# New Taxonomies for Limited Dependent Variables Models

Biørn, Erik and Wangen, Knut R.
University of Oslo

17. July 2012

# NEW TAXONOMIES FOR LIMITED DEPENDENT VARIABLES MODELS

ERIK BIØRN

*Department of Economics, University of Oslo,*
*P.O. Box 1095 Blindern, 0317 Oslo, Norway.*

E-mail: erik.biorn@econ.uio.no

KNUT R. WANGEN

*Department of Health Management and Health Economics, University of Oslo,*
*P.O. Box 1089 Blindern, 0317 Oslo, Norway*

E-mail: k.r.wangen@medisin.uio.no

ABSTRACT: We establish a 'map' for describing a wide class of Limited Dependent Variables models much used in the econometric literature. The classification system, or language, is an extension of Amemiya's typology for tobit models and is intended to facilitate communication among researchers. The class is defined in relation to distributions of latent variables of an arbitrarily high dimension; the region of support can be divided into an arbitrary number of subsets, and the observation rules in each subset can be any combination of the observed, censored, and missing status. Consistent labeling is suggested at different levels of detail.

KEYWORDS: Limited dependent variables. Latent variables. Censoring. Truncation. Missing observations.

JEL CLASSIFICATION: C16, C24, C25, C34, C35, C51

## 1. Introduction

In the time before Amemiya (1984) it may have been less than obvious how a study like *"Application of a threshold regression model to household purchase of automobiles"* (Dagenais, 1975) was related to *"Censored regression model with unobserved stochastic censoring thresholds"* (Nelson, 1977). By Amemiya's account they are closely related and he labeled both as 'Type II tobit models'. His classification system for tobit models quickly became standard in the econometrics literature.

However, Amemiya's scope were limited in the outset: *"My review of the empirical literature suggests that roughly 95 % of the econometric applications of Tobit models fall into one of ... five types"* (p. 4). Although a typology based on previous *empirical* literature can be useful for review purposes, it has, from a theoretical perspective, at least two disadvantages: the empirical literature existing at any time is limited by currently available computing resources, and, since the number of possible types of Limited Dependent Variables (LDV) models is infinite, the empirical literature will never cover all cases.

We suggest a classification system which fills these lacunae. We also extend the model class to allow for both censored and missing variables. Any LDV model within the class can be described in a compact and consistent manner. We take the discussion up to the point of demonstrating how likelihood functions can be represented, but refrain from discussing typical inference issues.

## 2. Definitions and notation

2.1. **Latent variables, subsets, and observation rules.** Our general framework has three *basic elements*. The *first* is a vector of latent stochastic variables, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)$, defined over the Euclidian space, $\mathbb{R}^N$. The *second* is a partition of $\mathbb{R}^N$ into $I$ subsets, denoted as $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)$, so that

$$\bigcup_{i=1}^{I} \alpha_i = \mathbb{R}^N, \quad \alpha_i \cap \alpha_j = \emptyset \;\forall\; i \neq j, \tag{1}$$

The *third* is a register of observation rules. In each subset, each variable $\eta_n$ $(n = 1, \ldots, N)$, has one among three possible observational statuses: *observable*, *censored*, *missing*, indicated by $o$, $c$, $m$. The observation rule for subset $i$, $r_i$, is a 'word' with $N$ letters indicating the observational statuses for all latent variables (confer Hein, 2002, for an introduction to formal languages). These rules are collected in the tuple $\boldsymbol{r} = (r_1, \ldots, r_I)$. An example is one latent variable, *censored* below a threshold value, $\theta$, and observed above, *i.e.*, $N = 1$, $I = 2$. Then, for subsets $\alpha_1 = \{\eta_1 \in \mathbb{R}^1; \eta_1 \leq \theta\}$ and $\alpha_2 = \{\eta_1 \in \mathbb{R}^1; \eta_1 > \theta\}$, the observation rules are $\boldsymbol{r} = (c, o)$. Changing the rule in $\alpha_1$ from $c$ to $m$, giving $\boldsymbol{r} = (m, o)$, a model sometimes called a univariate *truncated* model emerges.

The observation rules for a few standard models are illustrated in the first column of Table 1. Amemiya's Type II is a bivariate model $(N = 2)$ with two subsets defined by the value of one of the latent variables. One variable is censored in both subsets, as only the sign is assumed observed, the other is censored in one subset and observed in the other. The subsets then become $\alpha_1 = \{(\eta_1, \eta_2) \in \mathbb{R}^2; \eta_1 \leq \theta\}$ and $\alpha_2 = \{(\eta_1, \eta_2) \in \mathbb{R}^2; \eta_1 > \theta\}$ with observation rules $\boldsymbol{r} = (cc, co)$.

2.2. **Coding of observations.** Data can be generated in three steps: first, realizations of the latent variables are drawn; second, each realization is assigned to one subset $\alpha_i$

with observation rule $r_i$; third, depending on the latter, an observation is recorded, say in a computer readable file. Let realization $t$ of $\boldsymbol{\eta}$ be denoted as $\boldsymbol{\eta}_t = (\eta_{1t}, \ldots, \eta_{Nt})$. Corresponding to $\boldsymbol{\eta}_t$ we define, conceptually, a vector of *observable stochastic variables*, $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Nt})$, *regardless of the observation status*. Then, letting the observation rule for realization $t$ be denoted $R_t$, we can define any observation as a pair, $(\boldsymbol{y}_t, R_t)$. Examples are given in Table 2.

Consider first the *univariate* case: If in subset $i$ the variable is observable, then $y_{1t} = \eta_{1t}$ is the obvious definition; if it is censored, a suggested observability convention may be $y_{1t} = i$, while if it is missing we let $(\boldsymbol{y}_t, R_t) = \Lambda$, $\Lambda$ representing an empty string. For cases with $N > 1$ the extension is straightforward, and for missing variables ($r_i = m, mm, mmm, \ldots$) we correspondingly define $(\boldsymbol{y}_t, R_t) = \Lambda$.

Whereas the number of potential observation rules, *i.e.*, possible selections of the $(o, c, m)$ triple, is, in general, $3^N$, it may be convenient to reduce the number of coding rules actually employed to $2^N + 1$, since when only some latent variables are missing, we can choose the same coding for missing variables as for censored ones. In for example a bivariate model with observational status $R_t = om$ in subset $i$ ($\eta_1$ observable, $\eta_2$ missing), we may treat $\eta_2$ as censored and code observations as $R_t = oc$, $\boldsymbol{y}_t = (\eta_{1t}, i)$. In this way, the set of observation rules used in coding can be condensed from $R_t \in \{mm, mc, mo, cm, cc, co, om, oc, oo\}$ to $R_t \in \{mm, cc, co, oc, oo\}$ . This notation allows us to present likelihood functions in a compact manner.

## 3. Likelihood function: Examples

3.1. **A univariate sub-class.** Let the density function of $\eta_1$, with parameter vector $\boldsymbol{\gamma}$, be $f(\eta_1, \boldsymbol{\gamma})$. Assuming that the subsets are defined as continuous intervals, all bounded by a pair of thresholds, collected in $\boldsymbol{\theta}_i = (\underline{\theta}_i, \overline{\theta}_i)$, we have

$$\alpha_i = \{\eta_1 \in \mathbb{R} : \underline{\theta}_i \leq \eta_1 < \overline{\theta}_i\}, \qquad i = 1, \ldots, I. \tag{2}$$

The probability that a realization of $\eta_1$ belongs to subset $i$ is denoted

$$\mathcal{F}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}) = \int_{\underline{\theta}_i}^{\overline{\theta}_i} f(\eta_1, \boldsymbol{\gamma}) d\eta_1, \qquad i = 1, \ldots, I. \tag{3}$$

Aggregating probabilities over all subsets that have the same observation rule, we have

$$\mathcal{F}_z = \textstyle\sum_{i:r_i=z} \mathcal{F}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}), \quad z = o, c, m, \quad \mathcal{F}_o + \mathcal{F}_c + \mathcal{F}_m = 1.$$

Suppose we have a set of observations, a sample, $T$. The likelihood for observation $t$, $t \in T$, takes different forms depending on the value of $R_t$:

$$\mathcal{L}_t(y_t, R_t) = \begin{cases} \dfrac{f(y_t, \boldsymbol{\gamma})}{\mathcal{F}_o + \mathcal{F}_c} & \text{if } R_t = o, \\ \dfrac{\mathcal{F}(\boldsymbol{\theta}_{y_t}, \boldsymbol{\gamma})}{\mathcal{F}_o + \mathcal{F}_c} & \text{if } R_t = c. \end{cases} \tag{4}$$

Let $T_i \subseteq T$ denote the subset of observations that falls in subset $i$. Then the likelihood for the full observation set, valid for any univariate model within the class, can be written as

$$\mathcal{L} = \prod_{i=1}^{I} \prod_{t \in T_i} \mathcal{L}_t(y_t, R_t). \tag{5}$$

If a covariate vector $\boldsymbol{x}$ is included, so that for instance $\mathsf{E}(\eta_1 | \boldsymbol{x}) = \boldsymbol{a}_1 \boldsymbol{x}$, where $\boldsymbol{a}_1$ is a coefficient vector, we extend $f(\cdot)$ to $f(\eta_1, \boldsymbol{\gamma}; \boldsymbol{a}_1, \boldsymbol{x}) = f_\epsilon(\eta_1 - \boldsymbol{a}_1 \boldsymbol{x}, \boldsymbol{\gamma})$, where $f_\epsilon(\epsilon_1, \boldsymbol{\gamma})$ is the density of $\epsilon_1 = \eta_1 - \mathsf{E}(\eta_1 | \boldsymbol{x})$.

3.2. **An $N$-variate sub-class.** Assume that $\boldsymbol{\eta}$ follows an $N$-variate distribution with density $f(\boldsymbol{\eta},\boldsymbol{\gamma})$, which can be modified to accommodate covariates. Subset $i$ is defined by

$$\boldsymbol{\alpha}_i = \{(\eta_1,\ldots,\eta_N)\in\mathbb{R}^N : \underline{\theta}_{ni}\leq\eta_n<\overline{\theta}_{ni},\quad n=1,\ldots,N\}; \tag{6}$$

the vectors of bounds by $\boldsymbol{\theta}_{ni}\equiv(\underline{\theta}_{ni},\overline{\theta}_{ni})$; and the index set of variables by $\mathcal{N}\equiv\{1,\ldots,N\}$. Let $\mathcal{A}_i$ and its complement $\mathcal{A}_i^*$ be any set containing the indices of the variables which, *in subset* $i$, are observed and non-observed (*i.e.*, censored or missing), respectively. Let correspondingly, $\boldsymbol{\eta}$ and the set of $\boldsymbol{\theta}_{ni}$ for subset $i$, be partitioned into

$$\begin{aligned}\boldsymbol{\theta}_{\mathcal{A}i}&\equiv\{\boldsymbol{\theta}_{ni}:n\in\mathcal{A}_i\}, & \boldsymbol{\theta}_{\mathcal{A}*i}&\equiv\{\boldsymbol{\theta}_{ni}:n\in\mathcal{A}_i^*\},\\ \boldsymbol{\eta}_{\mathcal{A}i}&\equiv\{\eta_n:n\in\mathcal{A}_i\}, & \boldsymbol{\eta}_{\mathcal{A}*i}&\equiv\{\eta_n:n\in\mathcal{A}_i^*\}.\end{aligned}$$

The number of $\mathcal{A}_i$ sets is, for all $i$, $2^N$, of which $N_p\equiv\binom{N}{p}$ contain $p$ observed and $N-p$ non-observed variables ($p=0,\ldots,N$). The prototype element in the likelihood function *for any observability status* in subset $i$, characterized by $\mathcal{A}_i$, can then be defined as:

$$F_{\mathcal{A}*i}(\boldsymbol{\eta}_{\mathcal{A}i},\boldsymbol{\theta}_{\mathcal{A}*i},\boldsymbol{\gamma})\equiv\textstyle\int_{\boldsymbol{\eta}_{\mathcal{A}*i}\in\boldsymbol{\theta}_{\mathcal{A}*i}} f(\boldsymbol{\eta},\boldsymbol{\gamma})d\boldsymbol{\eta}_{\mathcal{A}*i},\qquad i=1,\ldots,I. \tag{7}$$

Here integration goes across the *non-observable variables*, making the result a function of their interval bounds. For subsets with all, respectively no, variables observed, we have in particular: $F_{\mathcal{A}*i}(\boldsymbol{\eta}_{\mathcal{A}i},\boldsymbol{\theta}_{\mathcal{A}*i},\boldsymbol{\gamma})$ equals $f(\boldsymbol{\eta},\boldsymbol{\gamma})$ for $\mathcal{A}_i=\mathcal{N}$ and equals $\int_{\boldsymbol{\eta}\in\boldsymbol{\theta}_i} f(\boldsymbol{\eta},\boldsymbol{\gamma})d\boldsymbol{\eta}\equiv\mathcal{F}(\boldsymbol{\theta}_i,\boldsymbol{\gamma})$ for $\mathcal{A}_i=\emptyset$, $\mathcal{F}(\boldsymbol{\theta}_i,\boldsymbol{\gamma})$ being the subset $i$ probability, satisfying $\sum_{i=1}^N\mathcal{F}(\boldsymbol{\theta}_i,\boldsymbol{\gamma})=1$.

If no variable is missing in any subset, we can then, letting $t$ index observation and $\boldsymbol{y}_{\mathcal{A}it}=\boldsymbol{\eta}_{\mathcal{A}it}$, generalize (4) to

$$\mathcal{L}_t(\boldsymbol{y}_t)=\begin{cases}f(\boldsymbol{y}_t,\boldsymbol{\gamma}) & \text{if } \boldsymbol{\eta}_t\in\alpha_i,\ \mathcal{A}_i=\mathcal{N},\\ F_{\mathcal{A}*i}(\boldsymbol{y}_{\mathcal{A}it},\boldsymbol{\theta}_{\mathcal{A}*i},\boldsymbol{\gamma}) & \text{if } \boldsymbol{\eta}_t\in\alpha_i,\ \mathcal{A}_i\subset\mathcal{N},\ \mathcal{A}_i^*\subset\mathcal{N},\\ \mathcal{F}(\boldsymbol{\theta}_i,\boldsymbol{\gamma}) & \text{if } \boldsymbol{\eta}_t\in\alpha_i,\ \mathcal{A}_i^*=\mathcal{N}.\end{cases} \tag{8}$$

Let, in general, $\mathcal{F}_{NM}(\boldsymbol{\theta},\boldsymbol{\gamma})$, $\mathcal{F}_{SM}(\boldsymbol{\theta},\boldsymbol{\gamma})$, $\mathcal{F}_{AM}(\boldsymbol{\theta},\boldsymbol{\gamma})$ denote the subset probabilities aggregated over those subsets where, respectively, no variable, some variables, and all variables are missing. If $\mathcal{F}_{AM}(\boldsymbol{\theta},\boldsymbol{\gamma})$ is empty, so that all $\mathcal{A}_i^*$ contain *at least some censored variables*, we can either modify (8) by rescaling all elements by the factor $[\mathcal{F}_{NM}(\boldsymbol{\theta},\boldsymbol{\gamma})+\mathcal{F}_{SM}(\boldsymbol{\theta},\boldsymbol{\gamma})]^{-1}$ or, if it is desirable to curtail the sample by omitting observation sets with some observations missing, thus ensuring that all $\mathcal{A}_i^*$ included contain *censored variables only*, rescale by the factor $[\mathcal{F}_{NM}(\boldsymbol{\theta},\boldsymbol{\gamma})]^{-1}$.

Letting $t\in\mathcal{A}(p,r)_i$ symbolize that observation $t$ in subset $i$ belongs to selection $r$ among those having $p$ observed variables ($r=1,\ldots,N_p$), the prototype expression for the likelihood function, in which (8), or a modification, can be inserted, then becomes

$$\mathcal{L}=\prod_{i=1}^{I}\prod_{p=0}^{N}\prod_{r=1}^{N_p}\prod_{t\in\mathcal{A}(p,r)_i}\mathcal{L}_t(\boldsymbol{y}_t). \tag{9}$$

## 4. Label systems for models

Our most detailed label system refers directly to the observation rules. With this system, three univariate models considered in Maddala (1983, Section 6.8) as examples of 'friction models', can be represented by $\boldsymbol{r}=(c,c,c,m)$, $\boldsymbol{r}=(c,c,c)$, and $\boldsymbol{r}=(o,c,o)$. The second has a link to both the standard probit, $\boldsymbol{r}=(c,c)$, and the ordered probit, $\boldsymbol{r}=(c,\ldots,c)$. So does the first, but, as explained above, missing $\boldsymbol{\eta}$ variables give rise to distinctly different

likelihood functions. Maddala's grouping of these models undoubtedly makes sense. However, when describing them in terms of our observation rules, their key differences and their relationships to models usually not labeled as 'friction models' emerge more clearly.

This detailed classification is fully flexible with regard to subsets and dimensions. Amemiya's bivariate types, Type II and Type III, can be represented by $\boldsymbol{r} = (cc, co)$ and $\boldsymbol{r} = (cc, oo)$, respectively. We can also describe the 'tobit-like model' $\boldsymbol{r} = (oo, co)$ which is neither Type II nor Type III, and the model $\boldsymbol{r} = (cc, co, oo)$ which belongs as much to Type II as to Type III. Similarly we can label Amemiya's trivariate types, Type IV and Type V, by $\boldsymbol{r} = (cco, ooc)$ and $\boldsymbol{r} = (cco, coc)$, respectively. We can also describe the related models $\boldsymbol{r} = (ooo, ooc)$, $(ooo, occ)$, $(ooo, ccc)$, $(ooc, ccc)$, $(occ, ccc)$ which remain unclassified in Amemiya's typology, and the model $\boldsymbol{r} = (cco, ooc, coc)$ which belongs as much to Type IV as to Type V.

A less detailed label system can be obtained by counting the number of subsets for each observation rule. All univariate models can be labeled in the format $o(\cdot)c(\cdot)m(\cdot)$, the letters indicating observation rules and the following arguments the number of subsets. The univariate censored and the univariate truncated can be labeled $o(1)c(1)m(0)$ and $o(1)c(0)m(1)$, respectively, or by suppressing the non-occurring observation rules, simply as $o(1)c(1)$ and $o(1)m(1)$. For multivariate models, the description can be simplified further by ignoring the order of letters and regarding the string of letters as a product so that $cc = c^2$, $oo = o^2$, or $commom = co^2m^3$. This allows us to label Amemiya's Type V as $oc^2(2)$.

Finally, taking a bird's-eye view on all the models we have discussed, we suggest the general notation $OCM(N, I)$, $OCM$ indicating inclusion of observed, censored and missing variables, and $(N, I)$ the dimension and the number of subsets, as before. If a model does not involve all three observation rules, omitting letters in OCM may be shorter and more informative: We can let $OCM(1, 2)$ describe the univariate censored, the univariate truncated, and the probit, or we can use the respective labels $OC(1, 2)$, $OM(1, 2)$, and $C(1, 2)$ instead. In this notation, Amemiya's bivariate and trivariate models emerge as $OC(2, 2)$ and $OC(3, 2)$, respectively.

The choice of detail may depend on the context. Li (2011) estimates a four-dimensional model where a selection mechanism concerns two variables, each censored into two categories, which determine the observation status for two other variables. There are four subsets with distinct observation rules, and we would label it as $OC(4, 4)$, $o^2c^2(1)oc^3(2)c^4(1)$, or $\boldsymbol{r} = (oocc, cocc, occc, cccc)$. His more general selection mechanisms with, say, $\tau_3$ and $\tau_4$ categories for the two variables, can be described as $OC(4, \tau_3 \cdot \tau_4)$, but here the more detailed labels seem less practical, at least in verbal communication.

## 5. Concluding remarks

The taxonomies suggested in this paper apply to a large and frequently used class of econometric models and define precise relations between 'observed', 'censored', and 'missing' variables. Albeit it has been recognized for decades that members of this class have common features, previous descriptions of the class have been implicit and deliberately incomplete. Being applicable to models of any dimension of the latent variables, containing any number of subsets, and any combinations of observation rules, our classification system is complete. It is suitable for both parametric and non-parametric densities of the latent variables.

The progress towards a deepened understanding of LDV models still goes on. Notably, Schnedler (2005) presents theoretical results applicable *'to an almost arbitrary censoring problem'*. This makes likelihood estimation more accessible to applied econometricians. So does the open source estimation software package offered by Toomet and Henningsen (2008), who discuss estimation of tobit Types II and V and sketch how the package can be expanded to include more general models. The communication between workers in the various branches of the LDV model community may benefit from a common, shorthand, and precise description of models. We believe our taxonomies can serve this purpose.

Another use is in teaching: In contemporary textbooks 'censoring', 'selection', 'incomplete observation', 'defective data' and 'incidental truncation' are frequently occurring terms. Although the meaning within a single book usually is sufficiently clear, it may be less obvious how to generalize these terms to other models. With our classification system at hand, the whole class of models can be presented through a few simple examples and straightforward induction.

## References


Amemiya, T. (1984): Tobit models: A survey. *Journal of Econometrics* 24, 3–61.

Dagenais, M.G. (1975): Application of a threshold regression model to household purchases of automobiles. *Review of Economics and Statistics* 57, 275–285.

Hein, J.L. (2002): *Discrete Structures, Logic, and Computability, Second Edition*. Sudbury, MA: Jones and Bartlett Publishers.

Li, P. (2011): Estimation of sample selection models with two selection mechanisms. *Computational Statistics and Data Analysis* 55, 1099–1108.

Maddala, G.S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Nelson, F.D. (1977): Censored regression models with unobserved, stochastic censoring thresholds. *Journal of Econometrics* 6, 309–327.

Schnedler, W. (2005): Likelihood estimation for censored random vectors. *Econometric Reviews* 24, 195–217.

Toomet, O. and Henningsen, A. (2008): Sample selection models in R: Package sampleSelection. *Journal of Statistical Software* 27, 1–23.


Table 1. Classification of some standard models

| | $\boldsymbol{r}$ | Intermediate detail | $OCM(N,I)$ |
|---|---|---|---|
| Amemyia's Type II | $(cc, co)$ | $cc(1)co(1)$ | $OC(2,2)$ |
| Amemyia's Type III | $(cc, oo)$ | $cc(1)oo(1)$ | $OC(2,2)$ |
| Amemyia's Type IV | $(cco, ooc)$ | $oc^2(1)o^2c(1)$ | $OC(3,2)$ |
| Amemyia's Type V | $(cco, coc)$ | $oc^2(2)$ | $OC(3,2)$ |

Table 2. Coding of observations, $(\boldsymbol{y}_t, R_t)$, in different subsets. Examples

| | $\alpha_1$ | $\alpha_2$ |
|---|---|---|
| Univariate Censored (Tobit), $\boldsymbol{r} = (o, c)$ | $(\eta_{1t}, o)$ | $(2, c)$ |
| Univariate Truncated, $\boldsymbol{r} = (o, m)$ | $(\eta_{1t}, o)$ | $\Lambda$ |
| Amemyia's Type II, $\boldsymbol{r} = (cc, co)$ | $\big((1,1), cc\big)$ | $\big((2, \eta_{1t}), co\big)$ |