Statistics

1974

ESTIMATING BOUNDED REGRESSAND FUNCTIONS AND THE HYBRID MULTIVARIATE PROBIT APPLICATION

Paul K. Kau *

1. INTRODUCTION

In applied economic research, it is often found that the value which an economic variable can take is inherently bounded within certain range or bounded at one end. Among these bounded variables, some take continuous values within the lower and upper bounds while other can only take certain discrete values. A variable measured with "ratio" or "percent", for instance, will take any continuous values between 0 and 1 or 0 and 100. However, a binary variable such as whether a farmer purchases a particular piece of farm machinery or not, or whether a consumer chooses a particular brand or not, will take only a dichotomous values of either 0 or 1 in terms of probabilities.

When a bounded economic variable is designated as a dependent variable in a regression function, a researcher encounters many methodological problems and difficulties in statistical estimation. Both the dichotomous regressand problem and the bounded continuous regressand problem have caught great deal of attention among economists and econometricians in the past and many studies and alternative approaches have been discussed. Among the former

---

* Paul K. Kau is an Associate Professor of Economics at the Alcorn State University.

Contributed paper, AAEA meetings, College Station, Texas, Aug. 1974.

the Linear Probability Function [[illegible]], Logit [[illegible]], Discriminant Analysis [2, [illegible]], and Multivariate Probit [[illegible], 7]. Alternative models for the latter includes the Twin Probability Function model [5] and Hybrid Multivariate Probit [14]. Given the problems and situation stated above, the objectives of this paper are: (1) to discuss the major methodological problems encountered in estimating various bounded regressand functions and some possible alternatives, (2) to suggest a more generalized model in which the dichotomous regressand problem can be treated as a special case of bounded regressand problem, and (3) to compare the statistical properties and applicability of these models.

2. METHODOLOGICAL CONSIDERATIONS

Economists are often interested in estimating relationship of a dependent variable which reflects the behavior or decision of an economic unit in response to a set of economic stimuli. When the dependent variable is dichotomous, the application of the linear probability function model— a most commonly used model to dichotomous regressand problem encounters several serious difficulties: First, the variance of the disturbance-term varies according to the value of each different observation and is heteroscedastic [5, p.249]. Thus, the ordinary classical assumption of homoscedasticity is violated. Although the ordinary least squares estimator is still an unbiased it is no longer an efficient one. Second, the expected value of the regressand Y may sometimes become greater than 1 or less than 0 [4,5]. This is probably a more serious problem than the first one because it is inconsistent with the definition of a dichotomous regressand and the interpretation of the expectation of Y as a probability.

Application of Aitken's generalized least squares will overcome the heteroscedasticity problem. However, further difficulties will be faced if $E(Y) > 1$ or $E(Y) < 0$, because the weighting factor $\frac{1}{\sqrt{E(Y)[1-E(Y)]}}$ [9, p.208] will become indeterminable, or if the variance-covariance matrix is singular. To handle the problems of $E(Y) > 1$ or $E(Y) < 0$, various approaches have been suggested, including from the simpler method of truncating Y values [6, 8] to more complicated methods of logit [11], restricted least squares with quadratic programming [13], and Zellner and Lee's joint estimation techniques [15].

Even with the two above problems solved, it has been argued that the dichotomous decision or behavior implies that there exists a "breaking point" or "threshold" in the dimension of explanatory variables and the threshold phenomenon is commonly observed in all human behavior and reaction. Therefore, a model which incorporates threshold concept and possesses a threshold functional form is more appropriate [illegible]. The multivariate probit analysis used to estimate parameters in a threshold decision model appeared to overcome these statistical shortcomings mentioned above and possesses several desirable properties from methodological point of view {7}.

However, economists are sometimes interested in acquiring information, not only as to whether or not an event occurs (dichotomous) but also the magnitude of occurrence if it does occur. If Y is designated as dependent variable, then such Y is bounded at a lower limit of 0. The statistical difficulty encountered in this kind of bounded continuous regressand problem is that there will be a "concentration of points" at the limited end and thus the use of ordinary regression will cause a biased estimation {4}. An alternative approach to this type of problem is the twin probability function model which employs two steps in estimation process. An obvious weakness of this approach is that since there is no device to restrict the range of Y, some expected value of Y may excess the bounded range. The other weakness as Goldberger pointed out, is that the approach does not directly attempt to fit E(Y) to Y and takes no explicit account of the estimation of E(Y) {5}.

3. A GENERALIZED MODEL

A more appropriate approach to the bounded regressand problem is the hybrid multivariate probit analysis (or sometimes refered as Tobit) [4,5]. The hybrid probit model is a hybrid of probit analysis and multiple regression and therefore, is a more generalized model in terms of its applicability.

In hybrid probit analysis, the value of dependent variable Y is determined as:

$$Y = \begin{cases} L, \text{ if } I - I^* < L \\ I - I^*, \text{ if } I - I^* \geq L \end{cases} \qquad (1)$$

where L is the lower limit of Y; I is an index number which is assumed to be a linear combination of all X's. That is,

$$I = B_1X_1 + B_2X_2 + \ldots\ldots + B_mX_m \qquad (2)$$

I^* is defined to be a critical value of I and plays the role of disturbance forces. While the distribution of I^* in probit analysis is assumed to be N(5,1), it is assumed to be $N(0,\sigma^2)$ in hybrid probit model.

The term $(I - I^*)$ represents the net magnitude which determines the value of Y. Therefore, it is natural to set $Y = L$ if $(I - I^*)$ is less than its lower limit. But if $(I - I^*)$ is equal to or greater than its lower limit, Y is set exactly equal to that value. Following a parallel statistical theories and mathematical derivation, the conditional expectation of Y can be derived to be [14]:

$$E(Y|I,L) = L\cdot[1 - F(\frac{I-L}{\sigma})] + I\cdot F(\frac{I-L}{\sigma}) + \sigma\cdot f(\frac{I-L}{\sigma}) \qquad (3)$$

It can be seen that if Y is a dichotomous regressand and has the lower limit L = 0 and can only take a positive value 1, equation (1) will reduce to:

$$Y = \begin{cases} 0, \text{ if } I < I^* \\ 1, \text{ if } I \geq I^* \end{cases}$$

which is the exact specification used in probit analysis [5,15] Furthermore, in the case where L = 0, equation (3) also reduces to $E(Y|I) = I\cdot F(\frac{I}{\sigma}) + \sigma\cdot f(\frac{I}{\sigma})$ which is exactly the same formula Goldberger obtained [5, p.254] because L = 0 was the condition specified in his discussion. Hence, it is clear that the hybrid probit is a more general model in which probit analysis is merely a special case where the bounded regressand Y has the lower limit L = 0, and take only one positive value Y = 1 for any Y > 0.

The hybrid probit model just discussed also possesses desirable statistical properties. Using a given set of data, it was found that the locus of the estimated values derived from equation (2) which is called maximum likelihood line is linear [4]. The locus of expected values obtained from equation (3) which may be called hybrid probit curve is non-linear. If, for purposes of comparison, one were to apply ordinary regression methods to obtain a third line-- linear regression least-squares line, it can be found that there exists a certain pattern of relationships among these three different approaches of estimation. The ML-line, in general, asymptotically approaches the HP-curve at one end and the RLS-line from the other end of the sample range. Since the multiple regression approximates this non-linear relationship with a linear function, its estimation is found to be fairly accurate and acceptable only in the central range of the values of the sample, leaving both ends with possible large discrepancies. Therefore, it is conceivable that the application of multiple regression should be avoided where there is a concentration of points at the upper or the lower limit of the dependent variable.

To test the applicability and statistical performance of the model, a set of panel data[1/] contains 676 Illinois corn farmers was used to estimate the relationship between the percent of corn sold to country elevators (as dependent variable) and a set of selected farm variables. All the

---

1/ These data were collected as part of research conducted under the North Central Regional Marketing Committee NC-104, and were permitted to use from University of Illinois.

signs of estimated coefficients are consistent with both theoretical consideration and empirical observation. The estimation yielded a fairly satisfactory results with two variables, namely dryer ownership and convenience showing as extremely significant determinants (above 99% of significance). Another two variables, price importance and the number of price-checks, also are significant at above 95% level. Since the hybrid probit estimates $\hat{h}_i$ are the estimators of the corresponding parameters B_i/σ, the B_i's in equation (2) can be derived by using the relation of $\widehat{B_i} = (\hat{\sigma})(\widehat{h_i})$. The converted $\widehat{B_i}$'s are also shown in the table.

In order to compare the differences between hybrid probit estimates and ordinary regression estimates, multiple regression analysis was also conducted by using the same set of data and variables. As can be seen in the table, the signs and levels of significance for each estimated coefficient from multiple regression are rather compatible with those resulted from hybrid probit. However, a remarkable difference was found in the size of estimated coefficients. The values of all least-squares estimates $(\hat{b}_i)$ appeared to be slightly smaller than those estimated from hybrid probit $(\hat{h}_i)$. The statistical reason is that since Y has a lower limit of 0 and can not take any negative value, there will be a concentration of points at Y = 0. As a result, the estimated regression relation will necessarily be forced within the first quadrant and the $\widehat{b_i}$'s will be under-estimated because of the absence of negative Y's in the sample [14, 5].

4. SUMMARY

From this study it is implied that when the value of a regressand is inherently bounded, the use of ordinary regression analysis is, in general, inappropriate. The application of multiple regression to dichotomous regressand function encounters problems of heteroscedasticity and $E(Y) > 1$ or $E(Y) < 0$. Alternatives developed to overcome these difficulties are either ignore or insufficient to incorporate threshold concept into estimation process. Multivariate probit model avoids the statistical pitfalls mentioned above and is particularly appropriate for problems where the dependent variable is dichotomous and where threshold effect is evident in decision or behavior process.

Application of ordinary regression to bounded regressand problem should be avoid if possible, for the "concentration of points" will lead to a biased estimation. Although there are models applicable to this kind of situation, the hybrid probit is more appropriate due to its adoptability and statistical properties. It is a generalized model for estimating all type of bounded regressand problems of which dichotomous regressand is merely a special case.

RESULTS OF ESTIMATION FROM HYBRID MULTIVARIATE PROBIT

AND MULTIPLE REGRESSION MODELS

Variable 1/	Hybrid Probit Maximum Likelihood Estimates ($\hat{h}_i$)	Standard Error of $\hat{h}_i$	Converted $\hat{B}_i$	Multiple Regression Estimates $\hat{b}_i$	Standard Error of $\hat{b}_i$
Corn Storage Capacity	0.0000039	0.0000026	0.00019	0.00016	0.0001
Number of Price Checks	-0.0842	0.0367	-4.009	-3.071	1.292
Truck Capacity	0.0078	0.0055	0.371	0.268	0.213
Ownership of Dryer	-0.2635	0.0890	-12.524	-11.004	3.439
Loyalty Important in Selecting an Outlet	0.0970	0.0946	4.619	3.861	3.676
Price Important in Selecting an Outlet	-0.1706	0.0871	-8.143	-6.977	3.366
Convenience Important in Selecting an Outlet	0.2235	0.0931	10.643	8.667	3.587
Grain was Sold at Harvest	0.0703	0.0811	3.348	2.041	3.142
Reciprocal of Estimated Standard Deviation	0.0211	0.0007			

1/ The dependent variable is the percent of corn sold to country elevators by each respondent in the survey in 1971.

REFERENCES

1. Berkson, J., "Application of the Logistic Function to Bio-Assey," Journal of the American Statistical Association, Vol. 39, pp. 357-365.

2. Cooley, W. W. and Lohnes, P. R., Multivariate Procedures for the Behavioral Sciences, New York, John Wiley and Sons, 1962.

3. Cragg, John G., "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," Econometrica, Vol. 39, No. 5, September 1971, pp. 829-844.

4. Dagenais, M. G., "A Threshold Regression Model," Econometrica, Vol. 37, April 1969, pp.193-203.

5. Goldberger, Arthur S., Econometric Theory, John Wiley and Sons, Inc., New York, 1964.

6. Hallberg, M. C., "Projecting the Size Distribution of Agricultural Firms-- An Application of a Markov Process with Non-Stationary Transition Probabilities," American Journal of Agricultural Economics, May 1969, pp.289-302.

7. Hill, L. and P. Kau, "Multivariate Probit as a Research Tool," Chapter IV, Farm/Ranch Input Research--Yesterday, Today and Tomorrow, North Central Regional Research Publication No. 215, sponsored by Economic Research Service, The United States Department of Agriculture, 1973.

8. Hill, L. D., "Use of Weighted Regression in Estimating Models of Binary Choice," Canadian Journal of Agricultural Economics, November 1970, pp.64-72.

9. Johnston, J., Econometric Methods, McGraw-Hill Book Company, Inc., New York, 1963.

10. Kau, Paul K., "A Threshold Model of the Grain Dryer Purchasing Decisions of Illinois Farmers," Unpublished Ph.D. Dissertation, Department of Agricultural Economics, University of Illinois, 1971.

11. Kau, Paul K. and Lowell Hill, "A Threshold Model of Purchasing Decisions," Journal of Marketing Research, August 1972, pp. 264-270.

12. Ladd, G. W., "Linear Probability Models Functions and Discriminant Analysis," Econometrica, Vol. 34, No. 4, 1966, pp.873-885.

13. Lee, T. C., "Estimation of Regression Parameters with the Predicted Dependent Variable Restricted in a Certain Range: Comment," American Journal of Agricultural Economics, November 1970, pp.613-614.

14. Tobin, James, "Estimation of Relationships for Limited Dependent Variable," Econometrica, Vol. 26, January 1958, pp. 24-36.

15. Tobin, James, "The Application of Multivariate Probit Analysis to Economic Survey Data," Cowles Foundation Discussion Paper, No. 1, July 1955.

16. Zellner, A., and T.H. Lee, " Joint Estimation of Relationships Involving Discrete Random Variables," Econometrica, April 1965, pp. 382-394.