Limited Dependent Variables

1. What if the left-hand side variable is not a continuous thing spread from minus infinity to plus infinity? That is, given a model $Y = f(X, \beta, \varepsilon)$, where
   a. *Y* is bounded below at zero, such as wages or height;
   b. Or, *Y* is bounded above at 100,000 such as top-coded income in BLS data;
   c. Or, *Y* is discrete, taking on just the values 0 or 1, such as in yes/no responses to survey questions, or bankruptcy indicators for firms;
   d. Or, *Y* is discrete, but ordinal and many-valued, such as data that have been grouped into ranges (often the case for income data), or responses to survey questions about the depth of agreement (strongly agree, somewhat agree, don't agree at all);
   e. Or, *Y* is discrete and many-valued, but not ordinal, such as transit mode choices (did you take the bus, drive a car, bicycle or walk?).
2. Let us begin with the simplest case, *Y* is a zero-one binary variable, reflecting the answer to a yes/no question, coded 1 for 'yes', 0 for 'no'.
   a. Consider an OLS regression of *Y* on *X*. If greater values of *X* are associated with higher probabilities of 'yes', then the OLS regression coefficient on *X* will reflect this as a positive coefficient. Indeed, the coefficient will have a mean value equal to the marginal impact of *X* on the probability that *Y* is a 'yes'. Given that, we call this the *linear probability model*. OLS regressions for a linear probability model deliver consistent estimates of coefficients for an underlying model where
      i. $P[Y = 1] = X\beta + \varepsilon$. To see this, think about a model where *P[Y=1]=0.50* for *X=1* and *P[Y=1]=0.60* for *X=2*. In this case, 50% of cases will have *Y=1* and 50% will have *Y=0* at *X=1*, but 60% will have *Y=1* and 40% will have *Y=0* at *X=2*. The regression wants to find the conditional expectation of *Y* given *X*, and given the 0/1 coding of *Y*, this conditional expectation is the probability we are seeking.
   b. This linear probability model is unsatisfactory, though, for at least two reasons:
      i. The disturbances are not distributed 'nicely'. In the example above, the disturbances are evenly split between the values [-½, ½] at *X=1* and take on the values [-0.60 0.40] 40% and 60% of the time, respectively, at *X=2*. Although they are everywhere mean-zero, they are **heteroskedastic** because their distribution is different at different values of *X*. (Different distribution with same mean usually implies different variance.)
      ii. The model is *aesthetically* unappealing because we are estimating a model which we know could not generate the data. In the data, *Y* are not linear in *X*, but we are estimating a model in which *Y* is linear in *X*.
   c. The solutions to both these problems come with a single strategy: write down a model which could conceivably generate the data, and try to estimate its parameters. The cost of doing this is that we typically have to write out an explicit distribution for the disturbances (which we don't have to do when we use OLS and asymptotic variances and tests).

3. Binary choice models: Logits and Probits.
   a. *Latent variable* approach. Assume that there is some latent (unobserved) variable which determines the observed variable *Y*, and which behaves nicely (usually, linearly).
   b. $Y_i^* = X_i\beta + \varepsilon_i,\ \varepsilon_i \sim f(\varepsilon)$

      $Y_i = 1 if\ Y_i^* \geq 0$
   c. Here, *f* is some nice distribution for the disturbance terms, perhaps normal, perhaps not. The linear probability model would likely not satisfy any a priori distribution, and since this distribution is the same for each and every *i*, the linear probability model would certainly suffer from heteroskedasticity.
4. The **method of maximum likelihood** says: "find the parameters which maximise the probability of seeing the sample that you saw".
   a. For a general model $Y_i = X_i\beta + \varepsilon_i,\ \varepsilon_i \sim f(\varepsilon)$, the probability of seeing a whole set of *N* observation is really just the probability of seeing that set of disturbances. Consequently, maximum likelihood is a method of moments approach, because maximising likelihood implies solving a first-order condition, and we solve the first order condition by plugging in residuals for disturbances.
      i. For any *i*,

         $$P[Y = Y_i, X = X_i] = P[\varepsilon = \varepsilon_i] = P[\varepsilon = Y_i - X_i\beta] = f\left(Y_i - X_i\beta\right)$$

         and since we know *f*, and the disturbances are independent, the probability of seeing the whole set of observations, known as the *likelihood* is:
      ii. $$L = P[Y_1,...,Y_N, X_1,..., X_N; \beta] = \prod_{i=1}^{N} f\left(Y_i - X_i\beta\right)$$
      iii. The method of maximum likelihood says, 'choose β by maximising *L*'.
      iv. The likelihood above is expressed over disturbances, but once we choose a β, we are working with residuals. So, implementation of this maximisation involves substituting residuals for disturbances.
      v. Products are a drag to work with, so it you take the log, it turns into a sum:
      vi. $$\ln L = \ln P[Y_1,...,Y_N, X_1,..., X_N; \beta] = \sum_{i=1}^{N} \ln f\left(Y_i - X_i\beta\right)$$
      vii. If *f* were the normal density function, then $f(\varepsilon) = \exp\left(-\varepsilon^2 / \sigma^2 2\right) / \sigma\sqrt{2\pi}$
      viii. $$\ln f(\varepsilon) = -\ln\sigma - \frac{1}{2}\ln 2\pi + \left(-\varepsilon^2 / 2\sigma^2\right),$$
      ix. $$\ln L = -N\ln\sigma - \frac{N}{2}\ln 2\pi - \sum_{i=1}^{N}\left(Y_i - X_i\beta\right)^2 / 2\sigma^2$$, and the first-order condition for maximisation is

x. $\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{N} X_i \left(Y_i - X_i \beta\right) = 0$, aka, $X'e = 0$

(1) this expression is a moment condition, because the first order condition requires that a particular weighted average of disturbances equals zero. We solve it by substituting residuals for disturbances.

xi. So, if you know *f*, you can typically do ML. ML with normality on linear models yields the OLS estimator.

5. Consider the binary choice latent variable model as a maximum likelihood problem.

a. $Y_i^* = X_i \beta + \varepsilon_i,\ \varepsilon_i \sim f(\varepsilon)$
$Y_i = 1 if\ Y_i^* \geq 0$

b. Define the cumulative density function *F* dual to *f* as $F(v) = \int_{-\infty}^{v} f(u) \partial u$, so that *F(v)* gives the probability of the disturbance being less than *v*.

c. What is the probability of seeing any particular *Y*,*X* pair?

d. $P[1, X_i] = P[X_i \beta + \varepsilon_i \geq 0] = P[\varepsilon_i \geq -X_i \beta] = 1 - P[\varepsilon_i \leq -X_i \beta]$
$= 1 - F\left(-X_i \beta\right)$, and

e. $P[0, X_i] = P[X_i \beta + \varepsilon_i < 0] = P[\varepsilon_i < -X_i \beta]$
$= F(-X_i \beta)$, which sum to zero.

f. **ASSUME** that *f* is symmetric, so that a lot of minuses disappear. If *f* is symmetric, then $F\left(-X_i \beta\right) = 1 - F\left(X_i \beta\right),$
$1 - F\left(-X_i \beta\right) = F\left(X_i \beta\right)$

g. The likelihood is thus given by

h. $L = \prod_{i=1}^{N} F\left(X_i \beta\right)^{Y_i} \left(1 - F\left(X_i \beta\right)\right)^{1-Y_i}$, And the log-likelihood is

i. $\ln L = \sum_{i=1}^{N} Y_i \ln F\left(X_i \beta\right) + \left(1 - Y_i\right) \ln\left(1 - F\left(X_i \beta\right)\right).$

j.
$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{N} Y_i X_i \frac{\partial \ln F\left(X_i\beta\right)}{\partial\left(X_i\beta\right)} - \left(1-Y_i\right) X_i \frac{\partial \ln\left(1-F\left(X_i\beta\right)\right)}{\partial\left(X_i\beta\right)}$$
$$= \sum_{i=1}^{N} \frac{Y_i}{F\left(X_i\beta\right)} X_i \frac{\partial F\left(X_i\beta\right)}{\partial\left(X_i\beta\right)} - \frac{1-Y_i}{1-F\left(X_i\beta\right)} X_i \frac{\partial F\left(X_i\beta\right)}{\partial\left(X_i\beta\right)}$$
$$= \sum_{i=1}^{N} X_i \left(\left(\frac{Y_i}{F\left(X_i\beta\right)} - \frac{1-Y_i}{1-F\left(X_i\beta\right)} \right) f\left(X_i\beta\right) \right) = 0$$

i. This is the moment condition for the binary discrete choice maximum likelihood problem. It is analogous to *X'e=0* in OLS.

k. If the probabilities are linear in *X*, then $F(X_i\beta) = X_i\beta,\ f(X_i\beta) = 1$. In this case, the first-order condition on the likelihood function becomes

l.
$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{N} X_i \left(\left(\frac{Y_i}{X_i\beta} - \frac{1-Y_i}{1-X_i\beta} \right) \right)$$
$$= \sum_{i=1}^{N} X_i \left(Y_i - X_i\beta\right) = 0$$

m. That is why we call the OLS approach to this model the *linear probability model*.

i. It results in a linear optimisation problem which has a linear solution.

6. PROBIT model.

a. Denote the standard normal density as $\phi(u) = \exp\left(-u^2/2\right)/\sqrt{2\pi}$, and denote cumulative density of the standard normal density as $\Phi(v) = \int_{-\infty}^{v} \phi(u)\partial u$.

b. So, the probability of seeing a particular pair, *Y,X*, is one or the other of these
$$P[1, X_i] = P[X_i\beta + \varepsilon_i \geq 0] = P[\varepsilon_i \geq -X_i\beta] = 1 - P[\varepsilon_i \leq X_i\beta]$$
$$= 1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)$$

c. $$P[0, X_i] = P[X_i\beta + \varepsilon_i < 0] = P[\varepsilon_i < -X_i\beta] = \Phi\left(\frac{X_i\beta}{\sigma}\right)$$

d. These sum to zero.

e. The likelihood is thus given by

f. $L = \prod_{i=1}^{N} \Phi\left(\frac{X_i\beta}{\sigma}\right)^{Y_i} \left(1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right)^{1-Y_i}$, And the log-likelihood is

g. $$\ln L = \sum_{i=1}^{N} Y_i \ln \Phi\left(\frac{X_i\beta}{\sigma}\right) + (1-Y_i)\left(1 - \ln \Phi\left(\frac{X_i\beta}{\sigma}\right)\right).$$

h. Here, β, σ are not identified, because scaling these parameters by any scalar does not affect the likelihood.

i. So, one must impose a restriction. Typically, we either impose that one element of β is 1, or that $\sigma = 1$. They are equally valid identifying restrictions, so it is up to convenience to choose. Let us choose $\sigma = 1$.

i. First-order conditions are

j.
$$\frac{\ln L}{\partial \beta} = \sum_{i=1}^{N} Y_i X_i \frac{\partial \ln \Phi(X_i\beta)}{\partial (X_i\beta)} - (1-Y_i) X_i \frac{\partial \ln\left(1-\Phi(X_i\beta)\right)}{\partial (X_i\beta)}$$
$$= \sum_{i=1}^{N} \frac{Y_i}{\Phi(X_i\beta)} X_i \frac{\partial \Phi(X_i\beta)}{\partial (X_i\beta)} - \frac{1-Y_i}{1-\Phi(X_i\beta)} X_i \frac{\partial \Phi(X_i\beta)}{\partial (X_i\beta)}$$
$$= \sum_{i=1}^{N} X_i \left(\left(\frac{Y_i}{\Phi(X_i\beta)} - \frac{1-Y_i}{1-\Phi(X_i\beta)}\right)\phi(X_i\beta)\right)$$

k. These first-order conditions are nonlinear in the parameters, and so the solution must be found by iteration.

l. The difficult thing here in finding a solution is that the parameters are inside the normal pdf, which has an explicit analytic form, and inside the normal cdf, which does not have a close analytic form.

m. The picture for this is that which maps y-hat into y. y-hat maps onto a **number line**. For that reason, sometimes, these are called 'index models'. If the index y-hat is less than zero, then y is zero. The probit puts a pdf onto the density of ε, $\varepsilon \leq -X_i\beta$ corresponds to y-hat less than zero.

7. LOGIT model

a. The logit model is a simple alternative to probit model which has a similar distribution for the disturbance terms, but is much easier to solve.

b. Consider the logistic cumulative distribution function, denoted with a capital lambda, $F(v) = \Lambda(v) = \frac{\exp(v)}{1+\exp(v)}$. The associated probability density function is given by $\frac{\partial F(v)}{\partial v} = \frac{\partial \Lambda(v)}{\partial v} = \Lambda(v)\left(1-\Lambda(v)\right)$.

c. If we substitute these into the first-order condition for the likelihood function,

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{N} X_i \left(\left(\frac{Y_i}{\Lambda(X_i\beta)} - \frac{1-Y_i}{1-\Lambda(X_i\beta)} \right) \Lambda(X_i\beta)\left(1-\Lambda(X_i\beta)\right) \right)$$

$$= \sum_{i=1}^{N} X_i \left(\left(\frac{Y_i\left(1-\Lambda(X_i\beta)\right) - \left(1-Y_i\right)\Lambda(X_i\beta)}{\Lambda(X_i\beta)1-\Lambda(X_i\beta)} \right) \Lambda(X_i\beta)\left(1-\Lambda(X_i\beta)\right) \right)$$

$$= \sum_{i=1}^{N} X_i \left(\left(Y_i - Y_i\Lambda(X_i\beta)\right) - \left(\Lambda(X_i\beta) - Y_i\Lambda(X_i\beta)\right) \right)$$

$$= \sum_{i=1}^{N} X_i \left(Y_i - \Lambda(X_i\beta)\right)$$

d. You can see how this might be easier to solve with a computer. No ratios. Everything in there can be expressed analytically.

8. Interpretation of Parameters

a. The parameter β tells you the marginal effect of *X* on *F*. *F* gives the expectation of *Y* given *X*, but because it is nonlinear in its argument, β does not give the derivative of *Y* on *X*. In particular,
$E[Y|_{X=X_i}] = F(X_i\beta)$

b. $$\frac{\partial E[Y|_{X=X_i}]}{\partial} = \frac{\partial F(X_i\beta)}{\partial X} = \beta f(X_i\beta)$$

c. So, the marginal effect depends on *X*. In a probit, the marginal effect is given by $\beta\phi\left(X_i\beta\right)$, and in a logit, the marginal effect is given by
$\beta\Lambda(X_i\beta)\left(1-\Lambda(X_i\beta)\right) = \beta P_i\left(1-P_i\right)$, where *P* is the probability for *i*.

9. Multinomial choice with ordered choices

a. Consider a latent variable model with *J* ordered choices given by

$$Y_i^* = X_i\beta + \varepsilon_i,\ \varepsilon_i \sim f(\varepsilon)$$

b. $$Y_i = 1 if\ 0 < Y_i^* \le \mu_1$$

$$Y_i = 2 if\ \mu_1 < Y_i^* \le \mu_2$$

...

$$Y_i = J\ if\ \mu_{J-1} < Y_i^*$$

c. This is extremely similar to the models above, except that there is an additional unknown set of parameters.

d. Assume that *f* is the **normal** probability density function. (You could assume any pdf, as long as its form is known.)

e. Then, the probabilities of the various outcomes are given by:

$$P[Y_i = 0 \,|_{X=X_i}] = \Phi(-X_i\beta)$$

$$P[Y_i = 1 \,|_{X=X_i}] = \Phi(\mu_1 - X_i\beta) - \Phi(-X_i\beta)$$

$$P[Y_i = 2 \,|_{X=X_i}] = \Phi(\mu_2 - X_i\beta) - \Phi(\mu_1 - -X_i\beta)$$

....

$$P[Y_i = J \,|_{X=X_i}] = 1 - \Phi(\mu_{J-1} - X_i\beta)$$

f. Given that these are the probabilities, the marginal effects of *X* are given by differentiating these probabilities with respect to *X*.

g. These probabilities can be crammed into a log-likelihood function, which can be maximised with respect to the parameters $\beta, \mu_1, ..., \mu_{J-1}$.

10. Multinomial choice with unordered choices.

a. Probabilities of doing various choices are modelled directly. Here, we assume that *Y* could take on the values *j=0,...,J*. Here, there are going to be *J-1* latent variables which generate *J-1* probabilities for *Y* being observed in the *J* possible categories. Logit distributions are easy to work with here.

b. $P[Y_i = j \,|_{X=X_i}] = \dfrac{\exp(X_i\beta^j)}{1 + \sum_{j=1}^{J} \exp(X_i\beta^j)}$. If *J=1*, then this reverts to the binary logit. Given this structure, the probability that *Y* and *X* are observed together is

c. $P[Y = Y_i \,|_{X=X_i}] = \sum_{j=0}^{J} d_{ij} P[Y = j \,|_{X=X_i}]$ With *d* a dummy equal to 1 if *Y* is *j*.

d. The log likelihood is given by

e. $\ln L = \sum_{i=1}^{N} \sum_{j=0}^{J} d_{ij} \ln P\left[Y_i = j \,|_{X=X_i}\right]$ And the first-order condition is

f. $$\frac{\partial \ln L}{\partial \beta^j} = \sum_{i=1}^{N} X_i \left(d_{ij} - P\left[Y_i = j \,|_{X=X_i}\right]\right) = \sum_{i=1}^{N} X_i \left(d_{ij} - P_{ij}\right), \; j = 1, ..., J$$

i. This simple derivative is a consequence of the use of the logit function.

g. Chapters 21 and 22 in Greene are very good on limited dependent variables.