Section 13 Limited Dependent Variables

What is a limited dependent variable?

- Our standard assumption of an error term that is normally distributed conditional on the regressors implies that the dependent variable can be (with positive probability) any real number, positive or negative.
- **Limited dependent variables** are dependent variables that have limited ranges: usually either discontinuous or range bounded. There are many models of LDVs based on what the limitations are:
  - **0/1** dependent variables (dummies) by probit and logit
  - **Ordered** dependent variables by ordered probit and logit
  - **Categorical** dependent variables (with more than two categories) by multinomial logit
  - **Truncated** dependent variables by Heckman's procedure
  - **Censored** dependent variables by tobit
  - **Count** (integer) dependent variables by Poisson regression
  - **Hazard** (length) dependent variables by hazard models
- Because of the limited ranges of the dependent variable, the standard additive normal error is not tenable for these models. Instead we must model the probability of various discrete outcomes.
- LDV models are usually estimated by maximum likelihood, given the assumed distribution of the conditional probabilities of various outcomes.

Binary dependent variables

- For binary dependent variable: $E(y_i \mid x_i) = \Pr[y_i = 1 \mid x_i]$.
- **Linear probability model**: using OLS with a binary dependent variable
  - We can model, as usual in OLS, $E(y_i \mid x_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$.
  - Show graph of $\Pr[y_i = 1 \mid x_i]$ as linear function of $x\beta$.
  - However, we can't just stick a normal error term onto this function. If we write $y_i = E(y_i \mid x_i) + e_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + u_i$, then since y_i is either zero or one, e_i can only take on the values $1 - (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki})$ and $0 - (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki})$.
  - If $E(y_i \mid x_i) = \Pr[y_i = 1] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} \equiv x_i\beta$, then the error term must have a conditional Bernoulli distribution with $\Pr[u_i = 1 - (x_i\beta)] = x_i\beta$, and $\Pr[u_i = -(x_i\beta)] = 1 - (x_i\beta)$.

- Sums of random variables with the Bernoulli distribution do converge to normal, so the coefficient estimates will still be asymptotically normal.
- However, the immediate problem with this is that the linear function $\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$ will not lie in the range [0, 1] that is required for probabilities for all values of *x*.
- This problem is mirrored by the fact that the predicted values of *y* for some observations is likely to be outside [0, 1], which does not make sense as a prediction of Pr[y_i = 1 | *x*].
  - Show diagram of straight-line prediction of probability and possibility of predictions outside of [0, 1].
- Finally, there is heteroskedasticity in the model as the variance of the Bernoulli error term is $\left[-\left(x_i\beta\right)\right]\left[1-x_i\beta\right]$, which varies with *x*.
  - This is easily accommodated with robust standard errors.
- Bottom line on linear probability model:
  - Simple
  - Probably OK as long as *x* is close to sample means, so that predicted Pr[y_i = 1 | *x*] stays in [0, 1].
  - Not the best model when the dependent variable is binary.

- **Logit (logistic) and probit regression**
  - These are the standard models when the dependent variable is binary. They differ only in the assumed distribution of the error term and are in practice virtually equivalent.
  - Structure of the models: $E\left(y_i \mid x\right) = \Pr[y_i = 1 \mid x] = G\left[\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}\right]$, where *G* is a cumulative probability function that is either
    - Logistic: $G\left(z\right) = \Lambda\left(z\right) = \dfrac{1}{1+e^{-z}} = \dfrac{e^z}{1+e^z}$ for the logit model or
    - Normal: $G\left(z\right) = \Phi\left(z\right) = \int_{-\infty}^{z} \phi\left(\zeta\right) d\zeta = \int_{-\infty}^{z} \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\zeta^2}\, d\zeta$ for probit.
  - Draw graph of cumulative distribution function and show interpretation of *z* and implied probability of *y* = 1.
    - Compare to linear probability model's assumption of linear relationship between *z* and probability.
    - Show how actual data points would look on this graph.
  - Estimation of probit and logit models:
    - These models are always estimated by (nonlinear) maximum likelihood.
    - The (discrete) density function of y_i conditional on x_i is $f\left(y_i \mid x_i, \beta\right) = \left[G\left(x_i\beta\right)\right]^{y_i}\left[1-G\left(x_i\beta\right)\right]^{\left(1-y_i\right)}$, $y_i = 0, 1$, which can be

rewritten less compactly (but more intuitively) as

$$\Pr[y_i = 1 \mid x_i, \beta] = G(x_i\beta),$$
$$\Pr[y_i = 0 \mid x_i, \beta] = 1 - G(x_i\beta).$$

- The likelihood function, assuming that all observations in the sample are IID, is $L(\beta; y, x) = \prod_{i=1}^{N} \left[G(x_i\beta)\right]^{y_i} \left[1 - G(x_i\beta)\right]^{(1-y_i)}$.
- The likelihood maximization is always done in terms of the log-likelihood function:

$$\ln L(\beta; y, x) = \sum_{i=1}^{N} \left[y_i \ln\left[G(x_i\beta)\right] + (1 - y_i) \ln\left[1 - G(x_i\beta)\right]\right].$$

- This function can be evaluated for any choice of β. By searching over the parameter space for the value of β that maximizes this value, we can calculate the logit or probit coefficient estimator as the β that leads to the highest value of the likelihood function.
- Maximum likelihood estimators are known to be consistent, asymptotically normal, and asymptotically efficient under broadly applicable conditions.
- Calculating the standard errors of the coefficient estimators is complicated, but is handled by Stata. The asymptotic covariance matrix of any MLE is the inverse of the "information matrix": $\text{cov}\left(\hat{\beta}\right) = \left[I(\beta)\right]^{-1} = \left[-E\left[\frac{\partial^2 \ln L(\beta; Y, X)}{\partial\beta\partial\beta'}\right]\right]^{-1}$. The information matrix involves the expected values of the matrix of second partial derivatives of the log-likelihood function with respect to the β parameters. It can be approximated for the sample numerically to get an estimated covariance matrix for the parameter vector.
- Hypothesis tests in this, as in any ML model, are easiest as likelihood ratio tests: $2\left[\ln L_u - \ln L_r\right] \sim \chi^2_q$. Stata test command also works and does a Wald test: $t = \frac{\hat{\beta}_j - c}{se\left(\hat{\beta}_j\right)} \sim t_{(N-K)}$ where the *t* distribution is asymptotic.
- Goodness of fit:
  - Fraction predicted correctly:
    - If you take the prediction of y_i to be 1 if $G\left(x_i\hat{\beta}\right) > 0.5$ and zero otherwise, then you get a prediction of zero or one for each y_i. The fraction predicted correctly is just what it sounds like.
  - Pseudo-R^2:

- In the spirit of the usual R^2, this is
  $1-\frac{\ln L\left(\hat{\boldsymbol{\beta}};x,y\right)}{\ln L\left(\boldsymbol{\beta}_Z;x,y\right)}, \boldsymbol{\beta}_Z' \equiv \left(\bar{y},0,0,\ldots,0\right).$
- [Note: This formula is very strange and looks upside down, but it's not. The reason it looks weird is because we are taking the ratio of logs (we usually subtract them). Because (with a discrete dependent variable) the likelihood function is a product of probabilities, it is always less than one. This means that the logs are negative, with the denominator being more negative than the numerator. This, an improvement in fit increases the likelihood in the numerator by decreasing its absolute value, making the ratio smaller and the R^2 value closer to one.]
- This ratio is the likelihood function with the best parameter estimate divided by the likelihood function if we just predict each y by the sample proportion of y values that are one.

- Interpretation of β in probit and logit regressions:
  - In the usual OLS model, $\beta_j = \frac{\partial E\left[y_i \mid x_i\right]}{\partial x_j}$, which is what we are interested in knowing.
  - In probit or logit model, $\beta_j = \frac{\partial z}{\partial x_j}$ is not in useful units because z has no direct interpretation.
    - Use graph to demonstrate β as horizontal movement
  - What we're interested in knowing (for a continuous regressor x) is $\frac{\partial \Pr\left[y=1\right]}{\partial x_j} = \frac{d\Pr\left[y=1\right]}{dz}\frac{\partial z}{\partial x_j} = G'\left(z\right)\beta_j = g\left(z\right)\beta_j,$ where g is the probability density function associated with the cumulative distribution function G.
    - Graphical interpretation: β measures horizontal movement due to unit change in x; $G'(z)$ measures effect of unit horizontal movement on probability of $y = 1$.
    - They have the same sign, so tests of $\beta = 0$ are equivalent to tests of $\frac{\partial \Pr\left[y=1\right]}{\partial x_j} = 0$.

- For logit, $g(z)=\frac{e^z}{\left(1+e^z\right)^2}=\Lambda(z)\left[1-\Lambda(z)\right]=G(z)\left[1-G(z)\right]$.
    - The results of the logit model is often expressed in terms of odds ratios:

      $$\Pr\left[y_i=1\mid x_i\right]=\Lambda\left(x_i\beta\right)=\frac{e^{x_i\beta}}{1+e^{x_i\beta}}$$
      $$\left(1+e^{x_i\beta}\right)\Lambda\left(x_i\beta\right)=e^{x_i\beta}$$
      $$\Lambda\left(x_i\beta\right)=e^{x_i\beta}-\Lambda\left(x_i\beta\right)e^{x_i\beta}=\left(1-\Lambda\left(x_i\beta\right)\right)e^{x_i\beta}$$
      $$e^{x_i\beta}=\frac{\Lambda\left(x_i\beta\right)}{1-\Lambda\left(x_i\beta\right)}=\frac{\Pr\left[y_i=1\mid x_i\right]}{\Pr\left[y_i=0\mid x_i\right]}=\text{"odds ratio"}$$

    - β_j is the effect of x_j on the "log odds ratio"
- For probit, $g(z)=\phi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$.
- Because they are density functions, $g(z) \geq 0$ for all z, so the "partial effects" $\frac{\partial \Pr\left[y=1\right]}{\partial x_j}$ have the same sign as β_j.
- For dummy regressors, we are interested in $\Pr\left[y=1\mid x_j=1\right]-\Pr\left[y=1\mid x_j=0\right]$.
- In Stata: **probit** reports the coefficients and **dprobit** (which is still supported but no longer official) reports the partial effects. The regression is identical for each.
    - Note that the partial effects depend on z and thus on x. You can specify the values at which to evaluate the partial effects in dprobit with the default being at the means.
    - Partial effects of dummy variables are reported (by default) as difference in probabilities above, with other variables at means.
- In Stata: **logit** reports coefficients and **logistic** reports the "odds-ratio" $e^{\hat{\beta}_j}$. (This is really the proportional effect of the variable on the odds ratio, not the odds ratio itself.)
    - If x_{ji} increases by one, $e^{x_i\beta}$ increases to $e^{x_i\beta+\beta_j}=e^{x_i\beta}e^{\beta_j}$, so $e^{\hat{\beta}_j}$ measures the estimated proportion by which a one-unit change in x_{ji} changes the odds ratio.
    - Interpretation can be tricky:
        - All e^{β} values are positive.
        - A zero effect means that $\beta = 0$ and $e^{\beta} = 1$.

        - A variable that reduces the odds ratio has a $\beta < 1$.
        - A variable that increases the odds ratio has a $\beta > 1$.
        - Example: If $e^{\beta j} = 2$ and the initial probability p of $y = 1$ for this observation is .2, (so the initial odds-ratio $p/(1 - p)$ is (.2) / (.8) = 0.25), then a one-unit increase in x_j multiplies the odds ratio by $e^{\beta j} = 2$, making it 0.5, which means that the probability of $y = 1$ has increased from 0.2 to 0.333 = 0.5/(1 + 0.5).
        - If we do the same example for an observation with an initial $p = 0.5$, then the initial odds ratio is 1, the unit increase in x_j multiplies it by 2, making the new odds ratio 2, and thus the probability has increased from 0.5 to 2/(1 + 2) = 0.667.
    - Post-estimation commands after probit and logit are very useful to get predictions (predict gives probability $y = 1$ by default, not xb, and margins can be used to get predictions at specific values of x)
        - See probit/logistic postestimation help file for details
        - Important for project
- Reliability of probit and logit estimators
    - Omitted-variable bias
        - This is more of a problem in probit and logit models because a coefficient of an included variable can be inconsistent even when it is uncorrelated with the omitted variable
    - Heteroskedasticity
        - Again, more of a problem in probit and logit because the standard MLE based on an assumption of homoskedasticity is inconsistent.
        - You can use the White robust estimator for the covariance ("robust standard errors"), but you are calculating a valid standard error for a coefficient that does not converge to the true parameter value, so it is of less utility than in OLS,
    - How to deal with these issues?
        - Be careful about omitted variables
        - Try to specify the model in a scaled way that makes variance as constant as possible

Discrete-choice dependent variables

- What if there are more than two choices?
  - Instead of the dependent variable being whether someone attends Reed or not, it could be whether someone attends Reed ($y = 3$), attends another private college (2), attends a public college (1), or doesn't attend college at all ($y = 0$).
  - This would be four choices rather than two.
  - This is an "unordered-choice model:" There is no obvious order to these choices. If we define y as above and say that changes in characteristics of the individual (not of the choices) x (say, higher SAT) that make y more likely to move from 0 to 1, we can't also be confident that these changes in x are more likely to make y move from 1 to 2 or from 2 to 3.
- **Multinomial (polytomous) logit model** (Greene 6/e, section 23.11)
  - $\Pr[y_i = j \mid x_i] = \dfrac{e^{x_i \beta_j}}{\sum_{m=1}^{M} e^{x_i \beta_m}}$, where there are M distinct choices. This model has $M(k + 1)$ β parameters, but only $(M - 1)(k + 1)$ of them are unique because the sum of the probabilities must be one. (If an increase in family income raises the probabilities that you will choose $y = 2$, 3, and 4, it must lower the probability of choosing $y = 1$ by an equivalent amount. Thus, $\beta_{i,1}$ can be determined from $\beta_{i,2}$, $\beta_{i,3}$, and $\beta_{i,4}$. Where the second subscript refers to the choice and the first to the independent variable.) We usually normalize by setting the vector $\beta_1 = 0$, which makes the numerator of the probability fraction 1 for choice 1.
  - In the multinomial logit model, $\ln\left(\dfrac{\Pr[y_i = j \mid x_i]}{\Pr[y_i = 1 \mid x_i]}\right) = x_i \beta_j$. The coefficients thus can be interpreted as the effect of x on the log odds ratio.
  - **Independence of irrelevant alternatives** assumption is implicit in multinomial logit model
    - It shouldn't matter for the coefficients of the attending-Reed equation whether one adds attending Lewis & Clark as a special case of attending a private college (making 5 alternatives) or not
    - This assumption may not be reasonable in some cases, making the model inappropriate.
  - Multinomial logit models can be estimated by maximum likelihood methods. In Stata, use **mlogit**.
- Related models:
  - **Conditional logit model**: The x variables relate to properties of the choices instead of or in addition to the individual. (Not clogit in Stata; that's something else.)

- **Nested logit model**: Decisions are nested. For example, decision whether to attend college, then if attending whether to attend Reed, another private college, or a public. In Stata, use **nlogit**.
- **Multinomial probit**: Same thing with normal rather than logistic function. ***Very*** time-consuming to estimate, so it's not used often.

Ordered dependent variables

- Many variables are ***ordinal*** in nature: we know that 4 is bigger than 3 and that 3 is bigger than 2, but we don't know the 4 is the same amount bigger than 3 as 3 is bigger than 2.
  - Examples would include bond ratings, opinion-survey responses, academic actions, and perhaps grades and SAT scores.
- We can think of the ordinal dependent variable y as representing levels of the outcomes of some underlying latent variable y^*.
  - We assume that $y_i^* = x_i\beta + e_i$, and that we observe the ordinal choice y_i:

$$y_i = \begin{cases} 1 & \text{if } y_i^* \le \mu_1, \\ 2 & \text{if } \mu_1 < y_i^* \le \mu_2, \\ 3 & \text{if } \mu_2 < y_i^* \le \mu_3, \\ \vdots & \\ M & \text{if } \mu_{M-1} < y_i^*. \end{cases}$$

  - If the error term is normal, then we can use **ordered probit** to estimate the β vector and the thresholds μ corresponding to the different levels of the variables.
  - **Ordered logit** is used when the error term follows the logistic distribution.
  - Ordered probit/logit involves estimating the β vector *and* the threshold values μ_1 through μ_{M-1} by maximum likelihood.
  - If we normalize the model to give the error term unit variance (divide y and x by the standard deviation of error), then we have

$$\begin{aligned} &\Pr[y_i = 1 \mid x_i] = \Phi(\mu_1 - x_i\beta) \\ &\Pr[y_i = 2 \mid x_i] = \Phi(\mu_2 - x_i\beta) - \Phi(\mu_1 - x_i\beta) \\ &\Pr[y_i = 3 \mid x_i] = \Phi(\mu_3 - x_i\beta) - \Phi(\mu_2 - x_i\beta) \\ &\vdots \\ &\Pr[y_i = M \mid x_i] = 1 - \Phi(\mu_{M-1} - x_i\beta). \end{aligned}$$

  - The likelihood function is $L(\beta, \mu; y, x) = \prod_{i=1}^{n} \left[\sum_{m=1}^{M} I(y_i = m) \Pr[y_i = m \mid x_i; \beta, \mu] \right]$, where $I(y_i = m)$ is an indicator function that is one if the condition is true and the probability is given by the formulas above. The likelihood function is maximized by searching over alternative values of β and μ to find those that maximize.

- Show Greene (6/e) Figure 23.4 from p. 833.

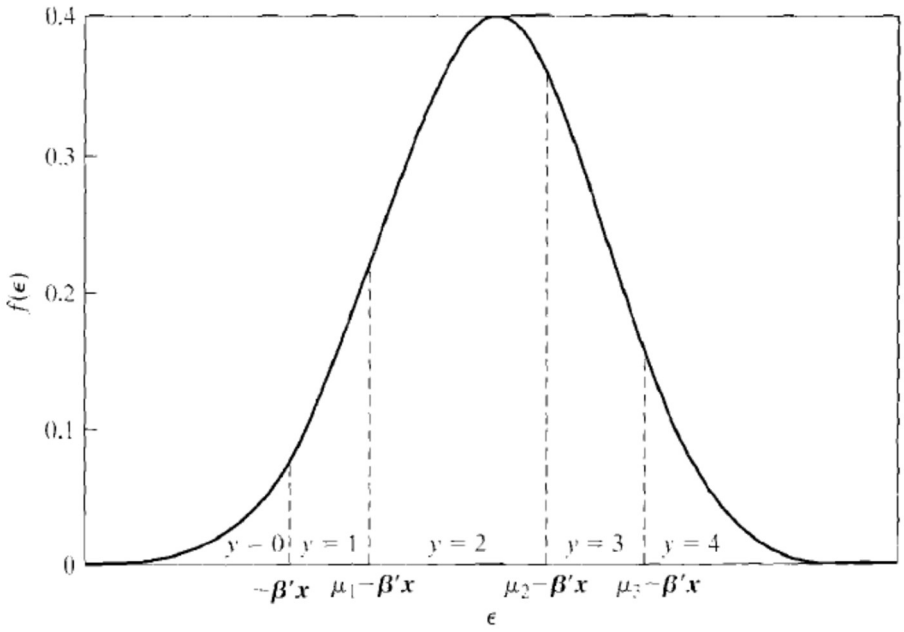

FIGURE 23.4 Probabilities in the Ordered Probit Model.

- Partial effects: what does β_j mean?
  - As in the standard probit and logit models, β_j is the derivative of the unobserved y^* with respect to x_j.
  - We can derive marginal effects of x_j on the probabilities of y being each value as

$$\frac{\partial \Pr[y_i = 1 \mid x_i]}{\partial x_{j,i}} = -\phi(\mu_1 - x_i\beta)\beta_j,$$

$$\frac{\partial \Pr[y_i = 2 \mid x_i]}{\partial x_{j,i}} = [\phi(\mu_1 - x_i\beta) - \phi(\mu_2 - x_i\beta)]\beta_j,$$

$$\vdots$$

$$\frac{\partial \Pr[y_i = M \mid x_i]}{\partial x_{j,i}} = \phi(\mu_{M-1} - x_i\beta)\beta_j.$$

  - In Stata, you can use the **margins** command after oprobit to get marginal effects.
    - Margins, dydx(*) predict (outcome(#1)) will calculate the effect of a one-unit change in each regressor (*) on the probability of outcome #1. Leave off the predict () to get all.
  - **Predict** gives predicted probabilities *of each lev*el for each observation.
    - Predict probs* creates new variables starting with probs with probabilities of each outcome. Can also restrict to get only some with outcomes() option.

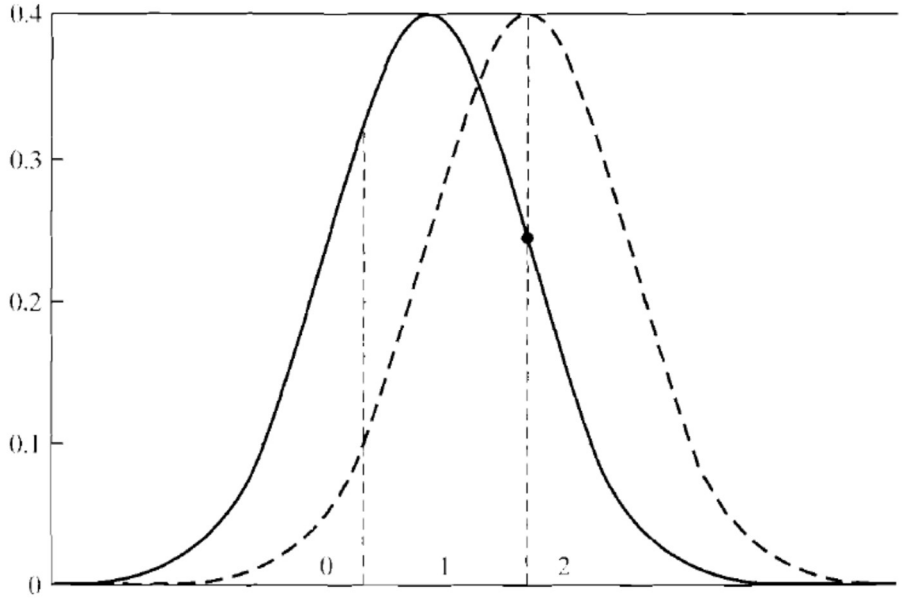

FIGURE 23.5 Effects of Change in *x* on Predicted Probabilities.

Count dependent variables

- Count dependent variables can only take on non-negative integer values.
  - Normal distribution is not a plausible choice.
  - Poisson distribution is often used for count models:
    - $\Pr[y_i = m \mid x_i] = \frac{e^{-\lambda_i}\lambda_i^m}{m!}$.
    - Poisson distribution has mean and variance both equal to λ_i, so $E[y_i \mid x_i] = \lambda_i$
    - In **Poisson regression** we model $\lambda_i = e^{X_i\beta}$.
    - The log-likelihood function is $\ln L = \sum_{i=1}^{n}\left[-e^{x_i\beta} + y_i x_i \beta - \ln(y_i!)\right]$ and we estimate as usual by maximizing this function.
    - Interpretation of coefficients: $\frac{\partial E[y_i \mid x_i]}{\partial x_j} = \lambda_i \beta_j = e^{x_i\beta}\beta_j$.
  - Poisson regression is implemented in Stata by the `poisson` command.
- Limitation of Poisson regression
  - The fact that the conditional mean and conditional variance of the Poisson distribution are the same is restrictive.
  - If it doesn't fit your data well, then a more general model might be appropriate.

- The most common alternative is the **negative binomial regression model**, which is implemented as nbreg in Stata.

Tobit, censored, and truncated regression models

- These three models are *very* easy to confuse!
  - All involve situations where we have no observations from some region of the (usually normal) distribution.
  - Example: Sometimes we have corner solutions in economic decisions: many people choose to consume zero of many commodities. (This is the tobit model.)
  - Example: Sometimes surveys are "top-coded" with the maximum response being 50,001 or something like that. (This is censored regression.)
  - Example: If the dependent variable is duration until death of patients after treatment, some patients will not yet have died. (Another censored regression.)
  - Example: Some events sell out, meaning that the observed values of demand are censored at the site capacity. (Yet another censored regression.)
  - Example: Sample consists only of people with values of y below a limit c. (This is truncated regression model.)
- **Tobit estimator for corner solutions**
  - Suppose that some finite fraction α of observations choose zero, but those choosing positive quantities follow the remainder of the normal distribution (lopping off the left-end α of probability).
  - Why can't we just use OLS?
    - Like linear probability model, we ignore restrictions on the distribution of u and we predict values < 0.
  - Why can't we just use the observations with $y_i > 0$?
    - This would imply selection on u because we'd be more likely to eliminate observations with $u < 0$.
  - Why can't we use ln(y)?
    - Observations with $y = 0$ would have $\ln(y) = -\infty$.
  - We can model this with a latent variable $y_i^* = \mathbf{x_i}\boldsymbol{\beta} + u_i$ as a latent underlying variable with a normal distribution and $y_i = \begin{cases} y_i^*, \text{ if } y_i^* \geq 0, \\ 0, \text{ otherwise} \end{cases}$ as the observed outcome.
    - This variable has a **censored distribution** with finite probability of a zero outcome but otherwise normally distributed over the positive values.

- The conditional density of y is

  $f(y_i \mid x_i) = \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x_i}\boldsymbol{\beta}}{\sigma}\right)$ for $y > 0$, and

  $\Pr[y_i = 0 \mid x_i] = 1 - \Phi(\mathbf{x_i}\boldsymbol{\beta} / \sigma)$.

- This density is the basis for the **tobit** estimator of the vector β.
  - Tobit maximizes (over β, σ) the log-likelihood function:

    $$\ln L(\boldsymbol{\beta}, \sigma; y, x) = \sum_{i:y_i=0} \ln\left[1 - \Phi\left(\frac{\mathbf{x_i}\boldsymbol{\beta}}{\sigma}\right)\right] + \sum_{i:y_i>0} \ln\left[\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x_i}\boldsymbol{\beta}}{\sigma}\right)\right]$$

- The limit value (zero here, but it could be some value c) must be specified.
- Can also have distributions that are censored above, or both above and below (perhaps the share of merlot in total wine consumption), where some people choose zero and some choose one).

o Interpreting tobit coefficients

- There are two expected values of interest in the tobit model:
  - **"Conditional (on $y_i > 0$) expectations"**

    $E[(y_i \mid x_i) \mid y_i > 0] = E[y_i \mid y_i > 0, x_i]$

    - Draw graph showing censorship at 0 and density function over $y_i > 0$ = $f(y_i \mid y_i > 0) = \frac{\phi(y_i)}{1 - \Phi(y_i)}$.
    - Remarkable and useful property of standard normal distribution: $E[z \mid z > c] = \frac{\phi(c)}{1 - \Phi(c)}$.
    - $y_i > 0$ iff $u_i > -x_i\beta$ and u_i is (by assumption) distributed normally with mean 0 and variance σ^2. Thus u_i/σ is standard normal and $E\left[\frac{u_i}{\sigma} \middle| \frac{u_i}{\sigma} > c\right] = \frac{\phi(c)}{1 - \Phi(c)}$.
    - Conditional on x, $E(x\beta) = x\beta$, so

      $$\begin{aligned} E(y_i \mid y_i > 0, x_i) &= x_i\beta + E(u_i \mid u_i > -x_i\beta, x_i) \\ &= x_i\beta + \sigma E\left(\frac{u_i}{\sigma} \middle| \frac{u_i}{\sigma} > -\frac{x_i\beta}{\sigma}, x_i\right) \\ &= x_i\beta + \sigma\frac{\phi\left(\frac{x_i\beta}{\sigma}\right)}{\Phi\left(\frac{x_i\beta}{\sigma}\right)}, \end{aligned}$$

      where we use the properties that $\phi(-z) = \phi(z)$ and $1 - \Phi(-z) = \Phi(z)$.

- We define the **inverse Mills ratio** as $\lambda(c) = \frac{\phi(c)}{\Phi(c)}$.
- Then $E(y_i \mid y_i > 0, x_i) = x_i\beta + \sigma\lambda\left(\frac{x_i\beta}{\sigma}\right)$ is the "conditional expectation" of *y* given that *y* is positive.

- **"Unconditional (on *y* > 0) expectation"** (which is still conditional on *x*) $E[y_i \mid x_i]$:

$$\begin{aligned} E[y_i \mid x_i] &= 0 \cdot \Pr[y_i = 0 \mid x_i] + E[y_i \mid y_i > 0, x_i] \cdot \Pr[y_i > 0 \mid x_i] \\ &= E[y_i \mid y_i > 0, x_i] \cdot \Pr[y_i > 0 \mid x_i] \\ &= \left[x_i\beta + \sigma\lambda\left(\frac{x_i\beta}{\sigma}\right)\right] \cdot \Pr\left[\frac{u_i}{\sigma} > \frac{-x_i\beta}{\sigma}\right] \\ &= \left[x_i\beta + \sigma\frac{\phi\left(\frac{x_i\beta}{\sigma}\right)}{\Phi\left(\frac{x_i\beta}{\sigma}\right)}\right]\Phi\left(\frac{x_i\beta}{\sigma}\right) \\ &= \Phi\left(\frac{x_i\beta}{\sigma}\right)x_i\beta + \sigma\phi\left(\frac{x_i\beta}{\sigma}\right). \end{aligned}$$

- Interpretation of β_j?
  - In the usual OLS model, $\beta_j = \frac{\partial E[y_i \mid x_i]}{\partial x_j}$.
  - Here,

$$\begin{aligned} \frac{\partial E[y_i \mid y_i > 0, x_i]}{\partial x_j} &= \beta_j + \sigma\frac{\partial\lambda(x_i\beta/\sigma)}{\partial(x_i\beta/\sigma)}\frac{\partial(x_i\beta/\sigma)}{x_j} \\ &= \beta_j + \beta_j\frac{\partial\lambda(x_i\beta/\sigma)}{\partial(x_i\beta/\sigma)}. \end{aligned}$$

    - By quotient rule, $\frac{\partial\lambda(c)}{\partial c} = \frac{\Phi(c)\phi'(c) - \phi(c)\Phi'(c)}{[\Phi(c)]^2}$.
    - But $\Phi'(c) = \phi(c)$ by definition and using the definition of the normal density function, $\phi'(c) = -c\,\phi(c)$, so

$$\begin{aligned} \frac{\partial\lambda(c)}{\partial c} &= \frac{-c\Phi(c)\phi(c) - [\phi(c)]^2}{[\Phi(c)]^2} \\ &= -c\lambda(c) - [\lambda(c)]^2 \\ &= -\lambda(c)[c + \lambda(c)]. \end{aligned}$$

- Therefore,

$$\frac{\partial E[y_i \mid y_i > 0, x_i]}{\partial x_j} = \beta_j \left\{ 1 - \lambda\left(\frac{x_i\beta}{\sigma}\right)\left[\frac{x_i\beta}{\sigma} + \lambda\left(\frac{x_i\beta}{\sigma}\right)\right]\right\}.$$

  - The expression in braces is between 0 and 1, so the effect of x_j on the conditional expectation of y is of the same sign as β_j but smaller magnitude.
  - Testing $\beta_j = 0$ is a valid test for the partial effect being zero.

- Given that $E[y_i \mid x_i] = E[y_i \mid y_i > 0, x_i] \cdot \Pr[y_i > 0 \mid x_i]$,

$$\frac{\partial E[y_i \mid x_i]}{\partial x_j} = \frac{\partial E[y_i \mid y_i > 0, x_i]}{\partial x_j} \cdot \Pr[y_i > 0 \mid x_i] + \frac{\partial \Pr[y_i > 0 \mid x_i]}{\partial x_j} \cdot E[y_i \mid y_i > 0, x_i].$$

  - $\dfrac{\partial \Pr[y_i > 0 \mid x_i]}{\partial x_j} = \dfrac{\partial \Phi\left(\frac{x_i\beta}{\sigma}\right)}{\partial \xi_j} = \dfrac{\beta_j}{\sigma}\phi\left(\dfrac{x_i\beta}{\sigma}\right).$
  - $\dfrac{\partial E[y_i \mid y_i > 0, x_i]}{\partial x_j} = \beta_j \left\{ 1 - \lambda\left(\dfrac{x_i\beta}{\sigma}\right)\left[\dfrac{x_i\beta}{\sigma} + \lambda\left(\dfrac{x_i\beta}{\sigma}\right)\right]\right\}.$
  - So (with all Φ, ϕ, and λ functions evaluated at $x_i\beta/\sigma$)

$$\begin{aligned}\frac{\partial E[y_i \mid x_i]}{\partial x_j} &= \beta_j \Phi(\,) \left\{ 1 - \lambda(\,)\left[\frac{x_i\beta}{\sigma} + \lambda(\,)\right]\right\} + \frac{\beta_j}{\sigma}\phi(\,)\left[x_i\beta + \sigma\lambda(\,)\right] \\ &= \beta_j \left\{ \Phi(\,) - \phi(\,)\left[\frac{x_i\beta}{\sigma} + \lambda(\,)\right] + \phi(\,)\left[\frac{x_i\beta}{\sigma} + \lambda(\,)\right]\right\} \\ &= \beta_j \Phi\left(\frac{x_i\beta}{\sigma}\right).\end{aligned}$$

- Doing tobit estimation in Stata
  - `tobit depvar indvars , ll(0)` does tobit with zero lower censorship
    - `ul( )` option specifies possible upper point of censorship
  - After estimation, can use the predict command to generate some useful series:
    - `predict , pr(0, .)` gives the predicted probability that each observation is not censored, $\Pr[y_i > 0 \mid x_i] = \Phi\left(\dfrac{\mathbf{x_i\beta}}{\sigma}\right).$
    - `predict , e(0, .)` gives the predicted value of each observation conditional on not being censored, $E(y_i \mid y_i > 0, x_i)$

- predict , ystar(0, .) gives the unconditional predicted value of each observation $E(y_i \mid x_i)$.

  - margins is used after tobit to get partial effects
    - The options correspond to those of predict:
      - Margins dydx(*) , pr(0,) gives effect of unit change in all variables (*) on probability that $y > 0$: $\frac{\partial \Pr[y > 0]}{\partial x}$
      - Margins dydx(*) , e(0,) gives effect of unit change in all variables on expected value condition on not being censored $\frac{\partial E[y \mid y > 0, x]}{\partial x}$
      - Margins dydx(*) , ystar(0,) gives effect of unit change in all variables on unconditional expected value $\frac{\partial E(y^*)}{\partial x} = \frac{\partial E[\max(y, 0)]}{\partial x}$

- **Censored regression** (top-coding problems, unexpired duration models, etc.)
  - We have data on the *x* variables for all observations, but have no observations on *y* for those at one end (or both ends) of the distribution.
    - If $y > c$, then we observe *c*.
  - Let $y_i = \mathbf{x_i}\boldsymbol{\beta} + u_i$, where u_i is homoskedastic normal. We don't observe *y* but instead observe $w_i = \min(y_i, c_i)$, where c_i is a known constant that can vary with *i*.
  - Note difference from tobit model: In tobit a finite fraction of people chose the limit value. Here they chose something continuous outside the limit but we simply do not observe it.
    - This means that we don't have to model the censorship as part of the choice, rather only account for it in the estimation based on our flawed data.
  - For uncensored observations, we have the usual distribution of *y*:

    $$f(w_i \mid x_i) = f(y_i \mid x_i) = \frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right).$$

  - For censored observations,

    $$\begin{aligned}\Pr[w_i = c_i \mid x_i] &= \Pr[y_i \geq c_i \mid x_i] \\ &= \Pr[e_i \geq c_i - x_i\beta \mid x_i] \\ &= 1 - \Phi\left(\frac{c_i - x_i\beta}{\sigma}\right).\end{aligned}$$

  - So likelihood function is same as the tobit model, as is estimation.

- However, in the censored regression case we don't need to worry about people *choosing* the limit value, we only worry about *observing* it. Thus, β_j is the effect of x_j on y, period. We don't need to hassle with the marginal effects calculations as in the tobit model. Consequently, we can use the Stata tobit command and just neglect the margins command afterward.

- **Truncated regression models**
  - Truncated regression differs from censored regression in that neither y nor x is observed for observations beyond the limit point. Thus, we cannot use these data points at all, making the tobit estimator impossible to calculate.
    - This is a sample problem again, but truncation of the sample (all variables) is more severe than censorship of a single variable because we have less (no) information about the missing observations.
    - In the censored model, we can use the x values of the censored observations to determine what kinds of observations will be in the censored range. In the truncated model, we don't have that information.
  - Truncated regression model
    - $y_i = \beta_0 + x_i\beta + \upsilon_i,$
      $u_i \mid x \sim N(0, \sigma^2).$
    - IID assumption is violated:
      - We observe (x_i, y_i) only if $y_i \geq c_i$, where the truncation threshold can vary with i and can depend on x_i.
    - The conditional density function of y_i given that it is in the sample ($> c_i$) is $g(y_i \mid x_i, c_i) = \dfrac{f(y_i \mid x_i, \beta, \sigma_e^2)}{1 - F(c_i \mid x_i, \beta, \sigma_e^2)} = \dfrac{\phi\left(\dfrac{y_i - x_i\beta}{\sigma_e}\right)}{1 - \Phi\left(\dfrac{c_i - x_i\beta}{\sigma_e}\right)}, \quad y_i \geq c_i.$
    - The Φ function in the denominator is the probability that observation i is not censored, given x_i and c_i. We divide by this to redistribute the truncated amount of probability over the remaining density function.
    - The log-likelihood function is just the log of this density summed over all the observations in the sample.
    - OLS in this case would give slope estimates that are biased toward zero.

- **Incidental truncation and sample selection**
  - Sample selection does not bias OLS estimators unless the selection criterion is related to u. So selection based exclusively on x or on something outside the model that is uncorrelated with u does not present a problem.
  - "Incidental truncation" occurs when we observe y for only a subset of the population that depends not on y but on another variable, but the other variable is correlated with u.

- The primary (only?) example in the literature is y = ln(wage offer), which is observed only for people who work.
- But people who have unusually low wage offers (given their other characteristics) are less likely to work and therefore more likely to be truncated, so the variable determining truncation (work status) is correlated with the error term of the wage equation.

$y_i = \mathbf{x_i}\boldsymbol{\beta} + u_i$

- $s_i = \begin{cases} 1, \text{ if } \mathbf{z_i}\boldsymbol{\gamma} + v_i > 0, \\ 0, \text{ otherwise.} \end{cases}$
  - s_i is a sample indicator that is one for observations for which we observe y and zero otherwise.
  - We assume that $E(u_i \mid \mathbf{x_i}, \mathbf{z_i}) = 0$ and x_i is a strict subset of z_i.
  - We also assume that v is a standard normal that is independent of z, but that it may be correlated with u.
  - $E(y_i \mid \mathbf{z_i}, v_i) = \mathbf{x_i}\boldsymbol{\beta} + E(u_i \mid \mathbf{z_i}, v_i) = \mathbf{x_i}\boldsymbol{\beta} + E(u_i \mid v_i)$.
    - Let $E(u_i \mid v_i) = \rho v_i$ with ρ being a parameter of their joint normal distribution (related to the correlation).
    - This means that
      $E(y_i \mid \mathbf{z_i}, v_i) = \mathbf{x_i}\boldsymbol{\beta} + \rho v_i$,
      $E(y_i \mid \mathbf{z_i}, s_i) = \mathbf{x_i}\boldsymbol{\beta} + \rho E(v_i \mid \mathbf{z_i}, s_i)$.
  - Since our sample is the set of observations for which $s = 1$, we need the expected value of y conditional on $s = 1$, and by logic similar to that used in the tobit model, $E(v_i \mid z_i, s_i = 1) = \lambda(z_i\gamma)$, where λ is the inverse Mills ratio ϕ/Φ.
  - Thus, $E(y_i \mid z_i, s_i = 1) = \mathbf{x_i}\boldsymbol{\beta} + \rho\lambda(\mathbf{z_i}\boldsymbol{\gamma})$.
- We can't observe the λ term unless we know γ. The **Heckit** estimator is a two-step estimation procedure for estimating first γ, then β.
  - The selection variable s follows a probit model:
    $\Pr[s_i = 1] = \Pr[\mathbf{z_i}\boldsymbol{\gamma} + v_i \geq 0] = \Pr[v_i \geq -\mathbf{z_i}\boldsymbol{\gamma}]$
    $= \Pr[v_i \leq \mathbf{v_i}\boldsymbol{\gamma}] = \Phi(\mathbf{z_i}\boldsymbol{\gamma})$.
  - Thus, we estimate the **sample-selection equation** as a probit of s on $\mathbf{z}$, using all of the observations (because we don't need to observe y for this equation and we observe $\mathbf{z}$ for all observations).
  - We then compute the estimated inverse Mills ratio for each observation as $\hat{\lambda}_i = \lambda(\mathbf{z_i}\hat{\boldsymbol{\gamma}}) = \dfrac{\phi(\mathbf{z_i}\hat{\boldsymbol{\gamma}})}{\Phi(\mathbf{z_i}\hat{\boldsymbol{\gamma}})}$.

- We can then estimate β by running OLS on $y_i = \mathbf{x_i}\boldsymbol{\beta} + \rho\hat{\lambda}_i + u_i$ using only the observations for which *y* is observed. The inclusion of the estimated inverse Mills ratio on the right-hand side corrects the bias due to sample selection and makes the β estimates consistent and approximately normal.
- Testing ρ = 0 with a standard *t* test is a valid test for whether there was sample selection.

- Note that the regular OLS standard errors are incorrect because they assume that λ is exactly known. There will be error in estimating λ by $\hat{\lambda}$, so this error needs to be taken into account in calculating the reliability of $\hat{\beta}$.
  - Stata command `heckman` computes heckit estimator either by full maximum likelihood or by the two-step estimation method. This will correct the standard errors.
- In order to apply this model reliably, there must be at least one variable that determines sample selection that does not affect *y*.
  - In the wage equation, it is usually assumed that family variables such as number of children would not affect the wage offer but would affect a person's choice of whether or not to accept it and work.