Why and When to Use Tobit Analysis

Cynthia Fraser

and

Yoram Wind

May 1986

Working Paper 86-2

Cynthia Fraser is Associate Professor and Research Fellow, The International Trade Institute and Department of Marketing, Kansas State University, Manhattan, Kansas 66506. Yoram Wind is Director and The Lauder Professor at The Warton School, University of Pennsylvania, Philadelphia, PA 19174.

I. Introduction

Market researchers are often interested in examining both probability (of purchase ownership, use) and extent (number of units purchased, frequency of purchase, price paid, etc;) of consumer behavior. For example, we may be concerned with estimation the effects of promotion, advertising, family income, and education on both probability of purchase of a microwave oven and price those purchasing will pay. We would like to be able to (i) infer the effects of predictor variables on both probability and extent of purchase, (ii) classify prospective purchasers into "likely" and "unlikely" groups; and (iii) accurately predict extents of purchase. "Limited dependent variables," complicate analysis because we have observations on independent variables for the total sample, but no observations on extent of purchase for the subsample of nonpurchasers, as shown in Figure 1.

Several approaches are available for dealing with limited dependent variable problems. To infer predictor variable effects on probabilities of purchase and to subsequently classify potential purchasers into likely and unlikely groups, Discriminant Analysis, or nonlinear probit or logit regression may be utilized. For inference of the effects of predictor variables on extents of purchase and subsequent prediction of extents of purchase, Ordinary Least Squares (OLS) may be utilized on either the full sample of purchasers and nonpurchasers (for whom we assume extents of purchase are equal to zero) or on the subsample of purchasers. Alternatively, Tobit regression (Tobin, 1958) may be utilized to simultaneously (i) infer predictor variable influence on both probabilities and extents of purchase, (ii) classify prospective customers, and (iii) predict extents of purchase. Choice between these alternatives depends upon assumptions concerning the nature of the data and the major research objectives at hand. No single approach is preferable under all conditions.

In this paper we briefly examine the assumptions underlying these alternative procedures and compare performance of Ordinary Least Squares, Discriminant Analysis, and Tobit regression when applied to simulated data generated under a variety of known conditions. First, we briefly examine the approaches to be compared. Then, the experimental design and performance criteria are described. Finally results of the experiment are reported and implications for treatment of limited dependent variables discussed.

II. Three Alternative Approaches for Dealing With Limited Dependent Variables

Limited Models

Ordinary Least Squares

Given data containing a limited dependent variable, one is tempted to set extents of purchase to zero among non-purchasers and simply regress extent of purchase against predictors (such as marketing mix variables, demographic, or lifestyle measures) via Ordinary Least Squares. Individuals with predicted extents of purchase which fall below some minimum (e.g., zero, the lowest possible price, etc.) may be classified as non-purchasers. The weakness inherent in such a procedure lies in the concentration of zero values "observed" in the non-purchase sub-sample. This will cause the fitted regression line to be too flat (Goldberger, 1964); consequently, coefficient estimates will be biased small.

Twin Linear Model

A second approach consists of using OLS to examine the influences of predictors on extents of purchase (within the subsample of purchasers) and to predict extents of purchase, and (ii) Discriminant Analysis to investigate predictor variable effects on probabilities of purchase and to classify potential customers. This is analogous to Goldberger's Twin Linear Model (1964), with the substitution of linear Discriminant Analysisfor binary dependent variable linear regression for classification purposes.

There are two basic problems with this approach: lack of efficiency and probable specification bias. Lack of efficiency stems from the fact that we ignore information on extents of purchase in Discriminant Analysis, and we discard the subsample of non-purchasers in the OLS regression. In addition, heteroskedasticity associated with the dummy dependent variable in Discriminant Analysis reduces efficiency of the resulting parameter estimates.

Specification bias arises when predictor variable effects on probabilities of purchase are nonlinear. Discriminant analysis assumes that (i) group memberships are nonstochastic, (ii) predictor variable values are normally distributed within groups, and (iii) "probability" of purchase is a linear function of the independent variables, which is not likely. More reasonable are the assumptions that (i) purchases are probabilistic and (ii) a given marginal change in probability of purchase is more difficult to obtain when probability is close to either zero or one. To illustrate this latter point, consider the effect of family lifecycle and advertising on the probability or life insurance purchase. The majority of heads of households with children to support and mortgages to pay, possess life insurance; in this case, heavy advertising will have virtually no effect on probability of ownership. Conversely, the effect of such advertising on newly married couples should be substantial. Thus, in the first case, probabilities of ownership are almost certain, and, as a result, increases in marketing efforts presumably have little effect, while, in the latter case, in which probabilities are at moderate levels, marketing effects on purchases are potentially more dramatic.

Because of the truncated nature of observed extents of purchase (equal to zero for non-purchasers) and probable nonlinear functional form, linear models are conceptually inappropriate for analysis of data containing limited dependent variables.

Nonlinear Models

Tobit

If we are solely concerned with probability of purchase, nonlinear probit and logit regression are appropriate analytical procedures. If extent of purchase is also a concern, Tobit regression (Tobin, 1958) is appealing, in that it allows simultaneous examination of both probability and extent of purchase. We will focus discussion on the Tobit model. (An excellent discussion of probit and logit models is contained in Hanushek and Jackson, 1977.)

The Tobit model, which has been applied to marketing data by Parsons (1981) and Elrod and Winer (1980), is a hybrid of probit and OLS regressions. It assumes that for each individual, there exists an index (e.g., of desire or intent to purchase) which is a linear function of the predictors:

$$I_t = \underset{\sim}{X}_t' \underset{\sim}{\beta} \qquad (1)$$

Where I_t = (1 x 1) index of intent/desire to purchase for person t;

$\underset{\sim}{X}_t'$ = (1 x K) vector of observations on K independent variables for person t;

and $\underset{\sim}{\beta}$ = (K x 1) vector of coefficients.

If this index exceeds the individuals "threshold," purchase occurs. Extent of purchase is also a function of the predictors, through the index. Thus, the greater the intent/desire to purchase, the greater the extent of purchase:

$$\begin{aligned} y_t &= 0 \text{ if } I_t < I_t^* \\ &= I_t - I_t^* \text{ if } I_t > I_t^* \end{aligned} \qquad (2)$$

where y_t = (1 x 1) dependent variable representing extent of purchase by person t;

I_t^* = (1 x 1) critical threshold for person t.

Note that each person may have a different threshold value. Thus, if advertising is a predictor variable, for example, more advertising may be required to push one person over his threshold than that required to induce anotner person's purchase.

Since individual thresholds differ, at any given index value, there will be both a concentration of zeros (for nonpurchasers) and a distribution of positive extents of purchase (for purchasers). Probability of purchase, given a particular index value, is

$$\text{Prob } \{y>0|I\} = \text{Prob } \{I^*<I|I\} = F(I/\sigma)$$

and

$$\text{Prob } \{y= 0|I\}= \text{Prob } \{I^*> I|I\}= 1-F(I/\sigma), \qquad (3)$$

where $F(I_t/\sigma)$= value of the standard normal cumulative distribution at I/σ . Expected extent of purchase, given a particular index value is

$$E(y_t|I_t) = I\ F(I_t/\sigma) + \sigma f(I_t/\sigma), \qquad (4)$$

where $f(I_t/\sigma)$ = value of the standard normal density distribution at I_t/σ . Estimation of $\underset{\sim}{\beta}$ and σ is accomplished through maximum likelihood, since the functional form is non-linear.

III. The Simulation Experiment

Experimental Design

Perrormance of the three alternative approaches was examined in the context of a simulation experiment. Since true parameters are unknown in real datasets, use of simulated datasets allows exploration of the effects of a number of data components on performance of the estimators being examined. In this experiment, a full factorial design was utilized to generate 8 unique and varied datasets, each of which was replicated 100 times. Datasets differed with respect to (i) number (T) of observations in each sample, and (ii) the underlying model of the "true" buyer behavior world and (iii) the relative size of the error component. In all cases, two predictor variables were used.

Two sample size levels were utilized (T = 30, 120). To generate the matrices X or standard multivariate normal independent variable observations, the International Mathematical and Statistical Library (IMSL) Subroutine GGNSM was utilized.

Two underlying models of the purchase world were utilized to generate the datasets, one of which corresponded to a "Tobit world" and one of which corresponded to a "Twin Linear world." In the Tobit world, the T observations on 2 predictors were drawn from a common sample X. Dependent variable observations y on extent of purchase were generated by (i) subtracting standard normal random errors I* from the product $\underset{\sim}{I} = X\underset{\sim}{\beta}$, and (ii) setting $y_t = 0$ if the difference $I_t - I_t^*$ was negative:

$$\begin{aligned} y_t &= I_t - I_t^* \text{ if } I_t - I_t^* \geq 0 \\ &= 0 \qquad \text{if } I_t - I_t < 0, \end{aligned} \tag{5}$$

where $I_t^* \sim N(0,\sigma^2)$.

In the "Twin Linear" world, (T/2 x 2) purchase and nonpurchase data matrices X_0 and X_1 were generated separately with unique mean vectors:

$$\underset{\sim}{\mu}'_0 = [0 \; 0]; \quad \underset{\sim}{\mu}'_1 = \beta' R, \tag{6}$$

where R = correlation matrix between predictors, equal to .1 on the off-diagonal. Dependent variable observations $\underset{\sim}{y}$ on extent of purchase were set equal to zero in the first subsample and made a linear function of predictors in the second subsample:

$$\underset{\sim}{y}_1 = X_1\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}, \qquad (7)$$

where

$$\varepsilon \sim N(0,\sigma^2).$$

To manipulate the amount of error present on the data, two values were utilized: 1, .2. Coefficients $\underset{\sim}{\beta}$ were selected so that $\underset{\sim}{\beta}'\underset{\sim}{\beta}$ =1:

$$\beta_k = ½ , \qquad (8)$$

in which case the signal to noise ratio was equal to one of two levels

$$(\underset{\sim}{\beta}'\ \underset{\sim}{\beta}/\ \sigma^2 = 1,\ 25).$$

IMSL Subroutines RLMUL and ODNORM were utilized to obtain OLS regression, and Discriminant Analysis Coefficient estimates. A proprietary program, LIMDEP, was utilized to obtain Tobit coefficient estimates.

Performance Criteria

In the assessment of relative performance of the three approaches, we focus on estimation, prediction, and classification.

To compare "quality" of the coefficient estimates produced, the trace of the mean square error matrix was utilized:

$$\text{MSE }(\hat{\underset{\sim}{\beta}}) = \hat{\sigma}^2 \text{ tr}\Sigma^{-1} + E(\ \hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}\ )\ 'E(\ \hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}\ ). \qquad (9)$$

Since mean square error is the sum of variances and squared bias of the estimates, this statistic constitutes a composite measure of the inefficiency and biasedness of coefficients produced by the alternative estimators. We examine the average of mean square errors (over on hundred replications).

In order to evaluate predictive capability of the three approaches, hold-out samples (equal in size to samples used for coefficient estimation) were utilized to calculate mean square prediction errors of extents of purchase (y) among the subsamples of purchasers:

$$\text{MSPE } \underset{\sim}{y} = (\hat{\underset{\sim}{y}} - \underset{\sim}{y})\ '(\hat{\underset{\sim}{y}} - \underset{\sim}{y})\ /T_1 \qquad (10)$$

where T_1 = number of observations in the subsample of purchasers. As before, we examine the average of mean square prediction errors(across one hundred replications).

To assess classification capabilities, observations in the hold-out sample were classified by each of the three models into purchase and non-purchase groups. We examine the average of percents correctly (over one hundred replications).

Results of the Experiment

The averages of mean square error statistics are presented in Table 1 by sample size (T), signal to noise ratio (S/N) and model of the purchasing world for each estimator. Boxed entries denote superior performance at a 95 percent level of confidence. To assess the relative performances and sensitivities of the alternative estimators to data characteristics analysis of variance were used to analyze the experimental data. Results are shown in Table 2.

Mean Square Coefficient Error

Overall, Ordinary Least Squares on the subsample of purchasers (OLSS) produced truer coefficient estimates than other estimators. When models producing the data are accounted for, however, OLSS is superior only when the Twin Linear world is true. When a Tobit model generated data, Tobit reproduced coefficients best. Ordinary least squares on the full sample was never best and was equal only when the underlying model was weak (signal-to-noise low).

Signal-to-noise variation accounts for 35% of the variation in performances. When models were weak (error high), little difference in performance exists between OLSS and OLSF; when error is low, OLSS and Tobit perform better then OLSF.

Prediction of Extents of Purchase

Generally, OLSS produced the most accurate predictions of extent, although Tobit and Ordinary Least Squares on the full sample (OLSF) were equally accurate among large samples (T=120), weak models (S/N = 1), and Tobit model assumptions.

Signal-to-noise variation accounted for 79 percent of the variations in performance. OLS-S benefits most from strengthened models (reduced error), out performing Tobit and OLSF when error is small.

Classification

Results indicate that both Discriminant Analysis (DA) and Tobit consistently classify a larger proportion of the hold-out sample than does OLSF. This pattern prevails, regardless of sample size, signal-to-noise ratio, or true model of the purchase world. Performance of both Tobit and DA improves as models strengthen.

Summary

The biggest difference between estimators is observed in inference. When the Twin Linear world is true, the corresponding estimator, Ordinary Least Squares on the purchasing sample, is more accurate than Tobit; when the Tobit model reflects the data structure, Tobit is the most accurate estimator. Discriminant analysis is clearly least accurate. A similar pattern emerges when prediction accuracy is assessed.

For classification, Tobit and Discriminant Analysis are equally preferred. Both are more accurate than simple regression (OLSF).

IV. Conclusions

We have examined three alternative approaches for use in cases where a proportion of observations on the dependent variable are missing. This limited dependent variable problem occurs frequently in cases where extents of purchase are observed among purchasers (users, owners), but are essentially zero for nonpurchasers. Choice among the analytical procedures examined depends largely upon the underlying model which is valid.

These results are not terribly surprising. That the Twin Linear and Tobit models compare favorably to Ordinary Least Squares is to be expected, since the former two approaches were designed to improve upon deficiencies encountered when OLS was utilized to analyze data with limited dependent variables. When either model is appropriately specified, it performs well.

The choice between a Twin Linear approach and Tobit depends upon accurate choice of the model having generated the data. In rare instances, consumer behaviors (usage, purchase, etc;) might be non-stochastic. Physical differences differentiate users and non-users of contact lenses, for instance. In most cases, however, consumer behaviors are probabilistic. When one assumes that behaviors may be influenced by marketing efforts, one implicitly assumes tnat those behaviors are probabilistic. Consequently, logic suggests that a Tobit-type model generates the consumer behaviors that marketing researchers are typically interested in examining.

If a Tobit specification is appropriate, one should clearly not use a Twin Linear approach for inference. Coefficient estimates from least squares and discriminant analysis are significantly less accurate. One should restrict use of discriminant analysis to classification.

Table 1. Marginal Criteria Values

	Sample Size		Signal-to-noise		Model Generating Data		
Logarithm of mean square coefficient error	30	120	1	25	Tobit	Twin Linear	Overall
OLS-S:	-2.3	-3.3	-1.5	-4.2	-2.5	-3.1	-2.8
Tobit:	-1.7	-2.9	-1.5	-3.1	-3.1	-1.4	-2.3
OLS-F:	-1.3	-1.5	-1.4	-1.5	-1.3	-1.6	-1.4
DA:	-0.6	-0.6	-0.5	-0.7	-0.8	-0.4	-0.6
Logarithm ot Mean Square Prediction Error							
OLS-S:	-1.6	-1.5	-0.3	-2.8	-1.5	-1.6	-1.5
Tobit:	-0.7	-1.1	0.1	-1.9	-1.3	-0.4	-0.9
OLS-F:	-0.9	-0.9	-0.2	-1.6	-1.2	-0.6	-0.9
Percent Correctly Classified							
Tobit:	0.76	0.80	0.73	0.83	0.79	0.77	0.78
OLS-F:	0.64	0.64	0.60	0.68	0.66	0.62	0.64
DA:	0.78	0.80	0.74	0.84	0.83	0.75	0.79

References

Elrod, Terry and Rüssel S. Wimer (1980), "An Empirical Comparison of Aggregation Criteria for Developing Market Segments," Journal of Marketing

Goldberger, A. (1964). Econometric Theory. New York: John Wiley & Sons.

Hanushek, E. and J. Jackson (1977), "Models with Discrete Dependent Variables," Chapter 7, Statistical Methods for Social Scientists. New York: Academic Press.

IMSL Library, Edition 8. (1980). Houston, Texas: IMSL.

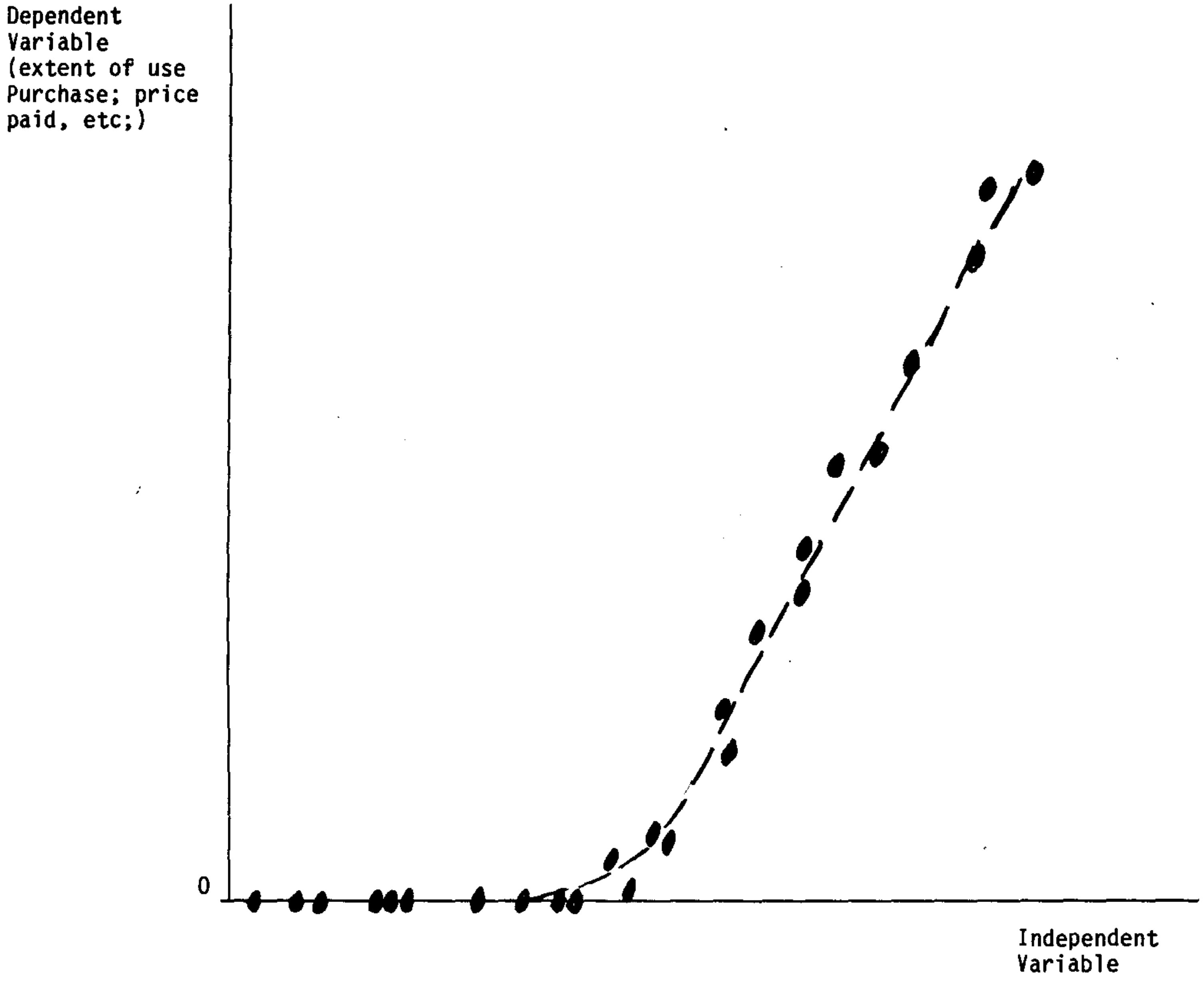
Tobin, J (1958), "Estimation of Relationships for Limited Dependent Variables," Econometrica 25, 24-36.

Table 2. Analyses of Variance in Criteria

Criteria:	mean square coefficient error			mean square prediction error			average percent correctly classified		
Source	SS	DF	F	SS	DF	F	SS	DF	F
Estimator	22.8	3	115.6**	2.2	2	7.8**	.11	2	21.5**
Estimator *T	4.4	4	16.7**	.3	3	.8	.00	3	.5
*S/N	19.1	4	72.8**	24.9	3	58.9**	.06	3	7.5**
*world	7.3	4	27.9**	2.4	3	5.8*	.02	3	2.3
Error	1.1	16		1.7	12		.03	12	
Total:	54.7			31.6			.22		
Model F:			54.4**			19.3**			6.72**
R^2:	.98			.95			.86		

**Significant at a ninty-nine percent level of confidence.
*Significant at a ninety-five percent level of confidence.

Hypothetical Data Containing a
Limited Dependent Variable
Figure 1.


Dependent
Variable
(extent of use
Purchase; price
paid, etc;)
Independent
Variable