NBER TECHNICAL PAPER SERIES

THE EFFECT OF IGNORING HETEROSCEDASTICITY ON ESTIMATES OF THE TOBIT MODEL

Charles C. Brown

Robert Moffitt*

Technical Paper No. 27

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02138

January 1983

*On leave, University of Maryland, 1981-1982. The authors would like to thank the University of Maryland Computer Science Center for research support, and Arthur Goldberger for helpful comments. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #27
January 1983

THE EFFECT OF IGNORING HETEROSCEDASTICITY ON ESTIMATES OF THE TOBIT MODEL

ABSTRACT

We consider the sensitivity of the Tobit estimator to heteroscedasticity. Our single independent variable is a dummy variable whose coefficient is a difference between group means, and the error variance differs between groups. Heteroscedasticity biases the Tobit estimate of the two means in opposite directions, so the bias in estimating their difference can be significant. This bias is not monotonically related to the true difference, and is greatly increased if the limit observations are not available. Perhaps surprisingly, the Tobit estimates are sometimes more severely biased than are OLS estimates.

Charles Brown
Department of Economics
University of Maryland
College Park MD 20742
(301) 454 4182

Robert Moffitt
Department of Economics
University of Wisconsin
Madison WI 53706
(608) 263 6724

I. INTRODUCTION

In two recently published papers, Hurd (1979) and Arabmazar and Schmidt (1981) have examined the sensitivity of Tobit estimates to the presence of heteroscedasticity. Hurd found the estimates to be quite sensitive, but Arabmazar and Schmidt, pointing out that Hurd examined only the case in which the limit observations are not used, found considerably less sensitivity.. However, Arabmazar and Schmidt examined only the sensitivity of the estimated mean of the normal distribution to heteroscedasticity, whereas, in economics, we are almost always concerned with regression coefficients. In this paper, we report the results of examining the sensitivity to heteroscedasticity of the Tobit estimator of the coefficient of a single dummy independent variable (i.e., a difference between means), and find results considerably less reassuring than Arabmazar and Schmidt. Essentially this is because estimates of the two means are biased in opposite directions, making the coefficient more sensitive to heteroscedasticity than a single mean. We also examine the effect on the coefficient estimate of excluding the limit observations (we find this increases the bias); of using OLS (we find, surprisingly, that sometimes OLS is better); and of using the Greene (1981) procedure (we find it is also sometimes better, though less often than OLS). We present our main simulations in the first section below; then compare them with truncated estimators in the next section; and conclude with a brief summary.[1]

II. THE EFFECT OF HETEROSCEDASTICITY IN THE TOBIT MODEL

The Tobit model we wish to investigate is the following:

$$
\begin{aligned}
y_j^* &= \alpha_1(1 - D_j) + \alpha_2 D_j + \varepsilon_{1j}(1 - D_j) + \varepsilon_{2j} D_j \\
y_j &= y_j^* \quad \text{if} \quad y_j^* \geq 0 \\
&= 0 \quad \text{if} \quad y_j^* < 0
\end{aligned} \tag{1}
$$

where j indexes the observations and D_j is a dummy variable equal to zero for N_1 observations and equal to one for N_2 observations. We assume that $\varepsilon_{1j} \sim N(0, \sigma_1^2)$ and $\varepsilon_{2j} \sim N(0, \sigma_2^2)$. This model is equivalent to one in which there is an intercept and a single independent variable, the dummy variable D_j, with coefficient $\alpha_2 - \alpha_1$ -- the model as described in the introduction. The log likelihood function that must be maximized with respect to α_1, α_2, σ_1, and σ_2 is the following:

$$
\begin{aligned}
L = &- \frac{1}{2} \sum_{N_1^+} \left(\frac{y_j - \alpha_1}{\sigma_1} \right)^2 - \frac{1}{2} \sum_{N_2^+} \left(\frac{y_j - \alpha_2}{\sigma_2} \right)^2 \\
&- N_1^+ \log \sigma_1 - N_2^+ \log \sigma_2 + N_1^0 \log F\left(\frac{-\alpha_1}{\sigma_1}\right) + N_2^0 \log F\left(\frac{-\alpha_2}{\sigma_2}\right)
\end{aligned} \tag{2}
$$

where F is the normal c.d.f. and where N_i^+ is the number of positive observations in subsample $i(i = 1, 2)$ and N_i^0 is the corresponding number of zero observations. The function is maximized at the solutions to the four normal equations:

$$
\frac{1}{s_i^{*2}} \sum_{N_i^+} (y_j - a_i^*) - \frac{N_i^0}{s_i^*} \frac{f(-a_i^*/s_i^*)}{F(-a_i^*/s_i^*)} = 0 \qquad i = 1, 2 \tag{3}
$$

$$
\frac{1}{s_i^{*3}} \sum_{N_i^+} (y_j - a_i^*)^2 - \frac{N_i^+}{s_i^*} + N_i^0 \frac{a_i^*}{s_i^{*2}} \frac{f(-a_i^*/s_i^*)}{F(-a_i^*/s_i^*)} = 0 \qquad i = 1, 2 \tag{4}
$$

where f is the unit normal density function. The consistency of the estimators of the model has been proven by Amemiya (1973). However, if the model is assumed to have a single error term distributed $N(0, \sigma^2)$, the log likelihood function is:

$$L = -\frac{1}{2} \sum_{i=1,2} \sum_{N_i^+} \left(\frac{y_j - \alpha_i}{\sigma} \right)^2 - \sum_{i=1,2} N_i^+ \log \sigma$$

$$+ \sum_{i=1,2} N_i^0 \log F(-\alpha_i/\sigma) \qquad (5)$$

The three normal equations for the three coefficient estimators, a_1, a_2, and s, are the following:

$$\frac{1}{s^2} \sum_{N_i^+} (y_j - a_i) - \frac{N_i^0}{s} \frac{f(-a_i/s)}{F(-a_i/s)} = 0 \qquad i = 1,2 \qquad (6)$$

$$\frac{1}{s^3} \sum_{i=1,2} \sum_{N_i^+} (y_j - a_i)^2 - \sum_{i=1,2} \frac{N_i^+}{s} + \sum_{i=1,2} N_i^0 \frac{a_i}{s^2} \frac{f(-a_i/s)}{F(-a_i/s)} = 0 \qquad (7)$$

It is clear that the estimators in equation (6) will not be those of a_i^* in equation (3). Each equation in (6) will generate a solution value for an a_i given a value of s which satisfies all three equations (6)-(7). But these values cannot be those obtained for the a_i^* in equations (3) since the latter are each a function of the two different estimators $s_i^* \neq s$. And since s_i^* are consistent estimators of the σ_i, the s_i^* cannot equal s. Moreover, it can be seen from inspection that the difference $(a_1 - a_2)$, which is the main object of interest, will bear no linear relationship to $(a_i^* - a_2^*)$ in equation (3) as a result of the nonlinearity of the normal distribution. Consequently the difference in means will also be inconsistently estimated.

The question we wish to consider is the magnitude of the asymptotic bias under various values of the true parameters. The true parameters will determine the mean and variance of the y_j and the N_i in equations (6)-(7) in large samples and hence will determine the estimated values of a_1, a_2, and s. Letting $N = N_1 + N_2$ be the total sample size, note that in large samples we will have:

$$\text{plim} \sum_{N_i^+} (y_j - a_i) = N_i^+ (\bar{y}_i - a_i) \qquad i = 1, 2$$

$$\text{plim} \sum_{N_i^+} (y_j - a_i)^2 = N_i^+ [V_i + (\bar{y}_i - a_i)^2] \qquad i = 1, 2$$

$$\text{plim} \quad N_i^+/N = F(\frac{\alpha_i}{\sigma_i}) P_i \qquad i = 1, 2$$

$$\text{plim} \quad N_i^0/N = [1 - F(\frac{\alpha_i}{\sigma_i})] P_i \qquad i = 1, 2$$

where

$$\bar{y}_i = \alpha_i + \sigma_i \lambda_i$$

$$V_i = \sigma_i^2[1 - (\frac{\alpha_i}{\sigma_i})\lambda_i - \lambda_i^2]$$

$$\lambda_i = \frac{f(\alpha_i/\sigma_i)}{F(\alpha_i/\sigma_i)}$$

and where P_i is the fraction of the sample in category i (Johnson and Kotz, 1972). Dividing equations (6)-(7) through by N and using the above large-sample values we can rewrite the normal equations as:

$$\frac{1}{s^2} F(\frac{\alpha_i}{\sigma_i}) (\bar{y}_i - a_i) - \frac{1}{s} [1 - F(\frac{\alpha_i}{\sigma_i})] \frac{f(-a_i/s)}{F(-a_i/s)} = 0 \qquad (8)$$

$$\frac{1}{s^3} \sum_i F(\frac{\alpha_i}{\sigma_i}) P_i [V_i + (\bar{y}_i - a_i)^2] - \sum_i F(\frac{\alpha_i}{\sigma_i}) P_i \frac{1}{s}$$

$$+ \sum_i (1 - F(\frac{\alpha_i}{\sigma_i})) P_i \frac{f(-a_i/s)}{F(-a_i/s)} \frac{a_i}{s^2} = 0 \qquad (9)$$

Rewritten in this fashion the equations show implicitly the relationship of the estimated values of a_1, a_2, and s to the underlying values α_1, α_2, σ_1, σ_2, and the P_i ($P_1 = 1 - P_2$). Note that the sample proportions, the P_i, do not directly affect the estimates of the a_i in equations (8). Rather, they enter indirectly through equation (9), where they act as weights determining the value of s.

There are very few analytic relationships between the true parameters and the estimated parameters derivable from equations (8)-(9). Therefore we shall assume various values for the true parameters and solve equations (8)-(9) for the estimates of a_1, a_2, and s for each assumed set, with our main interest focusing on a_1 and a_2 and their difference in relation to the true difference. Some normalization is necessary, so we set $\sigma_1 = 1$ throughout.[2] We then allow α_1 and α_2 to each take on the values -1.0, 0.0, and 1.0, and we allow σ_2 to take on the values 0.5, 0.8, 1.25, and 2.0. We also vary the P_i by .25, .50, and .75. These ranges generate a wide range of truncation points and hence generate samples with wide ranges of truncation percentages. Since it is the implied truncation percentage that will be the most important underlying driving force in the size of the bias magnitude, these ranges should span the situations faced by most data analysts.

The results are shown in Table 1 under the "Tobit" columns. The table shows the difference between the estimated ($a_1 - a_2$) and the true ($\alpha_1 - \alpha_2$) for each value of σ_2, with $P_1 = P_2 = 0.5$ for the time being. When $\sigma_2 = 1$, there is no bias since the variances are equal; this case is therefore not shown. Several patterns appear in the table. (1) The bias is positive if $\sigma_1 > \sigma_2$ and negative if $\sigma_2 > \sigma_1$, and it grows with the absolute value of the difference in the variances.

This arises because our simulated values of s (not shown in the table) are always weighted averages of σ_1 and σ_2, and hence always fall between them.[3] We find, without exception, that if $\sigma_2 > s > \sigma_1$, then $a_2 > \alpha_2$ and $a_1 < \alpha_1$, giving a negative bias in $a_1 - a_2$. The opposite occurs when $\sigma_1 > s > \sigma_2$. (2) There is no simple relation between the size of $\alpha_1 - \alpha_2$ and the amount of bias. Rather, it depends upon the amount of truncation involved. Either holding α_1 constant and decreasing α_2 (hence increasing $\alpha_1 - \alpha_2$) or holding α_2 constant and decreasing α_1 (hence decreasing $\alpha_1 - \alpha_2$), the bias grows in absolute value. In both of these cases, the amount of truncation is increased, distorting the estimates of the untruncated means even further. (3) For the same reason, holding $\alpha_1 - \alpha_2$ constant, the bias grows as the level of each falls.

The relationship between the extent of bias and the degree of truncation thus depends on how the latter is changed. Increasing truncation by decreasing α_1 or α_2 increases the bias, but increasing truncation by changing σ_2 is more complicated. Moving σ_2 away from 1.0 (i.e., away from σ_1) increases the bias, whether truncation is increasing ($\alpha_2 < 0$ and $\sigma_2 < 1$ or $\alpha_2 > 0$ and $\sigma_2 > 1$) or decreasing ($\alpha_2 > 0$ and $\sigma_2 < 1$ or $\alpha_2 < 0$ and $\sigma_2 < 1$).

The more important question is when the absolute amount of the bias is large. The table confirms what has already been hinted at: it is large when the values of α_1 and α_2 are low and hence the truncation percentages are large. For example, six of bias values in the table are .99 or greater in absolute value. Of these, three have 98 percent zeros for one of the subsamples, and the other three have 84 percent zeros for one of the subsamples.[4] Thus the truncation is extreme. If the two subsamples each have at least 30 percent positives, the largest biases in the table are -.60 and -.67, which occur when the true difference is fairly large (one or two) and when the σ_2 is at its largest value. Hence the largest biases seem to occur when the truncation is extreme or when σ_2 is double σ_1. In other cases, the bias does not seem large at all. If at least 50 percent of the two subsamples are positive, and if σ_2 is 0.8 or 1.25, the bias never exceeds .12 in absolute value, and is generally less than .10. Recall that these numbers are in units relative to $\sigma_1 = 1$.

The effect of different sample proportions are not shown in the table. We found that different proportions have uniform effects when the true $\alpha_1 - \alpha_2 = 0$. In this case, increasing the proportion of the sample in the category with the larger variance always decreases the bias. In general, however, with unequal alphas, there is no regular pattern to the effects of sample proportions on the bias. However, letting P_1 equal .25 or .75 does not radically alter the entries in Table 1.

The table also shows the OLS estimates of the equation (estimated on the positive values only). Surprisingly, the Tobit estimator does not always do better than OLS. OLS is sometimes a bit better, and occasionally very much better (less than 50 percent of the Tobit bias). The occasions where this occurs are again where the truncation percentages are high: where α_1 and α_2 are low or where the variances are double one another. Apparently the distortion in the shape of the normal distribution resulting from the equal-variance assumption is great enough in these cases to make a simple comparison of truncated means a superior estimate of the untruncated means than Tobit.

We also computed the estimates suggested by Greene (1981) for the Tobit model. Greene has shown that if the independent variables are jointly normally distributed, asymptotically unbiased estimates of the coefficients can be obtained by dividing the OLS coefficients (estimated on the entire sample) by the mean sample proportion positive.[5] Again, like OLS, the Greene estimator occasionally showed less bias than the Tobit estimator, but in fewer cases than OLS. There was some tendency for it to do better than Tobit again in extreme cases of truncation, but this pattern was much less apparent than in OLS.

III. COMPARISON WITH PREVIOUS STUDIES

A previous study of heteroscedasticity in the Tobit model by Hurd (1979) differs from ours primarily in that Hurd examined estimates using only the truncated sample (i.e., zeros excluded), whereas we have considered the complete sample. Since Hurd found that heteroscedasticity effects are much larger than ours, we have repeated our analysis using only the truncated sample. The points made in the previous section in regard to the inconsistency of estimators in the presence of heteroscedasticity all apply equally in this case, but the likelihood function differs. In the truncated case, each individual probability density is conditioned upon being positive and hence is divided by the probability of being positive (see Hurd).[6]

Table 2 shows the biases in the Tobit estimators repeated from Table 1 along with the biases obtained when using the truncated sample only. As the table shows, the biases in the truncated sample are indeed much larger than those in the complete sample. In ratio terms, the bias in the truncated estimator is up to 10 times larger than the bias in the complete sample. Moreover, the truncated estimator is much more sensitive to the bias-inducing characteristics we noted in the last section. All the factors that make for a more serious bias -- larger percentages truncated, larger differences in the variances holding constant α_1 and α_2, and so on -- have a much greater absolute effect on the truncated estimators. In the worst case of bias in the complete sample, the bias is -1.07 compared to -5.24 for the truncated estimator, a large absolute difference. Apparently the extra statistical information on the shape of the distribution provided by the limit observations has a large effect in stabilizing the estimates.

We conclude two things from this examination. First, if the complete sample is available to the analyst, there is much less cause for concern from heteroscedasticity that if only the truncated sample is available. Second, the large difference in the two estimators suggests that, where possible, estimates be obtained with and without using the limit observations, as an informal test of specification. If the underlying distributional assumptions (normality with homoscedasticity) are correct, the estimators should be close. But if either of these assumptions fails, the "test statistic" of the difference in the truncated and complete sample estimators may be large. Fortunately, Table 2 suggests that the difference in the estimators (equal to the difference in the biases given in Table 2) will tend to be largest when the bias from using either is most pronounced.

Arabmazar and Schmidt (1981, p. 258) also found that including limit observations reduced the bias due to heteroscedasticity--so much so that they "conjecture . . . that moderate heteroscedasticity (say, variance differing by a factor of two) is not likely to cause substantial inconsistency unless the sample is heavily censored (say, more than half of the observations at the limit)." Our results are less comforting. For example, when $\alpha_1 = 0$, $\alpha_2 = 1$, and $\sigma_2 = 2$, sixty percent of the sample will have nonlimit observations but the estimate of $\alpha_1 - \alpha_2$ converges to -1.41 instead of -1.0. The bias of -0.41 would certainly be regarded as severe in some contexts; in any case, it is nearly double the bias from OLS for the same parameter values.

IV. CONCLUSIONS

We have considered the sensitivity of the Tobit estimate of the coefficient of an independent variable to heteroscedasticity. We conclude from our study that (1) the coefficient is inconsistently estimated when heteroscedasticity is present; (2) estimates of the coefficient (essentially a difference between means) is a good deal more sensitive than the estimate of a single mean as reported in Arabmazar and Schmidt (1981); (3) the amount of the bias is not monotonically related to the size of the true coefficient, but rather depends, in a complicated way, on the amount of truncation in the sample; (4) both the OLS and the Greene (1981) estimators sometimes do better than Tobit when heteroscedasticity is present, with OLS a bit better of the two; (5) excluding the limit observations as Hurd (1979) did greatly increases the bias in the Tobit coefficient; and (6) an informal specification test is suggested in which estimates on both the truncated and untruncated samples are obtained.

NOTES

1. Nelson (1979) also simulated heteroscedasticity biases, although again only with a single mean. See also Peterson and Waldman (1981) for an example of a Tobit model estimated with (and without) correction for heteroscedasticity. We should note the separate literature examining the effect of non-normality on the Tobit estimators (Arabmazar and Schmidt, 1982; Goldberger, 1980; Olsen, 1982) and the general specification test provided by Nelson (1981).

2. The necessity for normalization can be seen by noting that doubling both the true parameters and the estimators in equations (8) and (9) will leave the equalities intact.

3. The weights on the two variances are complicated functions of the α_i and σ_i, but are approximately equal to the sample proportions, the P_i.

4. Specifically, 98 percent of the observations in Sample 2 are zero if $\alpha_2 = -1$, $\sigma_2 = .5$, and $\alpha_1 = 1, 0,$ or -1. For $\alpha_1 = -1$, $\sigma_2 = 2$, and $\alpha_2 = 1, 0,$ or -1, 84 percent of observations in Sample 1 are zero.

5. Greene also found empirically that his coefficient estimates on dummy variables like ours, though not normally distributed, were quite accurate.

6. Hurd's simulations differ from ours in several ways. For example, his single independent variable was continuous rather than dichotomous. He simulated by drawing pseudo random numbers on this variable and on the size of the variance for each value of the variable. He deleted any observation simulated to be a limit observation. He also summarized his results by multiple regression analysis, since he simulated many more observations than we have.

REFERENCES

Amemiya, T. "Regression Analysis When the Dependent Variable Is Truncated Normal." Econometrica 41 (November 1973): 997-1016.

Arabmazar, A. and P. Schmidt. "Further Evidence on the Robustness of the Tobit Estimator to Heteroscedasticity." Journal of Econometrics 17 (1981): 253-258.

__________. "An Investigation of the Robustness of the Tobit Estimator to Non-Normality." Econometrica 50 (July 1982): 1055-1064.

Goldberger, A. "Abnormal Selection Bias." Working Paper 8006. Madison: SSRI, University of Wisconsin, 1980.

Greene, W. "On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model." Econometrica 49 (March 1981): 505-513.

Hurd, M. "Estimation in Truncated Samples When There Is Heteroscedasticity." Journal of Econometrics 11 (1979): 247-258.

Johnson, N. and S. Kotz. Continuous Multivariate Distributions. New York: Wiley, 1972.

Nelson, F. "The Effect of and a Test for Misspecification in the Censored-Normal Model." Social Science Working Paper 291, California Institute of Technology, 1979.

__________. "A Test for Misspecification in the Censored Normal Model." Econometrica 49 (September 1981): 1317-1330.

Olsen, R. "Distributional Tests for Selectivity Bias and a More Robust Likelihood Estimator." International Economic Review 23 (February 1982): 223-240.

Peterson, D. and D. Waldman. "The Treatment of Heteroscedasticity in the Limited Dependent Variable Model." Mimeographed, University of North Carolina, Chapel Hill, 1981.

Tobin, J. "Estimation of Relationships for Limited Dependent Variables." Econometrica 26 (January 1958): 24-36.

TABLE 1

BIAS IN ESTIMATES OF a_1-a_2*

			$\sigma_2 = 0.5$		$\sigma_2 = 0.8$		$\sigma_2 = 1.25$		$\sigma_2 = 2.0$	
α_1	α_2	α_1-α_2	Tobit	OLS	Tobit	OLS	Tobit	OLS	Tobit	OLS
1	1	0	.03	.26	.02	.12	- .04	- .17	- .19	- .73
1	0	1	.19	-.11	.07	- .35	- .09	- .71	- .33	-1.31
1	-1	2	1.02	-.90	.26	-1.09	- .22	-1.42	- .60	-2.00
0	1	-1	.16	.77	.07	.64	- .09	.34	- .41	- .22
0	0	0	.24	.40	.10	.16	- .12	- .20	- .49	- .80
0	-1	1	1.01	-.39	.25	- .59	- .22	- .91	- .67	-1.48
-1	1	-2	.51	1.50	.21	1.36	- .26	1.07	-1.06	.51
-1	0	-1	.50	1.13	.20	.89	- .26	.53	-1.05	- .07
-1	-1	0	.99	.34	.26	.14	- .28	- .18	-1.07	- .76

*Numbers shown are values of $(a_1-a_2) - (\alpha_1-\alpha_2)$. $P_1 = P_2 = 0.5$, $\sigma_1 = 1$. The fraction of observations above zero is $F(\alpha_i/\sigma_i)$:

α_i/σ_i	$F(\alpha_i/\sigma_i)$
0	.50
0.5	.69
0.8	.79
1.0	.84
1.25	.89
2.00	.98

TABLE 2

COMPARISON OF BIASES IN TRUNCATED AND COMPLETE SAMPLES[a]

			σ_2=0.5		σ_2=0.8		σ_2=1.25		σ_2=2.0	
α_1	α_2	$\alpha_1-\alpha_2$	C	T	C	T	C	T	C	T
1	1	0	.03	.36	.02	.19	-.04	-.29	-.19	-1.49
1	0	1	.19	1.66	.07	.46	-.09	-.44	-.33	-1.62
1	-1	2	1.02	3.99	.26	.91	-.22	-.71	-.60	-1.99
0	1	-1	.16	.68	.07	.35	-.09	-.57	-.41	-3.19
0	0	0	.24	1.48	.10	.45	-.12	-.57	-.49	-2.97
0	-1	1	1.01	3.82	.25	.89	-.22	-.72	-.67	-2.75
-1	1	-2	.51	1.27	.21	.63	-.26	-1.02	-1.06	-5.49
-1	0	-1	.50	1.40	.20	.63	-.26	-.99	-1.05	-5.39
-1	-1	0	.99	3.22	.26	.84	-.28	-.95	-1.07	-5.24

[a]C = Complete sample, T = truncated sample. Biases defined as in Table 1, $P_1 = P_2 = .5$, $\sigma_1 = 1$.