Annals of Economic and Social Measurement, 5/4, 1976

ON A GENERAL COMPUTER ALGORITHM FOR THE ANALYSIS OF MODELS WITH LIMITED DEPENDENT VARIABLES

BY FORREST D. NELSON*

Several econometric models for the analysis of relationships with limited dependent variables have been proposed including the probit, Tobit, two-limit probit, ordered discrete, and friction models. Widespread application of these methods has been hampered by the lack of suitable computer programs. This paper provides a concise survey of the various models; suggests a general functional model under which they may be formulated and analyzed; reviews the analytic problems and the similarities and dissimilarities of the models; and outlines the appropriate and necessary methods of analysis including, but not limited to, estimation. It is thus intended to serve as a guide for users of the various models, for the preparation of suitable computer programs, for the users of those programs; and, more specifically, for the users of the program package utilizing the functional model as implemented on the NBER TROLL system.

INTRODUCTION

Economic relationships involving limited dependent variables are receiving widespread attention in the Econometrics literature. Much of the discussion has focused on methodology with only scattered application to real problems, the one exception being the qualitative variable problem frequently treated with logit and probit analysis. Since potential applications for these models abound, it is likely that the scarcity of computer programs and their limited dissemination is partly responsible for the infrequency of empirical studies using them. In turn, useable computer routines may be scarce because the models though similar in many respects are dissimilar enough so as to seem to require a separate algorithm for each model.

The purpose of this note is to suggest a general functional model which is readily adap'able to computer coding and flexible enough to fit a wide variety of limited dependent variable problems.[1] It should be emphasized that the model presented here is functional as opposed to theoretical. That is, it is not advocated as *the structural* model underlying any limited dependent variable relationship. Rather we suggest that many of the theoretical relationships may be reformulated to fit this functional model so that a single computer program may be used to analyze all of them.

The terminology "limited dependent variable" is used here to denote variables endogenous to some underlying economic relationship which are not continuously measurable (or observable) over the entire real line either directly or even after some transformation such as logarithms. Thus it applies to discrete (ordinal) variables, qualitative (non-ordinal) variables and to variables subject to threshold constraints such as non-negativity. Such discontinuities may result from

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

[1] Tom Johnson [1] presents a general discussion of many of the models but falls short of describing in detail a central model around which a computer algorithm can be constructed.

theoretical considerations, from physical constraints on the variable or simply from measurement difficulties.

The effect of the discontinuities on estimation is that when such a dependent variable enters the usual sort of regression model the properties of the implied disturbance term cannot satisfy the assumptions needed for least squares estimation. The alternative estimation method generally proposed is maximum likelihood. After a suitable choice for the distribution of the disturbances is specified the distribution of the limited dependent variable is derived and the likelihood function is constructed. This typically involves both probability density and distribution functions and yields non-linear normal equations so that iterative maximization algorithms, generally Newton–Raphson, are suggested for obtaining estimates. These procedures are of course straightforward but they may become quite expensive and time consuming if computer programs do not exist for the particular model being examined.

Section I of this paper presents a brief review of a number of limited dependent variable models. Such a survey will serve to motivate the types of models to be treated and highlight their similarities and dissimilarities. In Section II the functional model is introduced. It is of course possible to outline a completely general model but the aim here is for a model which may be easily implemented in a single computer algorithm. With this goal in mind reasonable restrictions on the model are imposed and many of the details needed for implementation are discussed. A final section outlines features which should be included in a general computer algorithm.

I. Review of Some Limited Dependent Variable Models

A. *Binomial Choice Models*[2]

In these models each measuring unit or individual is faced with the choice of one of two mutually exclusive alternatives and the choice made is thought to depend on some vector of exogenous variables. One way to formalize the choice mechanism is to view the decision maker as having associated with each alternative some preference function, say

$$\begin{aligned} I_{1i} &= f_1(X_i) + v_{1i} \\ I_{2i} &= f_2(X_i) + v_{2i}, \end{aligned}$$

and choosing that alternative which yields the higher preference. Assuming $f_j(X_i)$, $j = 1, 2$, is of the form $f_j(X_i) = \alpha_j' X_i$, alternative 2 is chosen if

$$\begin{aligned} I_{2i} > I_{1i} &\Rightarrow \alpha_2' X_i - \alpha_1' X_i + v_{2i} - v_{1i} > 0 \\ &\Rightarrow \beta' X_i + u_i > 0 \end{aligned}$$

[2] These models appear to have been first examined in the context of economics by Tobin [8] who outlined the method of estimation which he termed "probit regression analysis." Theil [7], among others, treated the same problem with "logit" analysis. The distinction between the two lies in the assumptions made regarding the distribution of the underlying disturbance.

where $\beta = \alpha_2 - \alpha_1$ and $u_i = v_{2i} - v_{1i}$. The model can be rewritten in the alternative form:

$$\begin{aligned} Y_i &= \beta' X_i + u_i \\ W_i &= 0 \qquad \text{if } Y_i < 0 \\ &= 1 \qquad \text{if } Y_i \geq 0 \end{aligned}$$

where Y_i is some latent (i.e., unobserved) variable and W_i is the observed dependent variable which indicates the choice made. Maximum likelihood estimation requires some assumption about the distribution of u_i. If that distribution is normal, i.e., $u_i \sim \text{IN}(0, \sigma^2)$, the likelihood function is given by

$$L(\beta, \sigma | W, X) = \prod_{w_i=0} P\left(\frac{-\beta' X_i}{\sigma}\right) \prod_{w_i=1} P\left(\frac{\beta' X_i}{\sigma}\right)$$

where $P(x)$ represents the standard normal cumulative density function, $P(x) = \int_{-\infty}^{x} 1/\sqrt{2\pi} \exp(-u^2/2)\, du$.

Unfortunately $\beta' X_i/\sigma$ is observationally equivalent to $(k\beta)' X_i/(k\sigma)$ where k is any positive constant so that σ is not estimable and β is estimable only up to a scale factor. Thus we estimate $\alpha = (1/\sigma)\beta$, say, which is equivalent to normalizing σ at unity.

An interesting related model is

$$\begin{aligned} Y_i &= \beta' X_i + u_i \\ W_i &= 1 \qquad \text{if } Y_i \leq Z_i \\ &= 0 \qquad \text{if } Y_i > Z_i \end{aligned}$$

where Z_i is some observed variable. A concrete example might be the estimation of a wage expectation function for say new labor force entrants. Expectations (Y_i) are not observable but we might argue that when faced with a job offer (that is an offered wage of Z_i) the entrant will accept the job ($W_i = 1$) only if that offer meets or exceeds his expectation. The appropriate likelihood function, under the assumption of normality, is given by

$$L(\beta, \sigma | W, X, Z) = \prod_{W_i=1} P\left(\frac{Z_i - \beta' X_i}{\sigma}\right) \cdot \prod_{W_i=0} 1 - P\left(\frac{Z_i - \beta' X_i}{\sigma}\right).$$

In this case σ is estimable because observations on Z_i provide information on the scale of Y_i.

In another variation on the same model Z_i is replaced by some constant threshold. If $\beta' X_i$ includes an intercept term then σ is again not estimable since $(c - \beta_0 - \beta' X)/\sigma$ is observationally equivalent to $(c - \alpha_0 - \alpha' X)/(k\sigma)$ where $\alpha_0 = k\beta_0 + (1-k)c$ and $\alpha = k\beta$. If that constant is also unknown and to be estimated the identification problem is further compounded and estimation will require some normalization on either β_0 or the threshold parameter.

B. *Multinomial Choice Models*[3]

An obvious generalization of the binomial choice model is to allow for more than two alternatives in the set of possible choices. Such models fit a large and important set of problems encountered in economics and are mentioned here for that reason. Regretably the functional model to be presented here cannot be used to analyze these models. This is the one class of limited dependent variable models, however, for which there seems to be wide dissemination of suitable computer programs. The approach used in these programs is logit analysis, a choice dictated in part by the fact that a specification of the underlying disturbance distributions such that the selection probabilities are of the logistic form leads to tractable likelihood functions, while almost every other choice of distributions leads to nearly insurmountable computational difficulties.

C. *Ordinally Discrete Dependent Variables*[4]

Another extension of the binary choice model is to allow for more than two alternatives but to require that those alternatives be ranked in some well defined order. Such models might arise when the magnitude of the observed dependent variable reflects the magnitude but not the scale of some underlying but unobserved dependent variable. As an example years of schooling might be a proxy measure for accumulated human capital but it may not be reasonable to assume that twice as much education implies twice as much capital. Alternatively the observed dependent variable may have the scale relevant to a particular relationship being examined but it may be measurable only in coarse discrete units.

In the case with unknown scale the model appears as:

$$\begin{aligned} Y_i &= \beta' X_i + u_i \\ W_i &= 1 \qquad \text{if } Y_i < \mu_1 \\ &= 2 \qquad \text{if } \mu_1 \leq Y_i < \mu_2 \\ &\dots \\ &= S-1 \qquad \text{if } \mu_{s-2} \leq Y_i < \mu_{s-1} \\ &= S \qquad \text{if } \mu_{s-1} \leq Y_i. \end{aligned}$$

If the u_i's are independently and normally distributed with mean zero the likelihood function is

$$\begin{aligned} L(\beta, \mu | X, W) = \prod_{W_i=1} P\left(\frac{\mu_1 - \beta' X_i}{\sigma}\right) \cdot \prod_{W_i=2} P\left(\frac{\mu_2 - \beta' X_i}{\sigma}\right) - P\left(\frac{\mu_1 - \beta' X_i}{\sigma}\right) \cdot \dots \\ \cdot \prod_{W_i=s-1} P\left(\frac{\mu_{s-1} - \beta' X_i}{\sigma}\right) - P\left(\frac{\mu_{s-2} - \beta' X_i}{\sigma}\right) \cdot \prod_{W_i=s} 1 - P\left(\frac{\mu_{s-1} - \beta' X_i}{\sigma}\right) \end{aligned}$$

As in the binomial choice model, σ is not identifiable and the set of thresholds μ_j and the intercept cannot all be estimated. After suitable normalization, for example $\sigma = 1$ and $\mu_1 = 0$ we can estimate β up to a multiplicative scale factor and

[3] Refer to McFadden [3] for a description of the most general multinomial model, an extensive bibliography of practical applications and a discussion of the estimation problems.

[4] See McKelvey [4] for a detailed discussion of the models.

the difference between the thresholds up to the same scale. Estimates of the μ_j's would represent the relative scale among the values taken on by the observed dependent variable.

When the scale of the variable W is known the model is the same except for replacing the unknown μ_j's with appropriate known constants and in this case σ is estimable.

D. *Truncated Dependent Variables*[5]

In many economic relationships the dependent variable is necessarily non-negative. Thus we might write the model as

$$\begin{aligned} W_i &= \beta' X_i + u_i \qquad \text{if RHS} > 0 \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Alternatively we might conceive of an unconstrained latent variable Y_i and reformulate the model as

$$\begin{aligned} Y_i &= \beta' X_i + u_i \\ W_i &= Y_i \qquad \text{if } Y_i \geq L_i \\ &= L_i \qquad \text{if } Y_i < L_i \end{aligned}$$

where the threshold of 0 has been replaced by a more general variable threshold and only X_i, W_i and L_i are observed. For independent normal u_i's the likelihood function is given by

$$L(\beta, \sigma | W, X, L) = \prod_{W_i = L_i} P\left(\frac{L_i - \beta' X_i}{\sigma}\right) \cdot \prod_{W_i > L_i} \frac{1}{\sigma} Z\left(\frac{W_i - \beta' X_i}{\sigma}\right)$$

where $Z(x)$ is the standard normal density function $(1/\sqrt{2\pi}) \exp(-x^2/2)$.

Examples of problems to which this model might be applied include consumer expenditure on some class of goods, which is constrained to be non negative, and interest rates paid by commercial banks on savings deposits, which are constrained by regulation Q not to exceed a certain rate fixed by the Federal Reserve. Note that for purposes of estimation alone the particular value assigned to W_i for limit observations is not used while the threshold value is. On the other hand for non limit observations the threshold value need not be known. Thus the model may under certain circumstances be utilized to estimate separately the two equations of the following disequilibrium market model:

$$\begin{aligned} D &= \beta_1' X_1 + u_1 \\ S &= \beta_2' X_2 + u_2 \\ Q &= \text{Min}\,(S, D) \\ & u_1 \text{ and } u_2 \text{ independent.} \end{aligned}$$

The observed variables are Q, X_1 and X_2 and we assume that X_1 and X_2 are independent of u_1 and u_2. For estimation of the demand equation D is the latent

[5] These models were investigated by Tobin [9] and have come to be called "Tobit" models.

variable and S the threshold with the roles reversed for estimation of the supply equation. We must, for the truncated model to apply, know which observations in a given sample correspond to demand (i.e. excess supply) and which correspond to supply. Furthermore information on this sample separation must be exogenous.[6]

Suppose that in the simple truncated dependent variable model the threshold is an unknown constant to be estimated with limit observations on W_i somehow distinguishable, though not equal to the threshold. Then direct maximization of the likelihood function with respect to β, σ and μ (the threshold) would lead to an estimate for μ of infinity. But this would be inconsistent with the model which specifies that the constant threshold must necessarily be less than or equal to the minimum observed value of W_i over the set of non-limit observations. Thus the maximum likelihood estimate of μ would be this minimum value of W_i and the other estimates would be obtained by maximizing the likelihood with respect to the other parameters holding μ fixed.

E. *Doubly Truncated Dependent Variables*

Some dependent variables of interest may be truncated both at high and at low values. The model[7] becomes

$$\begin{aligned} Y_i &= \beta' X_i + u_i \\ W_i &= L_{1i} \quad \text{if } Y_i < L_{1i} \\ &= Y_i \quad \text{if } L_{1i} \leq Y_i \leq L_{2i} \\ &= L_{2i} \quad \text{if } L_{2i} < Y_i \end{aligned}$$

and the likelihood function is given by

$$L(\beta, \sigma | W, X, L) = \prod_{W_i = L_{1i}} P\left(\frac{L_{1i} - \beta' X_i}{\sigma}\right) \cdot \prod_{W_i = Y_i} \frac{1}{\sigma} Z\left(\frac{W_i - \beta' X_i}{\sigma}\right)$$
$$\cdot \prod_{W_i = L_{2i}} 1 - P\left(\frac{L_{2i} - \beta' X_i}{\sigma}\right).$$

In some problems the intermediate or non-limit observations may also be unobserved. Provided the sample may still be separated into the three subsets of observations and the thresholds are known constants or observable variables, all parameters of the model are still estimable. The middle term in the likelihood function is replaced in this case by

$$\left[P\left(\frac{L_{2i} - \beta' X_i}{\sigma}\right) - P\left(\frac{L_{1i} - \beta' X_i}{\sigma}\right)\right],$$

and the model is seen to be a specific case of the ordered discrete variable model with known scale.

[6] See Maddala and Nelson [2] for a detailed discussion of disequilibrium market model estimation under these and other assumptions.

[7] See Rosett and Nelson [6] for a detailed treatment of this class of models.

An example of a problem to which this model has been applied is the demand for health insurance by people on medicare. A certain minimum coverage (the lower threshold) is provided to all participants. They may purchase supplemental insurance only up to some maximum which falls short of full coverage.

F. *Models of Friction*

Rosett [5] considered a model in which the dependent variable responds only to numerically large values of the exogenous variables. His model may be written as:

$$\begin{aligned} Y_i &= \beta' X_i + u_i \\ W_i &= Y_i - \alpha_1 \qquad \text{if } Y_i < \alpha_1 \\ &= 0 \qquad \text{if } \alpha_1 < Y_i < \alpha_2 \\ &= Y_i - \alpha_2 \qquad \text{if } \alpha_2 < Y_i. \end{aligned}$$

Denote the sample separation into the three subsets by three sets of integers Ψ_1, Ψ_2 and Ψ_3. The likelihood function is given by

$$L(\alpha_1, \alpha_2, \beta, \sigma | W, X) = \prod^{\Psi_1} \frac{1}{\sigma} Z\left(\frac{W_i + \alpha_1 - \beta' X_i}{\sigma}\right) \cdot \prod^{\Psi_2} P\left(\frac{\alpha_2 - \beta' x_i}{\sigma}\right) - P\left(\frac{\alpha_1 - \beta' x_i}{\sigma}\right)$$

$$\cdot \prod^{\Psi_3} \frac{1}{\sigma} Z\left(\frac{W_i + \alpha_2 - \beta' X_i}{\sigma}\right).$$

The model provides for a different intercept in the two sets of continuous observations. One might assume no difference in the intercepts by setting $W_i = Y_i$ in both extreme cases and deleting α_1 and α_2 from the corresponding terms in the likelihood. Going the other direction even the slope coefficients might be permitted to change between the two sets by appropriate modification of the model and the likelihood function.

Examples of problems to which this model might apply are changes in the holdings of some asset in response to changes in its price or rate of return and changes in wage offers by a firm in response to changes in market conditions.

II. A General Functional Model

Most of the limited dependent variable models may be specified, perhaps after reformulation, as

(i) a single regression equation relating a latent, i.e., not directly observable, endogenous variable to a stochastic function of some vector of exogenous variables, say $Y_i = f(X_i, \beta, u_i)$ and

(ii) a discontinuous mapping from the latent variable Y_i to an observable dependent variable W_i, say $W_i = g(Y_i, Z_i)$

The role played by the vector of exogenous variables Z will be discussed below. Observed variables include X_i, Z_i and W_i and parameters to be estimated include the vector β and perhaps parameters of the distribution of u_i and of the function g.

The functional form of both g and f must be known and constant over all observations. If the model is to conform to the various limited dependent variable models and be operationally feasible we will require certain restrictions on the form of these two functions. Consider first the function f. Since the estimation method to be used is maximum likelihood the distribution of the stochastic component must be specified. We will assume that the disturbance term u appears, perhaps after a suitable transformation, additively and follows an independent normal distribution with zero mean and constant variance.[8] Restrictions on the degree of non-linearity of f may also be desirable. The iterative maximization algorithms used for obtaining estimates generally require at least first and perhaps second derivatives. Thus if nonlinear specifications for f are to be allowed implementation will require a computer system with analytic differentiation capability, numerical derivatives or user supplied derivatives. Restricting f to be linear would avoid this problem but we will not impose that constraint here. The regression equation to be used in the model is thus of the form

$$(1) \qquad Y_i = f(X_i, \beta) + u_i, \qquad u_i \sim \mathrm{IN}\,(0, \sigma^2).$$

In the limited dependent variable models the mapping $W_i = g(Y_i)$ is necessarily discontinuous with the discontinuities appearing at well defined points, to be called thresholds, in the range of Y_i. Assume that there are $S-1$ threshold points and partition the range of Y_i into the S disjoint intervals. Then $g(Y_i, Z_i)$ may be written as

$$(2) \qquad g(Y_i, Z_i) = g_j(Y_i) \qquad \text{if } t_{ij-1} \le Y_i < t_{ij}, \qquad j = 1, \ldots, S$$

where t_{ij}, $j = 1, \ldots, S-1$ are the threshold points and t_{i0} and t_{is} are defined to be $-\infty$ and $+\infty$ respectively. The constraint $t_{ij-1 \le t_{ij}}$, $j = 1, \ldots, S$ must hold across all observations i but the threshold points need not be constant across observations. Any combination of the following specifications for the t_{ij}'s should be permissible:

(i) known numeric constants
(ii) observable variables (i.e. one of the variables in the vector Z_i)
(iii) constant but unknown parameters to be estimated.

The individual $g_j(Y_i)$'s, $j = 1, \ldots, s$ are of two basic types, to be called continuous and mass point as determined by the distribution of the random variable W_i within the relevant interval on Y_i.[9] A mass point $g_j(Y_i)$ specifies that *within* the jth interval of the range of Y_i W_i is a constant function of Y_i (i.e., independent of the level of Y_i). Typical specifications for mass point g_j's are

(i) $g_j(Y_i) = t_{ik}$ (where t_{ik} is one of the threshold points of the type (i) or (ii) as given above)

[8] The choice of distributions may of course be changed but is an integral part of the analysis and thus must be held fixed for implementation of the model. Note that the normal distribution leads to probit analysis for the binomial choice model and is the distribution suggested most often for extensions of the limited dependent variable models. A choice of the logistic (sech^2) distribution would lead to logit analysis for the binomial choice model.

[9] The terms mass point and continuous will be loosely applied to the subfunctions g_j, to the corresponding interval on Y_i and to the values taken on by W_i. What is implied in all cases is that, within some intervals of the range of Y_i, W_i is defined by g_j to be a constant so that its associated measure of probability is probability mass. In other intervals W_i is a continuous function of Y_i within that interval so that the appropriate measure of probability is its probability density.

(ii) $g_j(Y_i)=Z_{ik}$ (where Z_{ik} is some observable exogenous variable)

(iii) $g_j(Y_i)=c$ (some known constant).

Continuous $g_j(Y_i)$'s specify continuous and strictly increasing functions of Y_i *within* the corresponding interval on Y_i. The most common specification will be

$$g_j(Y_i)=Y_i.$$

We will in fact require that all continuous g_j's be of this form, delaying for the moment a discussion of the advantages and disadvantages of such a restriction.

Derivation of the likelihood function for the functional model is straightforward. We need first to derive the distribution of W_i. For mass point intervals we have

$$\begin{aligned}\Pr(W_i=g_j(Y_i)) &= \Pr(t_{ij-1}\leq Y_i<t_{ij})\\ &= \Pr(t_{ij-1}-f(X_i,\beta)\leq u_i<t_{ij}-f(X_i,\beta))\end{aligned}$$

which under the Normality assumption on u_i becomes

$$\Pr(W_i=g_j(Y_i))=P\left(\frac{t_{ij}-f(X_i,\beta)}{\sigma}\right)-P\left(\frac{t_{ij-1}-f(X_i,\beta)}{\sigma}\right)$$

where $P(\chi)$ is the standard normal cumulative density function. A general derivation of the density function for W_i over continuous intervals requires strong assumptions about the specification of continuous $g_i(Y_i)$'s. If these functions are strictly increasing (decreasing) over the relevant interval on Y_i then the inverse function

$$Y_i=g_j^{-1}(W_i)$$

exists and is differentiable so that the p.d.f. of continuous W_i, say $h(W_i)$, is given by

$$h(W_i)=J_j\frac{1}{\sigma}Z\left(\frac{g_j^{-1}(W_i)-f(X_i,\beta)}{\sigma}\right)$$

where J_j is the Jacobian of the transformation, $J_j=|\partial g_j^{-1}/\partial W_i|$, and Z is the standard normal density function. Construction of the likelihood requires knowledge of the sample separation. That is for each observation on W_i we must be able to determine the interval in which the corresponding unobserved value for Y_i lies.[10] For notational convenience define the subsets Ψ_j of integers $1,\ldots,n$, where n is the sample size, as

$$i\in\Psi_j \quad \text{if } t_{ij-1}\leq Y_i<t_{ij}, \quad i=1,\ldots,n, \quad j=1,\ldots,s$$

The likelihood function is given by

$$L(\theta|W,X,Z)=\prod_{i\in\Psi_1}A_{i1}\cdot\prod_{i\in\Psi_2}A_{i2}\cdot\ldots\cdot\prod_{i\in\Psi_s}A_{is} \tag{3}$$

[10] Determination of the sample separation is made by comparing, for each observation, W_i with each $g_j(Y_i)$. For mass point g_j's a matching of W_i and $g_j(Y_i)$ for some j determines that the observation corresponds to a value of Y_i in the jth interval. This leaves only the continuous observations to be classified but, as will be pointed out later, so long as we restrict continuous g_j's to be of the form $g_j(Y_i)=Y_i$ the knowledge that an observation on W_i is a continuous one is all that is required; we need not know to which continuous interval it belongs.

where θ is a vector of all parameters to be estimated and the A_{ij}'s are defined as

$$A_{ij} = P\left(\frac{t_{ij} - f(X_i, \beta)}{\sigma}\right) - P\left(\frac{t_{ij-1} - f(X_i, \beta)}{\sigma}\right)$$

if j corresponds to a mass point interval on Y_i and

$$A_{ij} = J_j \frac{1}{\sigma} Z\left(\frac{g_j^{-1}(W_i) - f(X_i, \beta)}{\sigma}\right)$$

if j corresponds to a continuous interval of Y_i.

It should now be clear why the restrictive specification $g_j(Y_i) = Y_i$ for continuous intervals was imposed. Such a restriction makes it easy to distinguish mass point from continuous intervals and permits all continuous observations to be grouped into a single subset, for purposes of estimation, since they all enter the likelihood in exactly the same form ($J_j = 1$ and $g_j^{-1}(W_i) = W_i$ for every continuous interval j.) Thus we can avoid a good deal of perhaps messy computer coding and additional user supplied information. Note too that this restriction creates difficulty with only one of the limited dependent variable models reviewed in section I, the friction model. But even this problem is easily surmounted by judicious use of dummy variables.

The friction model, with intercepts which differ in the two continuous intervals, is repeated here.

$$\begin{aligned} Y_i &= \beta' X_i + u_i \\ W_i &= Y_i - \alpha_1 \qquad \text{if } Y_i < \alpha_1 \\ &= 0 \qquad \text{if } \alpha_1 \le Y_i \le \alpha_2 \\ &= Y_i - \alpha_2 \qquad \text{if } \alpha_2 < Y_i. \end{aligned}$$

Reformulate the regression equation as

$$Y_i = \alpha_1 D_{i1} + \alpha_2 D_{i2} + \beta' X_i + u_i$$

where

$$\begin{aligned} D_{i1} &= -1 \text{ when } Y_i \text{ lies in the lower continuous interval} \\ &= 0 \text{ otherwise} \\ D_{i2} &= -1 \text{ when } Y_i \text{ lies in the upper continuous interval} \\ &= 0 \text{ otherwise.} \end{aligned}$$

The threshold structure is then written as

$$\begin{aligned} W_i &= Y_i \qquad \text{if } Y_i < \alpha_1 \\ &= 0 \qquad \text{if } \alpha_1 \le Y_i \le \alpha_2 \\ &= Y_i \qquad \text{if } \alpha_2 < Y_i. \end{aligned}$$

Note that the two continuous intervals on Y_i are not properly defined in this formulation but recall that for continuous intervals the threshold points do not appear in the corresponding terms in the likelihood function. Thus with regard to estimation the inconsistency is only transparent. The inconsistency could in fact be removed by redefining the two intervals as $Y_i < 0$ and $0 \leq Y_i$. But this would make the model more difficult to implement since then, without specifically accounting for the specification of $f(X_i, \beta)$, the intervals on Y_i wouid appear to either overlap or fail to exhaust the entire range of Y_i. Several other points are worth noting. In this model $\beta' X_i$ should not include an intercept term or identification problems among β_0, α_1 and α_2 will arise. The friction model is unique in that threshold parameters and parameters of the function f overlap. Finally, similar use of dummy variables can provide for slope coefficients which differ in the two continuous intervals while if all intercept and slope coefficients are the same the restriction on the specification of the continuous g_j's is satisfied without a reformulation using dummy variables.

III. Features of a Computational Algorithm

In this section we will discuss the specific details involved in a suitable computer program for the functional model. First the model is restated.

The functional model is defined as

$$(1) \qquad Y_i = f(X_i, \beta) + u_i$$

$$(2) \qquad W_i = g_j(Y_i) \qquad \text{if } t_{ij-1} \leq Y_i < t_{ij}, j = 1, \ldots, S$$

$$u_i = \text{In}\,(0, \sigma^2)$$

Y_i is a latent variable and W_i, the vector X_i and perhaps some vector Z_i are the observed variables. Parameters to be estimated include β and perhaps σ and/or some of the t_{ij}'s. The threshold points t_{i0} and t_{is} are defined as $-\infty$ and $+\infty$ respectively for all $i = 1, \ldots, n$ where n is the sample size. The remaining threshold points $t_{i1}, \ldots, t_{is-1}$ may be any of the following:

(i) known numeric constants

(ii) observable exogenous variables (one of the Z_{ik}'s)

(iii) constant but unknown parameters to be estimated

The $g_j(Y_i)$'s define W_i to be either a mass point observation or a continuous observation when the unobserved Y_i falls in the corresponding jth interval. Continuous $g_j(Y_i)$'s must be of the form

$$g_j(Y_i) = Y_i$$

while mass point $g_j(Y_i)$'s may be either

(i) known constants, i.e., $g_j(Y_i) = C$

or

(ii) observable exogenous variables, i.e., $g_j(Y_i) = Z_{ik}$

Furthermore the mass point $g_j(Y_i)$'s must be such that a comparison of W_i for each observation with each mass point g_j will determine uniquely a sample separation defined by the following subsets of the integers $1, \ldots, n$

$$\begin{array}{lll} i \in \Psi_j & \text{iff } W_i = g_j(Y_i) & \text{for mass point interval } j \\ i \in \Psi_0 & \text{iff } W_i \neq g_j(Y_i) & \text{for any mass point interval } j. \end{array}$$

Note that Ψ_j will be empty for any continuous interval j.

The components of the likelihood function were presented in Section II. Estimation involves maximization of the logarithm of the likelihood function. The normal equations obtained by setting the derivatives of log L with respect to each estimable parameter equal to zero will be nonlinear so that some iterative maximization algorithm is required. Experience has shown that the Newton–Raphson algorithm[11] works quite well on these models with fairly rapid convergence when starting from reasonable initial estimates. This algorithm does require both first and second derivatives which, though messy, are fairly easy to derive. Table 1 presents the components of the likelihood function corresponding to each type of interval on Y_i and the associated terms in the first and second derivatives of the log likelihood function. Several points should be noted. First the parameters to be estimated are denoted by the vector θ with elements θ_i. Secondly, the derivatives presented there make the following use of the chain rule: The terms in the log likelihood function involve the functions $P(A)$ and $Z(B)$, where A and B are representative arguments, and have the following derivatives:

$$\frac{\partial P(A)}{\partial \theta_i}=\frac{Z(A)}{P(A)}\cdot\frac{\partial A}{\partial \theta_i} \qquad \text{and} \qquad \frac{\partial Z(B)}{\partial \theta_i}=-Z(B)\cdot B\cdot\frac{\partial B}{\partial \theta_i}$$

We have carried the differentiation only this far, since the arguments A and B involve the unspecified function $f(X_i, \beta)$, and assume that the derivatives of these arguments can be readily obtained by some combination of user supplied derivatives, restrictions on the functional form of f and internal differentiation capability.[12] Finally, note that lower and upper mass point intervals have been distinguished in that table from interior mass point intervals since recognition of their simpler structure generally will achieve significant economies in computer time.

As was suggested by the discussion in Section I, not all parameters in the functional model are necessarily estimable. In particular σ can be estimated only if the observed variable W_i contains some information regarding the scale of the latent variable Y_i. In general any one of the following conditions on the model will be sufficient to permit estimation of σ.

(i) At least one continuous interval.
(ii) At least one threshold is an observable, varying threshold.
(iii) At least two threshold points are known constants.

If none of these conditions are met then estimation may proceed only after normalization of σ, e.g., $\sigma = 1$. If the model includes both threshold parameters to be estimated and an intercept term in the regression equation there will generally be an identification problem among this set of parameters—only the difference

[11] As was noted earlier the constraint $t_{ij-1} \leq t_{ij}$ must hold for $j = 1, \ldots, s$ and all $i = 1, \ldots, n$. If these thresholds include parameters to be estimated the constraint should be taken into account in the maximization algorithm. This is awkward to do however, in the general model since not all problems will require estimation of threshold parameters. There is no danger that straightforward application of Newton's method will produce estimates which violate the constraint since this would require taking logarithms of negative numbers. We therefore suggest using Newton's method with the provision of allowing some user control in the iterative process for handling those occasional problems in which the constraint causes difficulty.

[12] The TROLL system on which the author has implemented the functional model does have the internal capability of obtaining analytic derivatives. This feature is extremely useful for such simple functions as the arguments like A and B in that it renders unnecessary further restrictions of f or alternatively, heavy user input. On the other hand it cannot be used to avoid the programming of derivative calculation to the level presented in Table I without resulting in prohibitive computer time.

TABLE 1

Component of Likelihood Function and Corresponding First and Second Derivatives of the Log Likelihood

Type of interval to which an observation belongs	Corresponding component in likelihood	First derivative (w.r.t. θ_i)	Second derivative (w.r.t. θ_i and θ_k)
Lower mass point	$P(A_j)$	$\frac{Z(A_j)}{P(A_j)} \cdot A_j^i$	$\frac{Z(A_j)}{P(A_j)} \cdot [A_j^{ik} - A_j \cdot A_j^i \cdot A_j^k] - \frac{\partial P(A_j)}{\partial \theta_i} \cdot \frac{\partial P(A_j)}{\partial \theta_k}$
Interior mass point	$P(A_j) - P(B_j)$	$\frac{Z(A_j) \cdot A_j^i - Z(B_j) \cdot B_j^i}{P(A_j) - P(B_j)}$	$\frac{Z(A_i) \cdot [A_j^{ik} - A_j \cdot A_j^i \cdot A_j^k] - Z(B_j) \cdot [B_j^{ik} - B_j \cdot B_j^i \cdot B_j^k]}{P(A_j) - P(B_j)}$ $- \frac{\partial[P(A_j) - P(B_j)]}{\partial \theta_i} \cdot \frac{\partial[P(A_j) - P(B_j)]}{\partial \theta_k}$
Upper mass point	$1 - P(B_j)$	$\frac{-Z(B_j)}{[1 - P(B_j)]} \cdot B_j^i$	$\frac{-Z(B_j)}{[1 - P(B_j)]} \cdot [B_j^{ik} - B_j B_j^i B_j^k] - \frac{\partial[1 - P(B_j)]}{\partial \theta_i} \cdot \frac{\partial[1 - P(B_j)]}{\partial \theta_k}$
Continuous	$\frac{1}{\sigma} Z(C)$	$\langle -1/\sigma \rangle^{*} - C \cdot C^i$	$\langle\langle 1/\sigma^2 \rangle\rangle^{**} - (C^i \cdot C^k + C \cdot C^{ik})$

$$P(\chi) = \int_{-\infty}^{\chi} Z(u)\, du, \qquad Z(\chi) = \frac{1}{\sqrt{2\pi}} \exp(-\chi^2/2)$$

$$A_j = \frac{t_j - f(X, \beta)}{\sigma}, \quad B_j = \frac{t_{j-1} - f(X, \beta)}{\sigma}, \quad C = \frac{W - f(X, \beta)}{\sigma}, \quad A_j^i = \frac{\partial A_j}{\partial \theta_i}, \quad A_j^{ik} = \frac{\partial^2 A_j}{\partial \theta_i \partial \theta_k}, \text{ etc.}$$

* The term $-1/\sigma$ appears only for $\theta_i = \sigma$.

** The term $1/\sigma^2$ appears only for $i = k$, $\theta_i = \sigma$.

between each pair of parameters in the set can be estimated. Again a normalization is required on one parameter in this set.

The iterative maximization algorithm will require starting values for the parameters to be estimated. We have not been successful in obtaining a straightforward routine for selecting good starting values for all parameters in the functional model. Tobin [9], in the context of the Tobit model, suggested approximating the non-linear terms in the normal equations by some simple functions to allow analytic solution of those equations but this approach becomes quite difficult to implement in the more general functional model, especially if the regression equation is itself non-linear. Similarly some expansion of the normal equations with a low order truncation is also difficult to implement. In lieu of a general solution we offer the following suggestions for implementation on a case by case basis.

(a) If the model includes continuous intervals, least squares regression of W_i on X_i over just the subset of continuous observations will often provide satisfactory, though biased, starting values for the regression coefficients and for σ.

(b) For threshold parameters choose starting values such that the spacing between adjacent threshold points is proportional to the percentage of observations falling in each interval.

(c) In models with no continuous intervals and values for W_i which correspond ordinally to Y_i try a straightforward least squares regression of W_i on X_i for starting values for the regression coefficients.

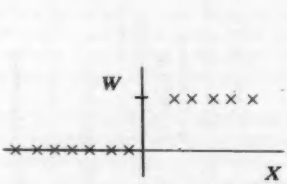
(d) for many data sets and if the iterative maximization algorithm is fairly stable, zero starting values for many of the parameters will generally suffice.

Generally parameter estimation is only part of the analysis to be performed on a given model. The remainder of this section discusses various other analyses which may often be desired and which are reasonably easy to implement in the functional model.

It is often quite informative to examine simple descriptive statistics, such as mean, variance and range, of various variables in the model both over all observations and over the subsets of observations corresponding to each interval on Y_i. Furthermore while such information may be of use by itself it can as well serve to detect or explain failures in the estimation process. To see this consider a simple binomial choice model with a single regressor variable. The likelihood function is given by

$$L(\alpha,\beta|W,X)=\prod_{W_i=0}P(-\alpha-\beta X_i)\cdot\prod_{W_i=1}[1-P(-\alpha-\beta X_i)].$$

Suppose that in a given set of data the observations are as pictured in the figure to the right. It is easy in this case to find values for α and β such that whenever $W_i=0\,(-\alpha-\beta X_i)$ is positive and when $W_i=1\,(-\alpha-\beta X_i)$ is negative. All observations can thus be perfectly classified on the basis of the mean



value $(\alpha+\beta X)$ for such α and β. In fact the likelihood is maximized as α and β tend to negative and positive infinity respectively. This failure in the estimation process could easily be predicted, in this simple model, by comparing the range of X_i within the two sets of observations. The same problem arises in this model with more than one exogenous variable and all the other models as well, suggesting that as a prelude to estimation one should always critically examine simple statistics, especially the range, of the exogenous variables within each subset of observations. In addition, even if the individual exogenous variables do overlap, there may be some combination which provides perfect classification of the observations. Such a situation is often difficult to detect until after the estimation process has failed. Performing the same analysis on $\hat{Y}_i = f(x_i, \hat{\beta})$ where $\hat{\beta}$ is the vector of regression coefficient estimates when the iterative maximization procedure began to diverge may often reveal the source of the problem.

Estimated classification probabilities (i.e., $\Pr(W_i = g_j(Y_i)|X_i, Z_i)$ or alternatively $\Pr(t_{ij-1} \leq Y_i < t_{ij}|X_i, Z_i)$) are often as important to the analysis as estimates of the parameters themselves. The expressions for obtaining them are given by the components of the likelihood function for mass point intervals and similar expressions for continuous intervals. In addition to their independent use they serve an important role in an examination of the estimation results analogous to residual analysis in least squares regression. They provide, for example, one measure of classification error. Let j^* be the interval in which an observation falls and $\hat{j}$ be the interval with largest associated classification probability. An observation may be viewed as being misclassified if $j^* \neq \hat{j}$.[13]

A variety of measures of "residuals" may be readily obtained. Using estimated coefficients to compute $\hat{Y}_i = f(X_i, \hat{\beta})$ we can obtain directly $\hat{u}_i = W_i - \hat{Y}_i$ for continuous observations. For mass point observations the estimated residual may be "bracketed" by $t_{ij} - \hat{Y}_i$ and $\hat{Y}_i - t_{ij-1}$. Another indicator of misclassification is given by a comparison of $\tilde{j}$ and j^* where $\tilde{j}$ the interval in which $\hat{Y}_i$ falls and, as before, j^* is the observed interval.

An important part of the analysis for a given problem might be the calculation of mean values for the observed dependent variables. These might be needed, for example, for prediction purposes or for the calculation of elasticities.[14] The expected value of W_i for given X_i (and Z_i) is

$$E(W_i|X_i, Z_i) = \int_{-\infty}^{\infty} \frac{1}{\sigma} Z\left(\frac{y - f(X_i, \beta)}{\sigma}\right) g(y)\, dy$$

$$= \sum_{j=1}^{s} \int_{t_{ij-1}}^{t_{ij}} \frac{1}{\sigma} Z\left(\frac{y - f(X_i, \beta)}{\sigma}\right) g_j(y)\, dy = \sum_{j=1}^{s} A_{ij}.$$

[13] Whether this is an appropriate measure of misclassification will depend on the model being examined. For example it may be a useful measure for the binomial choice model while in the ordinally discrete model, since the frequency of misclassification under this measure is easily altered by arbitrarily collapsing adjacent intervals, it may not be at all appropriate.

[14] If the prediction or elasticity is for a single individual or observation then the appropriate value for W to be used should be $\hat{W}_i = g(\hat{Y}_i)$. On the other hand if we need the mean predicted value or aggregate elasticity the appropriate value is $E(W_i|X_i, Z_i)$ as is given here.

For mass point intervals $g_j(y)$ is a constant so that the corresponding term in the expected value of W_i is

$$A_{ij} = g_j(Y_i) \cdot \left[P\left(\frac{t_{ij} - f(X_i, \beta)}{\sigma}\right) - P\left(\frac{t_{ij-1} - f(X_i, \beta)}{\sigma}\right) \right]$$

For continuous intervals, integration over the relevant range yields[15]

$$A_{ij} = f(X_i, \beta) \cdot \left[P\left(\frac{t_{ij} - f(X_i, \beta)}{\sigma}\right) - P\left(\frac{t_{ij-1} - f(X_i, \beta)}{\sigma}\right) \right]$$
$$- \sigma \left[Z\left(\frac{t_{ij} - f(X_i, \beta)}{\sigma}\right) - Z\left(\frac{t_{ij-1} - f(X_i, \beta)}{\sigma}\right) \right].$$

One could compute similar expressions for the variance of the deserved dependent variable,[16] but it would not be of much practical use. It is not useful, for example, in constructing confidence intervals about individual or mean predicted values of W_i. For these, one must return to the regression equation, if the model contains continuous intervals, and make probability statements about intervals around $f(X, \hat{\beta})$ or $f(X, \hat{\beta}) + u$ as would be done in the usual regression model but taking care to account explicitly for the threshold points. For mass point values, estimated selection probabilities themselves provide concise probability statements about occurrence or nonoccurrence.

Regarding tests of hypotheses about estimated coefficients, the use of maximum likelihood estimation provides straightforward solutions. The matrix of second derivatives of the log likelihood with respect to the coefficients being estimated, or at least an approximation to it, will generally fall directly out of the iterative maximization algorithm. Minus one times the inverse of this hessian matrix may be used as an asymptotic approximation to the covariance matrix of coefficient estimates. Square roots of diagonal elements provide estimates of standard errors and these as well as submatrices of variances and covariances can be used for a variety of hypothesis tests and confidence intervals.

[15] We have

$$A_j = \int_{t_{j-1}}^{t_j} y \frac{1}{\sigma} Z\left(\frac{y - f(X, \beta)}{\sigma}\right) dy = \int_{L_1}^{L_2} [f(X, \beta) + \sigma \chi] Z(\chi)\, d\chi$$
$$= f(X_1 \beta) \int_{L_1}^{L_2} Z(\chi)\, d\chi + \sigma \int_{L_1}^{L_2} \chi Z(\chi)\, d\chi$$

where

$$L_1 = (t_{j-1} - f(X, \beta)]/\sigma \qquad \text{and } L_2 = [t_j - f(X, \beta)]/\sigma.$$

Since

$$\chi Z(\chi) = -dZ(\chi)/d\chi, \qquad \int \chi Z(\chi)\, d\chi = -Z(\chi)$$

and we obtain

$$A_j = f(X, \beta) \cdot [P(L_2) - P(L_1)] - \sigma [Z(L_2) - Z(L_1)].$$

[16] Such an expression for the "Tobit" model with a lower threshold of zero, for example, would have the variance going to σ for large, positive $f(X_i, \beta)$ and to zero for large, negative $f(X_i, \beta)$.

References

[1] Johnson, Thomas. "Qualitative and Limited Dependent Variables in Economic Relationships," *Econometrica*, May, 1972.

[2] Maddala, G. S. and F. D. Nelson, "Maximum Likelihood Methods for Models of Markets in Disequilibrium," *Econometrica*, November, 1974.

[3] McFadden, Daniel. "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, P. Zarembka (Ed.), Academic Press, 1973.

[4] McKelvey, Richard D. and M. Zavoina, "A Statistical Model for the Analysis of Ordinal Level Dependent Variables," *J. Mathematical Sociology*, 1975.

[5] Rosett, Richard N. "A statistical Model of Friction in Economics," *Econometrica* 1959.

[6] Rosett, Richard N. and F. D. Nelson. "Estimation of the Two-Limit Probit Regression Model," *Econometrica*, January, 1975.

[7] Theil, Henri. *Principles of Econometrics*, chapter 12; Wiley, 1971.

[8] Tobin, J. "The Application of Multivariate Probit Analysis to Economic Survey Data," Cowles Foundation Discussion Paper No. 1, New Haven, Conn., 1955.

[9] Tobin, J. "Estimation of Relationships for Limited Dependent Variables," *Econometrica* 1958.