WORKING PAPERS

THE APPLICATION OF TOBIT AND PROBIT

ESTIMATION TO AGGREGATE DATA

Frederick I. Johnson

WORKING PAPER NO. 85

June 1983

BUREAU OF ECONOMICS
FEDERAL TRADE COMMISSION
WASHINGTON, DC 20580

TOBIT AND PROBIT WITH AGGREGATE DATA

Limited dependent variable models have been applied with great success to many economic analyses. Unfortunately, these models have one drawback which has limited their wider applicability: in general they can only be applied to disaggregated data. In this paper we explore the aggregation problem in two limited dependent variable models, Tobit and Probit. Our conclusion is remarkably sanguine: we find that, if the explanatory variables are themselves normally distributed, we can use mean aggregate data in Tobit and Probit estimation.

I. Tobit

A. With Disaggregated Data

In the Tobit model the distribution of the dependent variable is truncated. (The truncation may be envisioned as either an upper or lower limit. Here we treat it as the former.) Consequently, values of the dependent variable (y_{it}) in excess of the truncation point (v_{it}) are not observed. The model is characterized as follows:

$$(1)\qquad y_{it} = \begin{cases} X_{it}\beta + \varepsilon_{it} & \Longleftrightarrow \quad I_{it} = 1 \\ v_{it} & \Longleftrightarrow \quad I_{it} = 0 \end{cases}$$

where I_{it} is the switching condition (i.e. whether or not the upper constraint v_{it} is binding) and is defined as:

$$(2)\qquad I_{it} = \begin{cases} 1 & \Longleftrightarrow \quad X_{it}\beta + \varepsilon_{it} < v_{it} \\ 0 & \Longleftrightarrow \quad X_{it}\beta + \varepsilon_{it} \geq v_{it} \end{cases}$$

The subscripts i and t represent individual i at time t. X_{it} is a vector of explanatory variables, with β the vector of corresponding parameters, and ε_{it} is a disturbance term, $\varepsilon_{it} \sim N(0,\sigma^2)$. Equations (1) and (2) can be combined to form a single expression for y_{it}, from which its expectation can then be derived.

$$(3) \qquad y_{it} = I_{it}(X_{it}\beta + \varepsilon_{it}) + (1-I_{it})v_{it}$$

$$= v_{it} + I_{it}(X_{it}\beta - v_{it}) + I_{it}\varepsilon_{it}$$

$$(4) \qquad E(y_{it}) = v_{it} + (X_{it}\beta - v_{it})\ F\left(\frac{v_{it}-X_{it}\beta}{\sigma}\right) - \sigma\ f\left(\frac{v_{it}-X_{it}\beta}{\sigma}\right)$$

where $f(\cdot)$ is the standard normal density function and $F(\cdot)$ is the normal cumulative distribution function. The parameter β and the standard error σ can be estimated from equation (5):

$$(5) \qquad \min_{\beta\ \sigma} \sum_i \sum_t \{y_{it} - E(y_{it})\}^2$$

B. With Mean Aggregate Data

Suppose, however, that only mean data are available. That is, we have observations on y_t, X_t, and v_t (each of which is the mean, at time t, of individuals i). Under what circumstances can we estimate β and σ from the mean analogs of equations (4) and (5) (the same equations without the subscript i)?

$$(6)\quad E(Y_t) = \frac{1}{n}\sum E(y_{it})$$

$$= v_t + \frac{1}{n}\sum (X_{it}\beta - v_t)\, F\left(\frac{v_{it}-X_{it}\beta}{\sigma}\right) - \frac{\sigma}{n}\sum f\left(\frac{v_{it}-X_{it}\beta}{\sigma}\right)$$

$$= v_t + (X_t\beta - v_t)\, F\left(\frac{v_t-X_t\beta}{\sigma}\right) - \sigma\, f\left(\frac{v_t-X_t}{\sigma}\right)$$

if $(X_{it}\beta - v_{it}) = (X_t\beta - v_t)$, $i = 1,\ldots,n$

In general, if Tobit analysis is to be applied to mean data to estimate β and σ, then each of the individuals at time t must face:

(i) identical values for the explanatory variables $(X_{it} = X_t)$

(ii) identical truncation points $(v_{it} = v_t)$

Actually, despite the formidable conclusion we have just reached, it turns out that Tobit analysis can still be applied to aggregate data even when individuals are not essentially identical. That is, if the X_{it} and v_{it} are normally distributed we can still estimate the parameter vector β, but not the standard error σ.

Let

$$X_{it}\beta = X_t\beta + \delta_{it}$$

$$v_{it} = v_t + \eta_{it}$$

where δ_{it} and η_{it} are normal with zero mean. The model of equations (1) and (2) can be rewritten as:

$$(7)\quad y_{it} = \begin{cases} X_t\beta + \varepsilon_{it} + \delta_{it} & \Longleftrightarrow \quad I_{it} = 1 \\ v_t + \eta_{it} & \Longleftrightarrow \quad I_{it} = 0 \end{cases}$$

$$(8) \qquad I_{it} = \begin{cases} 1 & \iff X_t\beta + \varepsilon_{it} + \delta_{it} < v_t + \eta_{it} \\ 0 & \iff X_t\beta + \varepsilon_{it} + \delta_{it} \geq v_t + \eta_{it} \end{cases}$$

As before, these equations can be combined to solve for y_{it}:

$$(9) \qquad \begin{aligned} y_{it} &= I_{it}(X_t\beta + \varepsilon_{it} + \delta_{it}) + (1-I_{it})(v_t + \eta_{it}) \\ &= v_t + I_{it}(X_t\beta - v_t) + I_{it}\gamma_{it} + \eta_{it} \end{aligned}$$

where

$$\gamma_{it} = \varepsilon_{it} + \delta_{it} - \eta_{it}$$

and therefore

$$\gamma_{it} \sim N(0,\tau^2)$$

As in equation (4) we can take the expectation of y_{it} (noting that $E(\eta_{it})=0$) and obtain:

$$(10) \qquad E(y_{it}) = v_t + (X_t\beta - v_t)\, F\left(\frac{v_t - X_t\beta}{\tau}\right) - \tau\, f\left(\frac{v_t - X_t\beta}{\tau}\right)$$

Finally, since $E(y_{it})$ has the same value for all i:

$$(11) \qquad E(y_t) = E(y_{it})$$

Consequently, we can still use Tobit to estimate β even if we only have observations on the means y_t, X_t and v_t. However, the variance estimated, τ^2, is that of the composite term γ_{it}:

$$(12) \qquad \begin{aligned} \tau^2 &= \mathrm{var}(\gamma_{it}) \\ &= \mathrm{var}(\varepsilon_{it} + \delta_{it} - \eta_{it}) \\ &= \mathrm{var}\ \varepsilon_{it} + \mathrm{var}\ \delta_{it} + \mathrm{var}\ \eta_{it} + 2\mathrm{cov}(\varepsilon_{it},\delta_{it}) \\ &\qquad - 2\mathrm{cov}(\varepsilon_{it},\eta_{it}) - 2\mathrm{cov}(\delta_{it},\eta_{it}) \end{aligned}$$

Presumably, ε_{it} will be independent of either δ_{it} or η_{it}, in which case:

$$(13) \qquad \tau^2 = \text{var } \varepsilon_{it} + \text{var } \delta_{it} + \text{var } \eta_{it} - 2\text{cov}(\delta_{it}, \eta_{it})$$

Furthermore, it is quite possible that either δ_{it} or η_{it} will have zero variance. For example, each individual faces the same truncation point. Then

$$(14) \qquad \tau^2 = \text{var } \varepsilon_{it} + \text{var } \delta_{it}$$

And, of course, if both δ_{it} and η_{it} have zero variance, then

$$(15) \qquad \tau^2 = \sigma^2 \ (\equiv \text{var } \varepsilon_{it})$$

One should note that in order to estimate β it is necessary that τ be constant over time (or at least that τ be a known function of time). This in turn implies that the various constituents of τ, itemized in equation (12), themselves be constant (or well known functions of time).

C. Applications

Maddala and Nelson (1975) investigate the problem of Tobit analysis with aggregate data in the context of bank interest rates. The upper bound on permissable interest rates is set by Regulation Q. The pattern of interest rates in the absence of this regulation could be inferred by Tobit analysis if disaggregated data on individual banks were available. Maddala and Nelson show that if only data on mean interest rates (unweighted by the varying size of deposits at different banks) are available, then Tobit can still be used to estimate β and σ. In essence, the restrictions they impose are that all banks face identical values of the explanatory variables ($X_{it}=X_t$) and truncation points ($v_{it}=v_t$).

Johnson (1982) encounters the aggregation problem in estimating crop yields. Sales of the crop are limited by a quota, which suggests the application of Tobit. However, the observations are only on aggregate data: sales (s), area cultivated (a), and quotas (q). Is Tobit appropriate?

If we denote mean yields by μ, we have the following model of observed yields:

$$(19) \quad y_{it} = \begin{cases} \mu_{it} + \eta_{it} & \iff \quad \mu_{it} + \eta_{it} < v_{it} \\ v_{it} & \iff \quad \mu_{it} + \eta_{it} \geq v_{it} \end{cases}$$

where

$$y_{it} = s_{it}/a_{it}$$

$$v_{it} = q_{it}/a_{it}$$

While mean yields μ_{it} may vary from farm to farm, it is reasonable to assume that they are normally distributed about the group mean μ_t. As for the truncation point v_{it} it is not immediately apparent that all growers should face the same constraint. However, it turns out that if all the quotas are assigned efficiently (by which it is meant that all growers produce the last unit at the same marginal cost) then the v_{it} will be identical, at time t. Allowing for errors in calculating and assigning quotas efficiently implies variance in the v_{it}, but again it is reasonable to assume that such error is distributed normally. Consequently, the Tobit model described in equation (7) is appropriate.

II. Probit

The same conclusions extend also to the Probit model. This model is similar to Tobit, except that now we are concerned only with the expectation of I_{it}. As before

$$(16) \quad I_{it} = \begin{cases} 1 & \iff \quad X_{it}\beta + \varepsilon_{it} < v_{it} \\ 0 & \iff \quad X_{it}\beta + \varepsilon_{it} \geq v_{it} \end{cases}$$

$$(17) \quad E(I_{it}) = \Pr\{\varepsilon_{it} < v_{it} - X_{it}\beta\}$$

$$= F\left(\frac{v_{it} - X_{it}\beta}{\sigma}\right)$$

If observations are only available on the means X_t and v_t, and

$$X_{it}\beta = X_t\beta + \delta_{it}$$

$$v_{it} = v_t + \eta_{it}$$

then I_{it} can be written as in equation (8) and

$$(18) \quad E(I_{it}) = \Pr\{\varepsilon_{it} + \delta_{it} - \eta_{it} < v_t - X_t\beta\}$$

$$= F\left(\frac{v_t - X_t\beta}{\tau}\right)$$

and

$$(19) \quad E(I_t) = E(I_{it})$$

Note that while I_{it} can only take on values of 0 or 1, I_t can fall anywhere on the interval (0,1). Furthermore, by the Central Limit Theorem the means I_t are asymplotically normal if the I_{it} are independent and identically distributed (i.e. $X_{it}\beta - v_{it} = X_t\beta - v_t$, for all i at time t). This attribute has led to the use of mean data when the underlying distribution is not known. A prime example is the analysis of pesticide dosages on insect mortality, when all the insects are assumed to be drawn from the same population. The implication of equation (18) is that it is not necessary that the insects all be drawn from the same population: rather, it is sufficient that the distribution of insects from each population be normal, and that this distribution be stable over time.

III. Summary

Tobit and Probit models are formulated for observations on individuals. If the data consist only of observations on the mean, then in general neither Tobit nor Probit is appropriate except in the unusual event that each of the individuals has identical characteristics. However, if these characteristics differ from individual to individual, but the pattern of these characteristics is itself normal, then Tobit and Probit methods can be employed to estimate the mean ($X_t\beta$) but not the variance (σ^2).

REFERENCES

Johnson, Frederick I. (1982) "Estimating Supply Despite a Quota: Brazil's Supply of Sugar," Economic Inquiry XX(1) pp. 54-71.

Lee, Lung-Fei (1975) "Limited Information Estimation of Some Switching Regression Models," University of Rochester, Department of Economics Discussion Paper 75-20.

Maddala, G.S. and Forest D. Nelson (1975) "Specification Errors in Limited Dependent Variable Models," National Bureau of Economic Research Working Paper (Preliminary: not numbered).

Tobin, James (1958) "Estimation of Relationships for Limited Dependent Variables," Econometrica XXVI, pp. 24-36.