}\right)$$

with bias given by

$$B(\hat{\alpha}) \simeq -\alpha_o\left[1 - F\left(\frac{\alpha_o}{\lambda_o}\right)\right] + \lambda_o f\left(\frac{\alpha_o}{\lambda_o}\right) \tag{2.9}$$

This "first guess" of bias clearly depends on only α_o and λ_o. Differentiating yields[1]

$$\frac{\partial \beta}{\partial \alpha_o} = -\left[1-F(\alpha_o/\lambda_o)\right] \quad \text{and} \quad \frac{\partial \beta}{\partial \lambda_o} = f(\alpha_o/\lambda_o)$$

Thus the "bias" decreases with α_o and increases with λ_o. In addition it can be shown that the "bias" is always positive. Since (1-F) and λ_o are both positive, dividing B by $\lambda_o(1-F)$ does not change its sign. Thus (2.9) can be written as

$$\frac{B(\hat{\alpha})}{\lambda_o[1-F(\alpha_o/\lambda_o)]} = -\frac{\alpha_o}{\lambda_o} + \frac{f(\alpha_o/\lambda_o)}{1-F(\alpha_o/\lambda_o)} \tag{2.10}$$

---

[1]Note that $\frac{\partial F}{\partial \alpha_o} = f\cdot(1/\lambda_o)$ $\frac{\partial F}{\partial \lambda_o} = -f\cdot(\alpha_o/\lambda_o^2)$

$\frac{\partial f}{\partial \alpha_o} = -f\cdot(\alpha_o/\lambda_o^2)$ $\frac{\partial f}{\partial \lambda_o} = f\cdot(\alpha_o^2/\lambda_o^3)$

The function $f(x)/[1-F(x)]$ approaches x as $x \to +\infty$ and o as $x \to -\infty$, is always postive and is always greater than x. Thus the right hand side of (2.10) is non-negative. Finally, note from (2.8) and (2.9) that if the ratio α_o/λ_o is large, say greater than 4, the bias becomes negligible.

This comparison of E(y) and $E_o(y)$ thus suggests that $\hat{\alpha}$ (i.e. $\hat{\beta}_k$) will be biased upward. It does not provide as much information about the remaining (k-1) $\hat{\beta}_i$'s. Comparison of A and $A_o F(\frac{\alpha_o}{\lambda_o})$ suggests only that they will be collectively biased so as to make A on the average smaller than A_o. Some sampling experiments we have conducted confirm these analytical results.[1]

3. THE PROBLEM OF AGGREGATION

Consider the following micro-level model:

$$\begin{aligned} y_{it} &= \beta' X_t + u_{it} \quad \text{if RHS} < L_t \\ &= L_t \qquad\qquad \text{otherwise} \end{aligned} \tag{3.1}$$

$$u_{it} \sim IN(o, \sigma^2) \tag{3.2}$$

To motivate the model let y_{it} be the interest rate paid by bank i in time period t on time deposits and X_t a vector of observations at time t on a set of exogenous variables which affect savings account interest rates, and L_t the regulation Q ceiling on such rates.

Suppose there are N_t banks (i.e., i=1,2 ... N_t). If micro level data are available on all N_t banks during each period t (t=1,2... T), estimation of β and σ could be handled by the usual Tobit analysis. But if only

---

[1] More sampling experiments are being conducted to get further insights into the consequences of the heteroscedasticity problem.

aggregate measures of savings account interest rates are available, neither Tobit nor straightforward OLS are suitable. Suppose observed interest rates amount to the unweighted mean of interest rates paid by all N_t banks during each period t so that

$$y_t = \frac{1}{N_t} \sum_i y_{it} \tag{3.3}$$

Let n_t be the number of banks at period t paying an interest rate less than the ceiling rate L_t and re-order the observations within each period so that those banks are the first n_t. Then (3.3) becomes

$$\begin{aligned} y_t &= \frac{1}{N_t} \left[\sum_{i=1}^{n_t} (\beta' X_t + u_{it}) + (N_t - n_t) L_t \right] \\ &= \frac{n_t}{N_t} \beta' X_t + \left(1 - \frac{n_t}{N_t}\right) L_t + \frac{1}{N_t} \sum_{i=1}^{n_t} u_{it} \end{aligned} \tag{3.4}$$

If n_t were observable and if the last term on the right hand side of (3.4) had zero expectation, we could proceed by OLS. But neither are true. We might thus proceed by finding $E(y_t)$ and fitting

$$y_t = E(y_t) + w_t \tag{3.5}$$

by non-linear least squares method. Now $E(y_t) = L_t + E\left(\frac{n_t}{N_t}\right)(\beta' X_t - L_t) + E\left(\frac{1}{N_t} \sum_{i=1}^{n_t} u_{it}\right)$. Define the binomial variable

$$\begin{aligned} D_{it} &= 1 \quad \text{if } i \leq n_t \\ &= 0 \quad \text{if } i > n_t \end{aligned}$$

Then $\Pr(D_{it} = 1) = \Pr(\beta' X_t + u_{it} < L_t) = F\left(\frac{L_t - \beta' X_t}{\sigma}\right)$

$$= E(D_{it})$$

Where $F(\cdot)$ is the cdf of the standard normal distribution.

Now $n_t = \sum_{i=1}^{N_t} D_{it}$

$$E(\frac{n_t}{N_t}) = \frac{1}{N_t} N_t E(D_{it}) = F(\frac{L_t-\beta'X_t}{\sigma}) \tag{3.6}$$

Next, note that for each u_{it} such that $i \leq n_t$ we have

$$u_{it} < L_t - \beta'X_t$$

Thus this subset of n_t u_{it}'s follows a truncated conditional distribution given by

$$g(u_{it} \mid i \leq n_t) = \frac{1}{\sigma} f(\frac{u_{it}}{\sigma}) / F(\frac{L_t-\beta'X_t}{\sigma})$$

$$\text{For } -\infty < u_{it} < \frac{L_t-\beta'X_t}{\sigma}$$

$$= 0 \text{ otherwise.}$$

Where $f(\cdot)$ is the pdf of the standard normal. It is easily verified that

$$E(u_{it} \mid i \leq n_t) = -\sigma f(\frac{L_t-\beta'X_t}{\sigma}) / F(\frac{L_t-\beta'X_t}{\sigma}) \tag{3.7}$$

Thus

$$E(\frac{1}{N_t} \sum_{i=1}^{n_t} u_{it}) = E(\frac{n_t}{N_t}) \left[\frac{-\sigma f(\frac{L_t-\beta'X_t}{\sigma})}{F(\frac{L_t-\beta'X_t}{\sigma})} \right]$$

$$= -\sigma f(\frac{L_t-\beta'X_t}{\sigma}) \tag{3.8}$$

An alternative way of getting the same result is the following:

Define $u^*_{it} = u_{it}$ if $u_{it} < L-\beta'X_t$

$$= 0 \quad \text{otherwise.}$$

Then $E(u^*_{it}) = -\sigma f(\frac{L_t-\beta'X_t}{\sigma})$, and

$$E(\frac{1}{N_t} \sum_{i=1}^{n_t} u_{it}) = E(\frac{1}{N_t} \sum_{i=1}^{N_t} u^*_{it})$$

$$= \frac{N_t}{N_t} [-\sigma f(\frac{L_t-\beta'X_t}{\sigma})]$$

Using (3.6), (3.8) and (3.5) we get

$$y_t = L_t + F\left(\frac{L_t-\beta'X_t}{\sigma}\right)[\beta'X_t-L_t] - \sigma f\left(\frac{L_t-\beta'X_t}{\sigma}\right) + w_t \qquad (3.9)$$

where w_t now has zero expectation and is uncorrelated with each of the other terms on the right.

A much easier derivation of (3.9) is obtained by noting that

$$E(y_{it}) = F\left(\frac{L_t-\beta'X_t}{\sigma}\right)\beta'X_t - \sigma f\left(\frac{L_t-\beta'X_t}{\sigma}\right) + L_t\left[1 - F\left(\frac{L_t-\beta'X_t}{\sigma}\right)\right]$$

which yields (3.9) directly. Estimates of β and σ can be obtained using non-linear least squares in (3.9).[1] There is also the further problem of taking the heteroscedasticity of w_t into account.

4. CONCLUSIONS

In this paper we present some preliminary results on two problems commonly encountered in the analysis of limited dependent variable models: heteroscedasticity and aggregation. We are conducting some simulation studies to get insights on those aspects of these problems for which we have not been able to get analytical results. We are also investigating the other problems mentioned in the introduction and the performance of the non-linear least squares method suggested in (3.9).

---

[1] Collection of data and empirical estimation of this model are currently under progress.