POLI574 – Discrete Choice

Acknowleding a great debt to Matt Golder's notes,
themselves dependent on Train (2007)

Modeling Categorical Outcomes

- Dependent variable is unordered categories
  - Vote choice
  - Choice of policy instrument
  - Outcome of inter-state interactions (e.g. war, trade)

- OLS doesn't work, except LPM for 2 categories
- Logit/Probit are also for 2 categories
- Frequently two outcomes 'closer' together than to other outcomes (see 'IIA' later)
- Frequently nested choices or selection effects

But first... review Binary Dependent Variables

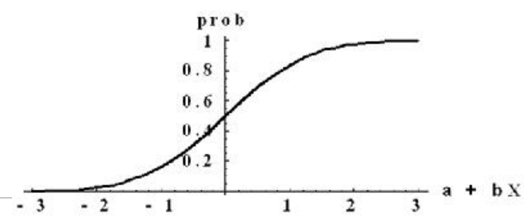
- Recall the linear probability model, which can be written as $P(y = 1|\boldsymbol{x}) = \beta_0 + \boldsymbol{x\beta}$
- An alternative is to model the probability as a function, $G(\beta_0 + \boldsymbol{x\beta})$, where $0<G(z)<1$
- This G just translates – or *squishes* -- the linear additive model into the 0 to 1 space

Logit

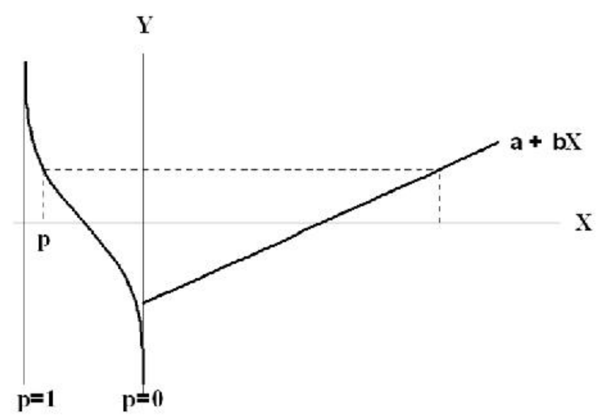
- A common choice for G(z) is the logistic function, which is the *cumulative distribution function* for a standard logistic random variable
- $G(x\beta) = \exp^{(x\beta)}/[1 + \exp^{(x\beta)}]$
  or $1/[1\text{-}\exp^{-x\beta}]$
- We're taking numbers from $-\infty$ to $+\infty$ and transforming those numbers using this cumulative distribution function

Binary Data – View 1 (CDF)

- View 1 – we compute a number that is a linear combination of our predictors, call it $y=\alpha+\beta\ x$. We then convert y into a probability p by using a cumulative distribution function (CDF).
  Our outcome is 1 with probability p.



Another CDF View



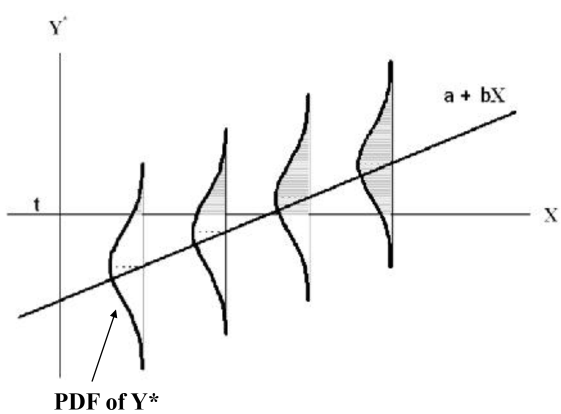
Binary Data – View 2 (Latent or Unobserved Variable)

- View 2 – we compute a number that is a linear combination of our predictors and then add an error term, call it

  $y^* = \alpha + \beta x + u$

  We then get an outcome of 1 if $y^* >= 0$, outcome 0 if $y^* < 0$. In this case, the probabilistic element is the error term u, and y^* is an <u>unobserved variable</u>.

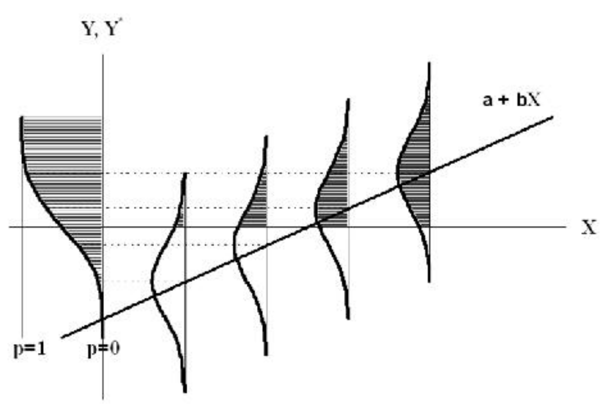
Binary Data – Unobserved Variable View



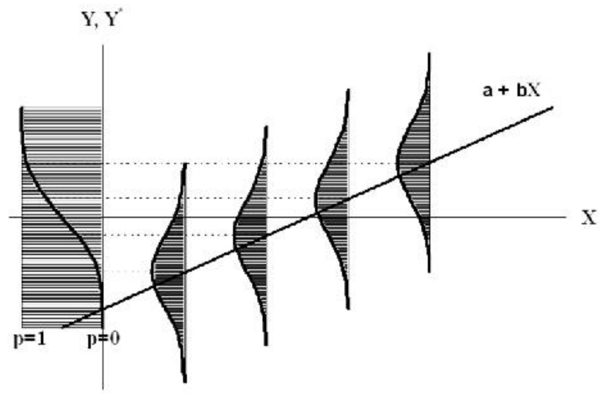
Comparing CDF and Latent Variable Views

- The two views are equivalent. Each one can be converted into the other, where the cumulative probability function (CDF) in view 1 matches the CDF of the distribution of *u* in view 2.

Combining the Two Views



Combining the Two Views



These are all NONLINEAR models

- The rate of change in the dependent var with respect to the independent var
  IS NOT CONSTANT
- So we have to estimate coefficients by
  trial and error
- So… maximum likelihood

12

Likelihood and Traditional Probability

- Theory of likelihood is the reverse of traditional probability theory
- Traditional theory: probability that we got this set of data given the TRUE parameter values
- In likelihood we're honest that we only have one set of data. So we talk about the 'likelihood' of each set of parameter values given the data we actually got
- What model (i.e. parameters) is most likely to have produced the data we collected?

Likelihood is a RELATIVE measure of uncertainty

- The *likelihood function* is a measure of the relative probability of all possible parameter values (i.e. estimates of the true model)
  - think of all possible parameter values. Whoah!
- So it gives us a mean (most likely parameter value) and a variance (how much more likely than others)
- The maximum of this function gives us an estimate of the mean of the parameter (vector)
- THIS APPLIES TO ALL POSSIBLE MODELS

14

Constructing a Likelihood (logit)

We assume a *data generating process*

- This applies to every observation
- For binary outcomes we assume they are generated by a Bernoulli distribution: $p_i^{y_i}(1-p_i)^{1-y_i}$
- Then we model *p*, the probability (our model), as a function of expalanatory variables: $p_i = g(x_i, \beta)$\
- For logit, let $p_i = \frac{1}{1+\exp(-x_i\beta)}$
- Now, since our observations are independent…
- The probability of all of the Y given one particular value of *p* (i.e. the model) is equal to the product of all the probabilities $\Pr(Y \mid p) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$
- So we combine these and get

$$\Pr(y \mid \beta) = \prod_{i=1}^{n}\left(\frac{1}{1+\exp(-x_i\beta)}\right)^{y_i}\left(1-\frac{1}{1+\exp(-x_i\beta)}\right)^{1-y_i}$$

15

Constructing a Likelihood Continued

From $$\Pr(y \mid \beta) = \prod_{i=1}^{n}\left(\frac{1}{1+\exp(-x_i\beta)}\right)^{y_i}\left(1-\frac{1}{1+\exp(-x_i\beta)}\right)^{1-y_i}$$

The theory of maximum likelihood says that the likelihood *function* $L(\beta|y)$ is proportional to this expression

So to get the log-likelihood that's easier to work with, we take the log of the expression and we get

$$\ln L(\hat{\beta} \mid y) = \sum_{i=1}^{n} - y_i \ln\left[1+\exp(-x_i\hat{\beta})\right] - (1-y_i)\ln\left[1+\exp\left(-x_i\hat{\beta}\right)\right]$$

We've gone from products to sums and from wanting to minimize something to maximizing this function

We plug in values for β, call them $\hat{\beta}$, and do an astronomical amount of simple arithmetic to get a log-likelihood for that set of estimates.

Then we use an algorithm to search for the set of estimates that maximizes this log-likelihood

Now, Multiple Outcomes

17

Notation (follows Golder)

- ***n*** individual cases (decision makers)
- ***J*** alternatives
- ***i and j*** are alternative outcomes
  - ***i*** <u>chosen</u> outcome (choice)
  - ***j*** all outcomes (alternatives)
- $\boldsymbol{\beta_j}$ is the set of coefficients for alternative ***j***
  (where one set is set to zero as the 'base category')
- ***X*** is still the linear-additive independent variables

$i \neq j \ \forall j$

18

Random Utility Model

- Differences in utility of alternatives result in choice / behaviour
- But a random component, so we get a predicted behaviour given characteristics of choices and choosers
- Probability of each outcome for each chooser
- Or: Proportion of each choice within population groups defined by combinations of characteristics

19

RUM

$$
\begin{aligned}
P_{ni} &= \text{Prob}(U_{ni} > U_{nj} \forall j \neq i) \\
&= \text{Prob}(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj} \forall j \neq i) \\
&= \text{Prob}(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj} \forall j \neq i)
\end{aligned}
$$

- This last step is weird
- It expresses the probability as:
  - *i* is chosen if the difference between the errors is less than the difference between the systematic difference in utilities
- Just like OLS in that the model minimizes the residuals – the ϵ
- Just like all MLE in that we choose a distribution for these errors
- Then to get probabilities we calculate the integral of these unobserved utilities
  - i.e. the probability that *i* is chosen is how much probability mass is below the threshold where the difference in the errors is more than the difference in the systematic portion of the utilities.

20

Differences in Utility

- As Golder says: "Only Differences in Utility Matter"
- Because utility is *unobserved* or 'latent', and we only know whether one alternative was chosen as opposed to another, we can only think of systematic influences as *relative*
- So the impact of a characteristic of a chooser (e.g. female) **is not** that it produces, on average, Θ_{n1} and Θ_{n2} and so on *U*tilities for the choices.
- Instead, it just tells us about the average *difference* in the utility of the two choices, i.e. $\Theta_2 - \Theta_1$
- Since we don't observe utility, that $\Theta_2 - \Theta_1$ is indeterminate, so we just set one of them to ZERO and interpret the Θ_i parameter as the difference in the utility of the i^{th} choice from the one choice for which we set all the Θ's to zero.

Logit Models for categorical outcomes

- Assume a distribution for the ϵ
- We actually use one that's mathematically convenient rather than substantively justified
  - Suffice to say it is a logistic dist. for choice btw any two alternatives
  - $\frac{e^{\tilde{\epsilon}_{nji}}}{1+e^{\tilde{\epsilon}_{nji}}}$, $where \tilde{\epsilon}^{\sim} = \epsilon_{nj} - \epsilon_{ni}$
- BIG assumption is that the unobserved part of the utility of one alternative is **independent** of the unobserved part of other alternatives (IIA, more later)
- Means you've got a good, well-specificed model: one that includes all systematic influences on the choices

Multiple Outcome Logit Choice Probabilities

- So the choice of one alternative by a chooser indicates that the error for each other choice was below
  $\epsilon_{ni} + V_{ni} - V_{nj}$
- With multiple choices, we need the probability that this is true for all $j \neq i$, which is the product of all of the cumulative distributions of the errors for all the non-chosen choices, relative to the distribution of the errors of *I*
  (that's roughly what Golder's eq. 16 says)
- That's the criterion analogous to 'least-squares' for OLS
- So the MNL choice probabilities are

$$P_{ni} = \frac{e^{x_{ni}\beta}}{\sum_j e^{x_{nj}\beta}}$$

- And the log likelihood is this over all choices and choosers

23

Two models, MNL and CoLogit

- Golder does Conditional Logit before Multinomial Logit
- Weird choice, but it makes a bit of sense
- I'm going to follow him

24

Conditional Logit

- Pure Conditional Logit involves only characteristics of choices
- Transportation models involved price, speed, comfort of each of modes of transport
- Notice that the *x* are subscripted by *nj*, meaning they are about the decision-maker *relative* to the alternatives

$$U_{nj} = V_{nj} + \epsilon_{nj}$$
$$= x_{nj}\beta + \epsilon_{nj}$$

- Like 'distance' from a party on policy, or a country's distance from potential allies or adversaries
- β **has no subscript** because the effect of this variable is constant across alternatives
  - E.g. 'distance' or higher price makes you less likely to choose something
  - Speed, comfort make choice more likely
  - Next page: same language as leader makes choice more likely distance from parties on corporate tax policy makes choice less likely

25

Conditional Logit in Stata

Vote Choice in Quebec, 2011

- `clogit choice samelang dist_corptax, group(id)`

```
Iteration 0:   log likelihood = -2197.0125
Iteration 1:   log likelihood = -2196.8142
Iteration 2:   log likelihood = -2196.8142

Conditional (fixed-effects) logistic regression   Number of obs   =       7428
                                                  LR chi2(2)      =      42.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -2196.8142                       Pseudo R2       =     0.0096

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    samelang |   .5009395   .0762589     6.57   0.000     .3514749    .6504041
dist_corptax |  -.0933343   .0478339    -1.95   0.051    -.1870869    .0004184
------------------------------------------------------------------------------
```

- Coefficients are change in log-odds of choosing an alternative, for one-unit change in the independent variable

26

Multinomial Logit (MNL)

$$U_{nj} = V_{nj} + \epsilon_{nj}$$

The systematic component of the utility function is given as:

$$V_{nj} = z_n \gamma_j$$

So, we have

$$U_{nj} = z_n \gamma_j + \epsilon_{nj}$$

- *z* is equivalent to *x* variables
- *γ (gamma)* is equivalent to β
- note that the *γ* are subscripted, so separate 'effects' of each *z* (characteristic) on each choice
  - E.g. **female** may have different effects on prob of choosing each party
  - Trade deficit may have a different effect on choice of trade war, unilateral tarriff reduction, bilateral negotiation, or multilateral trade negotiation
- MNL Choice Probabilities: $= \frac{e^{z_n \gamma_i}}{\sum_j e^{z_n \gamma_j}}$

27

MNL identification

- Attributes of choosers don't vary across alternatives
- So they can only create differences between alternatives
  - e.g. educ level can only make some parties more likely to be voted for
- Simple solution: set all coefficients for one alternative to zero
- Coefficients are always about the difference in choice probabilities between two of the choices
- As a decision-maker becomes more likely to choose one alternative, she is less likely to choose others
- This just works out to a different set of independent variables. The likelihoods are basically the same.

28

MNL is binary logits!

- MNL estimates the same parameters as a series of binary logits
- It's slightly more efficient (see Alvarez and Nagler)
- This is because of IIA
- Later, we'll talk about relaxing IIA

29

Digression: Don't estimate choice versus all others

- … unless you have a theoretical reason to
- Cautionary tale:
  IS BQ voting influenced by attitude to spending on Envrmt?

```
. logit vote4 sov  spend_EN

Logistic regression                               Number of obs   =        904
                                                  LR chi2(2)      =     243.17
                                                  Prob > chi2     =     0.0000
Log likelihood =  -430.4643                       Pseudo R2       =     0.2202

------------------------------------------------------------------------------
       vote4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sov |   2.455879    .179795    13.66   0.000     2.103487    2.808271
    spend_EN |   .1583196   .1679772     0.94   0.346    -.1709097    .4875489
       _cons |  -2.592755   .4599843    -5.64   0.000    -3.494307   -1.691202
------------------------------------------------------------------------------
```

- No effect of Environment attitudes?

30

MNL in Stata

```
. mlogit vote sov  spend_EN

Multinomial logistic regression                   Number of obs   =        904
                                                  LR chi2(8)      =     353.81
                                                  Prob > chi2     =     0.0000
Log likelihood = -1149.7148                       Pseudo R2       =     0.1333
 ------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Liberal      |
         sov |  -3.323585   .2804922   -11.85   0.000    -3.873339    -2.77383
    spend_EN |  -.0180076   .2244897    -0.08   0.936    -.4579993    .4219842
       _cons |   1.033057   .6140084     1.68   0.092    -.1703772    2.236492
-------------+----------------------------------------------------------------
Conservati~s |
         sov |  -3.063571   .2648611   -11.57   0.000    -3.582689   -2.544453
    spend_EN |  -.9299179   .2053723    -4.53   0.000     -1.33244   -.5273956
       _cons |   3.380755   .5476791     6.17   0.000     2.307323    4.454186
-------------+----------------------------------------------------------------
NDP          |
         sov |  -1.929275   .2010957    -9.59   0.000    -2.323415   -1.535135
    spend_EN |   .0810441   .1944765     0.42   0.677    -.3001229    .4622111
       _cons |   .8961252   .5389201     1.66   0.096    -.1601387    1.952389
-------------+----------------------------------------------------------------
Bloc_Quebe~s |  (base outcome)
-------------+----------------------------------------------------------------
Green_Party  |
         sov |  -1.252255   .3907826    -3.20   0.001    -2.018175   -.4863351
    spend_EN |   1.313919   .6122771     2.15   0.032     .1138781     2.51396
       _cons |  -4.987046   1.783435    -2.80   0.005    -8.482515   -1.491578
------------------------------------------------------------------------------
```


IIA - 1

Independence of/from Irrelevant Alternatives

- A **property** of the Multinomial Logit Model
  - It's built into the model by assumption
- Assumption about individual choosers: their own ratio of probabilities of two choices don't depend on other alternatives
- Classic example is Red Bus/Blue Bus from transp. mode choice

If $\frac{\Pr(Car)}{\Pr(RedBus)} = 1$, meaning $\Pr(Car) = \Pr(RedBus) = 0.5$

then an identical Blue Bus is introduced, we have to keep $\frac{\Pr(Car)}{\Pr(RedBus)} = 1$

so we get $\Pr(Car) = \Pr(RedBus) = \Pr(BlueBus) = 0.33$

- But we should have had
  $\Pr(Car) = 0.5, \quad \Pr(RedBus) = \Pr(BlueBus) = 0.25$
- This is a feature of **un**conditional probabilities

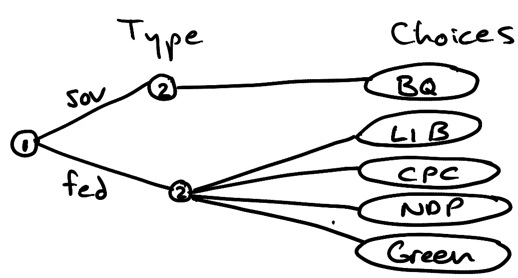
IIA - 2

- This is a feature of **un**conditional probabilities
- Which means it's about the errors:
  everything **un**explained about the choices
- Avoid IIA with good data, right model

- Or use a different model/estimator that relaxes IIA assumption

33

Nested Logit



- Nests or Levels or Stages: different vars. for different nests
- Based on Generalized Extreme Value distribution (GEV)
- Simple:
  Some stuff determines choices among nests,
  other stuff determines choices within nests $U_{nj} = W_{nk} + Y_{nj} + \epsilon_{nj}$
- So it avoids IIA, but IIA still holds at each level
- Nesting involves UNOBSERVED stuff

34

Nested Logit Probabilities

$$P_{nB_ki} = \frac{e^{w_{nk}\gamma+\lambda_k I_{nk}}}{\sum_{\iota=1}^{K} e^{w_{n\iota}\gamma+\lambda_\iota I_{n\iota}}} \times \frac{e^{x_{ni}\beta/\lambda_k}}{\sum_{j\in B_k} e^{x_{nj}\beta/\lambda_k}}$$

- Probability of choosing alternative *i* in nest *k* is

  prob of choosing nest *k* X prob of choosing *i* given choice of *k*
- I_{nk} is the 'inclusive value' for the nest for each person:
  the value of the nest, irrespective of which alternative is chosen
- And the γ_k is how independent (uncorrelated) are the errors for each alternative within a nest
- Note that in the lower-level probabilities, the $X\beta$ utilities for each alternative are divided by γ_k

35

Multinomial Probit

- Was hailed as the saviour…. But fell flat
  - Idea was to simulate different sets of alternatives and different characteristics of alternatives to get changes in probs
- Only rarely is it worthwhile
- Errors are multivariate normal so we can estimate correlation of unmeasured factors (*u*) between choices
- But only some of them

- I tried MNP on my Quebec CES data, and it took 45 minutes!
- That's with the "full covariance matrix" of the errors
- I could restrict it, but why not just use nested logit if we're going to theorize a structure

36

Censoring, Truncation, and Selection

37

Censoring, Truncation, and Selection

- *Censored* Data are when we don't have y values for some observations (X is known)
  - no wage for people who choose not to work
  - stadium attendance beyond capacity
  - duration models (something hasn't happened yet)
- *Truncated* Data are when don't have any data for a certain range of the dependent variable
  - no data on conflicts below (or above) a certain number of battle deaths
- *Selection* (or sample selection) is a form of truncation where there is a mechanism for inclusion in the sample that does not depend deterministically on y (though it can be related to y)
  - no indication of racist attitudes from REALLY racist people
  - e.g. no info on those who decline a survey
  - no sanctions in cases where they are judged unnecessary or unlikely to succeed

38

Censored and Truncated Likelihoods (from Kennedy)

The upper half of figure 16.1 illustrates why this is the case (ignore for now the lower half of this diagram). The relationship $y = \alpha + \beta x + \varepsilon$ is being estimated, where ε is a normally distributed error and observations with y values greater than k are not known. This could happen because y is the demand for tickets to hockey games and the arena on some occasions is sold out so that for these games all we know is that the demand for tickets is greater than k, the capacity of the arena. These unknown y values are denoted by small circles to distinguish them from known data points, designated by dots. Notice that for high values of x the known (dotted) observations below the (unconditional) expectation $E(y) = \alpha + \beta x$ are not fully balanced off by observations above $E(y) = \alpha + \beta x$, because some of these observations (the circled ones) are missing. This causes the resulting OLS regression line to be too flat, as shown by the dashed line.

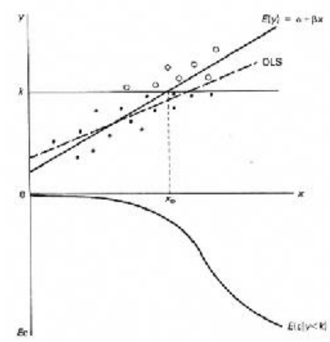


Figure 16.1
A limited dependent variable model

Samples with limited dependent variables are classified into two general categories, censored and truncated regression models, depending on whether or not the values of x for the missing y data are known.

(1) *Censored sample* In this case some observations on the dependent variable, corresponding to known values of the independent variable(s), are not observable. In figure 16.1, for example, the y values corresponding to the circled data points are not known, but their corresponding x values are known. In a study of the determinants of wages, for example, you may have data on the explanatory variables for people who were not working, as well as for those who were working, but for the former there is no observed wage.

(2) *Truncated sample* In this case values of the independent variable(s) are known only when the dependent variable is observed. In the example of the negative income tax experiment noted earlier, no data of any kind are available for those above the income threshold; they were not part of the sample.

39

Censored and Truncated Likelihoods (from Kennedy)

Technical Notes

16.1 Introduction

• The likelihood functions for censored and truncated samples are quite different. This can be illustrated with the help of figure 16.3, which graphs the density function of the error ε from figure 16.1. Consider a particular value x_3 of x. For y_3 to be observable, ε_3 must lie to the left of $k - \alpha - \beta x_3$; for y_3 unobservable, ε_3 must lie to the right of $k - \alpha - \beta x_3$. This result follows from the discussion of $E\varepsilon$ above.

Suppose first we have a censored sample. If x_3 corresponds to an observable y, then there will be a specific ε_3 and the likelihood for that observation is given by L_3 in figure 16.3, the height of the density function for ε at ε_3. But if x_3 corresponds to an unobservable (i.e., missing) value of y, we have no specific ε_3; all we know is that ε_3 must lie to the right of $k - \alpha - \beta x_3$. The likelihood of this observation is thus the probability that ε_3 exceeds $k - \alpha - \beta x_3$, given by the lined area in figure 16.3, and calculated as 1 minus the density function cumulated to the point $k - \alpha - \beta x_3$. The likelihood for each observation in the sample may be calculated in one of these two ways, depending on whether the y value is observed or unobserved. Multiplying together all of these likelihood expressions, some of which are densities and some of which are cumulative densities, creates the likelihood for the censored sample.

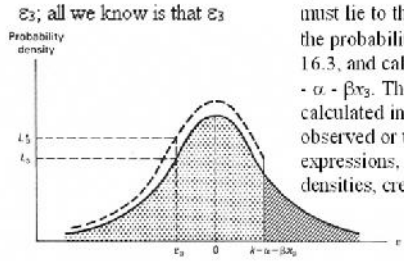


Figure 16.3
Explaining the likelihood for censored and truncated

40

Censored and Truncated Likelihoods (from Kennedy)

Suppose now we have a truncated sample. For every possible value of x_3 in the sample the associated error must come from the left of $k - \alpha - \beta x_3$ in figure 16.3. Consequently the lined area should not be viewed as part of the density of ε_3. Because of this, ε_3 can be viewed as being drawn from the truncated normal distribution given by the dashed curve in figure 16.3. This dashed curve is obtained by dividing the height of the original normal distribution by the dotted area, forcing the area under the dashed curve to equal 1. Thus the likelihood of the observation y_3 is given in figure 16.3 by L_3^T. Note that L_3^T is a complicated function of the data, consisting of the height of the normal density function at the observation (y_3, x_3), divided by that density function cumulated to the point $k - \alpha - \beta x_3$. Each observation will give rise to a different dashed curve from which the likelihood of that observation can be calculated. Multiplying together all these likelihood expression creates the likelihood function for the entire sample.

The Tobit Model for Corner responses

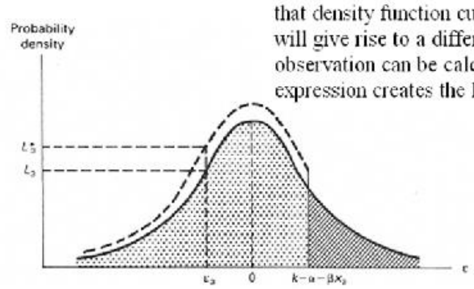


Figure 16.3
Explaining the likelihood for censored and truncated models

11

Sample Selection Corrections

- If a sample is truncated in a nonrandom way, then OLS suffers from selection bias
- Can think of as being like omitted variable bias, where what's omitted is how the observations were selected into the sample, so
- $E(y|z, s = 1) = \boldsymbol{x\beta} + \rho\lambda(\boldsymbol{z\gamma})$, where
- $\lambda(c)$ is the inverse Mills ratio: $\phi(c)/\Phi(c)$
  - the chance of being in the sample
  - ratio of point on stdnorm pdf to point on stdnorm cdf

Selection Correction (continued)

- We need an estimate of λ, so estimate a probit of *s* (whether *y* is observed) on $\boldsymbol{z}$
- These estimates of γ can then be used along with $\boldsymbol{z}$ to form the inverse Mills ratio
- Then you can just regress *y* on $\boldsymbol{x}$ and the estimated λ to get consistent estimates of $\boldsymbol{\beta}$
- See Berinsky article

44