Dealing With Zeros in Economic Data

Brad R. Humphreys*
University of Alberta, Department of Economics

April 4, 2013

Abstract

Zeros frequently appear in economic data. Zeros in economic data come from different sources and the source affects the choice of an estimator. This paper surveys the literature on dealing with zeros from the perspective of an applied economist, discusses alternative estimators, and offers some illustrative examples.

Keywords: zeros, limited dependent variable models, double hurdle model

JEL Code: C2, C3, C5

1 Introduction

This document represents my attempt to sort out the key issues and concepts related to the econometric analysis of data that contain a large number of zeros, from the perspective of an applied economist. People who work regularly with data, primarily survey data, constantly face the issue of determining the appropriate estimator for any specific case. Several years ago, I put a version of this document up on my personal web space so students in one of my graduate classes would have access to it. I then promptly forgot that I had made it available. Much to my surprise, I recently discovered that this document had been cited in several published papers. I guess that people have found this useful. It's a living document, and will be revised periodically. Please drop me an email if you have somehow found this paper and think it is useful.

Economic data, especially microeconomic data, frequently contain observations where some variable of interest is equal to zero for a number of the observations in the data set. In general, this phenomenon can called censored or truncated data. Censoring and truncation are different concepts. According to Greene (2003), *truncation* applies to a situation when sample data are drawn from a subset of a larger population, and is a characteristic of the distribution from which the subset of observations in a sample is drawn. *Censoring* is a defect in a sample that occurs when the value of a variable is only partially known.

The presence of many zeros can lead to a number of econometric problems when using Ordinary Least Squares to estimate the unknown parameters of a regression model. There are a number of econometric approaches to dealing with the problem of zeros. Most fall under the heading of limited dependent variables estimation methods. I deal with zeros regularly in many data sets I work with in my own research. Because of this, I have done a lot of reading about how to deal with this issue. This document attempts to make some sense out of the existing literature on how to deal with the problem of zeros in economic data. A comprehensive, if somewhat dated, formal treatment of this topic is the excellent survey paper by Amemiya (1984). The standard textbook treatment of limited dependent variable models is Maddala (1983).

*8-14 HM Tory, Edmonton, AB T6G 2H4 Canada; email: brad.humphreys@ualberta.ca; phone: 780-492-5143.

2 Concepts and Notation

Dealing with zeros in economic data requires an understanding of the difference between two related concepts: truncation and censoring. While many people think that these two terms are synonymous, they have different meanings. It is important to distinguish between these two concepts, because they represent different sources of observed zeros in economic data. The potential source of zeros in economic data affects estimator choice. The appropriate estimator to apply to censored data is different from the appropriate estimator to apply to truncated data.

2.1 Truncation

Greene (2003) presents an example of a truncated distribution. Consider the income distribution and the "typical upper affluent American" who makes $142,000 per year. In this example, the sub-set of interest is households with reported income over $100,000. This sub-set constitutes 2% of the US population. A sample drawn from the sub-set of households with reported household income over $100,000 per year represents a truncated distribution. A truncated distribution is a part of an untruncated distribution that is above or below some specified value.

2.2 Censoring

When a variable is censored, values in a certain range are all transformed to, or reported as, a single value. The process of top coding, where all values of a certain variable are given the same value, is an example of censoring. There are many other examples of censoring in economic data, including the household purchase of durable goods, the number of hours worked by females, and vacation expenditures. Censoring of explanatory variables is not a problem, but censoring of a dependent variable in a regression model leads to econometric problems.

Censoring is similar to truncation, but not identical. The primary difference between truncation and censoring is that in a truncated distribution, only the distribution above or below some value is relevant. The distribution itsself is not affected, but only a portion of it is used. When data are censored, the distribution that applies to the observed data is a mixture of discrete and continuous distributions.

Censored distributions are defined using latent variables. Suppose that some random variable y is censored. To understand the distribution of y, first transform y by a a variable y^*

$$\begin{array}{ll} y = 0 & \text{if } y^* \leq 0, \\ y = y^* & \text{if } y^* > 0. \end{array} \tag{1}$$

So the latent variable y^* can be negative or zero, but if it is less than zero, then the observed outcome is a zero in the data. If the latent variable is normally distributed ($y^* \sim N[\mu, \sigma^2]$), then the distribution that applies to the observed variable y when $y = 0$ is

$$Prob(y = 0) = Prob(y^* \leq 0) = \Phi\left(-\frac{\mu}{\sigma}\right) = 1 - \Phi\left(\frac{\mu}{\sigma}\right)$$

where $\Phi(\cdot)$ is the standard normal cumulative density function. The distribution that applies to observations where $y > 0$ has the density of y^*.

Amemiya (1984) points out that "[T]he model is called *truncated* if the observations outside a specified range are totally lost and *censored* if one can at least observe the exogenous variables."

2.3 Jones' (2000) Notation

Jones (2000) assumes two variables: d_i, a binary indicator variable and y_i a continuous variable. The variable d_i must be an either/or proposition; purchasing insurance or not purchasing insurance or buying snow skis or not buying snow skis. y_i is something like insurance co-pay spending or spending on skiing. d_i is associated with a vector of covariates x_1 and a vector of unknown parameters β_1; y_i is associated with a vector of covariates x_2 and a vector of unknown parameters β_2. Since d_i is an indicator, $y_i = 0$ when $d_i = 0$ and $y_i > 0$ when $d_i = 1$. So we have

$$\begin{aligned} d_i &= x_1\beta_1 \\ y_i &= x_2\beta_2. \end{aligned} \tag{2}$$

However, Jones make no mention of the error terms in either the binary indicator equation or the continuous variable equation in the chapter. These variables are clearly stochastic, and depend on some random component. This omission is curious.

The first issue addressed by Jones is the nature of the zeros; why do we observe $y_i = 0$? There are two options:

1. $y_i = 0$ represents an actual choice of non-consumption by the individual
2. $y_i = 0$ represents a non-observable response

Given these two observable variables, Jones defines two latent, or unobservable variables

$$\begin{aligned} y_{1i}^* &= x_{1i}\beta_1 + \varepsilon_1 \\ y_{2i}^* &= x_{2i}\beta_2 + \varepsilon_2 \end{aligned} \tag{3}$$

Why do we need two latent variables? Presumably because one is associated with the participation decision and the other with the consumption decision. But Jones is not clear on this point. The important point is that these two latent variables are used to distinguish hurdle models from sample selection models.

2.4 Jones' (1989) Notation

Jones (1989) writes down a "bivariate model" and identifies the following components

(a) *Observed Consumption*

$$y_i = d \cdot y_i^{**} \tag{4}$$

(b) *Participation Equation*

$$w = \alpha' z + \nu \tag{5}$$

and participation depends on the indicator variable so that

$$\begin{aligned} d &= 1 \text{ if } w > 0 \\ d &= 0 \text{ otherwise} \end{aligned}$$

(c) *Consumption Equation*

$$\begin{aligned} y_i^{**} &= max[0, y_i^*] \\ y_i^* &= \beta' x_i + u_i \end{aligned} \tag{6}$$

2.5 Amemiya's (1984) Notation

Amemiya (1984) surveys many limited dependent variables that are applicable to the problem of zeros. His notation is useful because it clearly points out the differences among these many estimators. Amemiya motivates the models with a household utility maximization model of expenditure on a durable good. In this model y is household expenditure on a durable good, y_0 is the price of the cheapest durable good, z is all other household expenditure and x is household income. The utility function is $U(y, z)$ and the budget constraint is $y + z \leq x$. There is also a boundary constraint $y \geq y_0$ or $y = 0$. Let y^* be the utility maximizing consumption of the durable when the boundary constraint does not bind, and express this as

$$y^* = \beta_1 + \beta_2 x + u$$

where u represents all unobserved factors in the utility function. The solution to this model is

$$\begin{aligned} y &= y^* \quad \text{if } y^* > y_0 \\ y &= 0 \quad \text{if } y^* \leq y_0 \end{aligned} \tag{7}$$

Assuming that u is a random variable and y_0 varies across households and is known, this model will generate both positive observations of y and observations piled up at zero. The likelihood function for n independent observations from the model is

$$L = \prod_0 F_i(y_{0i}) \prod_1 f_i(y_i) \tag{8}$$

where F_i is the distribution function of y_i^* and f_i is the density function of y_i^*.

In this set up, the Tobit model simply assumes that y^* is normally distributed and y_0 is identical across all households. Then the *Standard Tobit Model*, or Type 1 Tobit model is

$$\begin{aligned} y^* &= x_i'\beta + U \quad i = 1, 2, \ldots, n \\ y &= y^* \qquad\quad \text{if } y^* > 0 \\ y &= 0 \qquad\quad \text{if } y^* \leq 0 \end{aligned} \tag{9}$$

where u_i are i.i.d. drawings from $N(0, \sigma_u)$. y_i and x_i are observed in the sample, but the $y*_i$ are unobserved if $y_i^* \leq 0$. The likelihood function for this model is

$$L = \prod_0 \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \prod_1 \frac{1}{\sigma}\phi\left[\frac{y_i - x_i'\beta}{\sigma}\right] \tag{10}$$

where $\Phi(\cdot)$ is the standard normal distribution function and $\phi(\cdot)$ is the standard normal density function. This is a *censored* regression model, which means that x_is are observed even when $y_i = 0$ and $y^* \leq 0$.

Suppose instead that neither y_i nor x_i are observed when $y^* \leq 0$. In this case, the data are truncated, and the corresponding likelihood function for the truncated Tobit model is

$$L = \prod_1 \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)} \tag{11}$$

Amemiya also points out that the likelihood function of the standard Tobit model can be rewritten in terms of the Probit model and the truncated Tobit model

$$L = \prod_0 \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \prod_1 \Phi\left(\frac{x_i'\beta}{\sigma}\right) \times \prod_1 \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)}. \tag{12}$$

The first term on the right hand side of equation (12) is the Probit estimator and the second term is the truncated Tobit estimator, equation (11). Thus the Tobit model is equivalent to a Probit and a truncated regression model where the variables and the parameters in the participation and consumption equations are identical.

3 Alternate Econometric Approaches for Zeros

3.1 Binary Response Models

One simple way to deal with zeros is using binary response models – single equation models that only examine the participation decision. Jones (2000) uses the following notation to describe this case: consider a binary dependent variable y_i that is equal to one if the individual, indexed by i participates in some activity and is equal to zero if not. In sports economics, examples of this type of variable include the decision to attend a sporting event or participate in sport or exercise. If the outcome variable depends on a set of regressors or explanatory variables, x_i observable for each individual in the sample, then the conditional expectation of y_i is

$$E[y_i|x_i] = P(y_i = 1|x_i) = F(x_i). \tag{13}$$

To estimate the unknown parameters of Equation 13, a number of approaches can be taken. The simplest case is when $F(\cdot)$ is assumed to be a simple linear function, $x_i\beta$. In this case, the model is called the "linear probability model" and the unknown parameters β can be estimated by OLS.

The linear probability model is convenient, and may not be a bad option if $F(\cdot)$ is approximately linear over the values in x_i, in terms of the probabilities of observing $y_1 = 1$. White's approach to correcting for heteroscedasticity can easily be applied. However, the predicted probabilities generated by the linear probability model can lie outside the $[0, 1]$ interval, leading to logical inconsistencies. For this reason, many people choose to use nonlinear approaches to estimating Equation 13.

The nonlinear approaches are motivated by a latent variable interpretation. Under this interpretation, there is some unobservable latent variable y_i^* that represents, for example, the utility derived from some activity. We cannot observe y_i^*, but only observe the indicator variable y_i. The process for the generation of y_i is

$$\begin{array}{lll} y_i = 1 & iff & y_i^* > 0 \\ y_i = 0 & & otherwise. \end{array} \tag{14}$$

Furthermore, the latent variable approach makes the assumption that the latent variable y_i^* is related to a set of observable explanatory variables (x_i) that explain variation in the latent variable according to

$$y_i^* = x_i\beta + e_i \tag{15}$$

where β is a vector of unknown parameters to be estimated and e_i is an error term – a random variable that captures all other factors that affect y_i^*. For a symmetrically distributed error term

with a distribution function defined by $F(\cdot)$, the probability of observing an indicator variable y_i equal to one – that is observing a participant – is

$$P(y_i = 1|x_i) = P(y_i^* > 0|x_i) = P(e_i > -x_i\beta) = F(x_i\beta) \tag{16}$$

or in other words, it is the probability that the realization of the error term is larger than minus the value of the observable part of the equation, $e_i > -x_i\beta$. The common binary response variables are all based on Equation 16. For example, if e_i takes the standard normal distribution,

$$f(e_i) = \frac{1}{\sigma\sqrt{2\pi\sigma^2}} e^{-\frac{(e_i-\mu_e)^2}{2\sigma^2}}$$

where σ^2 is the variance of e_i and μ_e is the mean of e_i, then Equation 16 is the probit model. If e_i takes the standard logistic distribution, then Equation 16 is the logit model. Estimation of these models is by the method of maximum likelihood. Assuming the observations in the data are independent, a log likelihood function exists

$$logL = \sum_{i=1}^{n} [(1 - y_i)log(1 - F(x_i\beta)) + y_i log(F(x_i\beta))]$$

and this function can be numerically evaluated at values of β. Maximum likelihood estimators change the parameter vector, β and recalculate the log likelihood function until $logL$ is maximized.

3.2 Jones' (2000) Alternatives

Jones (2000) describes a "taxonomy" for dealing with limited dependent variables. This taxonomy is found in the section on *Limited Dependent Variables* under the subhead "Two-part, selectivity, and hurdle models" beginning on page 285. This section is difficult to understand at first glance, because the writing is *extremely unclear* and difficult to follow.

Basically, this literature mentions three types of econometric approaches to limited dependent variables and censoring. Everyone seems to have a different name for these three approaches, and there does not appear to be any consistency in names. In the 2000 *Handbook of Health Economics* chapter on health econometrics, Jones develops a taxonomy of sample selection models and a long discussion of the taxonomy. This process has been widely adopted in the literature, for example, by Madden (2008) which is almost readable, albeit riddled with typos in the equations describing the models. Jones (2000) explicitly mentions three specific approaches for dealing with the problem of zeros:

1. **Sample Selection Models** These are appropriate when $y_i = 0$ because of a non-observable response. Jones references Heckman (1979) as an example of this approach, which is the Heckman selectivity model (probit for the selection equation and OLS plus the inverse Mills ratio term for the participation equation), sometimes called the "heckit" model.

   "In the sample selection model, knowledge that $y_i = 0$ (as opposed to $d_i = 0$) is uninformative in estimating the determinants of the level of y_i."

   The sample selection model has a latent variable representation. This representation is

$$y_i = \begin{cases} y_{2i}^* & \text{iff } y_{2i}^* > 0, \\ \text{unobserved} & \text{otherwise} \end{cases} \tag{17}$$

To estimate a sample selection model (1) estimate a Probit for participation using the full sample, (2) compute the inverse Mills ratio from the fitted Probit results, and (3) estimate the consumption equation using OLS with the inverse Mills ratio as an explanatory variable for observations with positive consumption.

2. **Two-Part Models** These are appropriate when $y_i = 0$ is a genuine zero; that means the agent has chosen to consume none of it. In addition, two-part models are appropriate when the participation and consumption decisions are chronologically sequential.

   "In the two-part model, observations for which $y_1 = 0$ are uninformative in estimating the determinants of the level of $y_i|y_i > 0$)."

   "The two-part model is usually estimated by a logit or probit model for the probability of observing a positive value of y, along with OLS on the sub-sample of positive observations. Two-part models resemble the Heckman selectivity model, but do not include the inverse Mills ratio in the second part, and thus do not correct for selectivity. There is no latent variable representation for the two-part model. Instead it is motivated by a conditional mean independence assumption." This assumption is

$$E(y_i|y_i > 0, x_{2i}) = x_{2i}\beta_2 \tag{18}$$

   To estimate a sample selection model (1) estimate a Probit for participation using the full sample, and (2) estimate the consumption equation using OLS on observations with positive consumption.

3. **Hurdle Models** These are appropriate when $y_i = 0$ is a genuine zero; that means the agent has chosen to consume none of it. In addition, hurdle models are appropriate when the participation and consumption decisions are made simultaneously.

   "In hurdle models, the fact that $y_i = 0$ is used in estimation of β_2."

   Hurdle models also have a latent variable representation.

$$y_i = \begin{cases} y_{2i}^* & \text{iff } y_{2i}^* > 0 \text{ and } y_{1i}^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

   Jones discusses hurdle models in the context of a general model, the Box-Cox double hurdle model, that includes both the generalized tobit model and "two part" dependent variable models as special cases. This general model can be motivated by a two-variable latent variable set up where the observed dependent variable if interest, y_i, can be written in terms of two dependent variables: y_1^* and y_2^* where

$$\begin{aligned} & y_{ij}^* = x_{ij}\beta_j + \varepsilon_j \quad j = 1, 2 \\ & (\varepsilon_1, \varepsilon_2) \sim N(0, \Omega) \\ & \Omega = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma^2 \end{bmatrix} \end{aligned} \tag{20}$$

   and

$$
\begin{array}{lll}
y^*_{2i} = (y^\lambda - 1)/\lambda \quad for\ \lambda > 0 & & \\
 & iff\ y^*_{1i} > 0\ and\ y^*_{2i} > -\frac{1}{\lambda} & \\
y^*_{2i} = log(y_i) \qquad for\ \lambda = 0 & & \\
y^*_{2i} = 0 & otherwise &
\end{array}
\tag{21}
$$

The latent variables have a bivariate normal conditional distribution. In this specification participation can depend on both sets of regressors, x_{i1} and x_{i2} and there can be correlation between the error terms. The log-likelihood function for this set up is

$$
\begin{array}{l}
LogL = \sum_{y=0} log[1 - \Phi(x_1\beta_1, (x_2\beta_2 + \frac{1}{\lambda})/\sigma, \rho)] \\
+ \sum_{y>0} log\Phi[(x_1\beta_1 + (\frac{\rho}{\sigma})\,[\left(\frac{y^\lambda - 1}{\lambda}\right) - x_2\beta_2]/\sqrt{1-\rho^2} \\
+ \sum_{y>0}(\lambda - 1)log(y_i) + \sum_{y>0} log\left[\frac{1}{\sigma}\phi\left(\frac{\frac{y^\lambda - 1}{\lambda} - x_2\beta_2}{\sigma}\right)\right]
\end{array}
\tag{22}
$$

3.3 Amemiya's (1984) Alternatives

Amemiya (1984) draws a contrast between the standard Tobit model, or Tobit Type 1 model, and "generalized Tobit models" of four types. This distinction is based only on the likelihood functions, and not on any behavioral assumptions, so it is cleaner and clearer than the alternative approached described by Jones (2000).

Amemiya makes it clear that the Probit model

$$
\text{Probit } L = \prod_0 \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \prod_1 \Phi\left(\frac{x_i'\beta}{\sigma}\right)
$$

which can also be written

$$
\text{Probit } L = \prod_0 P(y^* \leq 0) \prod_1 P(y^* > 0)
$$

is part of the likelihood function of all five of the Tobit-type models he reviews. That means that the Probit model is common to all generalized Tobit, hurdle, and two-part models discussed by Jones. The only difference is what remains in the likelihood function after the Probit estimator is factored out.

Amemiya (1984) identifies the following generalized Tobit models

1. **Tobit Type 1 Models** This is the standard Tobit model

$$
\begin{array}{ll}
y^* = x_i'\beta + u & i = 1, 2, \ldots, n \\
y = y^* & \text{if } y^* > 0 \\
y = 0 & \text{if } y^* \leq 0
\end{array}
\tag{23}
$$

where u_i are i.i.d. drawings from $N(0, \sigma_u)$. y_i and x_i are observed in the sample, but the $y*_i$ are unobserved if $y_i^* \leq 0$. The likelihood function for this model is

$$
L = \prod_0 \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \prod_1 \frac{1}{\sigma}\phi\left[\frac{y_i - x_i'\beta}{\sigma}\right] \tag{24}
$$

2. **Tobit Type 2 Models** These models are defined by a consumer utility maximization model in terms of two latent variables: y_{1i}^* representing (unobserved) utility from participation and y_{2i}^* representing (observed) utility from consumption. Formally

$$\begin{array}{lll} y_{1i}^* = x_{1i}'\beta_1 + u_1, & y_{2i}^* = x_{2i}'\beta_2 + u_2 & \\ & y_{2i} = y_{2i}^* & \text{if } y_{1i}^* > 0 \\ & y_{2i} = 0 & \text{if } y_{1i}^* \leq 0 \end{array} \tag{25}$$

where (u_{1i}, u_{2i}) are realizations from an independent and identically distributed, mean zero, constant variance bivariate normal distribution. The variance of these two variables are σ_1^2 and σ_2^2 and the covariance between them is σ_{12}. By assumption, only the sign of y_{1i}^* is observed. y_{2i}^* is only observed when y_{1i}^* is positive. Also, variables in x_{1i} are observed for all i but variables in x_{2i} may not be observed when the utility of participation is negative ($y_{1i}^* \leq 0$.) These models usually contain an indicator variable for participation

$$\begin{array}{ll} w_{1i} = 1 & \text{if } y_{1i}^* > 0 \\ w_{1i} = 0 & \text{if } y_{1i}^* \leq 0. \end{array} \tag{26}$$

In this case, the observed variables in the sample are pairs (w_{1i}, y_{2i}). The likelihood function for this model is

$$L = \prod_0 P(y_{1i}^* \leq 0) \prod_1 f(y_{2i}|y_{1i}^* > 0) P(y_{1i}^* > 0) \tag{27}$$

where $f(\cdot|y_{1i}^* > 0)$ is the conditional density of y_{1i}^* given that the utility from participation is positive ($y_{1i}^* > 0$.) Note that this likelihood function can also be written

$$L = \prod_0 P(y_{1i}^* \leq 0) \prod_1 P(y_{1i}^* > 0) \times \prod_1 f(y_{2i}|y_{1i}^* > 0). \tag{28}$$

The first term on the right hand side of equation (28) is the Probit model for participation. The second term on the right hand side of equation (28) is the conditional density of the utility from consumption, given that the utility from participation is positive for the observations where positive consumption is observed.

Equation (27) can be written

$$L = \prod_0 P(y_{1i}^* \leq 0) \prod_1 \int_0^\infty f(y_{1i}^*, y_{2i}) dy_{1i}^* \tag{29}$$

where $f(\cdot, \cdot)$ is the joint density function for y_{1i}^* and y_{2i}^*. Amemiya (1994) shows that, using the definition of a conditional density function, this equation can be re-written

$$L = \prod_0 \left[1 - \Phi\left(\frac{x_i'\beta_1}{\sigma_1}\right)\right] \times \prod_1 \Phi\left[\frac{\left(\frac{x_{1i}'\beta_1}{\sigma_1} + \frac{\sigma_{12}}{\sigma_1\sigma_2}(y_{2i} - x_{2i}'\beta_2)\right)}{\left(\sqrt{1 - \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}}\right)}\right] \frac{1}{\sigma_2}\phi\left(\frac{y_{2i} - x_{2i}'\beta_2}{\sigma_2}\right) \tag{30}$$

This is the "full double hurdle model" derived by Jones (1992) that includes correlation between (u_{1i}, u_{2i}). Assuming no correlation between u_{1i} and u_{2i} ($\sigma_{12} = 0$) yields the Cragg model

$$L = \prod_0 \left[1 - \Phi\left(\frac{x_i'\beta_1}{\sigma_1}\right)\right] \prod_1 \Phi\left[\frac{x_{1i}'\beta_1}{\sigma_1}\right] \frac{1}{\sigma_2}\phi\left(\frac{y_{2i} - x_{2i}'\beta_2}{\sigma_2}\right) \tag{31}$$

where the first two terms on the right hand side are the Probit model for participation and the third term is a truncated normal regression model. Furthermore, assuming that all observations represent participants ($P(y_{1i}^* = 1$ and $y_{2i} = y_{2i}^*)$ and that y_{2i}^* is distributed normally with mean $x_{1i}'\beta_1$ and variance σ_1^2 yields the Tobit model

$$L = \prod_0 \left[1 - \Phi\left(\frac{x_i'\beta_1}{\sigma_1}\right)\right] \prod_1 \frac{1}{\sigma_1}\phi\left(\frac{y_{1i} - x_{1i}'\beta_1}{\sigma_1}\right) \tag{32}$$

3. **Tobit Type 3 Models** Tobit Type 3 models differ from Tobit Type 2 models only in that the latent variable y_{1t}^* can be observed when it is positive in the Type 3 model. Heckman's two-part model is in this class.

$$\begin{array}{ll} y_{1i}^* = x_{1i}'\beta_1 + u_1 & y_{2i}^* = x_{2i}'\beta_2 + u_2 \\ y_{1i} = y_{1i}^* & \text{if } y_{1i}^* > 0 \\ y_{1i} = 0 & \text{if } y_{1i}^* \leq 0 \\ y_{2i} = y_{2i}^* & \text{if } y_{1i}^* > 0 \\ y_{2i} = 0 & \text{if } y_{1i}^* \leq 0 \end{array} \tag{33}$$

where (u_{1i}, u_{2i}) are i.i.d drawings from a bivariate normal distribution with mean zero and constant variances σ_1^2 and σ_2^2 and covariance σ_{12}. By assumption, y_{1i}^* and y_{2i}^* are both observed when y_{1i}^* is positive; also, x_{1i} are observed for all i but x_{2i} may not be observed for $y_{1i}^* \leq 0$.

The likelihood function for this model is

$$L = \prod_0 P(y_{1i}^* \leq 0) \prod_1 f(y_{1i}, y_{2i}) \tag{34}$$

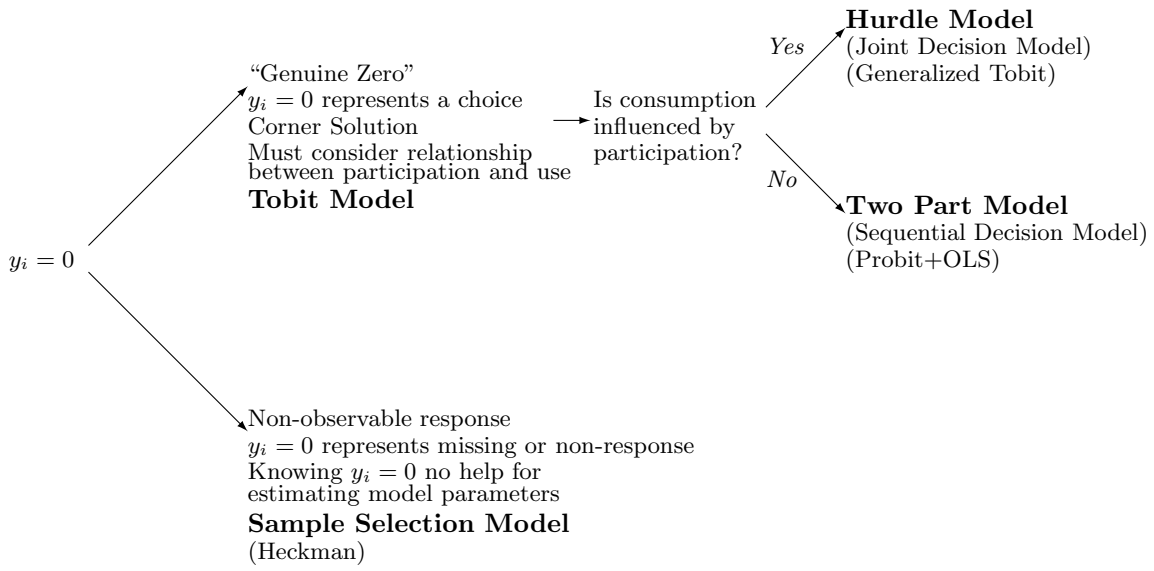
where $f(\cdot, \cdot)$ is the joint density of y_{1i}^* and y_{2i}^*. Since y_{1i}^* is observed when it is positive, all the parameters of the model are identifiable, including σ_2.

3.4 Discussion

The key decision facing any researcher working with a data set containing zeros is the choice of an estimation approach. The following decision tree summarizes the key elements of deciding on the appropriate estimator. Suppose that the variable of interest is y_i, and there are a large number of zero values for y in a given data set.

The first step is to determine why the zeros are present in the data. There are three alternatives: (1) The zeros represent a choice made by the agents in the survey; (2) the zeros represent either missing or non-response outcomes (3) the zeros represent a decision that the agent had no control over for some reason. The first alternative implies that the zero is the result of an optimal choice of some sort. For example, in the case of some discretionary leisure activity like gambling, zero spending on gaming for a household surveyed could be because zero expenditure on gambling is

Figure 1: Choosing an Estimator



optimal for that household. This outcome can be thought of as a corner solution to a constrained utility maximization problem. The second alternative implies that the individual decided to to respond to the question, or that the time frame of the survey was too short to include the behavior. For example, households purchase consumer durables like cars or televisions infrequently; a survey asking if a household purchased a car in the past month might contain many zeros. The third alternative is when the agent had no choice about the outcome. For example, in a data set containing information about the playing career of football (soccer) players, some players are observed to be "internationals" (played on their national team) and others are not. This is a binary choice model (1=international, 0=not) but the player had no part in the decision.

Under alternative (1), some individuals who are observed not participating could participate under some conditions. The price of the good may be high enough to generate a corner solution to the agents' utility maximization problem, for example, but at a lower price the individual might choose to participate. The second two do not involve any choice, so non-participants will always be observed not participating, no matter what happens.

4 Estimators

4.1 Tobit

The standard econometric method used in the case of censored dependent variables is the tobit model. The tobit model is typically written in terms of an index function (or latent variable)

$$\begin{aligned} y_i^* &= \beta' x_i + \epsilon_i \\ y_i &= 0 \qquad \text{if } y_i^* \leq 0, \\ y_i &= y_i^* \qquad \text{if } y_i^* > 0. \end{aligned} \tag{35}$$

Note that there are three different conditional mean functions that apply to the tobit model, one for each equation. For the latent variable

$$E[y_i^*] = \beta' x_i$$

For any observation randomly drawn from the population that could be censored or not censored

$$E[y_i|x_i] = \Phi\left(\frac{\beta' x_i}{\sigma}\right)(\beta' x_i + \sigma\lambda_i)$$

where

$$\lambda_i = \frac{\phi(\beta' x_i/\sigma)}{\Phi(\beta' x_i/\sigma)}.$$

4.1.1 Likelihood Functions

Several ways of writing the likelihood function for the Tobit model exist in the literature. Greene (2000) writes the log likelihood function for the tobit estimator as

$$logL = \sum_{y_i>0} -\frac{1}{2}\left[log(2\pi) + log(\sigma^2) + \frac{y_i - \beta' x_i}{\sigma^2}\right] + \sum_{y_i=0} log\left[1 - \Phi\left(\frac{\beta' x_i}{\sigma}\right)\right] \tag{36}$$

while Bockstael et al. (1990) write the Tobit likelihood function as

$$L = \prod_0 \left[1 - \Phi\left(\frac{\beta z_i}{\sigma_u}\right)\right] \prod_+ \frac{1}{\sqrt{2\pi\sigma_u^2}} \cdot e^{\left(\frac{(x_i - \beta z_i)^2}{-2\sigma_u^2}\right)}. \tag{37}$$

Since Tobit is a standard part of all current econometric packages, there is little need to write out this likelihood function explicitly – it does not need to be coded explicitly in any program. However, the form of the likelihood function is useful for understanding the relationship between the Tobit and double hurdle models.

4.2 Hurdle Models

"Hurdle model" is a generic term for a model applicable to situations where $y_i = 0$ is a "genuine" zero, in that the zero is the result of a utility maximizing choice, and the decision includes both a purchase/participation decision and a consumption decision. Jones' "full double hurdle" model and the Cragg model are examples of hurdle models.

4.3 Double Hurdle Models

Jones (1992) derives a "full double hurdle" model that includes dependence between the error term in the participation equation and the time equation. The primary difference between the Cragg model and this "full double hurdle" model is this correlation in the equation error terms. In Amemiya's notation, based on equation (26), $corr(u_1, u_2) \neq 0$ for a double hurdle model and $corr(u_1, u_2) = 0$ for the Cragg model.

4.3.1 Double Hurdle Model Likelihood Functions

The likelihood function for the "full double hurdle model" developed by Jones (1992) is

$$L = \prod_0 \left[1 - \Phi\left(X_1\beta_1, \frac{X_2\beta_2}{\sigma}, \rho\right)\right] \times \prod_+ \left[\Phi\left(\frac{X_1\beta_1 + \frac{\rho}{\sigma}(y - X - 2\beta_2)}{\sqrt{1-\rho^2}}\right)\frac{1}{\sigma}\phi\left(\frac{y - X_2\beta_2}{\sigma}\right)\right] \tag{38}$$

4.4 Cragg Model

The Cragg model is a "double hurdle" model in that there are two decisions, a purchase/participation decision and a consumption/time spent decision. The model was first developed by Cragg (1971). It assumes that the unobservable factors affecting the purchase/participation decision are uncorrelated with the unobservable factors affecting the consumption/time spent decision. The model has not been widely used in economics, except in the recreation and resource economics literature.

4.4.1 Cragg Model Likelihood Functions

Madden (2008) discussed the relative merits of double hurdle, selection, and two part models in the context of the decision to smoke and drink alcohol. Madden (2008) used the notation from Jones (1989), which is discussed in Section 2.4. Using this notation, Madden (2008) defined the Cragg model likelihood function as

$$L1 = \prod_{0} \left[1 - p(\nu > -\alpha' Z) p(u > -\beta' X)\right] \prod_{+} p(v > -\alpha' Z) p(u > -\beta' X) g(y | u > -\beta' X). \tag{39}$$

Bockstael et al. (1990) write the Cragg model likelihood function as

$$L = \prod_{0} \left[-\Phi\left(\frac{\beta z_i}{\sigma_u}\right)\right] \prod_{+} \frac{1}{\sqrt{2\pi\sigma_u^2}} \cdot e^{\left(\frac{(x_i - \beta z_i)^2}{-2\sigma_u^2}\right)} \tag{40}$$

Jones (1992) writes the Cragg model as a special case of the full double hurdle model. From equation (38), set $\rho = 0$ (ρ is the correlation between the participation equation error term and the time equation error term); this yields

$$L = \prod_{0} \left[1 - \Phi\left(X_1\beta_1, \frac{X_2\beta_2}{\sigma}\right)\right] \times \prod_{+} \left[\Phi\left(X_1\beta_1\right) \frac{1}{\sigma} \phi\left(\frac{y - X_2\beta_2}{\sigma}\right)\right] \tag{41}$$

4.5 Two Part Models

Two Part Models apply to the case where observed zeros represent corner solutions but the consumption decision does not depend on the participation decision. Two part models are independently estimated participation functions and consumption functions. The basic idea behind two part model is that the participation decision differs from the quantity decision in a fundamental way.

5 Empirical Examples

5.1 Garcia and Labeaga, *Oxford Bulletin*, 1996

In "Alternative approaches to modelling zero expenditure: An application to Spanish demand for tobacco," García and Labeaga (1996) examine different approaches for econometric modelling of zeros with an application to cigarette smoking in Spain. They discuss three reasons for zeros appearing in survey data:

1. In survey data with short recording periods (one week or one month, for example), purchase decisions could be made infrequently, generating zeros (infrequency of purchase)

2. Households may never consume the good under any circumstance (abstention)

3. Households would consumer the good, but at current income and prices, non is purchased during the recording period (corner solution).

This paper explores the difference between corner solutions and abstention. The paper formulates a demand function for cigarettes

$$y_i^* = \beta' X_i + \varepsilon_i$$

where y_i^* is a latent variable capturing utility generated from consuming cigarettes (they call it "notional demand" to distinguish this from "observed demand" y_i), X_i is a vector of explanatory variables, and ε_i is an unobservable random variable that captures all other factors that affect the decision about the number of cigarettes to smoke. In addition, the paper formulates an observability rule

$$I_i^* = \alpha' Z_i + \nu_i$$

where I_i^* is an unobservable indicator that determines participation in smoking. ν_i is an unobservable random variable that captures unobservable factors that affect the decision to smoke.

- **Double Hurdle Dependent Model**: If (ε_i, ν_i) are jointly distributed as a bivariate normal random variable with zero means and unit variances and a correlation coefficient of ρ, then the parameters of the two equations above can be estimated by evaluating the likelihood function

$$L_{DHD} = \prod_1 P(\nu_i > -\alpha' Z_i) P(\varepsilon_i > -\beta' X_i | \nu_i > -\alpha' Z_i) f(y_i | \varepsilon_i > -\beta' X_i, \nu_i > -\alpha' Z_i) \times \prod_0 (1 - P(\nu_i > -\alpha' Z_i) P(\varepsilon_i > -\beta' X_i | \nu_i > -\alpha' Z_i))$$

  where $f(\cdot)$ is the normal probability distribution function. Note that $P(\nu_i > -\alpha' Z_i)$ is the probability of observing a positive I_i^*, and $P(\varepsilon_i > -\beta' X_i | \nu_i > -\alpha' Z_i)$ is the probability of observing a positive y_i^* conditional on observing a positive I_i^*.

  The censoring mechanism for this model (the process which generates the zeros) is

$$y_i = 1(I_i^* = 1) \times max(y_i^*, 0)$$

  where $1(I_i^* = 1)$ is an indicator function for the occurrence of the event $I_i^* = 1$. That is, positive expenditure is observed if both no abstention occurs ($I_i^* = 1$) and if $y_i^* > 0$, there is no corner solution. However, zero expenditure could still be observed for someone at a corner solution.

- **Double Hurdle Independent Model**: If (ε_i, ν_i) are independent, then the likelihood function becomes

$$L_{DHD} = \prod_1 P(\nu_i > -\alpha' Z_i) P(\varepsilon_i > -\beta' X_i) f(y_i | \varepsilon_i > -\beta' X_i, \nu_i > -\alpha' Z_i) \times \prod_0 (1 - P(\nu_i > -\alpha' Z_i) P(\varepsilon_i > -\beta' X_i))$$

  This is the 'Cragg" model discussed above.

- **First Hurdle Dominance**: "Participation dominates consumption." Once an individual decides to participate, consumption takes place and there are no corner solutions. Statistically, this means $P(y_i^* \leq 0 | I_i^* = 1) = 0$, or in other words that the probability of observing an individual consuming no cigarettes ($P(y_i^* \leq 0)$ conditional on the first hurdle being met ($I_i^* = 1$) is zero. The first hurdle, in this case the decision to smoke or not to smoke, is dominant in that every person who has decided to be a smoker ($I_i^* = 1$) will be observed smoking in the sample. The likelihood function for this model is the Heckman generalized sample selectivity model

$$L_H = \prod_1 P(\nu_i > -\alpha' Z_i) f(y_i | \varepsilon_i > -\beta' X_i, \nu_i > -\alpha' Z_i) \times \prod_0 (1 - P(\nu_i > -\alpha' Z_i))$$

  To get from the double hurdle dependent model to the Heckman model, simply observe that $P(\varepsilon_i > -\beta' X_i) = 1$, implying that in the Heckman model the only important factor in determining cigarette consumption is the first hurdle.

- **Tobit** (Typical corner solution) Under the case where the first hurdle is irrelevant, the double hurdle model collapses to the familiar Tobit specification.

$$L_T = \prod_1 P(\varepsilon_i > -\beta' X_i) f(y_i | \varepsilon_i > -\beta' X_i) \times \prod_0 (1 - P(\varepsilon_i > -\beta' X_i)).$$

  The censoring mechanism that generates the zeros for the Tobit specification is

$$y_i = max(y_i^*, 0)$$

  Which means that when y_i is not observed it is replaced with a zero, and when it is observed it is equal to notional demand. All the zero expenditure observations are generated by non-expenditure, but nothing else is known about the decision not to purchase.

5.2 An Application to Sports Participation

5.2.1 Data

As an example of how to model zeros in a sport setting, consider the decision by individuals to participate in sport or physical activity. The Behavioral Risk Factor Surveillance System (BRFSS), is conducted annually by telephone to a random representative sample of the population over the age of 18 in states in the United States by the Center for Disease Control and Prevention (CDC). The survey collects data on preventative health factors, behavioral risk factors, and other economic and demographic characteristics and includes a rotating selection of modules, including one on participation in exercise and physical activity.

The survey asks about both frequency and duration of participation, which provides a relatively complete picture of self reported physical activity. The survey also asks questions about demographic factors like age, gender, race, ethnicity, and marital status, and questions about economic factors like income and labor market participation. This makes the BRFSS data an ideal setting for examining the economic determinants of physical activity. The physical activity module is not included in every survey year. We will use data from the 2000 BRFSS survey, which included a module about physical activity and exercise.

Figure 2: Participation in Physical Activity in Sample

```
 Physically |
     Active |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        908       19.51       19.51
          1 |      3,746       80.49      100.00
------------+-----------------------------------
      Total |      4,654      100.00
```

184,450 persons were surveyed in the 2000 BRFSS survey. We will work with a heterogenous sample consisting of only employed males between the ages of 30 and 35. This sample individuals should be relatively heterogenous. The data file `pe_data_zurich.dta` contains the data analyzed in this example. It contains 4,654 observations.

The 2000 BRFSS survey contained a module of questions on physical activity. The basic physical activity question in the BRFSS survey is

> *During the past month, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?*

Participation in physical activity can be defined using this survey question. The responses to this question are coded in the variable `active` in the data file. This variable is equal to one if an individual responded "yes" and zero if the individual responded "no" to the question. In this sample, about 80.5% of the individuals responded "yes" and about 19.5% responded "no." This is the binary dependent variable y_i in the notation developed by Jones (2000) in Section 3 that indicates if an individual is a participant ($y_i = 1$) or a non-participant ($y_i = 0$). Figure 2 shows the distribution of the sport participation variable, based on the Stata command `--tabulate--`.

BRFSS also contains detailed information about how frequently individuals in the survey participated in physical activities in the past month, and how much time the individuals spent in each activity on average. These data provide enough detail to construct an estimate of the number of times per week and minutes per week that each individual in the survey spent participating in physical activity. The variable `empw1` [(e)xercise (m)inutes (p)er (w)eek] contains an estimate of the time spent in physical activity in an average week for the individuals in the sample. There are two ways to examine the distribution of this variable: across all individuals and across only participants.

Figure 3 shows both, using the Stata command `--summarize--`. The fist command summarizes the distribution of the time spent variable for all individuals in the sample. Notice that the minimum value of this variable is zero, so it contains both participants and nonparticipants. The unconditional mean of the variable is 182.9, indicating that the average person in the sample spends about 3 hours per week participating in physical activity. However, this contains 908 individuals who do not participate. For participants, the time spent variable must be summarized only for participants. The second command summarizes the time spent variable for participants only. Note that 4654 observations are used to calculate the first summary statistics but just 3739 observations are used for the second. Conditional on participation, the average participant spends 228 minutes, nearly 4 hours per week, participating in physical activity. The average time spent in physical activity is longer because non-participants have been eliminated from the second summary command. Note

Figure 3: Time Spent in Physical Activity in Sample

```
. sum empw1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       empw1 |      4654    182.9351    435.4262          0      23760
. sum empw1 if empw1>0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       empw1 |      3739    227.7026    475.1941         .5      23760
```

Figure 4: Problems Identifying Participation

```
. tab active if empw1==0
 Physically |
     Active |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        908       99.23       99.23
          1 |          7        0.77      100.00
------------+-----------------------------------
      Total |        915      100.00
```

that the variance time spent participating in physical activity does not change much when the nonparticipants are removed from the calculation.

Survey data typically have problems that need to be cleaned up before they can be analyzed. In this case there are two data problems that will lead to issues when estimating an empirical model of time spent in physical activity. First, there are some discrepancies between individuals identified as non-participants from the participation question and the time spent question. Second, there are some outliers – extremely large values – in the time spent in physical activity.

Either because of coding errors, or because of mistakes made when answering the lengthy BRFSS survey questionnaire, the discrete indicator variable and the continuous variable are sometimes not codded consistently. In this case, some of the individuals who reported spending no time participating in physical activity in a week answered "yes" to the question about participation in physical activity. This can be seen by a simple tabulation of the values of the variable `active` when the participation variable `empw1` is equal to zero. This can be seen in Figure 4.

Recall from Figure 2 that 908 individuals reported not participating in physical activity in the last month. From Figure 4, when the time spent in participation variable `empw1` is equal to zero, there are 908 observations where `active` equal to zero and 7 observations where `active` is equal to 1. These 7 observations will cause problems in later estimation of the multivariate models, because the indicator variable and the time spent variable must agree in those settings, or the likelihood function will not be properly defined.

One of two changes needs to be made to fix this problem. Either the seven observations where `active` is equal to 1 must be changed to zero, or the seven observations where `active` is equal to 1 and `empw1` is not equal to zero must have a non-zero value filled in. Since we have no way to determine how much time those individuals actually spent participating, setting `active` equal to

zero is the best course of action here. Alternatively, those observations could be dropped.

The second data problem has to do with the amount of time spent participating in physical activity. From Figure 3, the maximum value of `empw1` in the sample is 23,760 minutes. That is almost 400 hours, an amount that cannot be possible. The fact that such a large value appears suggests that other problems may be present in the time spent variable. Figure 5 shows the right tail of the distribution of the time spent variable. 720 minutes is 12 hours spent participating in physical activity per week. That is just over 1.7 hours per day. Clearly the largest value is impossible - it is larger than the number of minutes in a week (10,800). 4800 minutes is more than 11 hours per day, seven days per week.

Determining where to trim the sample in this case depends on more than just what values might be considered "reasonable" based on the typical number of hours per week available to participate in physical activity. The empirical methods discussed below are maximum likelihood methods, and require that the statistical package used evaluate the likelihood function at all values in the sample. Some methods, especially the double hurdle models, are sensitive to outliers in that the presence of outliers in the data leads to less well behaved likelihood functions, and makes the estimation procedure settle on non-global maxima or to not converge to a local maxima.

In this case, I have removed all observations with an estimated number of minutes per week spent in physical activity greater than 720 minutes. From Figure 5, this removes slightly less than 3% of the observations from the sample.

Following Jones (2000), assume that this participation outcome depends on a set of regressors x. We will use a standard set of regressors used in the literature on the economic determinants of physical activity: income, marital status, and education. Income is measured in thousands of dollars, and there are no zeros because the sample has been restricted to individuals who are employed. Marital status is a dummy variable equal to one if the individual is married and zero otherwise. Education is a dummy variable equal to one if the individual is a college graduate and zero otherwise.

Table 1 summarizes the distribution of the explanatory variables used in this example, and the values of all variables after the sample was restricted to eliminate the outliers in terms of the time spent participating in physical activity variable. The values on the top panel are for the unrestricted sample; the values on the bottom are for the restricted sample, and for the seven recoded participation variables. Note that restricting the sample had very little effect on the means and variances, except for the time spent variable. Eliminating observations where time spent was more than 720 minutes per week reduced the average time spent to 146 minutes, from the 183 minutes per week reported above. The average time spent in physical activity for participants in the reduced sample was just over 183 minutes.

5.2.2 Analysis of Binary Response Variables

The simplest approach to analyzing participation in physical activity is to examine the determinants of participation using standard univariate limited dependent variable estimators. From Section 3.1 above, these are the probit, logit, and linear probability models. Lets have a look at the results that are generated by these three alternative estimators using the physical activity data described above.

The linear probability model (LPM) applies OLS to the model

$$y_i = x_i\beta + e_i$$

where y_i is an indicator variable for participation in physical activity, x_i is a vector of explanatory variables, β is a vector of unknown parameters to be estimated, and e_i is an unobservable random

Figure 5: Right Tail of Time Spent Variable

Estimated min. per week 2	Freq.	Percent	Cum.
720	37	0.80	97.16
748.75	1	0.02	97.19
750	9	0.19	97.38
770	2	0.04	97.42
780	1	0.02	97.44
810	3	0.06	97.51
840	16	0.34	97.85
900	20	0.43	98.28
960	9	0.19	98.47
990	2	0.04	98.52
1050	5	0.11	98.62
1080	2	0.04	98.67
1125	1	0.02	98.69
1198	1	0.02	98.71
1200	12	0.26	98.97
1260	10	0.21	99.18
1305	1	0.02	99.20
1350	3	0.06	99.27
1440	3	0.06	99.33
1500	3	0.06	99.40
1680	3	0.06	99.46
1800	8	0.17	99.63
1890	1	0.02	99.66
1920	1	0.02	99.68
2100	2	0.04	99.72
2160	1	0.02	99.74
2400	5	0.11	99.85
2520	1	0.02	99.87
2995	1	0.02	99.89
3360	2	0.04	99.94
3600	1	0.02	99.96
4800	1	0.02	99.98
23760	1	0.02	100.00
Total	4,654	100.00	

Table 1: Summary Statistics

Variable	Variable Name on File	Observations	Average	St.Dev.	Min	Max
Income	income	4654	56.60	33.02	5	115
Marital status	married	4654	0.583	0.493	0	1
Education	colgrad	4654	0.399	0.490	0	1
Income	income	4522	56.69	33.04	5	115
Marital status	married	4522	0.584	0.493	0	1
Education	colgrad	4522	0.402	0.490	0	1
Participation	active	4522	0.798	0.402	0	1
Time Spent	empw1	4522	145.99	154.7	0	720

Figure 6: Linear Probability Model Results

```
Linear regression                               Number of obs =    4522
                                                F(  3,  4518) =   60.95
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0359
                                                Root MSE      =  .39464
------------------------------------------------------------------------------
             |               Robust
      active |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0014191   .0001886     7.53   0.000     .0010494    .0017888
     married |  -.0324999   .0122228    -2.66   0.008    -.0564626   -.0085372
     colgrad |   .0936435   .0122064     7.67   0.000      .069713    .1175739
       _cons |   .6985196   .0134349    51.99   0.000     .6721805    .7248586
------------------------------------------------------------------------------
```

error terms that captures all other factors that affect the dependent variable. The LPM has several well known problems: it is heteroskedastic and it generates predictions outside the $(0, 1)$ interval. Figure 6 contains the Stata output from the linear probability model for participation in physical activity based on the 4522 observations from the 2000 BRFSS.

All three of the explanatory variables are highly significant. The model explains only about 3.5% of the observed variation in participation in physical activity in the sample. Note the use of the `robust` option, which invokes the White-Huber "sandwich" correction for heteroscedasticity in Stata. This needs to be used because the linear probability model is heteroscedastic. The variance of e_i observations where $y_i = 0$ is 0, while the variance for observations, while the variance of e_i observations where $y_i = 1$ is σ_e^2. Income is associated with an increase in the probability that an individual participates in physical activity. Being married reduces the probability of participating in physical activity. College graduates are more likely to participate in physical activity.

In this case, the predicted values from the model fall in the interval $\hat{y}_i \in [0.673, 0.955]$. None of the predicted values fall outside the $[0, 1]$ interval. So the commonly invoked problem with the linear probability model does not occur in this case. Although little of the observed variation in participation is explained, the model identifies three factors that explain variation in participation:

Figure 7: Probit Model

```
Probit regression                                 Number of obs   =       4522
                                                  LR chi2(3)      =     172.40
                                                  Prob > chi2     =     0.0000
Log likelihood = -2191.2367                       Pseudo R2       =     0.0378
------------------------------------------------------------------------------
      active |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0055617   .0007665     7.26   0.000     .0040593    .0070641
     married |  -.1234432    .045198    -2.73   0.006    -.2120296   -.0348568
     colgrad |   .3576935   .0491565     7.28   0.000     .2613485    .4540384
       _cons |   .4823051   .0459973    10.49   0.000     .3921519    .5724582
------------------------------------------------------------------------------
```

income, marital status, and education.

The LPM is widely used in economics, even with the advent of cheap and powerful desktop computers and widespread implementation of alternatives like probit and logit. I argue that the LPM shoud not be used at all. Why? the LPM has a very undesirable lage sample property: it is inconsistent. In a recent paper, Horrace and Oaxaca (2006) show that the LPM is biased and inconsistent. That means you cannot trust the estimates to converge to the underlying population parameters, even in large samples.

The next alternative for dealing with binary response variables is the probit model. Probit is a maximum likelihood estimator, and is commonly written

$$\text{Probit } L = \prod_{0} \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \prod_{1} \Phi\left(\frac{x_i'\beta}{\sigma}\right).$$

Probit models are estimated numerically - the econometric analysis program you are using evaluates this likelihood function using your data and initial values of the parameters, and changes the parameters until it finds a maximum on the likelihood surface. Figure 7 shows the probit results for the regression model described above.

The parameter estimates from probit have no direct interpretation. A positive parameter indicates that the probability of observing a one increases with that variable, and a negative value indicates that the probability of observing a one decreases with that variable. From Figure 7, people with higher income are more likely to participate in physical activity, married people are less likely to participate than single people, and college graduates are more likely to participate than those without a college degree. We cannot determine by how much the probability of participation changes from the probit parameter estimates. Note that these results are qualitatively identical to those from the LPM above, in terms of the sign and significance of the parameter estimates.

In order to get a better understanding of the size of the relationship, many people report marginal effects based on probit estimates. The marginal effect simply indicates how much the probability of observing a one increases with a one unit change in the explanatory variable. Figure 8 shows the marginal effects for this probit model.

Note that the probit marginal effects provide a more precise picture of the relationship between the explanatory variables and participation in physical activity. For each additional $1,000 in income, individuals are one tenth of one percent more likely to participate in physical activity.

Figure 8: Probit Model with Marginal Effects

```
Probit regression, reporting marginal effects           Number of obs =   4522
                                                        LR chi2(3)    = 172.40
                                                        Prob > chi2   = 0.0000
Log likelihood = -2191.2367                             Pseudo R2     = 0.0378
------------------------------------------------------------------------------
  active |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
  income |   .0015205   .0002082     7.26   0.000   56.6945   .001112  .001929
 married*|  -.0334402   .0121203    -2.73   0.006   .583591  -.057195 -.009685
 colgrad*|   .0947044   .0125084     7.28   0.000   .402034   .070188   .11922
---------+--------------------------------------------------------------------
  obs. P |   .7976559
 pred. P |   .8076823  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

Married people are 3% less likely to participate in physical activity than non-married people. College graduates are 9% more likely to participate in physical activity than those without a college degree.

5.2.3 Analysis of Censored Variables

The participation in physical activity data described above can also be used to illustrate different estimators that can be applied to censored variables. In this case, the variable `empw1`, minutes per week spent participating in physical activity, is a censored variable. It has properties of a discrete variable, in that non-participants have a value of zero and participants have a positive value for this variable, and of a continuous variable, in that participants also report their number of minutes per week participating, which is a continuous variable. There are three ways to empirically model time spent in physical activity: using Tobit, using the Heckman selectivity model, and using double hurdle models.

The Tobit model is a univariate approach. It estimates a single equation with the censored variable `empw1` as the dependent variable. The Tobit model was discussed in Section 4. Tobit models can be motivated using a standard latent variables approach, like shown in Equation (23). The likelihood function for the Tobit model is Equation (24). Figure **??** shows the results obtained when Tobit is used to estimate the regression model shown in Equation (24).

In this case, the dependent variable is the number of minutes spent participating in physical activity per week where non-participants have zero minutes reported. All three parameter estimates are statistically significant, suggesting that all three of these factors are associated with time spent participating in physical activity. Time spent increases with income, and married people spend about 26 minutes less participating in physical activity than non-married people (which in this case included singles, widowed people, and divorcees.) Recall that the Tobit model forces the effect of the explanatory variables on time spent and participation to be identical in sign and size. So the Tobit results on Figure **??** imply that the effect of income, marital status, and education have

Figure 9: Tobit Model Results

```
. tobit empw1 income married colgrad, ll(0);
Tobit regression                                  Number of obs   =       4522
                                                  LR chi2(3)      =      79.40
                                                  Prob > chi2     =     0.0000
Log likelihood = -24775.684                       Pseudo R2       =     0.0016
------------------------------------------------------------------------------
       empw1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .5998516   .0936154     6.41   0.000     .4163197    .7833835
     married |  -25.94937   5.846495    -4.44   0.000    -37.41136   -14.48738
     colgrad |    16.4301   6.137446     2.68   0.007     4.397704     28.4625
       _cons |    97.8811   6.112434    16.01   0.000     85.89774    109.8645
-------------+----------------------------------------------------------------
      /sigma |   183.4245   2.250029                      179.0134    187.8357
------------------------------------------------------------------------------
  Obs. summary:        915  left-censored observations at empw1<=0
                      3607     uncensored observations
                         0 right-censored observations
```

the same effect on the participation decision as on the quantity decision.

The double hurdle model is a multivariate approach that uses both participation and time spent, relaxing the key assumption made by the Tobit model. For this reason, these two variables (participation and time spent) must both agree, in that all nonparticipants must have zero minutes of time spent and all participants must have positive minutes of time spent. A double hurdle model has two equations that are estimated simultaneously: a participation equation with a discrete dependent variable and a quantity equation with a continuous, censored dependent variable. Figure 10 contains the results of applying the double hurdle estimator to the physical activity participation data. This was estimated using the user written STATA package `--dhurdle--`.[1]

Comparing Figures 9 and 9 illustrate the difference between the Tobit and Double hurdle models. The Tobit results indicate that marital status decreases both participation and time spent. However, the double hurdle results indicate that marital status only affects time spent, and that married people are no less likely to participate in physical activity than non-married people. Conditional on participation, married people spend about 25 minutes less per week participating in physical activity.

The Heckman model is also a multivariate approach for dealing with censored variables. It differs from the double hurdle model in that the fitted value from the first stage equation, the model describing participation in physical activity in this case, is added as an explanatory variable to the quantity equation, in this case the model describing the determination of the number of minutes spent participating in physical activity. Figure 11 contains the results generated when the Heckman model is applied to the physical activity participation data.

The Heckman model may require an identifying assumption. That means excluding one variable that appears in the participation equation from the quantity equation. In this case, I have imposed

[1] Available at `http://www.sml.hw.ac.uk/staffpages/somjaf/Stata/`.

Figure 10: Double Hurdle Model Results

```
. dhurdle empw1 income married colgrad,
sel(active = income married colgrad) technique(nr) mlmethod(d1) independent;

Double hurdle model                              Number of obs      =      4522
(model with selection and censoring)             Censored obs       =       915
                                                 Uncensored obs     =      3607
Log likelihood = -24766.73                       Independent errors
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
empw1        |
      income |   .4693371   .0981889     4.78   0.000     .2768905    .6617838
     married |  -25.25936   6.246232    -4.04   0.000    -37.50175   -13.01697
     colgrad |   11.55241   6.259515     1.85   0.065    -.7160164    23.82083
       _cons |   111.9538   7.254059    15.43   0.000     97.73606    126.1715
-------------+----------------------------------------------------------------
active       |
      income |   .0613327   .0129028     4.75   0.000     .0360436    .0866218
     married |  -.2322775    .285117    -0.81   0.415    -.7910966    .3265416
     colgrad |   4.021249   104.2219     0.04   0.969    -200.2499    208.2924
       _cons |   .1155949   .2963475     0.39   0.696    -.4652355    .6964253
-------------+----------------------------------------------------------------
    /lnsigma |   5.199103   .0127616   407.40   0.000     5.174091    5.224115
-------------+----------------------------------------------------------------
       sigma |   181.1098   2.311246                      176.636    185.6968
------------------------------------------------------------------------------
```

Figure 11: Heckman Selectivity Model Results

```
. heckman empw1 income married colgrad, select(active = income married colgrad)
Heckman selection model                         Number of obs      =      4522
(regression model with sample selection)        Censored obs       =       915
                                                Uncensored obs     =      3607
                                                Wald chi2(3)       =     18.69
Log likelihood = -25430.58                      Prob > chi2        =    0.0003
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
empw1        |
      income |   .1881117   .0982646     1.91   0.056    -.0044834    .3807068
     married |  -18.94534   5.422658    -3.49   0.000    -29.57356   -8.317131
     colgrad |  -14.38192   6.493637    -2.21   0.027    -27.10922    -1.65463
       _cons |   189.6873   13.49663    14.05   0.000     163.2344    216.1403
-------------+----------------------------------------------------------------
active       |
      income |   .0055623   .0007667     7.26   0.000     .0040597     .007065
     married |  -.1233691   .0452163    -2.73   0.006    -.2119915   -.0347467
     colgrad |   .3576765   .0491578     7.28   0.000     .2613289    .4540241
       _cons |   .4822331   .0460178    10.48   0.000     .3920398    .5724264
-------------+----------------------------------------------------------------
     /athrho |  -.0089489   .1627995    -0.05   0.956      -.32803    .3101323
    /lnsigma |   5.023923   .0117885   426.17   0.000     5.000818    5.047028
-------------+----------------------------------------------------------------
         rho |  -.0089486   .1627865                     -.3167496    .3005574
       sigma |   152.0064    1.79192                      148.5346    155.5594
      lambda |  -1.360249   24.74539                     -49.86033    47.13983
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =     0.00   Prob > chi2 = 0.9586
------------------------------------------------------------------------------
```

no identifying restriction on the Heckman model, to make the results comparable to the double hurdle model results. When no identification restriction is imposed, the participation decision is identified by the functional form of the first stage - i.e. based on the curvature of the probit model.

Figure 11 shows some significant differences between the double hurdle results and the Heckman model results. In particular, the effect of education in negative in the time spent equation for the Heckman model, but positive in the time spent equation for the double hurdle model. The primary difference between the double hurdle model and the Heckman model is the inclusion of a function of the predicted probability of participation (the "inverse Mills ratio") in the second stage of the Heckman model. The estimate of the parameter on this variable is identified as `lambda` on Figure 11. It is negative, but not statistically significant in this case. The Heckman results are sensitive to the inclusion of this variable, even though it is not statistically significant.

5.2.4 Nesting and Testing

As was discussed above, some of these censored regression models are nested. The Double hurdle model can be tested against the Tobit model using a standard likelihood ratio test, as the Tobit model is nested in the double hurdle model. That is, the Tobit model can be derived from the double hurdle model be restricting the parameters of the participation probit model to be equal to the parameters of the truncated regression time spent variable.

The test is straightforward. If LL^*_{DH} is the log likelihood value from the double hurdle model (equal to -24766.73 above) and LL^*_T is the log likelihood value from the Tobit model (equal to -24775.684) above, then the likelihood ratio test is

$$LR = -2(LL^*_{DH} - LL^*_T)$$

and the test statistic has a χ^2 distribution with degrees of freedom equal to the number of parameter restrictions made to get the Tobit model. In this case the number of restrictions is 3.

It is possible to test the applicability of the double hurdle model and the sample selection model to a given set of data. The main problem with this test is that the Heckman model is not nested in the double hurdle model. Vuong (1989) proposed a modified likelihood ratio test for non-nested maximum likelihood estimators, based on a transformed value of the log likelihood function, that can be applied to the non-nested sample selection and double hurdle models considered here. The Vuong test is based on the standard likelihood ratio statistic, using a simple transformation. Let LR_1 be the likelihood statistic formed from the difference between the value of the log likelihood function for the double hurdle model evaluated at its maximum and the Heckman model evaluated at its maximum given the data

$$LR_1 = LL^*_{DH} - LL^*_H.$$

The Vuong test is based on a transformation of the log likelihood values. The transformation used in the Vuong test is

$$w_n = \left(\frac{1}{n}\right)[LR_1]^2 - [\left(\frac{1}{n}\right)LR1]^2$$

where n is the number of observations. The test statistic for the Vuong test of non-nested maximum likelihood models is

$$\sqrt{n}\frac{LR_1}{w_n}$$

and Vuong (1989) shows that this test statistic has a standard normal distribution. If the value of the test statistic is greater in absolute value than a critical value from the standard normal distribution, then the double hurdle model fits these data better than the Heckman model. If the test statistic is smaller in absolute value than a critical value from the standard normal distribution, then the test cannot discriminate between the two models given the data.

The Vuong test is difficult to implement. I do not know of an implementation of this test in STATA or any other econometric software. Only a few papers in economics have used this test. One example is Tomlin (2000).

References

Amemiya, T. (1984). Tobit models: a survey. *Journal of Econometrics*, 24(1-2):3–61.

Bockstael, N. E., Strand Jr, I. E., McConnell, K. E., and Arsanjani, F. (1990). Sample selection bias in the estimation of recreation demand functions: an application to sportfishing. *Land Economics*, 66(1):40–49.

Cragg, J. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844.

García, J. and Labeaga, J. M. (1996). Alternative Approaches to Modelling Zero Expenditure: An Application to Spanish Demand for Tobacco. *Oxford Bulletin of Economics and Statistics*, 58(3):489–506.

Greene, W. (2003). *Econometric analysis*, volume 3. Prentice Hall Upper Saddle River, NJ.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, pages 153–161.

Horrace, W. C. and Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3):321–327.

Jones, A. (2000). Health econometrics. *Handbook of health economics*, 1:265–344.

Jones, A. M. (1989). A double-hurdle model of cigarette consumption. *Journal of Applied Econometrics*, 4(1):23–39.

Jones, A. M. (1992). A Note on Computation of the Double-Hurdle Model With Dependence With An Application to Tobacco Expenditure. *Bulletin of Economic Research*, 44(1):67–74.

Maddala, G. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press.

Madden, D. (2008). Sample selection versus two-part models revisited: The case of female smoking and drinking. *Journal of Health Economics*, 27(2):300–307.

Tomlin, K. M. (2000). The effects of model specification on foreign direct investment models: an application of count data models. *Southern Economic Journal*, 67(2):460–468.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.