19

Censored Data and Truncated Distributions

William Greene

Abstract

We detail the basic theory for regression models in which dependent variables are censored or underlying distributions are truncated. The model is extended to models for counts, sample selection models, and hazard models for duration data. Entry-level theory is presented for the practitioner. We then describe a few of the recent, frontier developments in theory and practice.

19.1 Introduction

The analysis of censoring and truncation arises not from a free-standing body of theory and economic/econometric modeling, but from a subsidiary set of results that treat a practical problem of how data are gathered and analyzed. Thus, we have chosen the title "Censored Data and Truncated Distributions" for this chapter, rather than the more often used rubric "Limited Dependent Variables" (see, e.g., Maddala, 1983) specifically to underscore the relationship between the results and this practical issue. The results that we examine here arise because otherwise ordinary data are censored between generation and observation. Likewise, truncation arises because of something the analyst or the sample-generating mechanism specifically does to the data-generating process that produces the data in hand. Formally, censored data arise through a transformation of a variable of interest, say y^*, through the many to one transformation $y = T(y^*)$. (It is the data on y^* that are censored.) Perhaps the most familiar example is the latent regression interpretation of binary choice; e.g., where y^* designates a one-dimensional representation of a voter's preferences and y denotes which of two parties the voter chooses in an election, so that $T(y^*) = 1(y^* > \alpha)$; an analogous representation might describe labor force participation y as a reflection of y^*, the difference between an underlying (and unobserved) reservation wage and an offered wage. Truncation likewise is a feature of the data-gathering (as opposed to -generating) mechanism. When data are drawn from a clearly defined subset of a larger population, the probability distribution that applies to the observed data will arise as a conditional distribution within that of the larger population – hence the "truncation" will usefully be analyzed in the framework of conditional probabilities. Consider, for example, modeling the probabilities of visits to recreation sites based only on individuals who visited those sites at least once. Likewise, we consider modeling family size by analyzing only families with at least one child. In this instance, while we might have interest in the characteristics of the population at large, $f(y^*)$, what we have direct access to via familiar tools to $f(y^*|T(y^*))$, the relationship between this and $f(y^*)$ remains to be established.

This chapter will survey the basic theory and a few recent developments in models based on censoring and truncation. It has numerous precedents, notably Maddala (1983) and Dhrymes (1986), as well as numerous more recent treatments such as Long (1997) and DeMaris (2004). Terra firma in this literature is the classical linear regression with normally distributed disturbances; indeed, most of the early development focused on this exclusively. Standard analyses examined the (undesirable) properties of least squares and the (more desirable) behavior of the maximum likelihood estimator. More recent treatments have examined less fragile specifications based, for example, on semiparametric specifications. We are also interested in models that extend beyond the linear regression platform, such as models for counts, ordered choice, and so on. We begin on terra firma, with a review of the firmly established results in the standard models. As noted, we are interested in more robust model specifications and estimators. We will also examine the special features of applications to panel data. This being an applied literature at its core, we will also be interested in the situations and modeling frameworks that give rise to problems of censoring and truncation.

We need to draw two distinctions to define the analytical arena of interest in this survey:

(a) The estimation and inference problem. Interest will be on a specific class of models, defined by the conditional density of a response variable y, conditioned on a set of variables **x** and unobservable characteristics, ε. The problems analyzed here arise from censoring, truncation, or selection with respect to y, not **x**, that is, ultimately, on the unobservables, ε. Since the model is defined with respect to the conditional distribution, problems, though they may apply to observed data on **x**, will not affect our estimation problem, since the conditional model will apply to the observations that remain. Problems such as they are will apply to analysis of the marginal distribution of x, but that will generally not be of concern here.

(b) It is important to make the distinction between censoring and truncation. Censoring is a feature of the data-gathering mechanism. Truncation, whether direct or indirect, is a characteristic of the population under study, and its relation to the population that has generated the data in hand. The distinction is occasionally loose. Indeed, the second condition can be created from the first. The most pedestrian example, long a staple of the pedagogical literature, is that in which the analyst holding a data set in which some observations are censored, discards the censored observations. The distribution of the uncensored data which remain in hand is truncated with respect to the population of interest. It is useful, as well, to draw a second distinction with respect to certain types of censoring – we will treat both types in this study. In certain cases, the data gathering process produces censoring. Greene (2003) suggests the example of ticket sales to sporting events, in which the actual latent demand is censored in translation to ticket sales because some events will sell out, that is, fill the facility to capacity. In other cases, the censoring is actually a natural part of the data generating mechanism. Duration data behave this way – when one observes spells of unemployment, for example, the survey period may end while some individuals under study remain unemployed. There is a possibly unwarranted assumption that were the survey period long enough, the spell would in fact, eventually end. But this need not be the case. We will consider the implied "split population" models below.

This survey proceeds as follows: section 19.2 will present results for truncated distributions. In terms of the received literature, this part of the theory is less often used. However, the central distributional results here are extended to produce the more common censored data models. These will be developed in section 19.3. Section 19.4 will present the central features of models of sample selection. Since Heckman's (1979) seminal work, a vast literature on this subject has appeared, and continues to draw a large amount of attention. We will present little more than a simple gateway to that literature. Section 19.5 presents some of the model extensions that are made possible by panel data. Some conclusions are drawn in section 6.

19.2 Truncation

In their pioneering study of income and education, Hausman and Wise (1977) make the strong distinction (as we do) between censored data which are "piled up" at a censoring point and truncation, which occurs when a relevant subset of the population which generates the data is *unobserved*. The foundation of this class of models, and our departure point, is a classical linear regression model with uncorrelated normally distributed disturbances,

$$y_i{}^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim N[0, \sigma^2], \quad i = 1, \ldots, N. \tag{19.2.1}$$

It follows, then, that the regression of y_i on $\mathbf{x}_i$ is $\mathrm{E}[y_i{}^* \mid \mathbf{x}_i] = \mathbf{x}_i{}'\boldsymbol{\beta}$. The log likelihood for this model is

$$\ln \mathrm{L} = \sum\nolimits_{i=1}^{N} \left[-\frac{1}{2}\ln 2\pi - \ln \sigma - \frac{1}{2}((y_i{}^* - \mathbf{x}_i'\boldsymbol{\beta})/\sigma)^2\right] \tag{19.2.2}$$

In this basic foundation, all the familiar properties (finite sample and asymptotic) apply to the usual least squares estimators, **b** and s^2. (All the results that will interest us here will be asymptotic, so we will ignore degrees of freedom corrections in what follows.)

Consider, then, analysis of the subset of the population defined by

$$\begin{aligned} &y_i = y_i{}^* \quad \text{if } y_i{}^* \geq 0 \\ &y_i \text{ is unobserved if } y_i^* < 0. \end{aligned} \tag{19.2.3}$$

(The choice of zero as the truncation point is innocent if $\mathbf{x}_i$ contains a constant term, which we assume here. The choice of lower truncation is a minor complication which we will deal with in passing below.) The truncation mechanism implies that for the observed data,

$$\varepsilon_i \geq -\mathbf{x}_i'\boldsymbol{\beta} \tag{19.2.4}$$

so the normal distribution assumed above is inappropriate. The regression is also inappropriate since, using known results for truncation in the normal distribution (Greene, 2003, ch. 22),

$$\begin{aligned} \mathrm{E}[y_i|\mathbf{x}_i] &= \mathrm{E}[y_i{}^*|\mathbf{x}_i, y_i{}^* \geq 0] = \mathbf{x}_i'\boldsymbol{\beta} + \mathrm{E}[\varepsilon_i|\varepsilon_i \geq -\mathbf{x}_i'\boldsymbol{\beta}] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \sigma\frac{\phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{1 - \Phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}. \end{aligned} \tag{19.2.5}$$

where $\phi(.)$ and $\Phi(.)$ are the standard normal density and cdf, respectively. If we write this as $\mathrm{E}[y_i \mid \mathbf{x}_i] = \mathbf{x}_i{}'\boldsymbol{\beta} + \sigma\lambda_i$ where

$$\lambda_i = \frac{\phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{1 - \Phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma)} = \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}{\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}, \tag{19.2.6}$$

we can see immediately that linear regression of y_i on $\mathbf{x}_i$ will omit a variable that is surely correlated with $\mathbf{x}_i$ (See Heckman, 1979). (The variable λ_i is called the inverse Mills ratio.) The implication is that linear least squares regression of y_i on $\mathbf{x}_i$ will produce a biased and inconsistent estimator of $\boldsymbol{\beta}$. (An early thread of the literature on this model considered the possibility of *nonlinear* regression of y_i on $\mathbf{x}_i$ which would produce consistent estimators of $\boldsymbol{\beta}$ and σ. The NLS estimator here would be demonstrably inefficient (compared to MLE), very inconvenient, and not robust to any violations of the model assumptions. So, we will not consider it any further.) The magnitude and direction of the bias in the least squares estimator will be data dependent, so little can be said analytically. For reasons that will be suggested shortly, the often observed empirical regularity is that the least squares estimator in this setting is *attenuated* (biased toward zero), approximately by the relationship

$$\text{plim } \mathbf{b} \approx \boldsymbol{\beta}[1 - a\lambda(a) - \lambda(a)^2] \tag{19.2.7}$$

where a would be approximated by $-\bar{\mathbf{x}}'\boldsymbol{\beta}/\sigma$ (see Greene, 1983). The bracketed term is strictly bounded by zero and one, so we expect $\mathbf{b}$ to be attenuated as an estimator of $\boldsymbol{\beta}$. (An exact result due to Cheung and Goldberger (1984), which parallels this, states that if $\text{E}[\mathbf{x}_i \,|\, y_i]$ is linear in y_i, then plim $\mathbf{b} = \boldsymbol{\beta}\tau$ for some proportionality constant τ. The condition is unlikely to hold in practice – most models contain dummy variables, for example – but it does provide a commonly observed approximation.)

Estimation of the parameters can be accomplished by maximum likelihood. We write the log likelihood function for the untruncated case as

$$\ln \text{L} = \sum\nolimits_{i=1}^{N} \ln\left[\frac{1}{\sigma}\phi\left(\frac{y_i{}^{*} - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]. \tag{19.2.8}$$

The density for the truncated random variable must be scaled to integrate to one over the range $\varepsilon_i > -\mathbf{x}_i{}'\boldsymbol{\beta}$, so for the truncated case,

$$\ln \text{L} = \sum\nolimits_{i=1}^{N} \ln\left[\frac{(1/\sigma)\phi((y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma)}{\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)}\right]. \tag{19.2.9}$$

Maximization of this log likelihood is fairly straightforward – it is preprogrammed into several widely used commercial software packages. The analytical first and second derivatives are very cumbersome (e.g., Wooldridge (2002, p. 526)) but are made vastly simpler by Olsen's (1978) transformation, which is a useful device for many models of this sort. Let $\theta = 1/\sigma$ and $\boldsymbol{\gamma} = (1/\sigma)\boldsymbol{\beta}$. Then, the log likelihood function and its derivatives become

$$\begin{aligned}
\ln L &= \sum\nolimits_{i=1}^{N} -\frac{1}{2}\ln 2\pi + \ln\theta - \frac{1}{2}(\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})^2 - \ln\Phi(\mathbf{x}_i'\boldsymbol{\gamma}),\\
\frac{\partial \ln L}{\partial \gamma} &= \sum\nolimits_{i=1}^{N}(\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})\mathbf{x}_i - \lambda_i\mathbf{x}_i,\\
\frac{\partial \ln L}{\partial \theta} &= \sum\nolimits_{i=1}^{N}[-(\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})y_i + (1/\theta)],\\
\frac{\partial^2 \ln L}{\partial \gamma \partial \gamma'} &= \sum\nolimits_{i=1}^{N} -\delta_i\mathbf{x}_i\mathbf{x}_i', \quad 0 < \delta_i = 1 - (\mathbf{x}_i'\boldsymbol{\gamma})\lambda_i - \lambda_i^2 < 1,\\
\frac{\partial^2 \ln L}{\partial \gamma \partial \theta} &= \sum\nolimits_{i=1}^{N}\mathbf{x}_i y_i,\\
\frac{\partial^2 \ln L}{\partial \theta^2} &= \sum\nolimits_{i=1}^{N}[-y_i^2 - (1/\theta)^2].
\end{aligned} \tag{19.2.10}$$

After estimation of $\boldsymbol{\gamma}$ and θ, the original parameters are recovered from $\sigma = 1/\theta$ and $\boldsymbol{\beta} = (1/\theta)\boldsymbol{\gamma}$. The asymptotic covariance matrix for the estimators of $(\boldsymbol{\beta},\sigma)$ is derived from that for $\boldsymbol{\gamma}$ and θ via the delta method

$$Asy.Var[(\hat{\boldsymbol{\beta}}', \hat{\sigma})'] = \mathbf{G} \times Asy.Var[(\hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\theta}})'] \times \mathbf{G}', \mathbf{G} = \begin{bmatrix} \frac{1}{\theta}\mathbf{I} & \frac{-1}{\theta^2}\boldsymbol{\gamma} \\ \mathbf{0}' & \frac{-1}{\theta^2} \end{bmatrix}. \tag{19.2.11}$$

For later reference, we note in $\partial^2 \ln L/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'$ the appearance of $\delta_i = 1 - a_i\lambda_i - \lambda_i{}^2$. This quantity appears at various points in the analysis of models with censoring and truncation, and derives from

$$\text{Var}[\varepsilon_i|\mathbf{x}_i, \varepsilon_i \geq -\mathbf{x}_i'\beta] = \sigma^2\delta_i. \tag{19.2.12}$$

As (it has been shown elsewhere, for example, as in Maddala (1983)) we have that $0 < \delta_i < 1$, and it follows that the truncation has the effect of reducing the variation in the truncated population.

Since this "truncated regression model" is also a nonlinear regression, the slopes (derivatives of the conditional mean function) are not equal to the parameters. Returning to the conditional mean function, we find that $\text{E}[y_i \,|\, \mathbf{x}_i] = \mathbf{x}_i{}'\boldsymbol{\beta} + \sigma\lambda_i$. Differentiating with respect to $\boldsymbol{\beta}$ and using the results we have above, we find (not surprisingly) that

$$\frac{\partial E[y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta}\delta_i \tag{19.2.13}$$

Note that the approximate result for the least squares estimator mimics this result for the true marginal effects.

This set of results has been widely applied to models with continuous dependent variables, such as hours equations and earnings models in finance. Another common application of truncation modeling occurs in analysis of data on counts. A particular application is counts of site visits, taken on site; see Shaw (1988). Consider recreation site "*q*," and we are interested in the number of visits that individual *i* makes to that site in a given period (year, for example). Survey data taken on site that ask the respondent for numbers of visits are truncated by construction – since they are there to answer, the response must be at least one. The Poisson regression model is commonly used for this application. Under the assumptions just made, the appropriate model for on site responses would be

$$\begin{aligned} \text{Prob}[y_i = j] &= \frac{\exp(-\mu_i)\mu_i^j}{j!\text{Prob}[y_i \geq 1]}, \quad \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}) \\ &= \frac{\exp(-\mu_i)\mu_i^j}{j!\{1 - \text{Prob}[\text{y}_\text{i} = 0]\}} \\ &= \frac{\exp(-\mu_i)\mu_i^j}{j!\{1 - \exp(-\mu_i)\}}. \end{aligned} \tag{19.2.14}$$

As before, estimation is not complicated. But we do note that the force of the truncation is likely to substantially change the estimated coefficients. The marginal effects are obtained from

$$\mathrm{E}[y_i|\mathbf{x}_i] = \mu_i/[1 - \exp(-\mu_i)]. \qquad (19.2.15)$$

After some tedious algebra, we find

$$\frac{\partial E[y_i|\mathbf{x}_i]}{\partial x_i} = E[y_i \,|\, \mathbf{x}_i]\{1 + \mathrm{Prob}[y_i = 0 \,|, \mathbf{x}_i]E[y_i \,|\, \mathbf{x}_i]\}\boldsymbol{\beta} = \kappa_{\mathrm{i}}\boldsymbol{\beta}. \qquad (19.2.16)$$

It is unclear how this compares to the derivative of the original conditional mean, $\mu_i\boldsymbol{\beta}$.

Truncation of this form is straightforward to build into the model – assuming that the larger population can be characterized. We label this form of truncation "direct." It takes the form of a reduction in the range of variation of the observed variable of interest. As we've seen in the two examples described, building it into the regression model of interest, and into the likelihood for estimation purposes, is accomplished by using the laws of probability; if $y_i{}^*$ is the "untruncated" random variable and y_i is observed counterpart,

$$\mathrm{E}[y_i|\mathbf{x}_i] = \mathrm{E}[y_i{}^*|\mathbf{x}_i, y_i{}^* \text{ is in the observed range}] \qquad (19.2.17)$$

and

$$\ln f(y_i|\mathbf{x}_i) = \ln f(y_i{}^*|\mathbf{x}_i) - \ln[\mathrm{Prob}(y_i{}^* \text{ is in the observed range}|\mathbf{x}_i)] \qquad (19.2.18)$$

When these have known forms, modification of regression functions and the log likelihood function is straightforward. Note, however, that in terms of these marginal effects of interest, the attenuation result of the linear model is not general – even in the simple Poisson model, the magnitude of the marginal effects can change substantially.

19.3 Censored data and the censored regression model

In terms of received applications, censoring is much more common than truncation; applications can be found throughout and beyond all the social sciences. (There are numerous surveys, beginning with Maddala (1983) and more recently, Long (1997) and DeMaris (2004).) Here, we will establish a few of the essential elements of a model with censoring, then point toward some more elaborate specifications and methods of analysis.

As before, we depart from the classical normal, linear regression model,

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim N[0, \sigma^2], \quad i = 1, \ldots, N. \qquad (19.3.1)$$

In this setting, the observed data, y_i are obtained by a many to one transformation of $y_i{}^*$,

$$y_i = \Sigma_{j=1}^J d_j T_j(y_i{}^*) \tag{19.3.2}$$

where $T_j(y_i^*)$ partitions the range of y_i^* into J ranges and maps the values of y_i^* in the specific range into a specific value and d_j equals one if y_i^* falls in range j and zero otherwise; $d_j = 1[y_i^*$ is in range $j]$. The most familiar case [the tobit model, from Tobin (1958)[1]] has $J = 2$, where the first range is $-\infty$ to 0, which is mapped to 0 and the second range is 0 to ∞ where y_i^* is mapped to itself. (Thus, we formalize the simple case of censoring values below zero to zero.) Another familiar case with $J = 2$ is the same as the first, save that the second range is mapped to one – the probit model for binary choice. The case of sellouts at sporting events represents a case in which actual ticket sales are a censored version of true demand. Another form of the data generating mechanism which is not censoring but which produces precisely the same specification is the *corner solution model* (Wooldridge, 2002), in which, for example, zero emerges as the choice outcome in one circumstance while a continuous y_i^* emerges in another. The choice of insurance coverage that one chooses might be such a case – zero amounts to a specific choice, not a censored value of some latent negative value. In the model as stated, censoring may be incomplete, when one or more of the ranges is uncensored $(T_j(y_i^*) = y_i^*)$, or it may be complete, as in the binary choice model just mentioned.

For simplicity, we consider the simplest case first; censoring at zero a range of values. In order to form the quantities of interest in this model, we apply the laws of probability to the underlying regression model. Thus, the model that applies to the observed data in this case is

$$y_i = \max(0, y_i{}^*) \tag{19.3.3}$$

(that is, $d_1 = 1(y_i{}^* < 0), d_2 = 1(y_i{}^* > 0), T_1(y_i{}^*) = 0, T_2(y_i{}^*) = y_i{}^*$). The conditional mean function in this model is

$$\mathrm{E}[y_i|\mathbf{x}_i] = \mathrm{Prob}[y_i{}^* < 0|\mathbf{x}_i] \times 0 + \mathrm{Prob}[y_i{}^* \geq 0|\mathbf{x}_i]\mathrm{E}[y_i{}^*|\mathbf{x}_i, y_i{}^* \geq 0]. \tag{19.3.4}$$

We obtained the necessary parts in our discussion of truncation. Using the probability and conditional mean function obtained there, we have

$$\mathrm{E}[y_i|\mathbf{x}_i] = \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma) \times (\mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i). \tag{19.3.5}$$

(Note that in this partially censored data case, $\Phi(\mathbf{x}_i{}'\boldsymbol{\beta}/\sigma)$ is the probability attached to the uncensored region.) The conditional mean function for this model is noteworthy. Figure 19.1 shows the function for the standard case. Referring back to the linear specification for $y_i{}^*$, we see that $y_i{}^*$ and $\mathrm{E}[y_i^*|\mathbf{x}_i]$ can take either sign. However, $\mathbf{x}_i'\boldsymbol{\beta}$ cannot serve as the regression model for the observed y_i, which is either zero or positive. The function $\mathrm{E}[y_i|\mathbf{x}_i]$ given above is always positive, even

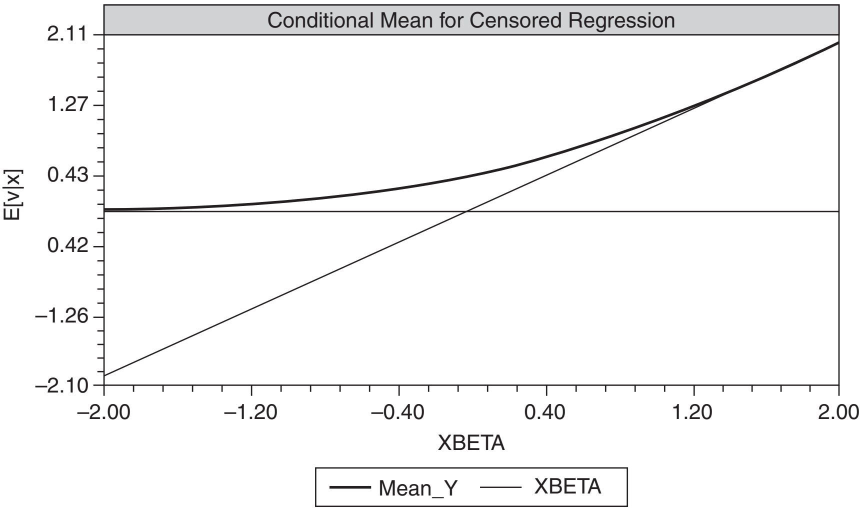

Figure 19.1 Conditional mean function for the censored regression model

when $\mathbf{x}_i'\boldsymbol{\beta}$ is negative. As in the truncation model examined earlier, the nonlinearity of the conditional mean function suggests that linear regression of y_i on $\mathbf{x}_i$ is unlikely to produce an estimate that resembles $\boldsymbol{\beta}$. Indeed, a surprising result emerges. Marginal effects are obtained by using our earlier results and, to some advantage, the Olsen transformation of the parameters;

$$\frac{\partial E[y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{\beta}. \tag{19.3.6}$$

That is, the partial effect in this model is equal to the coefficient times the probability attached to the noncensored region. Greene (1999, 2003) shows that this result extends to the "two-tailed" censoring model – that is below zero and above some positive value – and is not specific to the normal distribution but occurs regardless of the distribution of ε_i as long as it is continuous. On reflection, it should make sense. In the uncensored region, $\mathrm{E}[y_i|\mathbf{x}_i]$ responds to changes in $\mathbf{x}_i$ directly in measure $\boldsymbol{\beta}$, but in the censored region, we have a range of values for which changes in the value of $\mathbf{x}_i$ do not induce changes in y_i.

Faced with substantial censoring in the data, the researcher might be tempted simply to discard the "limit" observations and apply conventional techniques, for example, least squares, to the observations that remain. But, assembling the parts above, we see that the nonlimit observations are governed by the truncated regression model of the preceding section. This does not solve the problem; it merely moves it to a different modeling platform. Dionne *et al*. (1998) apply this principle to an "incomplete" panel of cost data on Canadian trucking firms. In their application, the specification is further complicated because the incompleteness of the data set results from "attrition," a form of sample selection that we consider in section 19.5.

19.3.1 Estimation and inference

Though linear least squares estimation of the tobit model is inappropriate, maximum likelihood estimation is no more difficult, and is preprogrammed in every contemporary econometrics computer program. The log likelihood is a nonstandard mixture of discrete and continuous parts;

$$\ln L = \sum\nolimits_{i=1}^{N} \ln\left[d_1(y_i^*)\Phi\left(\frac{-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) + d_2(y_i^*)\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]. \tag{19.3.7}$$

Amemiya (1978) showed that this nonstandard problem in maximum likelihood could be handled with standard techniques. (Again, Olsen's (1978) transformation proves extremely useful here.) Analysis of this log likelihood is, in fact, amenable to standard techniques, e.g., with inference based on the standard battery of LR, LM and Wald. The tobit model, like the truncated regression model and censored data models generally, is also amenable to the "missing data" treatment used to great advantage in the EM algorithm (Dempster, Laird, Rubin, 1977; Fair, 1977). Here, we note, if the censored observations were not censored, the appropriate estimator for $\boldsymbol{\beta}$ would be least squares. Given the actual data, we can compute the expectations of the missing data, as

$$\mathrm{E}[y_i^*|\mathbf{x}_i, y_i = 0] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma[-\phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma)/\Phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma)]. \tag{19.3.8}$$

The EM algorithm proceeds, with minor modification, by using this expression to compute the estimates for the missing observations, then using least squares based on the partially reconstructed sample. (This is the algorithm proposed in Fair (1978), though he did not treat it as an EM method.) Not surprisingly, the Bayesian MCMC estimator of the tobit model with data augmentation (see Chib, 1992) is, with trivial modification, the same computation.[2]

Construction of fit measures and predictions in this model are less straightforward than in the linear regression case. There is no counterpart to R^2 since one is not using OLS (with a constant term). Simply computing a prediction using $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is unsatisfactory since, for some of the sample, the linear predictor is being used to predict observations known to be zero, and none can legitimately be predicted to be less than zero. Likewise, the correlation between y_i and this prediction will give a misleading indication of how well the model fits the data. For prediction, the estimated conditional mean, $\hat{E}[y_i|\mathbf{x}_i] = \Phi(\mathbf{x}_i'\hat{\boldsymbol{\beta}})[\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{\sigma}\hat{\lambda}_i]$ makes more sense. Even with this predictor, however, summarizing the fit of the model to the data in an R^2-like measure is problematic because of the ambiguity of the limit observations. There is no consensus on how fit should be measured in this setting. Many contemporary researchers report the "pseudo-R^2."

$$\textit{pseudo-}R^2 = 1 - \ln L/\ln L_0 \tag{19.3.9}$$

where ln L is evaluated at the unrestricted maximum likelihood estimates and $\ln L_0$ is computed for a model which contains only a constant term. Whether this is truly useful as a fit measure is debatable as the log likelihood is not maximized to optimize "fit." It does have the virtues of lying between zero and one, and it does increase as variables are added to the model.[3]

19.3.2 Specification analysis

The corner solution interpretation of the model raises a question about the model. Under the assumptions already made, the probability that a corner solution emerges, i.e., $\text{Prob}[y_i^* < 0]$, has the same underlying specification as the regression model applied in the nonlimit case; in both cases, the index function in the density is $\mathbf{x}_i'\boldsymbol{\beta}$. One might be interested in whether the impact on the limit probability is different from that on the regression model given that it is not a limit case. To analyze this possibility, we write the log likelihood (using Olsen's transformation) for the corner solution model in the form

$$\begin{aligned}\ln L &= \sum\nolimits_{y_i=0} \ln \Phi(-\mathbf{x}_i'\boldsymbol{\gamma}) + \sum\nolimits_{y_i>0} \ln\{\theta\phi[(\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})]\} \\ &= \sum\nolimits_{y_i=0} \ln \Phi(-\mathbf{x}_i'\boldsymbol{\gamma}) + \sum\nolimits_{y_i>0} \ln \Phi(\mathbf{x}_i'\boldsymbol{\gamma}) \\ &\quad + \sum\nolimits_{y_i>0} \ln\{\theta\phi[(\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})]\} - \sum\nolimits_{y_i>0} \ln \Phi(\mathbf{x}_i'\boldsymbol{\gamma})\end{aligned} \tag{19.3.10}$$

Note that the second form is obtained simply by adding and subtracting the nonlimit probability. The first line is the log likelihood for a binary probit model for the probability of the corner solution. The second line is the log density for the observation conditioned on their having a nonlimit solution. It is also precisely the log density for the truncated regression model discussed in the preceding section. A natural specification test for whether the impact of the regressors is the same in the probability equation and in the conditional regression equation is a test of whether the coefficients in an independently estimated probit equation are the same as those in the truncated regression model for the nonlimit observations. Fin and Schmidt (1984) proposed a Lagrange multiplier test for this specification based on the results of the tobit model. A simpler computation which requires only that it be possible to compute the MLEs for all three models is the *LR* statistic

$$LR = 2\ [\ln L_{probit} + \ln L_{truncated\ regression} - \ln L_{tobit}]. \tag{19.3.11}$$

The test statistic will have a limiting chi-squared distribution with degrees of freedom equal to the number of variables in $\mathbf{x}_i$.

The preceding might be extended a step further to allow for different specifications in the probability equation in the regression. This produces a simple version of the *hurdle model* (Cragg, 1971). Estimation of this form of the model is quite simple, though again it requires estimation of the truncated regression model. Indeed, computation of the likelihood ratio statistic defined above actually requires fitting this hurdle model with the additional restriction that the regressor vectors are the same in the two equations. This is not required, of course. Two extensions of the hurdle model are also useful. Having bifurcated the model into the "participation" equation (the probability model) and the regression model, we are no longer required to specify a linear regression model for the "regression" equation. Jones (1994) analyzes a model of this sort in which the participation equation is a conventional probit model while the regression equation is a count (Poisson) model for smoking behavior. A second extension involves the underlying unobservables in the structural equations. A model which produces the hurdle log likelihood function

$$\begin{aligned} z_i^* &= \mathbf{w}_i' + u_i, \quad u_i \sim \mathrm{N}[0,1] \\ z_i &= d_1(z_i^*) = 1(z_i^* > 0) \quad \text{(a probit model)} \\ y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i | \varepsilon_i \sim \mathrm{N}[0,\sigma^2], \quad z_i = 1. \end{aligned} \tag{19.3.12}$$

The model considered so far includes the assumption that u_i and ε_i are uncorrelated (independent). If they are allowed to be correlated (bivariate normally distributed), then this form of the hurdle model produces the sample selection model that is discussed in section 19.4, below.

19.3.3 Heteroskedasticity

Since these models are typically employed with microeconomic data, two other specifications, heteroscedasticity (heterogeneity in scaling) and omitted heterogeneity (unobserved heterogeneity in the levels). In the linear regression model, conventional estimation and inference techniques are (more or less) robust to these failures of the model assumptions. Here, the estimators are not robust to any of these failures. (Nor, by and large, are they to any other failures of the model assumptions, which calls into question "robust" estimators. We turn to this issue below.)

Consider, first, a tobit model with heteroscedasticity. The modification of the model is straightforward. We define the model in terms of the log likelihood;

$$\ln L = \sum\nolimits_{i=1}^{N} \ln\left[d_1(y_i^*)\Phi\left(\frac{-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma_i}\right) + d_2(y_i^*)\frac{1}{\sigma_i}\phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma_i}\right)\right]. \tag{19.3.13}$$

Conventional ML (or Bayesian MCMC) estimation of the model parameters that ignores the heteroscedasticity is not robust to this failure of the model assumptions. Assuming that σ_i is a function of $\mathbf{x}_i$ (or variables that are correlated with $\mathbf{x}_i$), conventional estimators are not consistent, and nothing can be said about the magnitude or direction of the bias. There is no counterpart to White's robust, heteroscedasticity corrected estimator for the linear model either; the often cited Huber–White "sandwich" estimator, $\mathbf{H}^{-1}(\mathbf{G}'\mathbf{G})\mathbf{H}^{-1}$ where $\mathbf{H}$ is the negative of the inverse of the Hessian and $\mathbf{G}$ is the matrix (row by row) of first derivatives of ln L, does not solve the problem; it is merely a "robust" covariance matrix for an inconsistent estimator. (Robustness is a moot point.) Extension of the tobit model to allow for heteroskedasticity is straightforward, though it does require the analyst to specify the heteroskedasticity. For a model such as

$$\sigma_i = \sigma \times \exp(\mathbf{x}_i') \tag{19.3.14}$$

the log likelihood or posterior can simply be augmented to include the additional parameters. (We have written the scedastic function in terms of the same $\mathbf{x}_i$ that appears in the regression purely for convenience as will be clear below. Appropriately placed zeros in $\boldsymbol{\beta}$ and/or $\boldsymbol{\delta}$ can produced the desired different specifications.) With a formal specification in place, a test for heteroskedasticity in the tobit model can be based on the Wald or LR statistics by fitting the model with heteroskedasticity or by using an Lagrange multiplier statistic as shown in Greene (2003, p. 769). (Note that the ML statistic does not free the analyst from the necessity of specifying precisely what variables must appear in the scedastic function.) Partial effects in the model with heterosedasticity are (after some tedious algebra)

$$\frac{\partial E[y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \Phi(a_i)\boldsymbol{\beta} + \sigma_i\phi(a_i), \quad a_i = \mathbf{x}_i\boldsymbol{\beta}/\sigma. \tag{19.3.15}$$

For variables which appear in both the mean and variance components of the model, we see that both sign and magnitude of the partial effect can differ from those of the coefficients in $\boldsymbol{\beta}$. This suggests some care is called for in the interpretation of the estimated model components.

19.3.4 Unobserved heterogeneity

Unobserved heterogeneity in the tobit model that is uncorrelated with $\mathbf{x}_i$ is, surprisingly, benign. There is no need to prove the result analytically. If the model changes from

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim N[0, \sigma^2], \quad i = 1, \ldots, N. \tag{19.3.16}$$

to

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + c_i + \varepsilon_i, \varepsilon_i \sim N[0, \sigma^2], c_i \sim N[0, \tau^2], \quad i = 1, \ldots, N, \tag{19.3.17}$$

then the heterogeneity simply becomes part of the disturbance, which now has variance $\sigma^2 + \tau^2$. This simple result doesn't arise, for example, in the probit model because here, unlike the probit model, the sample data contain information on the scale of the latent $y_i{}^*$ whereas in the binary choice model, they do not.

19.3.5 Distribution

The specification of the tobit model, thus far, hangs crucially on the assumption of normality. How fragile the model is because of this is unknown; the only received results are (and will almost surely be) based on Monte Carlo studies of very limited generality. For better or worse, the normal distribution has provided the platform for nearly all the research on this model. One can, of course, specify an alternative distribution – we will explore how below. Of course, the resulting model is no less fragile than the censored normal model. A preferable alternative would be a less heavily parameterized, more robust estimator, such as Powell's (1981, 1984) least absolute deviations estimator. (See Melenberg and van Soest, 1996 for an application and Duncan, 1983, 1986; Newey, Powell and Walker, 1990; Lee, 1996; and Lee, 2002 for further theoretical development.)

Though estimation with an alternative model is computationally complicated, testing for the normality assumption remains worthwhile.[4] Several approaches have been devised, including a Hausman test that compares the robust LAD estimator to the tobit/normal estimator (Melenberg and van Soest, 1996), LM tests (Bera and Jarque, 1981, 1982) and conditional moment tests (Nelson, 1981; Chesher and Irish, 1987; and Pagan and Vella, 1989). The LM and conditional moment and LM tests require a set of residuals that contain information about the distribution – and nonnormality in particular. As noted above, the conventional residual, y_i – anything, has a built in problem whenever y_i equals zero. Chesher and Irish (1987) proposed the *generalized residual* for models such as this one. For the tobit (and many other models), the generalized residual can be computed as the derivative of the log-density with respect to the constant term, computed at the maximum likelihood estimators. Using the Olsen form of the log likelihood, we have

$$e_i = d_1 \frac{-\phi(-\mathbf{x}_i'\boldsymbol{\gamma})}{\Phi(-\mathbf{x}_i'\boldsymbol{\gamma})} + d_2(\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma}) \tag{19.3.18}$$

This residual has expectation and sample mean zero and accounts for the censoring.[5] A chi-squared test of the normality assumption (actually a test of whether the residual moments conform to what would be expected from a normal distribution) is computed using

$$LM = \mathbf{i}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{i} \tag{19.3.19}$$

where $\mathbf{i}$ is a column of ones and $\mathbf{M}$ is $N \times K + 3$, where each row contains

$$\mathbf{m}_i' = [e_i\mathbf{x}_i', b_i, e_i^3, e_i^4 - 3] \tag{19.3.20}$$

$$b_i = \frac{1}{2}\{d_1[(\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})^2 - 1] + d_2[\mathbf{x}_i'\boldsymbol{\gamma}\phi(-\mathbf{x}_i'\boldsymbol{\gamma})/\Phi(-\mathbf{x}_i'\boldsymbol{\gamma})]\} \tag{19.3.21}$$

(Pagan and Vella (1989) propose a variety of similar conditional moment tests for the tobit model.) Skeels and Vella (1999) have examined the behavior of this test in an extensive Monte Carlo study. The same style of specification test is extended to tests for the sample selection model examined in section 19.4 below by Vella (1992).

19.3.6 Other models with censoring

Censoring is found in many different types of applications. To suggest the range of possibilities, we note a few of them here. As in the tobit model above, the general approach to estimation and inference is generally to formulate the model in terms of the "latent" data, then deal with the censoring in the likelihood function or posterior density in the case of a Bayesian approach by using the basic laws of probability to modify the model.

The logical limit of the censoring model set out at the outset occurs when data are completely censored – none of the transformation functions $T(y_i^*)$ is one to one as it is in the uncensored region of the tobit model. Perhaps the most familiar case is the binary choice model noted at the outset,

$$\begin{aligned} y_i^* &= \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim N[0,\sigma^2], \quad i = 1, \ldots, N \\ y_i &= \sum\nolimits_{j=1}^{2} d_j T_j(y_i^*), d_1 = 1(y_i^* < 0), T_1(y_i^*) = 0, d_2 = 1 - d_1, T_2(y_i^*) = 1. \end{aligned} \tag{19.3.22}$$

A less extreme case is the *ordered probit model*, which maps ranges with unknown boundary points to the integers 0,1,...,*J*. The second equation in the structure above is

$$\text{Prob}[\mu_{j-1} < y_i^* \leq \mu_j] = \Phi[\mu_j - \mathbf{x}_i'\boldsymbol{\beta}] - \Phi[\mu_{j-1} - \mathbf{x}_i'\boldsymbol{\beta}], \quad \mu_j > \mu_{j-1}, \quad j = 0, \ldots, J, \tag{19.3.23}$$

with normalizations $\mu_{-1} = -\infty, \mu_0 = 0, \mu_J = +\infty$. Familiar applications include opinion measures, where the strength of opinions or preferences are expressed on a scale (usually zero to four). Another natural application (which remains to be explored at length) is self reported health status, such as the variable contained in Winkelmann (2004). In the ordered probit model, information about the scale of the dependent variable is lost – in the case of latent preference, it would have no meaning in any event. When data are censored to mask within range variation, the observed response may be *interval censored*. In Bhat (1994) a latent income variable is reported only in ranges. The structural model is identical to that of the ordered probit, except that the threshold parameters are known. This obviates the normalizations, and reveals the scaling information, so that an estimate of σ can be computed with the estimate of $\boldsymbol{\beta}$. As a consequence, the density for y_i is redefined to be

$$\text{Prob}[y_i = j] = \Phi\left(\frac{a_j - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{a_{j-1} - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right). \tag{19.3.24}$$

Each of these represents a method of modeling censoring in the context of the classical normal linear regression model. Two other leading cases of censored data are in models of counts and in duration data. In the count data model, we have the generic structure

$$\text{Prob}[y_i = j|\mathbf{x}_i] = f(j; \boldsymbol{\beta})$$

(The parameter vector may include other ancilliary parameters, such as the overdispersion parameter in the negative binomial model.) The most familiar case is the Poisson (loglinear) regression model,

$$\text{Prob}[y_i = j|\mathbf{x}_i] = \frac{\exp(-\mu_i)\mu_i^j}{j!}, \quad \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), \quad j = 0, 1, \ldots \tag{19.3.25}$$

Data may be censored at either end, though the leading case is *top coding*, in which the censoring takes the form of piling all values above a limit value into that value (see Terza, 1985). An example is Fair's (1978) study of extramarital affairs in which the reported count was censored at 12.[6] The censored Poisson model follows naturally from the definitions. For example, for censoring at upper limit C, we would have the model

$$\begin{aligned}\text{Prob}[y_i = j|\mathbf{x}_i] &= \frac{\exp(-\mu_i)\mu_i^j}{j!}, \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), j = 0, 1, \ldots, C-1,\\ \text{Prob}[y_i = C|\mathbf{x}_i] &= 1 - \sum\nolimits_{j=0}^{C-1} \frac{\exp(-\mu_i)\mu_i^j}{j!}.\end{aligned} \tag{19.3.26}$$

The conditional mean is altered in an expected fashion (see Greene, 2000);

$$\begin{aligned}E[y_i|\mathbf{x}_i] &= \mu_j - \sum\nolimits_{j=C}^{\infty} (j - C)\text{Prob}[y_i = j|\mathbf{x}_i]\\ &= \text{C} - \sum\nolimits_{j=0}^{C-1} (C - j)\text{Prob}[y_i = j|\mathbf{x}_i].\end{aligned} \tag{19.3.27}$$

The marginal effects also change;

$$\frac{E[y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \left[\sum\nolimits_{j=0}^{C-1} (j - C)(j - \mu_i)\text{Prob}[y_i = j|\mathbf{x}_i]\right]\boldsymbol{\beta}. \tag{19.3.28}$$

These can be substantially smaller than their uncensored counterparts, $\mu_i\boldsymbol{\beta}$.

The foregoing illustrate the effect of censoring on regression models, that is in models in which the conditional mean function and its derivatives is the central focus. A vast variety of other models, in which some variation of the regressand is masked by censoring, are all handled similarly and similar results emerge. Censoring, which masks variation brings predictable changes in the location of the mean, generally reduces marginal effects because in the censored region changes in the stimuli (independent variables) are not associated with changes in the response.

Another leading class of models in which censoring is an important feature is models of duration. In this setting, we model the length of time, t, from a "baseline" until a "transition" takes place (see Kiefer, 1985 for a survey). Familiar applications include the time until business failure, length of a spell of unemployment or the lengths of the intervals between children at the household level, or between wars at a global level. In all cases, what is typically of interest is not the length of time, but the *hazard rate*, which is roughly the probability that the transition takes place in interval t to $t + \Delta t$ given that it has not taken place up to time t. We consider a few of the formalities of hazard models to illustrate an extension of our class of censored data models.

For the random variable t, the time until an event occurs, $t \geq 0$, the density, cdf and *survival function* are denoted $f(t)$, $F(t)$ and $S(t) = 1 - F(t)$. The probability of an event occurring at or before time t is $F(t)$. The conditional probability that an event occurs in the interval t to $t + \Delta$ given that it has not occurred by time t is

$$\begin{aligned}h(t) &= \text{Prob(event occurs in time } t \text{ to } t + \Delta| \text{ event occurs after time } t)\\ &= \frac{F(t + \Delta) - F(t)}{1 - F(t)}.\end{aligned} \tag{19.3.29}$$

As $\Delta \to 0$, the function $[F(t+\Delta) - F(t)]/[\Delta(1 - F(t))]$ converges to *f(t)/S(t)*, which is called the *hazard function*, often denoted $\lambda(t)$. (This is not to be confused with λ_i as used in the preceding discussions, though there is clearly a relationship for the normal distribution.) Note that $\Delta\lambda(t)$ equals the probability sought, Prob $[t \leq T \leq t + \Delta | T \geq t]$. The hazard function is a descriptor of the probability distribution, as are the pdf and cdf. Indeed, we see the simple relationship $\lambda(t)S(t) = f(t)$. There are many different specifications that can be used to model the hazard for the duration variable *T*. The simplest is a function with "no memory;" that is, one with a constant hazard rate. For this model, we would have $\lambda(t) = \lambda$, a constant. It follows from the definition that the hazard follows the simple differential equation $\lambda(t) = -\mathrm{d}\ln S(t)/\mathrm{d}t$. The solution to $-\mathrm{d}\ln S(t)/\mathrm{d}t = \lambda$ is $S(t) = K\exp(-\lambda t)$, where *K* is the constant of integration. The boundary condition $S(0) = 1$ implies $K = 1$, which leaves $S(t) = \exp(-\lambda t)$. This is the exponential density,

$$f(t) = \lambda\exp(-\lambda t), \quad \lambda > 0, \quad t \geq 0. \tag{19.3.30}$$

This is the most basic hazard function model. Some other candidates are

$$\begin{aligned} \text{Weibull:} \quad & \lambda(t) = \lambda p(\lambda t)^{p-1}, p = 1 \text{ implies exponential,} \\ \text{log logistic:} \quad & \lambda(t) = \lambda p(\lambda t)^{p-1}/[1 + (\lambda t)^p], \\ \text{log normal:} \quad & \lambda(t) = \phi[-p\ln(\lambda t)]/\Phi[-p\ln(\lambda t)] \end{aligned} \tag{19.3.31}$$

Figure 19.2 shows the behavior of these hazard functions for a standard data set on strike duration (see Kennan, 1985).

Note that the hazard for the Weibull model declines monotonically – this is known as *negative duration dependence*. Over some ranges, the lognormal and log logistic have *positive duration dependence*, while the exponential model has no duration dependence.

The counterpart to the familiar regression models in this context would be the *accelerated failure time models*, in which the hazard function is modeled as a function of covariates. A familiar example is the loglinear model. For the Weibull model, this would be

$$\lambda(t|\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})p[\exp(\mathbf{x}'\boldsymbol{\beta})t]^{p-1} \tag{19.3.32}$$

Most data sets have incomplete observations. The observation consists of the time of the measurement and the indication that the transition (business failure, death, warranty exercise, next insurrection, next child) has not yet occurred. Such observations are censored at time *t*, the same as the censoring phenomenon observed earlier.

We now construct the log likelihood for a sample of duration data. For an uncensored observation, the contribution to the likelihood is the density. For a censored observation, it is the survival function. (Note that this is precisely the

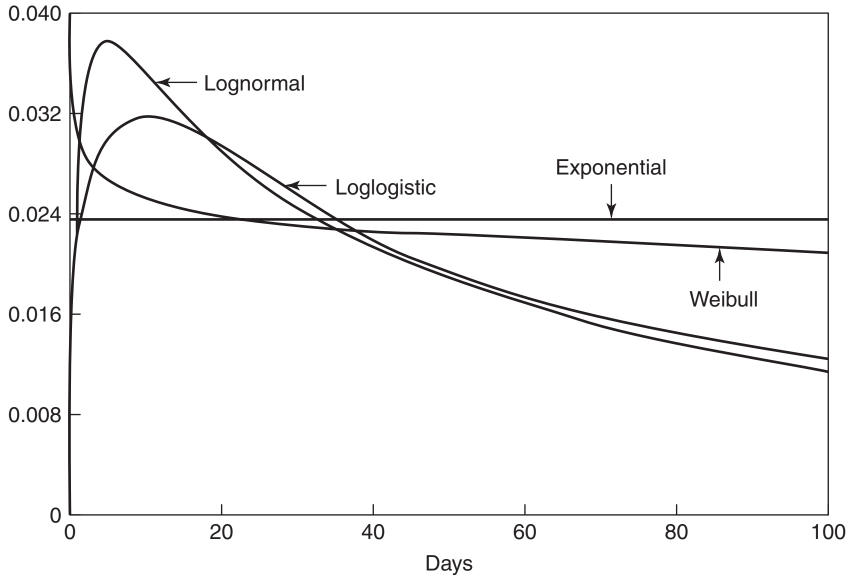

Figure 19.2 Hazard functions

format the likelihood takes for the regression model with right tail censoring that was discussed above.) Let d be a noncensoring indicator; $d = 0$ for a censored observation and $d = 1$ for an uncensored observation. We will also use the result noted earlier, $f(t|\mathbf{x}) = \lambda(t|\mathbf{x})S(t|\mathbf{x})$. Then, the log likelihood for a sample that contains both censored and uncensored observations is

$$\ln L = \sum\nolimits_{i=1}^{N} d_i \ln[\lambda(t_i|\mathbf{x}_i)] + \ln S(t_i|\mathbf{x}_i) \tag{19.3.33}$$

For the parametric models shown earlier, this is now a standard problem for maximum likelihood estimation and inference. To close the loop here, so to speak, we note that the preceding shows how different distribution could be used for a censored regression model. We used the normal distribution in our earlier discussion. This derivation shows how the exponential, Weibull and other models could be used. Moreover, to use this template to accommodate our standard model with left censoring at zero, we can simply use $-\ln t$ as the dependent variable (see Greene, 2000 for discussion).

19.4 Incidental truncation and sample selection

The results of the preceding sections have been extended to a "two-part" model that extends the hurdle model. Consider an observation mechanism that departs from the familiar regression model,

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim N[0, \sigma^2], \quad i = 1, \ldots, N. \tag{19.4.1}$$

and adds a "sample selection mechanism" to a binary probit model;

$$\begin{aligned} d_i^* &= \mathbf{z}_i'\boldsymbol{\alpha} + u_i \\ T(d_i^*) &= 1(d_i^* > 0) \\ T(y_i^*|d_i^*) &= y_i^* \quad \text{if } d_i^* > 0, y_i^* \text{ is unobserved otherwise.} \end{aligned} \tag{19.4.2}$$

This is a modification of the truncated regression model discussed in section 19.2, where $d_i^* = y_i^*$. Here, d_i^* is another variable in this two equation model. If u_i and ε_i are correlated, then the observed values of y_i^* are unusual compared to the full population. Hence we use the term "incidental truncation" for this specification. Applications of this sort of model abound in the literature, beginning with Heckman's pioneering work (e.g., 1979) on labor supply.[7] Some examples, in addition to this one, include analysis of returns in long time series of financial data ("survivorship" effects), analysis of program participation where observation at the end of the program is affected by attrition of the participants, count models of recreation site use, health care usage, and a vast catalog of other settings.

In all cases, it is the relationship between the unobservables in the models that exerts the impact on the estimation and inference procedures. Consider, in the model above, the standard case in which $(\varepsilon_i u_i)$ are bivariate normally distributed with correlation ρ. In the observed data, we will have

$$\begin{aligned} \mathrm{E}[y_i|\mathbf{x}_i, y_i \text{ is observed}] &= \mathrm{E}[y_i^*|\mathbf{x}_i, d_i^* > 0] = \mathrm{E}[y_i^*|\mathbf{x}_i, d_i = 1] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \mathrm{E}[\varepsilon_i|d_i = 1] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \mathrm{E}[\varepsilon_i|u_i > -\mathbf{z}_i'\boldsymbol{\alpha}]. \end{aligned} \tag{19.4.3}$$

From results for the bivariate normal distribution, this is

$$\begin{aligned} \mathrm{E}[y_i|\mathbf{x}_i, y_i \text{ is observed}] &= \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_\varepsilon\phi(-\mathbf{z}_i'\boldsymbol{\alpha})/[1 - \Phi(-\mathbf{z}_i'\boldsymbol{\alpha})] \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \kappa\lambda_i \end{aligned} \tag{19.4.4}$$

where $\lambda_i = \phi(\mathbf{z}_i'\boldsymbol{\alpha})/\Phi(\mathbf{z}_i'\boldsymbol{\alpha})$ is the inverse Mills ratio discussed earlier. Two conclusions follow from this derivation, before we consider estimation. First, by dint of the excluded variable, λ_i, it is clear that linear regression of y_i on $\mathbf{x}_i$ in the observed data will produce an inconsistent estimator of $\boldsymbol{\beta}$ if κ is not equal to zero (which we assumed) and if λ_i is correlated with $\mathbf{x}_i$, which is almost surely going to be the case, particularly if $\mathbf{z}_i$ and $\mathbf{x}_i$ have variables in common. To underscore the point, consider a modification of the model, known as the *treatment effects* model, where

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \delta d_i + \varepsilon_i, \varepsilon_i \sim N[0, \sigma^2], \quad i = 1, \ldots, N. \\ d_i^* &= \mathbf{z}_i'\boldsymbol{\alpha} + u_i, d_i = 1[d_i^* > 0] \end{aligned} \tag{19.4.5}$$

and $(y_i^*, \mathbf{x}_i)$ is observed for all cases. In an intriguing recent example [Dale and Krueger (1999)], consider the case in which $y_i{}^*$ is an income variable and d_i is an indicator of whether the individual attended an elite college. Clearly in this model, the "regressor" d_i is correlated with the disturbance ε_i, producing "simultaneous equations bias." With a bit of manipulation, we can recast this model as another example of our selection model – at least it shares the fundamental features. Returning to the original model, a second question arises; it is unclear whether $\boldsymbol{\beta}$ is even the quantity of interest. Using the device we used before, assume that $\mathbf{z}_i = \mathbf{x}_i$ (with appropriate zeros in $\boldsymbol{\beta}$ or $\boldsymbol{\alpha}$ as needed). Then, again using our earlier results, we find in this basic model,

$$\frac{\partial E[y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \beta - (a_i\lambda_i + \lambda_i^2)\alpha. \tag{19.4.6}$$

We conclude that, even after dealing appropriately with the estimation issues, some care is needed in interpreting the results.

There are two methods of estimating this model, *two-step* (not two-stage) least squares and maximum likelihood. The two-step method was proposed by Heckman (1979) (see also Greene, 1981, 2003). The logic of Heckman's method is strikingly simple. If λ_i were observed, ordinary least squares would provide a consistent (though not necessarily efficient) estimator of $(\boldsymbol{\beta}, \kappa)$. Since the parameters in λ_i can be consistently estimated by applying a binary probit model to the model for d_i, and $\mathbf{z}_i$ is observed, a "pointwise" consistent estimator of λ_i is obtained by using $\hat{\alpha}$ from the probit model. This is the first step of the two-step estimator. The second step is least squares regression of y_i on $\mathbf{x}_i$ and $\hat{\lambda}_i$. The conventionally estimated asymptotic covariance matrix for this least squares estimator is inappropriate for two reasons; first, the implied disturbance in the regression is heteroscedastic and, second, it does not account for the variation in the estimated parameter vector used to compute $\hat{\lambda}_i$ (see Murphy and Topel, 1985). Expressions for computing the appropriate covariance matrix appear in Heckman (1979) and Greene (1980, 2003). The treatment effects model is handled similarly. In this case, the counterpart to "λ_i" is the generalized residual from the probit model,

$$\hat{\lambda}_i = d_i \frac{\phi(\mathbf{z}_i'\boldsymbol{\alpha})}{\Phi(\mathbf{z}_i'\boldsymbol{\alpha})} + (1 - d_i) \frac{-\phi(-\mathbf{z}_i'\boldsymbol{\alpha})}{\Phi(-\mathbf{z}_i'\boldsymbol{\alpha})} \tag{19.4.7}$$

After estimation, a "test" for "selectivity" is based on the estimate of κ; a simple "t-test" of the significance of the coefficient on $\hat{\lambda}_i$ is equivalent to a test that ρ equals zero.

The second estimator is maximum likelihood. The log likelihood function for this model is constructed from the joint density for d_i and y_i for those observations for which y_i is observed. As usual, the Olsen transformation simplifies the notation;

$$\begin{aligned}\ln L = {} & \sum\nolimits_{d_i=1} \ln[\theta\phi(\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})] + \ln\Phi\left[\frac{\rho(\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma}) + \mathbf{z}_i'\boldsymbol{\alpha}}{\sqrt{1-\rho^2}}\right] \\ & + \sum\nolimits_{d_i=0} \ln\Phi[-\mathbf{z}_i'\boldsymbol{\alpha}].\end{aligned} \tag{19.4.8}$$

(There is yet another simplification possible by transforming ρ.) This is a complicated (because of ρ) but otherwise standard problem in maximum likelihood estimation. In addition to its theoretically greater efficiency, the MLE has another advantage over the Heckman two step estimator. The variable λ_i is a nonlinear function of z_i that is essentially linear in $\mathbf{z}_i'\boldsymbol{\alpha}$ over much of its range. This implies that if there is not much difference between $\mathbf{x}_i$ and $\mathbf{z}_i$ – in many applications they are the same – then there is the potential for serious multicollinearity in the augmented regression. Most researchers seek to accommodate this problem of "weak" identification by ensuring that there is at least one variable in $\mathbf{z}_i$ that is not in $\mathbf{x}_i$ and that has substantial variation.

We note an aspect of estimation here for the interested practitioner. The appearance of Heckman's "lambda" in the estimated selection equation has produced a temptation to augment other kinds of selection models likewise and thereby "take care of the selection problem." This form of the model is specific to the linear regression case. Notice, for example, that there is no inverse Mills ratio in the log likelihood for the model. Thus, for example, it is not appropriate to correct a Poisson regression model for selectivity by just adding an inverse Mills ratio to the index function in the model. See Terza (1998) and Greene (1995, 1997, 2000) for applications of sample selection corrections to the Poisson regression model. In these and other models, it is necessary to reconstruct the log likelihood function, somewhat similar to the form as it appears above.

The literature on selection models and treatment effects is vast and varied. This is an active and ongoing area of research in econometrics (see, for example, Angrist, 2001). The above discussion suggests only the most basic form of the model.

19.5 Panel data

Microeconomic data increasingly come in the form of extensive panel data sets, such as the National Longitudinal Surveys of Labor Market Experience (NLS), the German SocioEconomics Panel or the British Household Panel Survey (BHPS) which, among many others, contain rich multiple wave surveys of individual health and labor market behavior. Interesting response variables in these data sets, such as income, fertility and labor market experience, often come in the form of discrete, truncated, limited and otherwise range restricted variables to which the methods described here apply. We consider a few of the basic issues in analysis of panel data in the censoring and truncation models considered here. The issues are relatively common across modeling platforms, so to present the essential results, we will focus on the tobit model, and add some details about panel data and sample selection at the end of the section.

Thinking about incorporating individual heterogeneity in models such as the tobit model usually focuses on the two standard approaches, fixed and random effects. We modify the basic model to include the heterogeneity as

$$\begin{aligned} y_{it}^* &= \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_i \sim N[0, \sigma^2], \quad i = 1, \ldots, N, \\ y_{it} &= \max(0, y_{it}^*). \end{aligned} \tag{19.5.1}$$

Conventional wisdom about the model is guided by the linear model with individual heterogeneity. As we will see, some of that wisdom is useful, while some is not.

19.5.1 Estimating fixed effects models

The fixed effects model in the preceding specification allows correlation between α_i and $\mathbf{x}_{it}$. It is useful to digress briefly to explore the practical implication of the assumption, $\text{Cov}[\mathbf{x}_{it}, \alpha_i] \neq \mathbf{0}$. Suppose individual i is observed T_i times (where T_i may vary across individuals). Let $\mathbf{X}_i$ denote the $T_i \times K$ matrix of observations on the regressors and let $\mathbf{j}\alpha_i$ denote the $T_i \times 1$ column of observations (repeated) on the individual heterogeneity, α_i; $\mathbf{j}$ is a column of ones. Consider, then, the "estimator" of the covariance,

$$\begin{aligned} Est.Cov[\alpha_i, \mathbf{x}_{it}] &= \frac{\sum_{i=1}^N \sum_{t=1}^{T_i} \mathbf{x}_{it}\alpha_i}{\sum_{i=1}^N T_i} = \frac{\sum_{i=1}^N \alpha_i T_i (1/T_i) \sum_{t=1}^{T_i} \mathbf{x}_{it}}{\sum_{i=1}^N T_i} \\ &= \frac{\sum_{i=1}^N T_i \alpha_i \overline{\mathbf{x}}_i}{\sum_{i=1}^N T_i} = \sum_{i=1}^N w_i \alpha_i \overline{\mathbf{x}}_i \\ &\rightarrow Cov[\alpha_i, \overline{\mathbf{x}}_i]. \end{aligned} \tag{19.5.2}$$

(The weights in the sum, $w_i, 0 < w_i < 1, \sum_{i=1}^N w_i = 1$, accommodate an unbalanced panel. If T_i is the same for all i, then $w_i = 1/N$.) This suggests that the relationship between the invariant "effect" and the exogenous variables will be reflected in covariation between the effect and the group means. (We will employ this idea below with the "Mundlak (1978) correction" for the random effects model.) For reasons that will soon become clear, typically no distribution is assumed in the fixed effects model. The random effects model, in contrast, begins with an assumption that the effect, α_i and the data, $\mathbf{x}_{it}$ are uncorrelated. Also, it is typical to assume that the random effect is normally distributed with zero mean and constant variance. We will explore this issue in more detail below.

The fixed effects model is estimated by including in the model a set of N group dummy variables, $\mathbf{d}_\text{i} =$ the dummy variable indicating membership in group i. With this specification, the model becomes

$$\begin{aligned} y_{it}^* &= \Sigma_{i=1}^N d_{it}\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_i \sim N[0, \sigma^2], \quad i = 1, \ldots, N, \\ y_{it} &= \max(0, y_{it}^*). \end{aligned} \tag{19.5.3}$$

The log likelihood function for the tobit model with fixed effects is

$$\ln L = \sum\nolimits_{i=1}^{N} \sum\nolimits_{t=1}^{T_i} (1 - c_{it}) \ln \Phi(-\eta_i - \mathbf{x}_{it}'\boldsymbol{\gamma}) + c_{it} \ln[\theta\phi y_{it} - \eta_i - \mathbf{x}_{it}'\boldsymbol{\gamma})] \qquad (19.5.4)$$

where $c_{it}=1$ if $y_{it}>0$ and 0 otherwise and, as usual, we employ the Olsen transformation so that $\eta_i=\alpha_i/\sigma$. In practical terms, there are two problems with application of the fixed effects model in limited dependent variable models such as the tobit or truncated regression model. First, the number of individuals, N, is typically large, which implies that it is necessary to estimate a potentially very large number of parameters. In the linear model, this difficulty is handled by transforming observations to deviations from group means or by using first differences. In the Poisson model, there is a transformed likelihood that can be constructed that is free of the dummy variable coefficients. None of these approaches work here; since y_{it} is observed only after transformation, deviations of y_{it} from group means produces deviations in the transformations, not deviations in y_{it}^*. There is no transformation of the log likelihood that removes the dummy variable coefficients. In order to fit this model by maximum likelihood, it is necessary to estimate all $N+K+1$ parameters simultaneously.[8] This can, in fact be done – our example below includes estimates of 7,293 dummy variable coefficients – using the method described, e.g., in Greene (2005). Before turning to the theoretical shortcoming of the fixed effects estimator, we note one additional complication. It is easy to show that for any individual for which all observations are censored, the parameter η_i is inestimable. (For such an individual, the derivative of the log likelihood with respect to η_i is $\Sigma_t - \phi(-\eta_i - \mathbf{x}_{it}'\boldsymbol{\gamma})/\Phi(-\eta_i - \mathbf{x}_{it}'\boldsymbol{\gamma})$, which is always negative and hence cannot be equated to zero.) Note, finally, another shortcoming of the fixed effects model is that like the linear regression model, it is not estimable if $\mathbf{x}_{it}$ contains any time invariant regressors.

The practical issue has discouraged use of the fixed effects estimator.[9] However, the more vexing problem is the *incidental parameters problem* of the maximum likelihood estimator in the presence of fixed effects (Neyman and Scott, 1948). Note that in the log likelihood function above, the number of parameters increases with N – each individual specific constant term is estimated with T_i observations. Since T_i is fixed, one can expect a problem with consistency of the estimator. This is generally expected to introduce a "small sample bias" into the parameter estimator. The thinking on this issue has long been guided by some well established results on binary choice models. It has been shown analytically (Andersen, 1970; Abrevaya, 1997) that in the binary logit model, the MLE of $\boldsymbol{\beta}$ in the presence of fixed effects, is biased by a factor of two (plim $\hat{\beta}_{MLE} = 2\boldsymbol{\beta}$).

A long history of Monte Carlo work (for example, Greene, 2004) has suggested that the essentially the same result applies to the binary probit model – it has not been shown analytically. Analytic results for T greater than 2 have not been shown for any model, but, again, the Monte Carlo studies suggest, as intuition might also, that the bias in binary choice estimators diminishes as T increases, but relatively slowly – it remains substantial for T as large as 10. Until recently, analysis of this sort was limited to binary choice models, but it was, by and large, taken as a given (see, for example, Wooldridge, 2000) that similar results apply to other models. In fact, this appears not to be the case. Table 19.1 shows the results of an analysis of the tobit model under the specification,

$$\begin{aligned} y_{it}^* &= \alpha_i + x_{it}\beta + z_{it}\delta + \varepsilon_{it} \\ y_{it} &= \max(0, y_{it}^*) \end{aligned} \tag{19.5.5}$$

The two regressors are a continuous variable x_{it} and a dummy variable z_{it}. The R^2 in the latent regression is about .77 and about 40 percent of the observations are censored. The values in the table are the percentage biases against the known true values of the items shown; the true values of β, δ and σ were all one. The results are strongly at odds with the conventional wisdom. First, there is essentially no bias in the estimated slope parameters (far less then one percent), but there is some bias in the estimated marginal effects (at the data means), but not very much in view of what is known about the binary choice models. The results do suggest that estimated standard errors are biased downward somewhat. As noted, these results are not consistent with those for the binary choice models. They are consistent with the original Neyman and Scott results, who found that the bias in the MLE of σ^2 in the linear model was downward, by a factor of $(T-1)/T$.[10] Surprisingly, and in conflict with our intuition, the results above seem not to extend to the truncated regression. The same study produces the results in Table 19.2. Note, in this case, everything is biased toward zero, rather than away.

The end result would seem to be that estimation of fixed effects models with censoring and truncation presents no practical obstacle. The incidental parameters problem is to be reckoned with, but if the Monte Carlo results given here have any

Table 19.1 Tobit model: effect of group size on estimates

Estimate	*T* = *2*	*T* = *3*	*T* = *5*	*T* = *8*	*T* = *12*	*T* = *15*	*T* = *20*
β	0.67	0.53	0.50	0.29	0.098	0.082	0.047
δ	0.33	0.90	0.57	0.54	0.32	0.16	0.14
σ	− 36.14	− 23.54	− 13.78	− 8.40	− 5.54	− 4.43	− 3.30
ME$_x$	15.83	8.85	3.65	1.30	0.44	0.22	0.081
ME$_z$	19.67	11.85	5.08	2.16	0.89	0.46	0.27
S.E.(β)	− 32.92	− 19.00	− 11.30	− 8.36	− 6.21	− 4.98	0.63
S.E.(δ)	− 32.87	− 22.75	− 12.66	− 7.39	− 5.56	− 6.19	0.25

Table 19.2 Truncated regression model: behavior of the MLE/FE

Estimate	*T* = *2*	*T* = *3*	*T* = *5*	*T* = *8*	*T* = *12*	*T* = *15*	*T* = *20*
β	− 17.13	− 11.97	− 7.64	− 4.92	− 3.41	− 2.79	− 2.11
δ	− 22.81	− 17.08	− 11.21	− 7.51	− 5.16	− 4.14	− 3.27
σ	− 35.36	− 23.42	− 14.28	− 9.12	− 6.21	− 4.94	− 3.75
ME$_x$	− 7.52	− 4.85	− 2.87	− 1.72	− 1.14	− 0.94	− 0.67
ME$_z$	− 11.64	− 8.65	− 5.49	− 3.64	− 2.41	− 1.90	− 1.53

generality, then the IP problem in this setting is far less severe than in the binary choice case.

19.5.2 Estimating random effects models

In the random effects model, the heterogeneity is assumed to be uncorrelated with the regressors. This suggests an altogether different approach to estimation and inference. The conditional log likelihood in the presence of the random effect is

$$\ln L^C = \sum_{i=1}^{N} \ln \prod_{t=1}^{T_i} [\Phi(-\tau w_i - \mathbf{x}'_{it}\boldsymbol{\gamma})]^{1-c_{it}} [\theta\phi(\theta y_{it} - \tau w_i - \mathbf{x}'_{it}\boldsymbol{\gamma})]^{c_{it}} \tag{19.5.6}$$

where $\tau = \sigma_\alpha/\sigma$ and $w_i \sim \mathrm{N}[0,1]$. Estimation of the model entails estimation of the unknown parameters γ, θ and τ. Since the conditional log likelihood function includes the unobserved random effect, it cannot serve as the basis for estimation. The unconditional log likelihood function is

$$\ln L = \sum_{i=1}^{N} \ln \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} [\Phi(-\tau w_i - \mathbf{x}'_{it}\boldsymbol{\gamma})]^{1-c_{it}} [\theta\phi(\theta y_{it} - \tau w_i - \mathbf{x}'_{it}\boldsymbol{\gamma})]^{c_{it}} \phi(w_i) dw_i \tag{19.5.7}$$

Estimation of the random effects model can be done by Gauss–Hermite quadrature as designed by Butler and Moffitt (1982) or by Monte Carlo integration (Greene, 2000).

The random effects form of the model is much more manageable than the fixed effects form. Here, however, one trades the difficulty of the incidental parameters problem and the practical complication of time invariant regressors in the fixed effects case for the possibly unpalatable assumption that the effects are uncorrelated with the regressors in the random effects model. A path between the horns of this dilemma (see Wooldridge, 2005, for example) is suggested by the Mundlak idea outlined at the beginning of this section. Suppose in either the fixed or random effects specification, we project the unknown effect on the means of the time varying variables; then,

$$\begin{aligned} y^*_{it} &= \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \varepsilon_{it} \sim N[0, {}^2], \\ \alpha_i &= \bar{\mathbf{x}}'_i\pi + w_i, w_i \sim N[0, 1], \\ y_{it} &= \max(0, y^*_{it}). \end{aligned} \tag{19.5.8}$$

This produces a random effects model which can be estimated by either method mentioned above and in which, one hopes, the effect of correlation between the unobserved effects and the regressors, is picked up by the group means.

19.5.3 An application of fixed and random effects estimators

To illustrate a few of the models discussed above, we will fit and analyze the data used in Winkelmann (2004). This is an unbalanced panel survey of health care utilization of 27,326 German individuals. The sample contains 7,293 individuals observed from one to seven times in the panel. Counts for the group sizes are 1,525, 1,079, 825, 926, 1,051, 1,000 and 887 for $T_i = 1, \ldots, 7$, respectively. We have fit a model for household income as a function of age, education, marital status and whether there are children in the household. Descriptive statistics for the data are given in Table 19.3. The raw income data in the survey range from zero (a handful of observations) to about 2. We have "top coded" ("for privacy") the income variable at 0.35, thus censoring 12,369 observations, or 45.4 percent of the sample. Assuming that a linear regression model applies to the raw data, the tobit and truncated regression models should likewise be appropriate for the censored data.

Table 19.4 presents least squares and maximum likelihood estimates for several approaches.[11] The OLS estimates, compared to their ML counterparts, clearly

Table 19.3 Panel data on income and sociodemographic variables. N = 27,326

Variable	*Mean*	*Standard deviation*	*Minimum*	*Maximum*
Income	.288208986	.0754686019	0	0.35
Age	43.5256898	11.3302475	25	64
Education	11.3206310	2.32488546	7	18
Married	.758618166	.427929136	0	1
Children in household	.402730001	.490456267	0	1

Table 19.4 Estimates of model parameters

Estimator	*Constant*	*Age*	*Education*	*Married*	*Children*	*σ*
OLS Nonlimit Data	0.1772	–0.0006	0.004497	0.05341	–0.0018	0.0633
Logl = 20012.91	(0.0044)	(0.00005)	(0.00029)	(0.00126)	(0.0012)	
MLE Truncation	0.1699	–0.0008	0.00641	0.07070	–0.0011	0.0756
Logl = 21110.15	(0.0064)	(0.00007)	(0.0004)	(0.00177)	(0.0018)	
OLS All Data	0.1931	–0.0007	0.0073	0.0602	–0.01025	0.0698
Logl = 33965.22	(0.0031)	(0.00004)	(0.0002)	(0.0011)	(0.0010)	
MLE Tobit	0.1169	–0.00071	0.01599	0.09105	–0.0176	0.1117
Logl = 2745.94	(0.0058)	(0.00007)	(0.00037)	(0.0019)	(0.0018)	
Tobit Fixed Effects		0.02406	0.03043	0.1553	–0.0657	0.0832
Logl = 17957.33		(0.00027)	(0.00230)	(0.00365)	(0.0027)	
Tobit RE (B&M)	0.03662	0.00098	0.0180	0.07426	–0.0207	
Logl = 7133.42	(0.00697)	(0.00008)	(0.00047)	(0.00164)	(0.0015)	0.0706
σ_u0.09117						
Tobit RE (MSL)	0.03073	0.00119	0.01798	0.07345	–0.02103	
Logl = 7167.50	(0.00285)	(0.000034)	(0.00018)	(0.0008)	(0.0009)	0.0693
$\sigma_u = 0.09708$						
Tobit RE-Mundlak	0.1668	0.00905	0.01641	0.07107	–0.02223	
Logl = 8325.72	(0.0008)	(0.00015)	(0.00121)	(0.0020)	(0.0017)	0.0662
$\sigma_u = 0.08609$		–0.01041	–0.00220	0.01643	0.0119	
		(0.00018)	(0.0032)	(0.00319)	(0.0031)	

illustrate the attenuation effect noted earlier. The remaining results are for the tobit model. Comparing either the random effects or the fixed effects results to the restricted MLE, the difference in the log likelihoods strongly suggests that some model with unobserved heterogeneity is appropriate. As for choosing between the fixed and random effects models, there is no simple test with known properties. A Hausman test of the random effects alternative against the fixed effects null hypothesis would appear to be inappropriate. Whether the MLE slope estimator with fixed effects is consistent or not remains to be established – based on the Monte Carlo study, it appears to be consistent – but there is little doubt that the variance estimator for the MLE of $\boldsymbol{\beta}$ in the fixed effects model is inconsistent when T is small. Note, as well, that the sample standard deviation of the 7,293 estimated fixed effects (dummy variable coefficients) is 0.58 compared to a random effects estimate of the standard deviation of the effects of about 0.086 in the final set of results. There is far more variation in the fixed effects estimates, doubtless due to the small samples (one to seven observations) used to estimate them. The random effects estimator is consistent and efficient under the alternative hypothesis. The final set of results in the table use the Mundlak correction to accommodate correlation between the unobserved effects and the regressors. In the limited range of this study, these would probably be the preferred estimates.

19.5.4 Sample selection models for panel data

The development of methods for extending sample selection models to panel data settings parallels the literature on cross-section methods. It begins with Hausman and Wise (1979) who devised a maximum likelihood estimator for a two-period model with attrition – the "selection equation" was a formal model for attrition from the sample. The subsequent literature on attrition has formally drawn the analogy between attrition and sample selection in a variety of applications, such as Keane *et al.* (1988) and Nijman and Verbeek (1992). A formal "effects" treatment for sample selection was first suggested in complete form by Verbeek (1990), who formulated a random effects model for the probit equation and a fixed effects approach for the main regression. Zabel (1992) criticized the specification for its asymmetry in the treatment of the effects in the two equations, and for the likelihood that neglected correlation between the effects and regressors in the probit model would render the FIML estimator inconsistent. His proposal involved fixed effects in both equations. Recognizing the difficulty of fitting such a model (as noted above), he then proposed using the Mundlak correction. It is useful to lay out the model in full: (The original notation has been changed slightly to conform with the preceding.)

$$\begin{aligned}
&y_{it}{}^{*} = \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad \eta_i = \bar{\mathbf{x}}_i' + \tau w_i, w_i \sim N[0, 1] \\
&d_{it}^{*} = \theta_i + \mathbf{z}_{it}'\alpha + u_{it}, \quad \theta_i = \bar{\mathbf{z}}_i'\delta + \omega v_i, v_i \sim N[0, 1] \qquad (19.5.9) \\
&(\varepsilon_{it}, u_{it}) \sim N_2[(0, 0), (\sigma^2, 1, \rho\sigma)].
\end{aligned}$$

The "selectivity" in the model is carried through the correlation between ε_{it} and u_{it}. The resulting log likelihood is built up from the contribution of individual i,

$$
\begin{aligned}
L_i = & \int_{-\infty}^{\infty} \prod\nolimits_{d_{it}=0} \Phi[-\mathbf{z}_{it}'\boldsymbol{\alpha} - \bar{\mathbf{z}}_i' - \omega v_i]\phi(v_i)dv_i \\
& \times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \prod\nolimits_{d_{it}=1} \Phi\left[\frac{\mathbf{z}_{it}'\boldsymbol{\alpha} + \bar{\mathbf{z}}_i' + \omega v_i + (\rho/\sigma)\varepsilon_{it}}{\sqrt{1-\rho^2}}\right]\frac{1}{\sigma}\phi\left(\frac{\varepsilon_{it}}{\sigma}\right)\phi_2(v_i, w_i)dv_i dw_i \\
\varepsilon_{it} = & \, y_{it} - \mathbf{x}_{it}'\beta - \bar{\mathbf{x}}_i' - \tau w_i
\end{aligned}
$$

The log likelihood is then $\ln L = \Sigma_i \ln L_i$. The log likelihood is formidable, and does require integration in two dimensions for any selected observations. We do note, however, that the bivariate normal integration is actually the product of two univariate normals, because in the specification above, v_i and w_i are assumed to be uncorrelated. Vella (1998) notes, "given the computational demands of estimating by maximum likelihood induced by the requirement to evaluate multiple integrals, we consider the applicability of available simple, or two step procedures." Before we examine a few of those, we note that with simulation methods developed since this survey, the likelihood function above can be readily evaluated using relatively straightforward (and available) techniques. (Vella and Verbeek (1999) do suggest this in a footnote, but do not pursue it.) To show this, we note that the first line in the log likelihood is of the form $\mathrm{E_v}[\prod_{\mathrm{d}=0}\Phi(\ldots)]$ and the second line is of the form $\mathrm{E_w}[\mathrm{E_v}[\Phi(\ldots)\phi(\ldots)/\sigma]]$. Either of these expectations can be satisfactorily approximated with the average of a sufficient number of draws from the standard normal populations that generate w_i and v_i. The term in the simulated likelihood that follows this prescription is

$$
\begin{aligned}
L_i^S = & \frac{1}{R}\sum\nolimits_{r=1}^{R} \prod\nolimits_{d_{it}=0} \Phi\left[-\mathbf{z}_{it}'\boldsymbol{\alpha} - \bar{\mathbf{z}}_i' - \omega v_{i,r}\right] \\
& \times \frac{1}{R}\sum\nolimits_{r=1}^{R} \prod\nolimits_{d_{it}=1} \Phi\left[\frac{\mathbf{z}_{it}'\boldsymbol{\alpha} + \bar{\mathbf{z}}_i' + \omega v_{i,r} + (\rho/\sigma)\varepsilon_{it,r}}{\sqrt{1-\rho^2}}\right]\frac{1}{\sigma}\phi\left(\frac{\varepsilon_{it,r}}{\sigma}\right) \\
\varepsilon_{it,r} = & \, y_{it} - x_{it}'\beta - \bar{\mathbf{x}}_i' - \tau w_{i,r}
\end{aligned}
$$

Maximization of this log likelihood with respect to ($\boldsymbol{\beta}$, σ, ρ, $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, π, τ, ω) by conventional gradient methods is quite feasible. Indeed, this formulation provides a means by which the likely correlation between v_i and w_i can be accommodated in the model. Suppose that w_i and v_i are bivariate standard normal with correlation ρ_{vw}. We can project w_i on v_i and write

$$w_i = \rho_{vw}v_i + (1 - \rho_{vw}{}^2)^{1/2}h_i \qquad (19.5.12)$$

where h_i has a standard normal distribution. To allow the correlation, we now simply substitute this expression for w_i in the simulated (or original) log likelihood, and add ρ_{vw} to the list of parameters to be estimated. The simulation is then over the still independent normal variates, v_i and h_i.[12]

Notwithstanding the derivation above, much of the recent attention has focused on simpler two-step estimators. Building on Ridder (1990) and Verbeek and Nijman (1992) (see Vella, 1998, for numerous additional references), Vella and Verbeek (1999) propose a two-step methodology that involves a random effects framework similar to the one above. As they note, there is some loss in efficiency by not using the FIML estimator. But, with the sample sizes typical in contemporary panel data sets, that efficiency loss may not be large. As they note, their two-step template encompasses a variety of models including the tobit model examined in the preceding sections and the mover stayer model noted above.

The Vella and Verbeek procedures require some fairly intricate maximum likelihood procedures. Wooldridge (1995) proposes an estimator that, with a few probably but not necessarily innocent assumptions, can be based on straightforward applications of conventional, everyday methods. We depart from a fixed effects specification,

$$\begin{aligned} y_{it}^* &= \eta_i + \mathbf{x}_{it}'\beta + \varepsilon_{it}, \\ d_{it}^* &= \theta_i + \mathbf{z}_{it}'\alpha + u_{it}, \\ (\varepsilon_{it}, u_{it}) &\sim N_2[(0,0), (\sigma^2, 1, \boldsymbol{\rho}\sigma)]. \end{aligned} \tag{19.5.13}$$

Under the mean independence assumption $\mathrm{E}[\varepsilon_{it}|\eta_i, \theta_i, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT}, v_{i1}, \ldots, v_{iT}, d_{i1}, \ldots, d_{iT}] = \boldsymbol{\rho} u_{it}$, it will follow that

$$\begin{aligned} &\mathrm{E}[y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \eta_i, \theta_i, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT}, v_{i1}, \ldots, v_{iT}, d_{i1}, \ldots, d_{iT}] \\ &\quad = \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho u_{it}. \end{aligned} \tag{19.5.14}$$

This suggests an approach to estimating the model parameters, however it requires computation of u_{it}. That would require estimation of θ_i which cannot be done, at least not consistently – and that precludes simple estimation of u_{it}. To escape the dilemma, Wooldridge suggests Chamberlain's approach to the fixed effects model,

$$\theta_i = f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \cdots + \mathbf{z}_{iT}'\mathbf{f}_T + h_i. \tag{19.5.15}$$

With this substitution,

$$\begin{aligned} d_{it}^* &= \mathbf{z}_{it}'\alpha + f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \cdots + \mathbf{z}_{iT}'\mathbf{f}_T + h_i + u_{it} \\ &= \mathbf{z}_{it}'\alpha + f_0 + \mathbf{z}_{i1}'\mathbf{f}_1 + \mathbf{z}_{i2}'\mathbf{f}_2 + \cdots + \mathbf{z}_{iT}'\mathbf{f}_T + w_{it} \end{aligned} \tag{19.5.16}$$

where w_{it} is independent of $\mathbf{z}_{it}, t = 1, \ldots, T$. This now implies that

$$\begin{aligned} \mathrm{E}[y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \eta_i, \theta_i, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT}, v_{i1}, \ldots, v_{iT}, d_{i1}, \ldots, d_{iT}] &= \eta_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho(w_{it} - h_i) \\ &= (\eta_i - \rho h_i) + \mathbf{x}_{it}'\beta + \rho w_{it}. \end{aligned} \tag{19.5.17}$$

To complete the estimation procedure, we now compute T cross-sectional probit models (reestimating $\mathrm{f}_0, \mathrm{f}_1, \ldots$ each time), and compute $\hat{\lambda}_{it}$ from each one.

The resulting equation,

$$y_{it} = a_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho\hat{\lambda}_{it} + v_{it} \tag{19.5.18}$$

now forms the basis for estimation of $\boldsymbol{\beta}$ and ρ by using a conventional fixed effects linear regression with the observed data.

19.6 Recent developments

As with all areas in econometrics – one of the most active and heavily populated fields in economics – many researchers are extending the models we have discussed here in many directions. Space hardly allows even a cursory review of the literature. What follows is a small sampler chosen more or less haphazardly from the vast recent literature.

We note, first, consistent with other areas, recently developed simulation methods, the Gibbs sampler and Markov Chain Monte Carlo methods, have enabled researchers to extend classical methods into Bayesian frameworks. For example, Bayesian techniques have been developed for the sample selection model, including those by Li, Poirier and Tobias (2004) and Li (1998). The first of these examines a type of sample selection model sometimes called the *mover-stayer model*,

$$\begin{aligned}
&d_i^* = \mathbf{z}_i'\alpha + u_i, d_i = 1[d_i^* > 0]\\
&y_i|(d_i = 1) = \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_{i1}\\
&y_i|(d_i = 0) = \mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_{i0}\\
&\varepsilon_{i1}, \varepsilon_{i0} \sim N_2[(0.0), (\sigma_1^2, \sigma_0^2, \rho\sigma_1\sigma_0)], i = 1, \ldots, N
\end{aligned} \tag{19.6.1}$$

The name of the model derives from studies of migration, in which income is analyzed after migration or nonmigration. There are two intriguing aspects of the model that Poirier and Tobias examine. First, a crucial parameter, ρ, is not identified in the observed data. Second, the model specification suggests an interesting problem of predicting the outcome variable on the road not taken. (This theme figures prominently in the treatment effects literature, where the question of how the treatment would affect those not treated if they had taken it for example, job training, assistance, drug). The authors use the model to study incomes of high school students, some of whom drop out before their third year. Li (1998) examined a treatment effects model

$$\begin{aligned}
&y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \delta d_i + \varepsilon_i, \varepsilon_i \sim N[0, \sigma^2], i = 1, \ldots, N\\
&y_i = \max(0, \mathbf{x}_i'\boldsymbol{\beta} + \delta d_i + \varepsilon_i)\\
&d_i^* = \mathbf{z}_1'\alpha + u_i, d_i = 1[d_i^* > 0]
\end{aligned} \tag{19.6.2}$$

with the added complication that the outcome variable is censored. This precludes two-step least squares based estimation strategies, and mandates a likelihood based procedure instead. Li uses a Bayesian, MCMC procedure to estimate the parameters of the model. The technique is applied to a sample of times in default for firms who declare bankruptcy.

The strict normality assumptions that underlie the familiar tobit, probit, truncated regression and Heckman's sample selection model have perhaps attracted the most attention of contemporary researchers. Moon (1989) reconsidered the nonlinear least squares estimators mentioned earlier. The conditional mean function defined for the tobit model,

$$\mathrm{E}[y|\mathbf{x}] = \Phi(\mathbf{x}_i'\gamma)\sigma[\mathbf{x}_i'\gamma + \lambda(\mathbf{x}_i'\gamma)] \tag{19.6.3}$$

is amenable to nonlinear least squares. However, it is no less reliant on the normality assumption than is the likelihood function, so it has no advantage over the MLE and one shortcoming – it is less efficient. Moon (1989) examines ways to relax the assumptions to produce a more robust estimator that can be estimated by nonlinear least squares.

Many recent studies, both theoretical and applied, have proposed semiparametric estimators that rely on fewer or less stringent distributional assumptions. Powell (1984, 1986) is an early contribution. The censored least absolute deviations estimator (CLAD),

$$\hat{\boldsymbol{\beta}} = \arg\ \min_{\boldsymbol{\beta}} \textstyle\sum_{i=1}^{N} |y_i - \max(0, \mathbf{x}_i'\boldsymbol{\beta})| \tag{19.6.4}$$

is consistent even in the absence of normality – it requires only that the conditional median of y_i^* be zero. There are several practical problems in implementing the CLAD estimator, including the possibility of multiple optima. Bilias *et al.* (2000) proposed a bootstrapping method that they argue is better behaved. The programming problem is asymptotically equivalent to

$$\hat{\boldsymbol{\beta}}* = \arg\ \min_{\boldsymbol{\beta}} \textstyle\sum_{i=1}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| \times 1(\mathbf{x}_i'\boldsymbol{\beta}_0 > 0) \tag{19.6.5}$$

where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, that is, the parameter we are trying to estimate. We do note that, if we had the true $\boldsymbol{\beta}$ that we were trying to estimate, this minimization would be unnecessary. The authors suggest substituting Powell's consistent estimator for $\boldsymbol{\beta}_0$, then using a bootstrapping procedure to sharpen the estimator of $\boldsymbol{\beta}$. Chen and Khan (2000) further extend Powell's approach to allow for unspecified heteroscedasticity. Moon's (1989) proposed estimator is not unrelated to this, and he, as these authors do, takes the CLAD estimator as a benchmark for comparison. Honore (1992) suggests how the CLAD estimator can be extended to accommodate panel data models with fixed effects. (It is worth noting that these estimators focus on estimation of a constant multiple of $\boldsymbol{\beta}$. Without information about scaling, further computation of partial effects and/or predictions is not possible. Since the models are "robust" to heteroscedasticity, no such information will be forthcoming.) An empirical exploration of the semiparametric estimators is given in Chay and Honore (1998).

One extension of the semiparametric methods is to the panel data models of sample selection. Several studies have pursued this, including Kyriazidou (1997), Honore and Kyriazidou (2000) and Lee (2001). (See Vella (1998) for a lengthy survey of these and other semiparametric and nonparametric approaches to modeling selection.) In general, the recent applications have considered the assumptions under which first differences of y_{it} and x_{it} can be used for adjacent pairs of "selected observations". Kyriazidou's (1997) estimator builds on a fixed effects model,

$$\begin{aligned} &y^*_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \eta_i + \varepsilon_{it}, \varepsilon_{it} \sim N[0, \sigma^2], i = 1, \ldots, N, \text{observed when } d_{it} = 1. \\ &d^*_{it} = \mathbf{z}'_{it}\boldsymbol{\alpha} + \theta_i + u_{it}\ d_{it} = 1(d^*_{it} > 0). \end{aligned} \tag{19.6.6}$$

Minimal assumptions are made about the conditional distributions – that is, the point of the semiparametric approach. The estimator proceeds in two steps. At the first, a robust (semiparametric) estimator of $\boldsymbol{\theta}$ in the binary choice model is obtained (Manski's (1985, 1986, 1987) maximum score estimator or Klein and Spady's (1993) semiparametric estimator). At the second step, the estimator is weighted least squares using adjacent (both selected) observations,

$$\hat{\beta} = \left[\sum\nolimits_{i=1}^{N} D_i\hat{\psi}_i(\Delta\mathbf{x}_i)(\Delta\mathbf{x}_i)'\right]^{-1}\left[\sum\nolimits_{i=1}^{N} D_i\hat{\psi}_i(\Delta\mathbf{x}_i)(\Delta y_i)'\right] \tag{19.6.7}$$

where Δ creates the first differences of the observations, D_i equals 1 if $d_{it} = d_{i,t-1}$ (that is, if the two adjacent observations are both selected) and zero otherwise, and $\hat{\psi}_i$ is a weight that declines to zero as the magnitude of $|\mathbf{z}_{i,\,t}\hat{\alpha} - \mathbf{z}'_{i,\,t-1}\hat{\alpha}|$ increases; the author suggests a *kernel function* for the weight. Honore and Kyriazidou (2000) and Lee (2001) explore various aspects of this estimator. Note that the use of differences eliminates the- time-invariant effect from the equations, so it has the virtue of obviating any strong assumptions (such as random effects). On the other hand, using first differences removes any interesting time invariant independent variables, as well. Another interesting aspect of this class of estimators is that it allows the use of pairs of observations that are not adjacent in time. This *exchangeability* aspect is pursued at length in the papers mentioned. Rochina-Barrachina (1999) and Dustman and Rochina-Barrachina (2000) proposes a similar estimator based on differences of the selected observations.

The semiparametric approach has been applied to a variety of settings. Gurmu (1997) has used a hurdle/Poisson model with a semiparametric framework for unobserved heterogeneity in a model for the number of doctor visits in a sample of Medicaid patients. Lee (2004) also examined a count response variable; like Gurmu, he examined the number of doctor visits in a sample on health and retirement (see Lee (2004, p. 332) for discussion). In his study, the Poisson model is extended to accommodate an endogenous treatment effect, the amount of exercise. The treatment here is an ordinal variable – high, medium, low – so this model is a bit different (and somewhat more complicated) than the usual case in which the treatment is simply on or off, a binary variable. The hurdle model for counts in a study of health care outcomes is a frequent subject of analysis. Winkelmann (2004) (the source for the data in our application above) is another application.

The programming estimators considered here are "direct" estimators based on minimizing a particular criterion function, either the sum of absolute values or, in Moon's (1989) case, the sum of squares. A number of authors have approached the problem from the direction of moment based (GMM) estimation. Lee (2002) suggests an approach to estimation of the basic censored regression model, while Honore (2002) and Kyriazidou (1997, 2001) extend the model to the sample selection specification. In these cases, the estimators are highly robust, but at the high cost of limiting attention to the $T=2$ case. Research on this type of estimation methodology is ongoing.

The models discussed above are all static – there are no considerations of dynamic behavior thus far. That is a moot point in the cross-section variants of the models considered, but a crucial assumption of the panel data approaches described in section 19.5.[13] The issue of dynamics in panel data models is a vast literature in itself – at this late juncture, we eschew even a hint at a survey style list. The form in which dynamics should be introduced into the model (any model) is, itself, not a simple issue. Wooldridge (2005) proposes the following general specification for the tobit model (among several he considers) with unobserved time invariant effects.

$$\begin{aligned} y_{it} &= \max(0, \mathbf{x}_{it}'\boldsymbol{\beta} + g(y_{i,t-1})\rho + \alpha_i + \varepsilon_{it}) \\ \varepsilon_{it}&|\mathbf{x}_{it}, \alpha_i, y_{i,0}, y_{i,1}, \ldots, y_{i,t-1} \sim \mathrm{N}[0, \sigma^2]. \end{aligned} \tag{19.6.8}$$

where g(·) is some transformation of the lagged observed value – it will usually be $y_{i,t-1}$ itself – and $y_{i,0}$ is the observed initial condition. (We have changed the notation slightly to conform to ours, and limited attention to a single lagged value (as he actually does as well).) Wooldridge explores the conditions under which we may write the density for the observed variable (using the Olsen transformation as usual), as

$$\begin{aligned} \ln f(y_{i,t}|x_{i,t}, \alpha_i, y_{i,t-1}) = {} & 1[y_{i,t}=0] \ln \Phi\Big[\mathbf{x}_{i,t}'\boldsymbol{\gamma} + y_{i,t-1}\mu + \sigma_\alpha a_i\Big] \\ & + 1[y_{i,t}>0] \ln \theta\phi\Big[(\theta y_{i,t} - \mathbf{x}_{i,t}'\boldsymbol{\gamma} - y_{i,t-1}\mu - \sigma_\alpha a_i)\Big] \end{aligned} \tag{19.6.9}$$

We have isolated the standard deviation of α_i, and consistent with the normalization of the model by $1/\sigma$, what we have labeled σ_α above is actually $\mathrm{Var}[\alpha_1]^{1/2}/\sigma$. The crucial step in Wooldridge's analysis is the assumptions that allow projection of a_i on known information; he writes

$$a_i = \alpha_0 + \alpha_1 y_{i,0} + \mathbf{x}_i'\boldsymbol{\alpha}_1 + w_i \tag{19.6.10}$$

where $\mathbf{x}_i$ is (a bit ambiguously) defined to include some or all observations on $\mathbf{x}_{it}$, and w_i is normally distributed with zero mean and constant variance. (Asymptotics and other technical details may be found in Wooldridge's study.) Inserting the equation for a_i into the density for $y_{i,t}$, and summing the logs produces, as he notes, a "simple solution" to the initial conditions problem in a dynamic tobit model. The end result is a tobit random effects model, precisely the one we examined in section 19.5.[14]

19.7 Summary and conclusions

The preceding sections has outlined the basic modeling frameworks that are used in analyzing microeconomic data when the response variable is truncated, censored, or otherwise affected by transformation before being observed. The essential models for truncation, censoring and sample selection have provided the starting points for a vast array of applications and theoretical developments. The full set of results for the fully parametric models based on the normal distribution are well established. Ongoing contemporary research is largely focused on less parametric approaches, on panel data, and on different kinds of data generating mechanisms, such as models for counts and for discrete choices.

Appendix: LIMDEP Commands for Model Estimation

```
? Generic – File will be loaded from the File menu on the desktop
? Load ; File = Health.lpj $
Namelist ; xt = age,educ,married,hhkids$
? Censor the income data
Create ; income = hhninc;if(income > .35)income = .35$
Dstat ; rhs = income,xt$
Reject ; income >= .35$
? Pooled OLS nonlimit data
Regress ; lhs = income;rhs = one,xt$
? Pooled truncation using nonlimit data
Truncation ; lhs = income;rhs = one,xt;limit = .35;upper$
? Restore full sample
Sample ; All $
? Pooled OLS using full sample
Regress ; lhs = income;rhs = one,xt$
? Pooled tobit using full sample
Tobit;lhs = income;rhs = one,xt;limit = .35;upper$
? Tobit with fixed effects. Retain dummy variable coefficients
Tobit ; lhs = income;rhs = one,xt
      ;limit = .35 ;upper;pds = numobs;fem;parameters$
Sample ; 1-7293$
Create ; ai = alphafe$
Calc ; list ;sdv(ai)$$
Sample ; all$
? Tobit with random effects using quadrature
Tobit ; lhs = income ; rhs = one,xt ; limit = .35 ; upper;pds = numobs$
? For MSL program convert to a zero censored variable
Create ; income35 = .35 – income$
? Tobit with random effects using Monte Carlo integration.
? Need to reverse signs of coefficients and adjust constant
? Appropriate constant is .35-b0. Reported in Table 4.
```

```
Tobit ; lhs = income35;rhs = one,xt;pds = numobs
     ;rpm;fcn = one(n) ; Halton draws ; pts = 50 $
? Get group means for Mundlak correction
Matrix ; meanx = gxbr(xt,id)$
Create ; xbage = meanx(id,1)$
Create ; xbeduc = meanx(id,2)$
Create ; xbmarr = meanx(id,3)$
Create ; xbkids = meanx(id,4)$
? Random effects model with group means added to the model
Tobit ;lhs = income;rhs = one,xt,xbage,xbeduc,xbmarr,xbkids
     ;limit = .35;upper;pds = numobs$
```

Notes

1. The origin of the model's name, "tobit", is the subject of some speculation. Popular lore has it as a play on "Tobin's probit", in reference to Tobin (1958) and his model's connection to the probit (binary choice) model. However, a deeper look into the archives uncovers the same James Tobin's appearance as Tobit, the midshipman "with a mind like a sponge..." in Tobin's Columbia friend, Herman Wouk's (1951) classic work, *The Caine Mutiny*. (http://www.economyprofessor.com/theorists/jamestobin.php).
2. The differences between these estimators is illusory. In all cases, they are equivalent to gradient methods each using its own weighting matrix. Some, e.g., Newton's method, are more efficient (computationally) than others (e.g., the EM method).
3. Surprisingly, this fit measure has become a required standard in some fields in some journals. This fit measure bears only a slight connection to the fit of the model to the data, even in the linear regression model. For the linear model, a little algebra shows it to equal $\ln(s^2/s_0^2)/[\ln(s^2/s_0^2)+1+\ln 2\pi+\ln s_0^2)]$, which can be distressingly low even in models that have "excellent fit." Note that it is a function of the scale of the data. In a simple experiment, we used a random number generator to generate 1000 standard normal observations on x_i and ε_i, then, $y_i = x_i + \varepsilon_i$. Linear regression of y_i on x_i and a constant produces an R^2 of .5193 and a pseudo-R^2 of .20602. Multiplying y_i by 10 and repeating the exercise leaves R^2 unchanged (of course), but reduces the pseudo-R^2 to .09. To cite another example, in the author's experience, values of .02 appear to be routine in ordered probit models for which conventional prediction procedures based on the estimated model give the correct value for the dependent variable 90% of the time.
4. We note that, among the other shortcomings of most semiparametric estimators of the censored regression model, they are estimated "up to (an unknown) scale." Some are even robust to heteroscedasticity. This is not a virtue – it precludes prediction and estimation of partial effects.
5. The counterpart for the truncated regression model is $e_i = (\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma}) - \lambda_i$.
6. The dependent variable analyzed in Fair (1978) was a reported count that was censored in several ranges. The reported count variable was transformed to 0,1,2,3,(4–10) = 7, (anything else) = 12. Fair analyzed this count variable with the tobit model discussed above as if it were continuous, and treated the censoring as having occurred at the zero point. These data obviously fall more naturally into the corner solution interpretation (see Wooldridge, 2002). See Greene (2003, chapter 22) for a reanalysis of these data using the censored count data model suggested here.
7. Vella (1998) is a thorough, excellent survey of this topic recounted clearly from a practitioner's viewpoint.
8. Heckman and MaCurdy (1981) suggested an iterative procedure whereby, given initial estimates of the parameters, the dummy variable coefficients be estimated conditionally, one at a time each based on T_i observations, then with estimates of α_{i} in hand, the slopes be estimated, then back and forth until convergence. Because the Hessian is not block

diagonal – the parameter space cannot be partitioned – this procedure does not maximize the full log likelihood function. It can only be done directly, by "brute force."

9. The computational method for fittting the model with large numbers of dummy variables appears not to be widely known.
10. The exact expected value of the variance estimator in the linear model with fixed effects is easy to find with elementary matrix algebra – see any graduate-level textbook in econometrics, for example.
11. All computations reported were done using LIMDEP Version 8.0. Readers who wish to replicate (or extend) the results will find the data on the Journal of Applied Econometrics data archive website for 2004. They are also stored in the forms of an Excel$^{\text{TM}}$ spreadsheet and a LIMDEP project file on the author's website at http://www.stern.nyu.edu/~wgreene/Econometrics/healthcare.lpj (and .xls). The commands for LIMDEP are given in the appendix.
12. The estimator is contained in the current version of LIMDEP (Econometric Software, 2006).
13. There are very few strict time series applications of the models discussed in this chapter. Censoring and truncation are generally viewed as signature features of microeconomic (cross section and panel) data. However, Lee (1999, 2004) does consider a time series specification of the tobit model, extending it to a GARCH framework. This extension is, as one might expect, extremely complicated. Relevant applications remain forthcoming. Lee (1999) cites a number of natural candidates involving, for example, intervention in foreign exchange markets intended to limit movement of exchange rates.
14. In an application with more than a trivial number of periods and a substantial number of regressors, the expression for a_i is likely to have an excessive number of terms. As a useful approximation, one might want just to use the Mundlak approach, and replace the full set of vectors $\mathbf{x}_{it}$ with the group means of the time-varying variables.

References

Abrevaya, J. (1997) The equivalence of two estimators of the fixed effects logit model. *Economics Letters* **55**(1), 41–4.

Amemiya, T. (1978) The estimation of a simultaneous equation generalized probit model. *Econometrica* **46**(5), 1193–1205.

Andersen, E. (1970) Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* **32**, 283–301.

Angrist, J. (2001) Estimation of limited dependent variable models with binary endogenous regressors: simple strategies for empirical practice. *Journal of Business and Economic Statistics* **19**(1), 1–14.

Bera, A. and C. Jarque (1982) Model specification tests: a simultaneous approach. *Journal of Econometrics* **20**, 59–82.

Bera, A., C. Jarque and L. Lee (1981) Testing for the normality assumption in limited dependent variable models. Manuscript, Department of Economics, University of Minnesota.

Bhat, C. (1994) Imputing a continuous income variable from grouped and missing income observations. *Economics Letters* **46**(4), 311–20.

Bilias, Y., S. Chen and Z. Ying (2000) Simple resampling methods for censored regression quantiles. *Journal of Econometrics* **99**(2), 373–86.

Butler, J. and R. Moffitt (1982) A computationally efficient quadrature procedure for the one factor multinomial probit model. *Econometrica* **50**, 761–4.

Chay, K. and B. Honore (1998) Estimation of semiparametric censored regression models: an application to changes in black–white earnings inequality during the 1960s. *Journal of Human Resources* **33**(1), 4–38.

Chen, S. and S. Khan (2000) Estimating censored regression models in the presence of nonparametrtic multiplicative heteroscedasticity. *Journal of Econometrics* **98**(2), 283–316.

Chesher, A. and M. Irish (1987) Residual analysis in the grouped data and censored normal linear model. *Journal of Econometrics* **34**, 33–62.

Cheung, C. and A. Goldberger (1984) Proportional projections in limited dependent variable models. *Econometrica* **52**, 531–4.

Chib, S. (1992) Bayes regression for the Tobit censored regression model. *Journal of Econometrics* **51**, 79–99.

Cragg, J. (1971) Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* **39**, 829–44.

Dale, S. and A. Krueger (1999) Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables. Princeton University Industrial Relations Section Working Paper Number 409.

DeMaris, A. (2004) *Regression with Social Data: Modeling Continuous and Limited Response Variables*. New York: John Wiley and Sons.

Dempster, A., N. Laird and D. Rubin (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Dhrymes, P. (1986) Limited dependent variables. In Z. Griliches and M. Intriligator (eds), *Handbook of Econometrics*, vol. 2. Amsterdam: North-Holland.

Dionne, G., R. Gagne and C. Vanesse (1998) Inferring technological parameters from incomplete panel data. *Journal of Econometrics* **87**(2), 303–29.

Duncan, G. (1983) Sample selectivity as a proxy variable problem: on the use and misuse of Gaussian selectivity corrections. *Research in Labor Economics*, Supplement 2, 333–45.

Duncan, G. (1986) A Semiparametric censored regression estimator. *Journal of Econometrics* **31**, 5–34.

Dustman, C. and M. Rochina-Barrachina (2000) Selection correction in panel data models: an application to labour supply and wages. Discussion Paper No. 162, IZA, Bonn, Germany.

Econometric Software, Inc. (2006) *LIMDEP*, Version 9.0. Plainview, NY: Econometric Software, Inc.

Fair, R. (1977) A note on computation of the Tobit estimator. *Econometrica* **45**, 1723–1727.

Fair, R. (1978) A theory of extramarital affairs. *Journal of Political Economy* **86**, 45–61.

Fin, T. and P. Schmidt (1984) A test of the Tobit specification against an alternative suggested by Cragg. *Review of Economics and Statistics* **66**, 174–7.

Greene, W. (1981) Sample selection bias as a specification error: comment. *Econometrica* **49**, 795–8.

Greene, W. (1983) Estimation of limited dependent variable models by ordinary least squares and the method of moments. *Journal of Econometrics* **21**, 195–212.

Greene, W. (1995) Sample selection in the Poisson regression model. Working Paper No. EC-95–6, Department of Economics, Stern School of Business, New York University.

Greene, W. (1997) FIML estimation of sample selection models for count data. Stern School, Department of Economics, Working Paper 97–02.

Greene, W. (1999) Marginal effects in the censored regression model. *Economics Letters* **64**(1), 43–50.

Greene, W. (2000) *LIMDEP, User's Manual*. Plainview, NY: Econometric Software.

Greene, W. (2003) *Econometric Analysis*, 6th edn. Upper Saddle River: Prentice Hall.

Greene, W. (2004) Fixed effects and the incidental parameters problem in the tobit model. *Econometric Reviews* **23**(2), 125–48.

Gurmu, S. (1997) Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. *Journal of Applied Econometrics* **12**(3), 225–42.

Hausman, J. and D. Wise (1977) Social experimentation, truncated distributions, and efficient estimation. *Econometrica* **45**, 919–38.

Hausman, J. and D. Wise (1979) Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica* **47**(2), 1979, 455–573.

Heckman, J. (1979) Sample selection bias as a specification error. *Econometrica* **47**, 153–61.

Heckman, J. and T. MaCurdy (1981) A life cycle model of female labor supply. *Review of Economic Studies* **47**, 247–83.

Honore, B. (1992) Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* **60**, 533–65.

Honore, B. (2002) Non-linear models with panel data. Insititute For Fiscal Studies, CEMMAP, Working Paper CWP13/02.

Honore, B. and E. Kyriazidou (2000) Panel data discrete choice models with lagged dependent variables. *Econometrica* **68**(4), 839–74.

Honore, B. and E. Kyriazidou (2000) Estimation of tobit-type models with individual specific effects. *Econometric Reviews* **19**(3), 341–66.

Jensen, P., M. Rosholm and M. Verner (2001) A comparison of different estimators for panel data sample selection models. Manuscript, Department of Economics, CIM, CLS, Aarhus School of Business.

Jones, A. (1994) Health, addiction, social interaction and the decision to quit smoking. *Journal of Health Economics* **13**, 93–110.

Keane, M., R. Moffitt and D. Runkle (1988) Real wages over the business cycle: estimating the impact of heterogeneity with micro-data. *Journal of Political Economy* **96**(6), 1232–1265.

Kiefer, N. (ed.) (1993) Econometric analysis of duration data. *Journal of Econometrics* **28**(1), 1–169.

Klein, R. and R. Spady (1993) An efficient semiparametric estimator for discrete choice. *Econometrica* **61**, 387–421.

Kyriazidou, E. (1997) Estimation of a panel data sample selection model. *Econometrica* **65**(6), 1335–1364.

Kyriazidou, E. (2001) Estimation of dynamic panel data sample selection models. *Review of Economic Studies* **68**, 543–72.

Lee, L. (1999) Estimation of dynamic and ARCH tobit models. *Journal of Econometrics* **92**, 355–90.

Lee, L. (2004) Nonstandard dependent variables: some common structures of simulated specifications tests. Manuscript, Department of Economics, The Ohio State University.

Lee, M. (1996) *Method of Moments and Semiparametric Econometrics for Limited Dependent Variables*. New York: Springer-Verlag.

Lee, M. (2001) First difference estimators for panel censored selection models. *Economics Letters* **70**(1), 43–50.

Lee, M. (2002) *Panel Data Econometrics: Methods of Moments and Limited Dependent Variables*. New York: John Wiley.

Lee, M. (2004) Selection correction and sensitivity analysis for ordered treatment effect on count response. *Journal of Applied Econometrics* **19**(3), 323–37.

Li, K. (1998) Bayesian inference in a simultaneous equation model with limited dependent variables. *Journal of Econometrics* **85**(2), 387–400.

Li, M., D. Poirier and J. Tobias (2004) Do dropouts suffer from dropping out? Estimation and prediction of outcome gains in generalized selection models. *Journal of Applied Econometrics* **19**(2), 203–26.

Long, S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

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

Manski, C. (1985) Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics* **27**, 313–33.

Manski, C. (1986) Operational characteristics of the maximum score estimator. *Journal of Econometrics* **32**, 85–100.

Manski, C. (1987) Semiparametric analysis of the random effects linear model from binary response data. *Econometrica* **55**, 357–62.

Melenberg, B. and A. van Soest (1996) Parametric and semi-parametric modelling of vacation expenditures. *Journal of Applied Econometrics* **11**(1), 59–76.

Moon, C. (1989) A Monte Carlo comparison of semiparametric Tobit estimators. *Journal of Applied Econometrics* **4**(4), 361–82.

Mundlak, Y. (1978) On the pooling of time series and cross sectional data. *Econometrica* **56**, 69–86.

Murphy, K. and R. Topel (1985) Estimation and inference in two step econometric models. *Journal of Business and Economic Statistics* **3**, 370–9.

Nelson, F. (1981) A test for misspecification in the censored normal model. *Econometrica* **49**, 1317–1329.

Newey, W., J. Powell and J. Walker (1990) Semiparametric estimation of selection models. *American Economic Review* **80**, 324–8.

Neyman, J. and E. Scott (1948) Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Nijman, T. and M. Verbeek (1992) Nonresponse in panel data: the impact on estimates of the life cycle consumption function. *Journal of Applied Econometrics* **7**(3), 243–57.

Olsen, R. (1978) A note on the uniqueness of the maximum likelihood estimator in the Tobit model. *Econometrica* **46**, 1211–1215.

Pagan, A. and F. Vella (1989) Diagnostic tests for models based on individual data: a survey. *Journal of Applied Econometrics* **4**, Supplement, S29–S59.

Powell, J. (1984) Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* **25**, 303–25.

Powell, J. (1986) Censored regression quantiles. *Journal of Econometrics* **32**, 143–55.

Ridder, G. (1990) Attrition in multiwave panel data. In J. Hartog, E. Ridder and J. Theeuwes (eds), *Panel Data and Labor Market Studies*. Amsterdam: Elsevier.

Rochina-Barrachina, M. (1999) A new estimator for panel data sample selection models. *Annales d'Economie et de Statistique* **55/56**, 153–81.

Schafgans, M. (1998) Ethnic wage differences in Malaysia: parametric and semiparametric estimation of the Chinese–Malay wage gap. *Journal of Applied Econometrics* **13**(5), 481–504.

Shaw, D. (1988) "'On-Site Samples' Regression Problems of Nonnegative Integers, Truncation, and Endogenous Stratification." *Journal of Econometrics* **37**, 211–23.

Skeels, C. and F. Vella (1999) A Monte Carlo investigation of the sampling behavior of conditional moment tests in Tobit and probit models. *Journal of Econometrics* **92**(2), 275–94.

Terza, J. (1985) A Tobit type estimator for the censored poisson regression model. *Economics Letters* **18**, 361–5.

Terza, J. (1998) Estimating count data models with endogenous switching: sample selection and endogenous treatment effects. *Journal of Econometrics* **84**(1), 129–54.

Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica* **26**, 24–36.

Vella, F. (1992) Simple tests for sample selection bias in censored and discrete choice models. *Journal of Applied Econometrics* **7**(4), 413–22.

Vella, F. (1998) Estimating models with sample selection bias: a survey. *Journal of Human Resources* **33**(1), 127–69.

Vella, F. and M. Verbeek (1999) Two-step estimation of panel data models with censored endogenous variables and selection bias. *Journal of Econometrics* **90**, 239–63.

Verbeek, M. (1990) On the estimation of a fixed effects model selectivity bias. *Economics Letters* **34**, 267–70.

Verbeek, M. and T. Nijman (1992) Testing for selectivity bias in panel data models. *International Economic Review* **33**(3), 681–703.

Winkelmann, R. (2004) Health care reform and the number of doctor visits – an econometric analysis. *Journal of Applied Econometrics* **19**(4), 455–72.

Wooldridge, J. (1995) Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics* **68**(1), 115–32.

Wooldridge, J. (2002) *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, J. (2005) Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**(1), 39–54.

Wouk, H. (1951) *The Caine Mutiny*. New York: Alden Press.

Zabel, J. (1992) Estimating fixed and random effects models with selectivity. *Economics Letters* **40**, 269–72.