EC771: Econometrics, Spring 2008

Greene, Econometric Analysis (6th ed, 2008)

Chapters 21, 22:
Limited dependent variable models

We now consider models of *limited dependent variables*, in which the economic agent's response is limited in some way. The dependent variable, rather than being continuous on the real line (or half-line), is restricted. In some cases, we are dealing with *discrete choice*: the response variable may be restricted to a Boolean or binary choice, indicating that a particular course of action was or was not selected. In others, it may take on only integer values, such as the number of children per family, or

the ordered values on a Likert scale. Alternatively, it may appear to be a continuous variable with a number of responses at a threshold value. For instance, the response to the question "how many hours did you work last week?" will be recorded as zero for the non-working respondents. None of these measures are amenable to being modeled by the linear regression methods we have discussed. These models require greater computational effort to estimate, and they also pose challenges to the researcher in interpreting their findings.

We first discuss models of binary choice which may be estimated by binomial logit or probit techniques. The following section takes up their generalization to ordered logit or ordered probit in which the response is one of a set of values from an ordered scale. We then present techniques appropriate for truncated and censored data and their extension to sample selection models. We will not discuss models

of "count data" in which the response variable is the count of some item's occurrence for each observation. The methodology appropriate for these data is not a standard linear regression, since it cannot take into account the constraint that the data (and the model's predictions) can only take on non-negative integer values. Stata provides comprehensive facilities for modeling count data via *Poisson regression* and its generalization, the *negative binomial regression*; see `poisson` and `nbreg`, respectively.

Binomial logit and probit models

We first consider models of Boolean response variables, or *binary choice*. In such a model, the response variable is coded as 1 or 0, corresponding to responses of True or False to a particular question:

- Did you watch the seventh game of the 2004 World Series?

- Were you pleased with the outcome of the 2004 presidential election?

- Did you purchase a new car in 2005?

A behavioral model of each of these phenomena could be developed, including a number of "explanatory factors" (we should not call them regressors) that we expect will influence the respondent's answer to such a question. But we should readily spot the flaw in the *linear probability model*:

$$R_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i \qquad (1)$$

where we place the Boolean response variable in R and regress it upon a set of X variables. All of the observations we have on R are either 0 or 1. They may be viewed as the *ex post* probabilities of responding "yes" to the question posed. But the predictions of a linear regression model are unbounded, and the model

of Equation (**??**), estimated with `regress`, can produce negative predictions and predictions exceeding unity, neither of which can be considered probabilities. Because the response variable is bounded, restricted to take on values of $\{0,1\}$, the model should be generating a predicted *probability* that individual i will choose to answer Yes rather than No. In such a framework, if $\beta_j > 0$, those individuals with high values of X_j will be more likely to respond Yes, but their probability of doing so must respect the upper bound. For instance, if higher disposable income makes new car purchase more probable, we must be able to include a very wealthy person in the sample and still find that the individual's predicted probability of new car purchase is no greater than 1.0. Likewise, a poor person's predicted probability must be bounded by 0.

Although it is possible to estimate Equation (**??**) with OLS the model is likely to produce

point predictions outside the unit interval. We could arbitrarily constrain them to either 0 or 1, but this linear probability model has other problems: the error term *cannot* satisfy the assumption of homoskedasticity. For a given set of X values, there are only two possible values for the disturbance: $-X\beta$ and $(1-X\beta)$: the disturbance follows a Binomial distribution. Given the properties of the Binomial distribution, the variance of the disturbance process, conditioned on X, is

$$Var(u|X) = X\beta\ (1 - X\beta) \qquad (2)$$

There is no constraint to ensure that this quantity will be positive for arbitrary X values. Therefore, it will rarely be productive to utilize regression with a binary response variable; we must follow a different strategy. Before proceeding to develop that strategy, let us consider an alternative formulation of the model from an economic standpoint.

The latent variable approach

A useful approach to motivate such an econometric model is that of a *latent variable*. Express the model of Equation (**??**) as:

$$y_i^* = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i \qquad (3)$$

where y^* is an unobservable magnitude which can be considered the net benefit to individual i of taking a particular course of action (e.g., purchasing a new car). We cannot observe that net benefit, but can observe the outcome of the individual having followed the decision rule

$$\begin{aligned} y_i &= 0 \text{ if } y_i^* < 0 \\ y_i &= 1 \text{ if } y_i^* \geq 0 \end{aligned} \qquad (4)$$

That is, we observe that the individual did or did not purchase a new car in 2005. If she did, we observed $y_i = 1$, and we take this as evidence that a rational consumer made a decision that improved her welfare. We speak of

y^* as a *latent variable*, linearly related to a set of factors X and a disturbance process u. In the latent variable model, we must make the assumption that the disturbance process has a known variance σ_u^2. Unlike the regression problem, we do not have sufficient information in the data to estimate its magnitude. Since we may divide Equation (**??**) by any positive σ without altering the estimation problem, the most useful strategy is to set $\sigma_u = \sigma_u^2 = 1$.

In the latent model framework, we model the probability of an individual making each choice. Using equations (**??**) and (**??**) we have

$$\begin{aligned} Pr[y^* > 0|X] &= \\ Pr[u > -X\beta|X] &= \\ Pr[u < X\beta|X] &= \\ Pr[y = 1|X] &= \Psi(y_i^*) \end{aligned} \tag{5}$$

The function $\Psi(\cdot)$ is a cumulative distribution function (CDF) which maps points on

the real line $\{-\infty, \infty\}$ into the probability measure $\{0, 1\}$. The explanatory variables in X are modeled in a linear relationship to the latent variable y^*. If $y = 1$, $y^* > 0$ implies $u < X\beta$. Consider a case where $u_i = 0$. Then a positive y^* would correspond to $X\beta > 0$, and *vice versa*. If u_i were now negative, observing $y_i = 1$ would imply that $X\beta$ must have outweighed the negative u_i (and *vice versa*). Therefore, we can interpret the outcome $y_i = 1$ as indicating that the explanatory factors and disturbance faced by individual i have combined to produce a positive net benefit. For example, an individual might have a low income (which would otherwise suggest that new car purchase was not likely) but may have a sibling who works for Toyota and can arrange for an advantageous price on a new vehicle. We do not observe that circumstance, so it becomes a large positive u_i, explaining how $(X\beta + u_i) > 0$ for that individual.

The parameters of binary choice models may be estimated by maximum likelihood techniques.For each observation, the density of y_i given X_i may be written as:

$$f(y|X) = [\Psi(X_i\beta)]^{y_i} [1 - \Psi(X_i\beta)]^{1-y_i}, \; y_i = 0, 1 \tag{6}$$

This implies that the log-likelihood for observation i may be written as

$$\ell_i(\beta) = y_i \log [\Psi(X_i\beta)] + (1-y_i) \log [1 - \Psi(X_i\beta)] \tag{7}$$

and the log-likelihood of the sample is $L(\beta) = \sum_{i=1}^{N} \ell_i(\beta)$, to be numerically maximized with respect to the k elements of β.

The two common estimators of the binary choice model are the *binomial probit* and *binomial logit* models. For the probit model, $\Psi(\cdot)$ is the CDF of the Normal distribution function (Stata's `norm` function):

$$Pr[y = 1|X] = \int_{-\infty}^{X\beta} \psi(t)dt = \Psi(X\beta) \tag{8}$$

where $\psi(\cdot)$ is the probability density function (PDF) of the Normal distribution: Stata's `normden` function. For the logit model, $\Psi(\cdot)$ is the CDF of the Logistic distribution:*

$$Pr[y = 1|X] = \frac{\exp(X\beta)}{1 + \exp(X\beta)} \tag{9}$$

The $CDFs$ of the Normal and Logistic distributions are similar. The Logistic distribution has fatter tails, resembling a Student t distribution with seven degrees of freedom.† The two models will produce quite similar results if the distribution of sample values of y_i is not too extreme. However, a sample in which the proportion $y_i = 1$ (or the proportion $y_i = 0$) is very small will be sensitive to the

*The PDF of the Logistic distribution, which is needed to calculate marginal effects, is $\psi(z) = \exp(z)/[1 + \exp(z)]^2$.

†Other distributions, including non-symmetric distributions, may be used in this context. For example, Stata's `cloglog` command fits the complementary log-log model $Pr[y = 1|X] = 1 - \exp(\exp(-X\beta))$.

choice of CDF. Neither of these cases are really amenable to the binary choice model. If a very unusual event is being modeled by y_i, the "naïve model" that it will not happen in any event is hard to beat. The same is true for an event that is almost ubiquitous: the naïve model that predicts that everyone has eaten a candy bar at some time in their lives is quite accurate.

We may estimate these binary choice models in Stata with the commands `probit` and `logit`, respectively. Both commands assume that the response variable is coded with zeros indicating a negative outcome and a positive, non-missing value corresponding to a positive outcome (i.e., I purchased a new car in 2005). These commands do not require that the variable be coded $\{0,1\}$, although that is often the case. Because any positive value (including all missing values) will be taken as a positive outcome, it is important to ensure that missing

values of the response variable are excluded from the estimation sample either by dropping those observations or using an `if` *depvar* `<` `.` qualifier.

Marginal effects and predictions

One of the major challenges in working with limited dependent variable models is the complexity of explanatory factors' marginal effects on the result of interest. That complexity arises from the nonlinearity of the relationship. In Equation (**??**), the latent measure is translated by $\Psi(y_i^*)$ to a probability that $y_i = 1$. While Equation (**??**) is a linear relationship in the β parameters, Equation (**??**) is not. Therefore, although X_j has a linear effect on y_i^*, it will not have a linear effect on the resulting probability that $y = 1$:

$$\frac{\partial Pr[y = 1|X]}{\partial X_j} = \frac{\partial Pr[y = 1|X]}{\partial X\beta} \cdot \frac{\partial X\beta}{\partial X_j} = \Psi'(X\beta) \cdot \beta_j = \psi(X\beta) \cdot \beta_j.$$

The probability that $y_i = 1$ is not constant over the data. Via the chain rule, we see that the effect of an increase in X_j on the probability is the product of two factors: the effect of X_j on the latent variable and the derivative of the CDF evaluated at y_i^*. The latter term, $\psi(\cdot)$, is the probability density function (PDF) of the distribution.

In a binary choice model, the marginal effect of an increase in factor X_j *cannot* have a constant effect on the conditional probability that $(y = 1|X)$ since $\Psi(\cdot)$ varies through the range of X values. In a linear regression model, the coefficient β_j and its estimate b_j measures the marginal effect $\partial y/\partial X_j$, and that effect is constant for all values of X. In a binary choice model, where the probability that $y_i = 1$ is bounded by the $\{0,1\}$ interval, the marginal effect *must* vary. For instance, the marginal effect of a one dollar increase in disposable income on the conditional probability that $(y =$

1|X) must approach zero as X_j increases. Therefore, the marginal effect in such a model varies continuously throughout the range of X_j, and must approach zero for both very low and very high levels of X_j.

Binomial probit

When using Stata's `probit` command, the reported coefficients (computed via maximum likelihood) are b, corresponding to β in Equation (**??**). The alternative command `dprobit` may be used to display the marginal effect $\partial Pr[y = 1|X]/\partial X_j$, that is, the effect of an infinitesimal change in X_j.[‡] The `probit` command may be given without any arguments following a `dprobit` command in order to "replay" the probit results in this format. This does not

[‡]Since an indicator variable cannot undergo an infinitesimal change, the default calculation for such a variable is the discrete change in the probability when the indicator is switched from 0 to 1.

affect the z-statistics or p-values of the estimated coefficients. Given the nonlinear nature of the model, the `dF/dx` reported by `dprobit` will vary through the sample space of the explanatory variables. By default, the marginal effects are calculated at the multivariate point of means. They can be calculated at other points via the `at()` option.

Alternatively, you can use `mfx` to compute the marginal effects. If a `probit` estimation is followed by the command `mfx`, the `dF/dx` values (identical to those from `dprobit`) will be calculated. The `mfx` command's `at()` option can be used to compute the effects at a particular point in the sample space. The `mfx` command may also be used to calculate elasticities and semi-elasticities.

By default, the `dF/dx` effects produced by `dprobit` or `mfx` are the marginal effects for an average

individual. Some argue that it would be more preferable to compute the *average marginal effect*: that is, the average of each individual's marginal effect. In large samples, these two measures will give the same answer, but in samples of moderate size, they may not. Official Stata does not contain such a capability, but a useful `margeff` (average marginal effects) routine has been written by Tamus Bartus and available from SSC. provides such a facility for `probit`, `logit` and a number of other Stata commands discussed in this chapter (although not `dprobit`). Its `dummies()` option should be used to signal the presence of categorical explanatory variables. If some explanatory variables are integer variables, the `count` option should be used.

After estimating a probit model, the `predict` command may be used, with a default option

`p`, the predicted probability of a positive outcome. The `xb` option may be used to calculate the *index function* for each observation: that is, the predicted value of y_i^* from Equation (**??**), which is in z-units (those of a standard Normal variable). For instance, an index function value of 1.69 will be associated with a predicted probability of 0.95 in a large sample.

Binomial logit and grouped logit

When the Logistic CDF is employed in Equation (**??**) the probability (π_i) of $y = 1$, conditioned on X, is $\exp(X\beta)/(1 + \exp(X\beta)$. Unlike the CDF of the Normal distribution, which lacks an inverse in closed form, this function may be inverted to yield

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = X\beta. \tag{10}$$

This expression is termed the *logit* of π_i, with that term being a contraction of the *log of*

the odds ratio. The *odds ratio* reexpresses the probability in terms of the odds of $y = 1$. It is not applicable to microdata in which y_i equals zero or one, but is well defined for averages of such microdata. For instance, in the 2004 U.S. presidential election, the *ex post* probability of a Massachusetts resident voting for John Kerry according to `cnn.com` was 0.62, with a logit of $\log(0.62/(1-0.62)) = 0.4895$. The probability of that person voting for George Bush was 0.37, with a logit of -0.5322. Say that we had such data for all 50 states. It would be inappropriate to use linear regression on the probabilities *voteKerry* and *voteBush*, just as it would be inappropriate to run a regression on individual voter's *voteKerry* and *voteBush* indicator variables. In this case, the `glogit` (grouped logit) command may be used to produce weighted least squares estimates for the model on state-level data. Alternatively, the `blogit` command may be used to

produce maximum-likelihood estimates of that model on grouped (or "blocked") data. The equivalent commands `gprobit` and `bprobit` may be used to fit a probit model to grouped data.

What if we have microdata in which voters' preferences are recorded as indicator variables, for example `voteKerry = 1` if that individual voted for John Kerry, and *vice versa*? As an alternative to fitting a probit model to that response variable, we may fit a *logit* model with `logit`. This command will produce coefficients which, like those of `probit`, express the effect on the latent variable y^* of a change in X_j (see Equation (**??**). Similar to the earlier use of `dprobit`, we may use the `logistic` command to compute coefficients which express the effects of the explanatory variables in terms of the odds ratio associated with that explanatory factor. Given the algebra of the model, the odds ratio is merely $\exp(b_j)$ for the j^{th} coefficient estimated by `logit`, and may also be requested

by specifying the `or` option on the `logit` command. It should be clear that *logistic regression* is intimately related to the binomial logit model, and is not an alternative econometric technique to `logit`. The documentation for `logistic` states that the computations are carried out by calling `logit`.

As in the case of `probit`, the default statistic calculated by `predict` after `logit` is the probability of a positive outcome. The `mfx` command will produce marginal effects expressing the effect of an infinitesimal change in each X on the probability of a positive outcome, evaluated by default at the multivariate point of means. Elasticities and semi-elasticities may also be calculated. Bartus's `margeff` routine may be employed to calculate the average marginal effects over the sample observations after either `logit` or `logistic`.

Evaluating specification and goodness of fit

Since both the binomial logit and binomial probit estimators may be applied to the same model, you might wonder which should be used. The $CDFs$ underlying these models differ most in the tails, producing quite similar predicted probabilities for non-extreme values of $X\beta$. Since the likelihood functions of the two estimators are not nested, there is no obvious way to test one against the other.The coefficient estimates of `probit` and `logit` from the same model will differ algebraically, since they are estimates of (β_j/σ_u). While the variance of the standard Normal distribution is unity, the variance of the Logistic distribution is $\pi^2/3 = 3.290$, causing reported logit coefficients to be larger by a factor of about $\sqrt{3.29} = 1.814$. However, we often are concerned with the marginal effects generated by these models rather than their estimated coefficients. From the examples above, the magnitude of the marginal effects generated by `mfx` or Bartus's `margeff` routine are likely to be quite similar for both estimators.

Tests for appropriate specification of a subset model may be carried out, as in the regression context, with the `test` command. The test statistics for exclusion of one or more explanatory variables are reported as χ^2 rather than F-statistics due to the use of large-sample maximum likelihood estimation techniques. The other post-estimation commands: tests of linear expressions with `test` or `lincom`, and tests of nonlinear expressions with `nltest` or `nlcom` may be applied as they may in the context of `regress`.

How can we judge the adequacy of a binary choice model estimated with `probit` or `logit`? Just as the "ANOVA F" tests a regression specification against the *null model* in which all regressors are omitted, we may consider a *null model* for the binary choice specification to be $Pr[y = 1] = \bar{y}$. Since the mean of an indicator variable is the sample proportion

of 1s, it may be viewed as the unconditional probability that $y = 1$. We may contrast that with the conditional probabilities generated by the model that takes the explanatory factors X into account. Since the likelihood function for the null model can readily be evaluated in either the probit or logit context, both commands produce a likelihood ratio test.As mentioned above, the null model is hard to beat if $\bar{y}$ is very close to 0 or 1.

Although this likelihood ratio test provides a statistical basis for rejection of the null model versus the estimated model, there is no clear consensus on a measure of goodness of fit analogous to R^2 for linear regression. Stata produces a measure called `Pseudo R2` for both commands, and indeed for all commands estimated by maximum likelihood; see `maximize`. Let L_1 be the log-likelihood value for the estimated model, as presented on the estimation

output after convergence. Let L_0 be the log-likelihood value for the null model excluding all explanatory variables. This quantity is not displayed, but is available after estimation as `e(ll_0)`. The `LR chi2(`*k-1*`)` likelihood ratio test is merely $2(L_1 - L_0)$, distributed $\chi^2(k-1)$ under the null hypothesis that the explanatory factors are jointly uninformative.

If we rearrange the log-likelihood values, we may define the `pseudo R2` as $(1 - L_1/L_0)$ which like regression R^2 is on a (0,1) scale, with 0 indicating that the explanatory variables failed to increase likelihood and 1 indicating that the model perfectly predicts each observation. We cannot interpret this, as in the case of linear regression, as the proportion of variation in y explained by X, but in other aspects it does resemble an R^2 measure.Perfect prediction is not an inexorable result of adding more explanatory factors to the model, as it is in linear regression. In fact, perfect prediction may

inadvertently occur because one or more explanatory factors are perfectly correlated with the response variable. Stata's documentation in `probit` and `logit` discuss this issue, which Stata will detect and report when encountered in the data.

A number of alternative measures based on the predictions of the binary choice model have been proposed, but all have their weaknesses, particularly if there is a high proportion of 0s or 1s in the sample. A number of these measures may be computed by the `lstat`command. With a constant term included the binomial logit model will produce $\bar{\hat{y}} = \bar{y}$, as does regression: the average of predicted probabilities from the model equals the sample proportion $\bar{y}$. That is not guaranteed in the binomial probit model.

Ordered logit and probit models

We earlier discussed the issues related to the use of *ordinal variables*: those which indicate a ranking of responses, rather than a cardinal measure, such as the codes of a Likert scale of agreement with a statement. Since the values of such an ordered response are arbitrary, an ordinal variable should not be treated as if it was measurable in a cardinal sense and entered into a regression, either as a response variable or as a regressor. However, what if we want to model an ordinal variable as the response variable, given a set of explanatory factors? Just as we can use binary choice models to evaluate the factors underlying a decision without being able to quantify the net benefit of making that choice, we may employ a generalization of the binary choice framework to model an ordinal variable using *ordered probit* or *ordered logit* estimation techniques.

In the latent variable approach to the binary choice model, we observe $y_i = 1$ if the individual's net benefit is positive: i.e., $y_i^* > 0$. The

ordered choice model generalizes this concept to the notion of multiple thresholds. For instance, a variable recorded on a five-point Likert scale will have four thresholds. If $y^* \leq \kappa_1$, we observe $y = 1$. If $\kappa_1 < y^* \leq \kappa_2$, we observe $y = 2$. If $\kappa_2 < y^* \leq \kappa_3$, we observe $y = 3$, and so on, where the κ values are the thresholds. In a sense, this can be considered imprecise measurement: we cannot observe y^* directly, but only the range in which it falls. This is appropriate for many forms of microeconomic data that are "bracketed" for privacy or summary reporting purposes. Alternatively, the observed choice might only reveal an individual's relative preference.

The parameters to be estimated are a set of coefficients β corresponding to the explanatory factors in X as well as a set of $(I - 1)$ threshold coefficients κ corresponding to the I alternatives. In Stata's implementation of these

estimators via commands `oprobit` and `ologit`, the actual values of the response variable are not relevant. Larger values are taken to correspond to higher outcomes. If there are I possible outcomes (e.g., 5 for the Likert scale), a set of threshold coefficients or *cut points* $\{\kappa_1, \kappa_2, \ldots, \kappa_{I-1}\}$ is defined, where $\kappa_0 = -\infty$ and $\kappa_I = \infty$. Then the model for the j^{th} observation defines:

$$Pr[y_j = i] = Pr[\kappa_{i-1} < \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + u_j < \kappa_i]$$

where the probability that individual j will choose outcome i depends on the product $X\beta$ falling between cut points $(i-1)$ and i. This is a direct generalization of the two-outcome binary choice model, which has a single threshold at zero. As in the binomial probit model, we assume that the error is normally distributed with variance unity (or distributed Logistic with variance $\pi^2/3$ in the case of ordered logit).

Prediction is more complex in the ordered probit (logit) framework, since there are I possible predicted probabilities corresponding to the I possible values of the response variable. The default option for `predict` is to compute predicted probabilities. If I new variable names are given in the command, they will contain the probability that $i = 1$, the probability that $i = 2$, and so on.

The marginal effects of an ordered probit (logit) model are also more complex than their binomial counterparts, since an infinitesimal change in X_j will not only change the probability within the current cell (for instance, if $\kappa_2 < \widehat{y}^* \leq \kappa_3$), but will also make it more likely that the individual crosses the threshold into the adjacent category. Thus if we predict the probabilities of being in each category at a different point in the sample space (for instance, for a family with three rather than two children) we will

find that those probabilities have changed, and the larger family may be more likely to choose the j^{th} response and less likely to choose the $(j-1)^{st}$ response. The average marginal effects may be calculated with `margeff`.

Microeconomic research also makes use of response variables which represent unordered discrete alternatives, or *multinomial* models. We will not discuss those further. An excellent reference is the book by Long and Freese, 2006.

Truncated regression and Tobit models

We turn now to a context where the response variable is not binary nor necessarily integer, but subject to *truncation*. This is a bit trickier, since a truncated or censored response variable may not be obviously so. We must fully understand the context in which the data were generated. Nevertheless, it is quite important that

we identify situations of truncated or censored response variables. Utilizing these variables as the dependent variable in a regression equation without consideration of these qualities will be misleading.

Truncation

In the case of *truncation* the sample is drawn from a subset of the population so that only certain values are included in the sample. We lack observations on both the response variable and explanatory variables. For instance, we might have a sample of individuals who have a high school diploma, some college experience, or one or more college degrees. The sample has been generated by interviewing those who completed high school. This is a *truncated* sample, relative to the population, in that it excludes all individuals who have not completed high school. The characteristics of

those excluded individuals are not likely to be the same as those in our sample. For instance, we might expect that average or median income of dropouts is lower than that of graduates.

The effect of truncating the distribution of a random variable is clear. The expected value or mean of the truncated random variable moves away from the truncation point and the variance is reduced. Descriptive statistics on the level of education in our sample should make that clear: with the minimum years of education set to 12, the mean education level is higher than it would be if high school dropouts were included, and the variance will be smaller. In the subpopulation defined by a truncated sample, we have no information about the characteristics of those who were excluded. For instance, we do not know whether the proportion of minority high school dropouts exceeds the proportion of minorities in the population.

A sample from this truncated population cannot be used to make inferences about the entire population without correction for the fact that those excluded individuals are not randomly selected from the population at large. While it might appear that we could use these truncated data to make inferences about the subpopulation, we cannot even do that. A regression estimated from the subpopulation will yield coefficients that are biased toward zero—or *attenuated*—as well as an estimate of σ_u^2 that is biased downward.attenuation If we are dealing with a truncated Normal distribution, where $y = X\beta + u$ is only observed if it exceeds τ, we may define:

$$\begin{aligned} \alpha_i &= (\tau - X_i\beta)/\sigma_u \\ \lambda(\alpha_i) &= \frac{\phi(\alpha_i)}{(1 - \Phi(\alpha_i))} \end{aligned} \tag{11}$$

where σ_u is the standard error of the untruncated disturbance u, $\phi(\cdot)$ is the Normal density function (PDF) and $\Phi(\cdot)$ is the Normal CDF.

The expression $\lambda(\alpha_i)$ is termed the *inverse Mills ratio*, or IMR.

If a regression is estimated from the truncated sample, we find that

$$[y_i|y_i > \tau, X_i] = X_i\beta + \sigma_u\lambda(\alpha_i) + u_i \qquad (12)$$

These regression estimates suffer from the exclusion of the term $\lambda(\alpha_i)$. This regression is misspecified, and the effect of that misspecification will differ across observations, with a heteroskedastic error term whose variance depends on X_i. To deal with these problems, we include the IMR as an additional regressor. This allows us to use a truncated sample to make consistent inferences about the subpopulation.

If we can justify making the assumption that the regression errors in the *population* are Normally distributed, then we can estimate an equation for a truncated sample with the Stata

command `truncreg`.[§] Under the assumption of normality, inferences for the population may be made from the truncated regression model. The `truncreg` option `ll(#)` is used to indicate that values of the response variable less than or equal to `#` are truncated. We might have a sample of college students with *yearsEduc* truncated from below at 12 years. Upper truncation can be handled by the `ul(#)` option: for instance, we may have a sample of individuals whose income is recorded up to $200,000. Both lower and upper truncation can be specified by combining the options.

The coefficient estimates and marginal effects from `truncreg` may be used to make inferences about the entire population, whereas the results from the misspecified regression model should not be used for any purpose.

[§]More details on the truncated regression model with Normal errors are available in Greene, pp. 756–761.

Censoring

Let us now turn to another commonly encountered issue with the data: *censoring*. Unlike truncation, in which the distribution from which the sample was drawn is a non-randomly selected subpopulation, censoring occurs when a response variable is set to an arbitrary value above or below a certain value: the *censoring point*. In contrast to the truncated case, we have observations on the explanatory variables in this sample. The problem of censoring is that we do not have observations on the response variable for certain individuals. For instance, we may have full demographic information on a set of individuals, but only observe the number of hours worked per week for those who are employed.

As another example of a censored variable, consider that the numeric response to the question "How much did you spend on a new car

last year?" may be zero for many individuals, but that should be considered as the expression of their choice not to buy a car. Such a censored response variable should be considered as being generated by a mixture of distributions: the binary choice to purchase a car or not, and the continuous response of how much to spend conditional on choosing to purchase. Although it would appear that the variable *caroutlay* could be used as the dependent variable in a regression, it should not be employed in that manner, since it is generated by a censored distribution. Wooldridge (2002) argues that this should not be considered an issue of censoring, but rather a *corner solution problem*: the zero outcome is observed with positive probability, and reflects the "corner solution" to the utility maximization problem where certain respondents will choose not to take the action. But as he acknowledges, the literature has already firmly ensconced this problem as that of censoring. (p. 518)

A solution to this problem was first proposed by Tobin (1958) as the *censored regression* model; it became known as "Tobin's probit" or the *tobit* model.[¶] The model can be expressed in terms of a latent variable:

$$\begin{aligned} y_i^* &= X\beta + u \\ y_i &= 0 \text{ if } y_i^* \leq 0 \\ y_i &= y_i^* \text{ if } y_i^* > 0 \end{aligned} \tag{13}$$

As in the prior example, our variable y_i contains either zeros for non-purchasers or a dollar amount for those who chose to buy a car last year. The model combines aspects of the binomial probit for the distinction of $y_i = 0$ versus $y_i > 0$ and the regression model for $[y_i | y_i > 0]$. Of course, we could collapse all positive observations on y_i and treat this as a binomial probit

[¶]The term "censored regression" is now more commonly used for a generalization of the Tobit model in which the censoring values may vary from observation to observation. See the documentation for Stata's `cnreg` command.

(or logit) estimation problem, but that would discard the information on the dollar amounts spent by purchasers. Likewise, we could throw away the $y_i = 0$ observations, but we would then be left with a truncated distribution, with the various problems that creates.[‖] To take account of all of the information in y_i properly, we must estimate the model with the `tobit` estimation method, which employs maximum likelihood to combine the probit and regression components of the log-likelihood function. The log-likelihood of a given observation may be expressed as:

$$\begin{aligned} \ell_i(\beta, \sigma_u) \;=\; & I[y_i = 0] \log\left[1 - \Psi(X_i\beta/\sigma_u)\right] + \\ & I[y_i > 0] \log \psi\left[(y_i - X_i\beta)/\sigma_u\right] \\ & - \log(\sigma_u^2)/2 \end{aligned} \tag{14}$$

[‖]The regression coefficients estimated from the positive y observations will be attenuated relative to the tobit coefficients, with the degree of bias toward zero increasing in the proportion of "limit observations" in the sample.

where $I[\cdot] = 1$ if its argument is nonzero, and zero otherwise. The likelihood function, summing ℓ_i over the sample, may be written as the sum of the probit likelihood for those observations with $y_i = 0$ and the regression likelihood for those observations with $y_i > 0$.

Tobit models may be defined with a threshold other than zero. Censoring from below may be specified at any point on the y scale with the `ll(#)` option for *left censoring*. Similarly, the standard tobit formulation may employ an upper threshold (censoring from above, or *right censoring*) using the `ul(#)` option to specify the upper limit. This form of censoring, also known as *top coding*, will occur with a variable that takes on a value of "$x or more": for instance, the answer to a question about income, where the respondent is asked to indicate whether their income was greater than $200,000 last year in lieu of the exact amount.

Stata's `tobit` also supports the *two-limit tobit* model where observations on y are censored from both left and right by specifying both the `ll(#)` and `ul(#)` options.

Even in the case of a single censoring point, predictions from the tobit model are quite complex, since one may want to calculate the regression-like `xb` with `predict`, but could also compute the predicted probability that $[y|X]$ falls within a particular interval (which may be open-ended on left or right).** This may be specified with the `pr(a,b)` option, where arguments `a, b` specify the limits of the interval; the missing value code `(.)` is taken to mean infinity (of either sign). Another `predict` option, `e(a,b)`, calculates the expectation $y = EX\beta + u$ conditional on $[y|X]$ being in the a, b interval. Last, the `ystar(a,b)` option computes the prediction

**For more information see Greene, pp. 764–773.

from Equation (**??**): a censored prediction, where the threshold is taken into account.

The marginal effects of the tobit model are also quite complex. The estimated coefficients are the marginal effects of a change in X_j on y^* the unobservable latent variable:

$$\frac{\partial E(y^*|X_j)}{\partial X_j} = \beta_j \tag{15}$$

but that is not very useful. If instead we evaluate the effect on the observable y, we find that:

$$\frac{\partial E(y|X_j)}{\partial X_j} = \beta_j \times Pr[a < y_i^* < b] \tag{16}$$

where a, b are defined as above for `predict`. For instance, for left-censoring at zero, $a = 0, b = +\infty$. Since that probability is at most unity (and will be reduced by a larger proportion of censored observations), the marginal effect of X_j is attenuated from the reported

coefficient toward zero. An increase in an explanatory variable with a positive coefficient will imply that a left-censored individual is less likely to be censored. Their predicted probability of a nonzero value will increase. For a non-censored individual, an increase in X_j will imply that $E[y|y > 0]$ will increase. So, for instance, a decrease in the mortgage interest rate will allow more people to be homebuyers (since many borrowers' income will qualify them for a mortgage at lower interest rates), and allow prequalified homebuyers to purchase a more expensive home. The marginal effect captures the combination of those effects. Since the newly-qualified homebuyers will be purchasing the cheapest homes, the effect of the lower interest rate on the average price at which homes are sold will incorporate both effects. We expect that it will increase the average transactions price, but due to attenuation, by a smaller amount than the regression

function component of the model would indicate. The marginal effects may be computed with `mfx` or, for average marginal effects, by Bartus's `margeff`.

Since the tobit model has a probit component, its results are sensitive to the assumption of homoskedasticity. Robust standard errors are not available for Stata's `tobit` command, although bootstrap or jackknife standard errors may be computed with the `vce` option. The tobit model imposes the constraint that the same set of factors X determine both whether an observation is censored (e.g., whether an individual purchased a car) and the value of a non–censored observation (how much a purchaser spent on the car). Furthermore, the marginal effect is constrained to have the same sign in both parts of the model. A generalization of the tobit model, often termed the *Heckit* model (after James Heckman) can relax this constraint, and allow different factors

to enter the two parts of the model. This generalized tobit model can be estimated with Stata's `heckman` command.

Incidental truncation and sample selection models

In the case of truncation, the sample is drawn from a subset of the population. It does not contain observations on the dependent or independent variables for any other subset of the population. For example, a truncated sample might include only individuals with a permanent mailing address, and exclude the homeless. In the case of *incidental truncation*, the sample is representative of the entire population, but the observations on the dependent variable are truncated according to a rule whose errors are correlated with the errors from the equation of interest. We do not observe y because of the outcome of some other variable which generates the *selection indicator*, s_i.

To understand the issue of sample selection, consider a population model in which the relationship between y and a set of explanatory factors X can be written as a linear model with additive error u. That error is assumed to satisfy the zero conditional mean assumption. Now consider that we observe only some of the observations on y—for whatever reason—and that indicator variable s_i equals 1 when we observe both y and X and zero otherwise. If we merely run a regression on the observations

$$y_i = x_i\beta + u_i \tag{17}$$

on the full sample, those observations with missing values of y_i (or any of the elements of X_i) will be dropped from the analysis. We can rewrite this regression as

$$s_i y_i = s_i x_i\beta + s_i u_i \tag{18}$$

The OLS estimator b of Equation (**??**) will yield the same estimates as that of Equation

(**??**). They will be unbiased and consistent if the error term $s_i u_i$ has zero mean and is uncorrelated with each element of x_i. For the population, these conditions can be written

$$\begin{aligned} E(su) &= 0 \\ E[(sx_j)(su)] &= E(sx_j u) = 0 \end{aligned} \qquad (19)$$

because $s^2 = s$. This condition differs from that of a standard regression equation (without selection), where the corresponding zero conditional mean assumption only requires that $E(x_j u) = 0$. In the presence of selection, the error process u must be uncorrelated with sx_j.

Now let us consider the source of the sample selection indicator s_i. If that indicator is purely a function of the explanatory variables in X, then we have the case of *exogenous sample selection*. If the explanatory variables in X are uncorrelated with u, and s_i is a function of Xs, then it too will be uncorrelated with u, as will

the product sx_j. OLS regression estimated on a subset will yield unbiased and consistent estimates. For instance, if gender is one of the explanatory variables, we can estimate separate regressions for men and women without any difficulty. We have selected a subsample based on observable characteristics: e.g., s_i identifies the set of observations for females.

We can also consider selection of a random subsample. If our full sample is a random sample from the population, and we use Stata's `sample` command to draw a 10%, 20% or 50% subsample, estimates from that subsample will be consistent as long as estimates from the full sample are consistent. In this case, s_i is set randomly.

If s_i is set by a rule, such as $s_i = 1$ if $y_i \leq c$, then as we considered in discussing truncation OLS estimates will be biased and inconsistent. We

can rewrite the rule as $s_i = 1$ if $u_i \leq (c - x_i\beta)$, which makes it clear that s_i must be correlated with u_i. As shown above, we must use the truncated regression model to derive consistent estimates.

The case of *incidental truncation* refers to the notion that we will observe y_i based not on its value, but rather on the observed outcome of another variable. For instance, we observe an hourly wage when the individual participates in the labor force. We can imagine estimating a binomial probit or logit model that predicts the individual's probability of participation. In this circumstance, s_i is set to zero or one based on the factors underlying that participation decision:

$$y = X\beta + u \tag{20}$$

$$s = I[Z\gamma + v \geq 0] \tag{21}$$

where we assume that the explanatory factors in X satisfy the zero conditional mean assumption $E[Xu] = 0$. The $I[\cdot]$ function equals 1 if its argument is positive, zero otherwise. We observe y_i if $s_i = 1$. The *selection function* contains a set of explanatory factors Z, which must be a superset of X. For identification of the model, Z contains all X but must also contain additional factors that do not appear in X. The error term in the selection equation, v, is assumed to have a zero conditional mean: $E[Zv] = 0$, which implies that it is also independent of X. We assume that v follows a standard Normal distribution.

The problem of incidental truncation arises when there is a nonzero correlation between u and v. If both of these processes are Normally distributed with zero means, the conditional expectation $E[u|v] = \rho v$ where ρ is the correlation of u and v. From Equation (**??**),

$$E[y|Z, v] = X\beta + \rho v \tag{22}$$

We cannot observe v, but we note that s is related to v by Equation (**??**). Equation (**??**) then becomes

$$E[Y|Z,s] = X\beta + \rho E[v|\gamma,s] \qquad (23)$$

The conditional expectation $E[v|\gamma,s]$ for $s_i = 1$—the case of observability—is merely λ, the *inverse Mills ratio* defined above. Therefore we must augment equation (**??**) with that term:

$$E[y|Z,s=1] = X\beta + \rho\lambda(Z\gamma) \qquad (24)$$

If $\rho \neq 0$, OLS estimates from the incidentally truncated sample—for example, those participating in the labor force—will not consistently estimate β unless the IMR term is included. Conversely, if $\rho = 0$, that OLS regression will yield consistent estimates because it is the correlation of u and v which gives rise to the problem.

The IMR term includes the unknown population parameters γ, which must be estimated

by a binomial probit model

$$Pr(s = 1|Z) = \Phi(Z\gamma) \tag{25}$$

from the entire sample. With estimates of γ, we can compute the IMR term for each observation for which y_i is observed ($s_i = 1$) and estimate the model of Equation (**??**). This two-step procedure, based on the work of Heckman (1976) is often termed the *Heckit model*. Alternatively, a full maximum likelihood procedure can be used to jointly estimate the regression and probit equations.

The Heckman selection model in this context is driven by the notion that some of the Z factors for an individual are different from the factors in X. For instance, in a wage equation, the number of pre-school children in the family is likely to influence whether a woman participates in the labor force but should not be taken into account in the wage determination equation: it appears in Z but not X. Such factors

serve to identify the model. Other factors are likely to appear in both equations. A woman's level of education and years of experience in the labor force are likely to influence her decision to participate as well as the equilibrium wage that she will earn in the labor market.

Stata's `heckman` command will estimate the full maximum likelihood version of the Heckit model with the syntax

heckman *depvar varlist [if] [in]*, select(*varlist2*)

where *varlist* specifies the regressors in X and *varlist2* specifies the list of Z factors expected to determine the selection of an observation as observable. Unlike the `tobit` context, where the `depvar` is recorded at a threshold value for the censored observations (e.g., zero for those who did not purchase a car), the `depvar` should be coded as missing (`.`) for those observations

which are not selected.[††] The model is estimated over the entire sample, and an estimate of the crucial correlation ρ is provided, along with a test of the hypothesis that $\rho = 0$. If that hypothesis is rejected, a regression of the observed `depvar` on `varlist` will produce inconsistent estimates of β.[‡‡]

The `heckman` command is also capable of generating the *two-step* estimator of the selection model (Heckman, 1979) by specifying the `twostep` option. This model is essentially the regression of Equation (**??**) in which the inverse Mills ratio (IMR) has been estimated as the prediction of a binomial probit (Equation

[††] An alternative syntax of `heckman` allows for a second dependent variable: an indicator that signals which observations of *depvar* are observed.

[‡‡] The output produces an estimate of `/athrho`, the hyperbolic arctangent of ρ. That parameter is entered in the log-likelihood function to enforce the constraint that $-1 \leq \rho \leq 1$. The point and interval estimates of ρ are derived from the inverse transformation.

(**??**)) in the first step and used as a regressor in the second step. A significant coefficient of the IMR, denoted `lambda`, indicates that the selection model must be employed to avoid inconsistency. The `twostep` approach, computationally less burdensome than the full maximum likelihood approach used by default in `heckman`, may be preferable in complex selection models.

Bivariate probit and probit with selection

Another example of a limited dependent variable framework in which a correlation of equations' disturbances plays an important role is the *bivariate probit* model. In its simplest form, the model may be written as:

$$\begin{aligned} y_1^* &= X_1\beta_1 + u_1 \\ y_2^* &= X_2\beta_2 + u_2 \\ E[u_1|X_1, X_2] = E[u_2|X_1, X_2] &= 0 \\ var[u_1|X_1, X_2] = var[u_1|X_1, X_2] &= 1 \\ cov[u_1, u_2|X_1, X_2] &= \rho. \end{aligned}$$

The observable counterparts to the two latent variables y_1^*, y_2^* are y_1, y_2. These variables are observed as 1 if their respective latent variables are positive, and zero otherwise.

One formulation of this model, termed the *seemingly unrelated bivariate probit* model in `biprobit`, is similar to the seemingly unrelated regression model. As in the regression context, it may be advantageous to view the two probit equations as a system and estimate them jointly if $\rho \neq 0$, but it will not affect the consistency of individual probit equations' estimates.

However, one common formulation of the bivariate probit model deserves consideration here because it is similar to the selection model described above. Consider a two-stage process in which the second equation is observed conditional on the outcome of the first. For example, some fraction of patients diagnosed with

circulatory problems undergo multiple bypass surgery ($y_1 = 1$). For each of those patients, we record whether they died within one year of the surgery ($y_2 = 1$). The y_2 variable is only available in this context for those patients who are post-operative. We do not have records of mortality among those who chose other forms of treatment. In this context, the reliance of the second equation on the first is a issue of *partial observability*, and if $\rho \neq 0$ it will be necessary to take both equations' factors into account to generate consistent estimates. That correlation of errors may be very likely in that unexpected health problems that caused the physician to recommend bypass surgery may recur and cause the patient's demise.

As another example, consider a bank deciding to extend credit to a small business. Their decision to offer a loan can be viewed as $y_1 = 1$. Conditional on that outcome, the borrower will

or will not default on the loan within the following year, where a default is recorded as $y_2 = 1$. Those potential borrowers who were denied cannot be observed defaulting because they did not receive a loan in the first stage. Again, the disturbances impinging upon the loan offer decision may well be correlated (in this case negatively) with the disturbances that affect the likelihood of default.

Stata can estimate these two types of bivariate probit model with the `biprobit` command. The seemingly unrelated bivariate probit model allows $X_1 \neq X_2$, but the alternate form that we consider here only allows a single *varlist* of factors that enter both equations. In the medical example, this might include the patient's body mass index (a measure of obesity), indicators of alcohol and tobacco use, and age—all of which might both affect the recommended treatment and the one-year survival rate. With

the `partial` option, we specify that the partial observability model of Poirier, 1981 is to be estimated.

Binomial probit with selection

A closely related model to the bivariate probit with partial observability is the *binomial probit with selection* model. This formulation, first presented by Van de Ven and Van Praag has the same basic setup as Equation (**??**) above: the latent variable y_1^* depends on factors X, and the binary outcome $y_1 = 1$ arises when $y_1^* > 0$. However, y_{1j} is only observed when

$$y_{2j} = (X_2\gamma + u_{2j} > 0) \tag{26}$$

that is, when the selection equation generates a value of 1. This could be viewed, in the earlier example, as y_2 indicating whether the patient underwent bypass surgery. We observe the following year's health outcome only

for those patients who had the surgical procedure. As in Equation (**??**), there is a potential correlation (ρ) between the errors of the two equations. If that correlation is nonzero estimates of the y_1 equation will be biased unless the selection is taken into account. In this example, that suggests that focusing only on the patients who underwent surgery (for whom $y_2 = 1$) and studying the factors that contributed to survival will not be appropriate if the selection process is nonrandom. In the medical example, it is surely likely that selection is nonrandom in that those patients with less serious circulatory problems are not as likely to undergo heart surgery.

In the second example, we consider small business borrowers' likelihood of getting a loan, and for successful borrowers, whether they defaulted on the loan. We can only observe a default if they were selected by the bank to

receive a loan ($y_2 = 1$). Conditional on receiving a loan, they did or did not fulfill their obligations as recorded in y_1. If we only focus on loan recipients and whether or not they defaulted we are ignoring the selection issue. Presumably a well-managed bank is not choosing among loan applicants at random. Both deterministic and random factors influencing the extension of credit and borrowers' subsequent performance are likely to be correlated. Unlike the bivariate probit with partial observability, the probit with sample selection explicitly considers $X_1 \neq X_2$. The factors influencing the granting of credit and the borrowers' performance must differ in order to identify the model. Stata's `heckprob` command has a syntax similar to `heckman`, with a *varlist* of the factors in X_1 and a `select(`*varlist2*`)` option specifying the explanatory factors driving the selection outcome.