Applied Econometrics

Lecture 10

Nathaniel Higgins

ERS and JHU

8 November 2010

Qualitative dependent variables
Outline

- Review of binary dependent data models
  - Model logic
  - Maximum likelihood (helpful for lots of stuff in the future)
- Multiple-choice models
  - Multinomial logit
  - Ordered probit
- Homework review (if time)
- Start matching (if time)

Binary dependent variables

- Let's introduce y^* and say it represents the net happiness of attending JHU

Binary dependent variables

- Let's introduce y^* and say it represents the net happiness of attending JHU
- In terms of our variables:
  $y^* = X\beta + \epsilon$
- We call this type of model a *latent variables* model because y^* is unobservable

Binary dependent variables

- Let's introduce y^* and say it represents the net happiness of attending JHU
- In terms of our variables:
  $y^* = X\beta + \epsilon$
- We call this type of model a *latent variables* model because y^* is unobservable
- Individual attends if (unobservable) net happiness from attending is positive, i.e.
  $y^* = X\beta + \epsilon > 0$

Binbary dependent variables

- What is in our data:
  1. A bunch of stuff that we think might influence net happiness (the X data)
  2. Whether or not the person attended
- We say that

  Individual i attends ($y_i = 1$) if $y_i* > 0$

  Individual i does not attend ($y_i = 0$) if $y_i* < 0$

Binary dependent variables

- We modeled y_i* as a function of the observable data:

Binary dependent variables

- We modeled y_i* as a function of the observable data:
  Individual i attends ($y_i = 1$) if $y_i* = X_i\beta + \epsilon_i > 0$
  Individual i does not attend ($y_i = 0$) if $y_i* = X_i\beta + \epsilon_i < 0$

Binary dependent variables

- We modeled y_i* as a function of the observable data:
  Individual i attends ($y_i = 1$) if $y_i* = X_i\beta + \epsilon_i > 0$
  Individual i does not attend ($y_i = 0$) if $y_i* = X_i\beta + \epsilon_i < 0$
- This means that:
  $y_i = 1$ if $X_i\beta + \epsilon_i > 0$
  $y_i = 0$ if $X_i\beta + \epsilon_i < 0$

Binary dependent variables

- We modeled y_i* as a function of the observable data:
  Individual i attends ($y_i = 1$) if $y_i* = X_i\beta + \epsilon_i > 0$
  Individual i does not attend ($y_i = 0$) if $y_i* = X_i\beta + \epsilon_i < 0$
- This means that:
  $y_i = 1$ if $X_i\beta + \epsilon_i > 0$
  $y_i = 0$ if $X_i\beta + \epsilon_i < 0$
- Just move ϵ_i over the inequality to get:
  $y_i = 1$ if $X_i\beta > -\epsilon_i$
  $y_i = 0$ if $X_i\beta < -\epsilon_i$

Binary dependent variables

- We modeled y_i* as a function of the observable data:
  Individual i attends ($y_i = 1$) if $y_i* = X_i\beta + \epsilon_i > 0$
  Individual i does not attend ($y_i = 0$) if $y_i* = X_i\beta + \epsilon_i < 0$
- This means that:
  $y_i = 1$ if $X_i\beta + \epsilon_i > 0$
  $y_i = 0$ if $X_i\beta + \epsilon_i < 0$
- Just move ϵ_i over the inequality to get:
  $y_i = 1$ if $X_i\beta > -\epsilon_i$
  $y_i = 0$ if $X_i\beta < -\epsilon_i$
- We assume that ϵ is distributed according to either the normal or the logistic distribution

Binary dependent variables

- Forget about which we assume (or why we assume either) for the moment

Binary dependent variables

- Forget about which we assume (or why we assume either) for the moment
- Simply realize that ϵ is a random variable with a distribution function F

Binary dependent variables

- Forget about which we assume (or why we assume either) for the moment
- Simply realize that ϵ is a random variable with a distribution function F
- Take a step back and recall some facts about distribution functions
- Recall the interpretation of a distribution function evaluated at z: $F(z)$

Binary dependent variables

- Forget about which we assume (or why we assume either) for the moment
- Simply realize that ϵ is a random variable with a distribution function F
- Take a step back and recall some facts about distribution functions
- Recall the interpretation of a distribution function evaluated at z: $F(z)$

$$\begin{aligned} F(z) &= Pr(Z \leq z) \\ &= Pr(\text{random draw} \leq \text{fixed } z) \end{aligned}$$

- Now back to our model ...

Binary dependent variables

- We have:
  $y_i = 1 \quad \text{if} \quad X_i\beta > -\epsilon_i$
  $y_i = 0 \quad \text{if} \quad X_i\beta < -\epsilon_i$
- And we know that ϵ is a random variable with some distribution function F

Binary dependent variables

- We have:
  $y_i = 1 \quad \text{if} \quad X_i\beta > -\epsilon_i$
  $y_i = 0 \quad \text{if} \quad X_i\beta < -\epsilon_i$
- And we know that ϵ is a random variable with some distribution function F
- Look at the relationship $X_i\beta > -\epsilon_i$
- Try to make an analogy between $X_i\beta > -\epsilon_i$ and $F(z) = Pr(Z \leq z)$

Binary dependent variables

- We have:
  $y_i = 1 \quad \text{if} \quad X_i\beta > -\epsilon_i$
  $y_i = 0 \quad \text{if} \quad X_i\beta < -\epsilon_i$
- And we know that ϵ is a random variable with some distribution function F
- Look at the relationship $X_i\beta > -\epsilon_i$
- Try to make an analogy between $X_i\beta > -\epsilon_i$ and $F(z) = Pr(Z \leq z)$
- ϵ_i is the random variable (like "big" Z) and $X_i\beta$ is the specific value (like "little" z)

Binary dependent variables

- We know that $y_i = 1$ if $X_i\beta > -\epsilon_i$
- Which is the same as saying $y_i = 1$ if $-\epsilon_i < X_i\beta$

Binary dependent variables

- We know that $y_i = 1$ if $X_i\beta > -\epsilon_i$
- Which is the same as saying $y_i = 1$ if $-\epsilon_i < X_i\beta$
- Saying $-\epsilon_i < X_i\beta$ is the same as saying $\epsilon_i > -X_i\beta$

Binary dependent variables

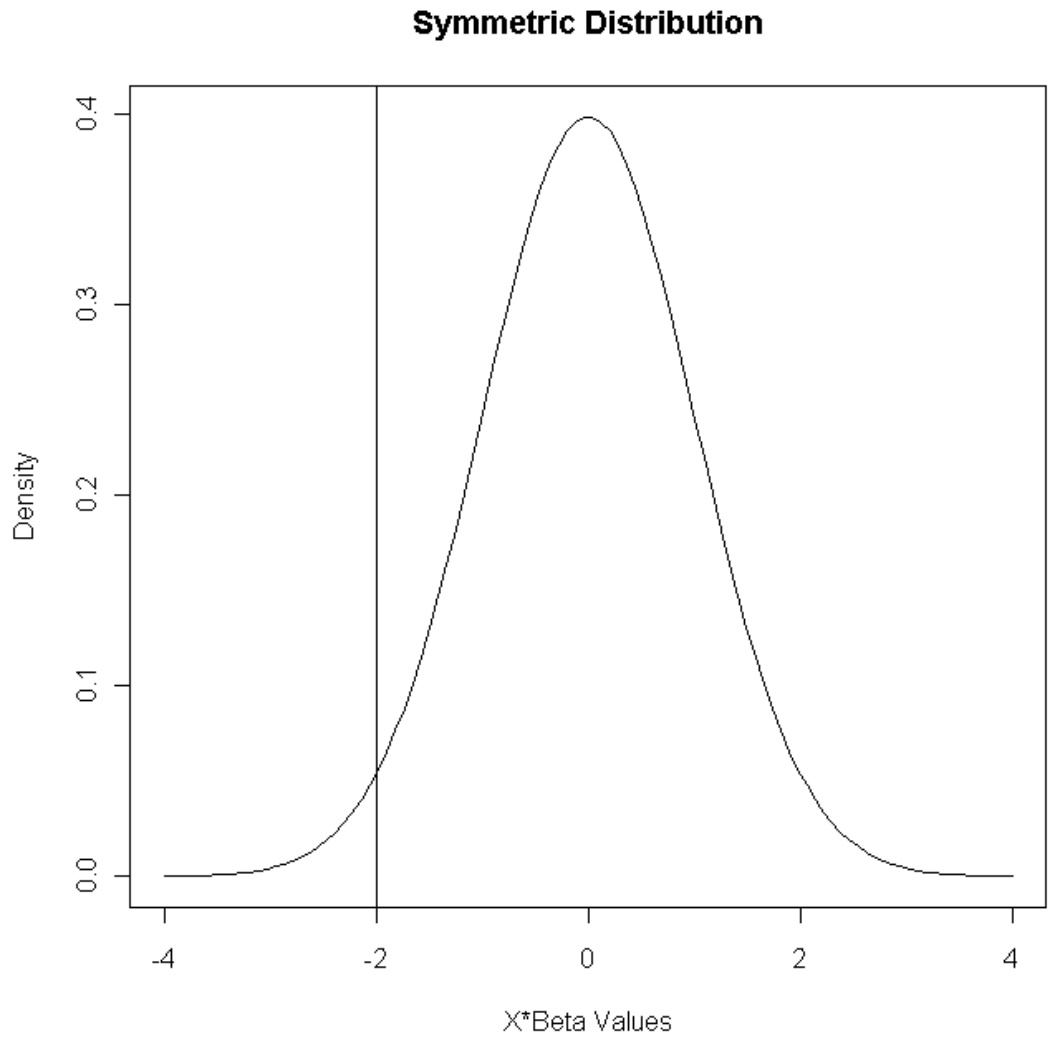
- We know that $y_i = 1$ if $X_i\beta > -\epsilon_i$
- Which is the same as saying $y_i = 1$ if $-\epsilon_i < X_i\beta$
- Saying $-\epsilon_i < X_i\beta$ is the same as saying $\epsilon_i > -X_i\beta$
- Now for the final piece of the puzzle

Binary dependent variables

- We want a function to describe the condition $\epsilon_i > -X_i\beta$

Binary dependent variables

- We want a function to describe the condition $\epsilon_i > -X_i\beta$

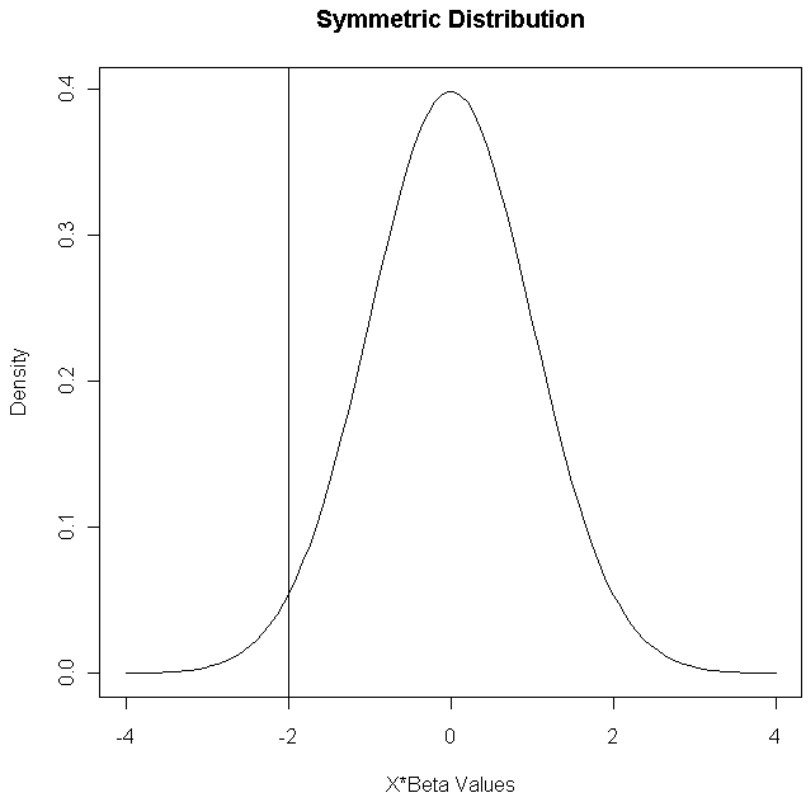


Binary dependent variables

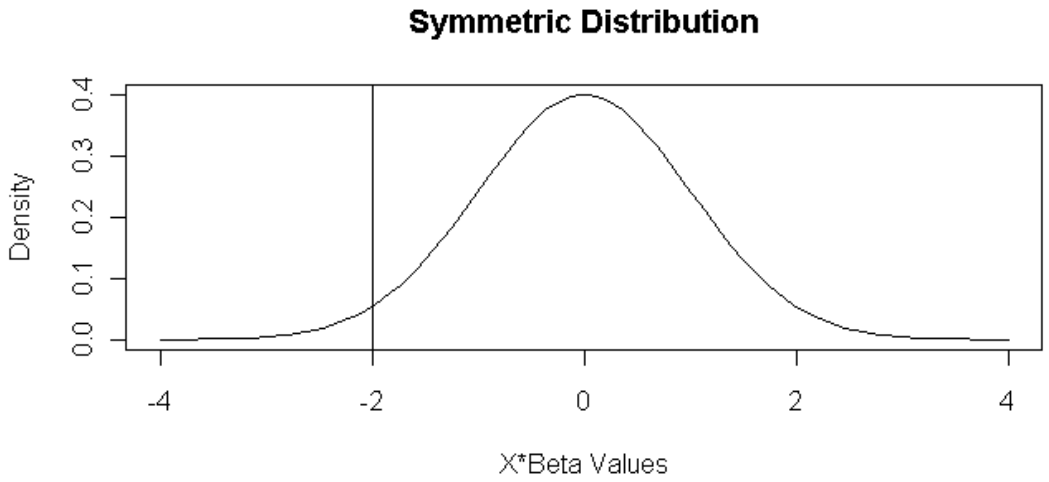
- Both the normal distribution and the logit distribution are symmetric
- The expected value of ϵ is 0

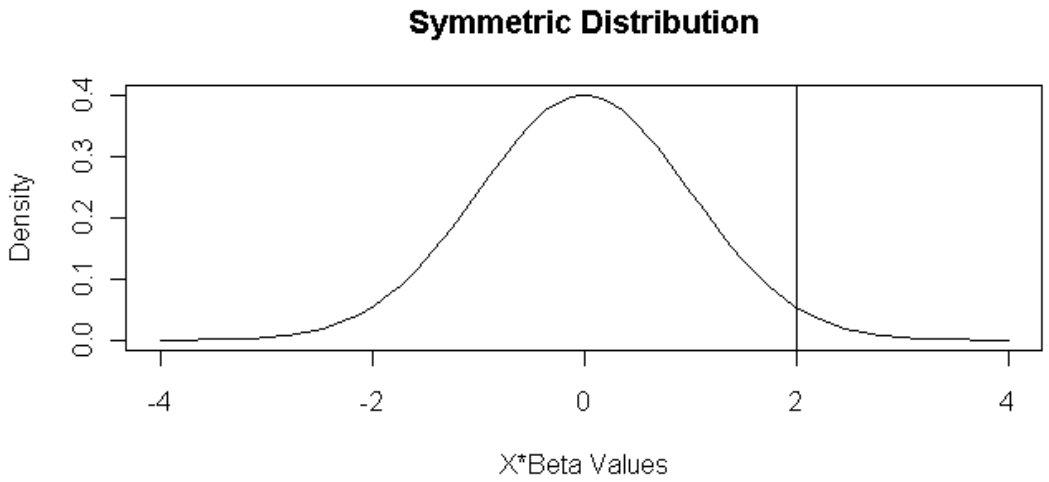
Binary dependent variables

- Both the normal distribution and the logit distribution are symmetric
- The expected value of ϵ is 0
- The probability that $\epsilon_i > -X_i\beta$ is the probability that ϵ falls somewhere to the right of $-X_i\beta$ (-2 in the example below)



Binary dependent variables





Binary dependent variables

- The probability that $\epsilon_i > -X_i\beta$ is *exactly* the same as the probability that $\epsilon_i < X_i\beta$ when we have a symmetric distribution centered around 0

Binary dependent variables

- The probability that $\epsilon_i > -X_i\beta$ is *exactly* the same as the probability that $\epsilon_i < X_i\beta$ when we have a symmetric distribution centered around 0
- Therefore:

$$
\begin{aligned}
y_i &= 1 \\
y_i* &= X_i\beta + \epsilon_i > 0 \\
&= X_i\beta > -\epsilon_i \\
&\quad Pr(X_i\beta > -\epsilon_i) \\
&= Pr(-X_i\beta < \epsilon_i) \\
&= Pr(\epsilon_i < X_i\beta) \\
&= F(X_i\beta)
\end{aligned}
$$

Binary dependent variables

- Put it all together:
- The probability that $y_i = 1$ is modeled as $F(X_i\beta)$, where F is either the normal distribution function or the logit distribution function

Binary dependent variables

- Put it all together:
- The probability that $y_i = 1$ is modeled as $F(X_i\beta)$, where F is either the normal distribution function or the logit distribution function
- If we assume that ϵ is distributed normally, $F(X_i\beta)$ is expressed like this:

$$\int \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(X_i\beta)^2}{2\sigma^2}\right]$$

Binary dependent variables

- Put it all together:
- The probability that $y_i = 1$ is modeled as $F(X_i\beta)$, where F is either the normal distribution function or the logit distribution function
- If we assume that ϵ is distributed normally, $F(X_i\beta)$ is expressed like this:

$$\int \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(X_i\beta)^2}{2\sigma^2}\right]$$

- So what the heck do we do with this?

Binary dependent variables

- We have a model that says

$$Pr(y_i = 1) = \int \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(X_i\beta)^2}{2\sigma^2}\right]$$

$$Pr(y_i = 0) = 1 - \int \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(X_i\beta)^2}{2\sigma^2}\right]$$

- Now that we have a model, don't forget our goal: We want to estimate β
- How do we estimate β?

Maximum likelihood

- This is where maximum likelihood comes in

Maximum likelihood

- This is where maximum likelihood comes in
- Instead of minimizing the sum of squared errors, we maximize something called the "likelihood" that represents, loosely speaking:

the probability of the data

Maximum likelihood

- This is where maximum likelihood comes in
- Instead of minimizing the sum of squared errors, we maximize something called the "likelihood" that represents, loosely speaking:

  the probability of the data

- Now it's time for the balls-in-a-hat example (done right this time)

Maximum likelihood example

- Got some balls in a hat here

Maximum likelihood example

- Got some balls in a hat here
- Let's go around the room and create some data by sampling

Maximum likelihood example

- Got some balls in a hat here
- Let's go around the room and create some data by sampling
- Let's have everyone choose a ball from the hat (put it back when you're done) and we'll record the data on the board
- (I'll add the data to our lecture slides before posting them)

Maximum likelihood example

- Got some balls in a hat here
- Let's go around the room and create some data by sampling
- Let's have everyone choose a ball from the hat (put it back when you're done) and we'll record the data on the board
- (I'll add the data to our lecture slides before posting them)
- But not just yet
- First, let's set up the likelihood function

Likelihood function

- I am telling you now that there are four balls in the hat
- Therefore the probability of any one of you pulling a red ball out of the hat is

$$\frac{something}{4}$$

- And we want to find that *something*

Likelihood function

- I am telling you now that there are four balls in the hat
- Therefore the probability of any one of you pulling a red ball out of the hat is

$$\frac{something}{4}$$

- And we want to find that *something*
- Think of that *something* as the parameter we're after
- So let's call it β

Likelihood function

- Let's set up a likelihood function in exactly the same way we set it up last time

Likelihood function

- Let's set up a likelihood function in exactly the same way we set it up last time
- We'll denote the likelihood function L

Likelihood function

- Let's set up a likelihood function in exactly the same way we set it up last time
- We'll denote the likelihood function L
- We want to write down the likelihood of observing the data that we do (the data we haven't generated yet)

Likelihood function

- Let's set up a likelihood function in exactly the same way we set it up last time
- We'll denote the likelihood function L
- We want to write down the likelihood of observing the data that we do (the data we haven't generated yet)

$$L(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) = \prod_{(y_i=1)} \left[\frac{\beta}{4}\right] * \prod_{(y_i=0)} \left[1 - \frac{\beta}{4}\right]$$

Likelihood function

- Let's set up a likelihood function in exactly the same way we set it up last time
- We'll denote the likelihood function L
- We want to write down the likelihood of observing the data that we do (the data we haven't generated yet)

$$L(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) = \prod_{(y_i=1)} \left[\frac{\beta}{4}\right] * \prod_{(y_i=0)} \left[1 - \frac{\beta}{4}\right]$$

- We can write this in a slightly easier way:

$$L(Y_1 = y_1, \ldots, Y_N = y_N) = \left[\frac{\beta}{4}\right]^{\text{NumRed}} * \left[1 - \frac{\beta}{4}\right]^{\text{NumBlack}}$$

- Let's do it!

Likelihood function

- Now it's time to graph the likelihood for all the possible values of β
- R script (R is just a bit faster to do this in than Stata, and this is just for demonstration purposes anyway)

R code

```
# How many red balls?
numred <-
# How many black balls?
numblack <-
# Some possible values of b
b <- seq(0,4,length.out=100)
# Set up the likelihood function
L <- ( (b^numred)/(4^numred) )*(
(1-(b/4))^numblack )
# Now plot the likelihood function
plot(b,L)
```

Likelihood function

- As you can see, the likelihood function has a maximum value
- The value of β that maximizes the likelihood function is the maximum likelihood estimator of β

Likelihood function

- As you can see, the likelihood function has a maximum value
- The value of β that maximizes the likelihood function is the maximum likelihood estimator of β
- We choose the value of β that makes the data we observe to be *most likely* to be our estimate of the true parameter β

Likelihood function

- You will often see the log-likelihood used instead of the likelihood function
- Don't get nervous
- It's really the same concept
- People use *lnL* instead of *L* because it's easier to compute
- Rather than a likelihood function *L*

$$L(Y_1 = y_1, \ldots, Y_N = y_N) = \prod P_i$$

- We use the log-likelihood

$$lnL(Y_1 = y_1, \ldots, Y_N = y_N) = \sum ln(P_i)$$

Homework review
If time

- Otherwise do this in TA session

Other important discrete data models

- Multinomial logit/probit
  - Discrete number of choices
  - More than just two choices
  - Choices cannot be placed in a uniformly agreed-upon order
  - Classic example: Transportation choice
- Ordered logit/probit
  - Discrete number of choices
  - More than just two choices
  - Choices *can* be placed in a uniformly agreed-upon order
  - Classic example: Grades (A > B > C > D > F)
- FYI: models also known as “polychotomous” dependent variables models

Other important discrete data models

- There are a few other models you may encounter but multinomial logit/probit and ordered logit/probit are the most common
- Others we may get a chance to mention today (and you will read about in the textbook) are
  - Conditional logit
  - Nested logit
  - Rank-ordered logit
- Generally speaking these models are relatively easy to understand if you understand the multinomial logit/probit and ordered logit/probit

Multinomial logit/probit
Transportation choice

- People choose how to commute to work

Multinomial logit/probit
Transportation choice

- People choose how to commute to work
- Many choices (can't possibly include all of them)
- But we can include most of the relevant choices

Multinomial logit/probit
Transportation choice

- People choose how to commute to work
- Many choices (can't possibly include all of them)
- But we can include most of the relevant choices
- In Washington DC the choices are:
  - Car
  - Train
  - Metro
  - Bus
  - Bike
  - Walk
  - Others?

Multinomial logit/probit
Transportation choice

- People choose how to commute to work
- Many choices (can't possibly include all of them)
- But we can include most of the relevant choices
- In Washington DC the choices are:
  - Car
  - Train
  - Metro
  - Bus
  - Bike
  - Walk
  - Others?
- Clearly we can't use OLS because these choices don't fall on a continuum
- Clearly we can't use probit or logit because there are too many choices to fit in one model

Multinomial logit/probit
Transportation choice

- We number the alternatives, but their numbers are not indicative of any ordering
  1. Car
  2. Train
  3. Metro
  4. Bus
  5. Bike
  6. Walk
  7. Others?
- Utility to an individual of choosing an alternative is usually thought of as a linear function of the characteristics of two things:
  1. The individual doing the choosing (since different people have different preferences)
  2. The characteristics of the choice

Multinomial logit/probit

- The standard multinomial logit/probit model can only identify coefficients that vary over individuals
- Therefore we cannot identify coefficients on things that very over choices but not over individuals
- There is a separate model for this

- **Matching:** It is the matching of personal characteristics and characteristics of the choices that eventually cause an individual to select a mode of transportation
- If there are seven choices then there will be seven utilities, one associated with each choice for each individual
- Each choice has its own error term
- The probability that a given individual in our data chooses a particular alternative is equal to the probability that the utility of that alternative to that individual is greater than all the other utilities to that individual
- How to model this?

Multinomial logit

- Seven is just too damn many choices to write everything out, so I'll do an example with three
- The model generalizes easily to any number of choices

Multinomial logit

- Seven is just too damn many choices to write everything out, so I'll do an example with three
- The model generalizes easily to any number of choices
- Suppose there are three choices: A, B, and C (just shorthand for car, metro, and walk)

Multinomial logit

- Seven is just too damn many choices to write everything out, so I'll do an example with three
- The model generalizes easily to any number of choices
- Suppose there are three choices: *A*, *B*, and *C* (just shorthand for car, metro, and walk)
- We must consider one of these cases to be the *base* case
- All coefficients of the model will be relative to a base case
- Let's make *C* the base case

Multinomial logit

- Seven is just too damn many choices to write everything out, so I'll do an example with three
- The model generalizes easily to any number of choices
- Suppose there are three choices: A, B, and C (just shorthand for car, metro, and walk)
- We must consider one of these cases to be the *base* case
- All coefficients of the model will be relative to a base case
- Let's make C the base case

$$Pr(\text{individual } i \text{ chooses } A \text{ relative to } C) = exp(X_i\beta_A)$$

$$Pr(\text{individual } i \text{ chooses } B \text{ relative to } C) = exp(X_i\beta_B)$$

Multinomial logit

- If the probability of choosing A over C is $exp(X_i\beta_A)$ and the probability of choosing B over C is $exp(X_i\beta_B)$ then:
- The probability of choosing A over B is $exp(X_i\beta_A)/exp(X_i\beta_B)$

Multinomial logit

- If the probability of choosing A over C is $exp(X_i\beta_A)$ and the probability of choosing B over C is $exp(X_i\beta_B)$ then:
- The probability of choosing A over B is $exp(X_i\beta_A)/exp(X_i\beta_B)$
- Likewise the probability of choosing B over A is $exp(X_i\beta_B)/exp(X_i\beta_A)$

Multinomial logit

- If the probability of choosing A over C is $exp(X_i\beta_A)$ and the probability of choosing B over C is $exp(X_i\beta_B)$ then:
- The probability of choosing A over B is $exp(X_i\beta_A)/exp(X_i\beta_B)$
- Likewise the probability of choosing B over A is $exp(X_i\beta_B)/exp(X_i\beta_A)$
- Combine these facts with the final identity that the sum of all possible outcomes is 1 and you get a fairly simple model

$$Pr(A) = \frac{exp(X_i\beta_A)}{1 + exp(X_i\beta_A) + exp(X_i\beta_B)}$$

$$Pr(B) = \frac{exp(X_i\beta_B)}{1 + exp(X_i\beta_A) + exp(X_i\beta_B)}$$

$$Pr(C) = \frac{1}{1 + exp(X_i\beta_A) + exp(X_i\beta_B)}$$

Multinomial logit

- So just like before we have the probabilities of all possible outcomes and we have a bunch of data

Multinomial logit

- So just like before we have the probabilities of all possible outcomes and we have a bunch of data
- Form the likelihood in an analogous way:

$$L(Y_1 = y_1, \ldots, Y_N = y_N) =$$

$$\prod_{(y_i=A)} Pr(A) \prod_{(y_i=B)} Pr(B) \prod_{(y_i=C)} Pr(C)$$

Multinomial logit

- So just like before we have the probabilities of all possible outcomes and we have a bunch of data
- Form the likelihood in an analogous way:

$$L(Y_1 = y_1, \ldots, Y_N = y_N) =$$

$$\prod_{(y_i=A)} Pr(A) \prod_{(y_i=B)} Pr(B) \prod_{(y_i=C)} Pr(C)$$

- Then just like before, we choose the estimate of β that maximizes the likelihood function given our data

Multinomial logit/probit

- In Stata we use `mlogit` or `mprobit`

Stata code

```
mlogit y x1 x2 ..., baseoutcome(base_y_num)
```

- Notice: to get predictions in this model you need to specify several new variables (as there are several alternatives)
- Example: `predict option1 option2 ...`

Multinomial probit

- In Stata we use `mprobit` or `mlogit`

Stata code

```
mprobit y x1 x2 ..., baseoutcome(base_y_num)
```

- Notice: to get predictions in this model you need to specify several new variables (as there are several alternatives)
- Example: `predict option1 option2 ...`

Multinomial logit/probit

- As with logit and probit, the coefficients are not marginal effects
- In general: the signs of the coefficients need not be the same as the signs of the marginal effects!
- Unlike with binary logit/probit, the coefficients can influence *every other option*
- Use either `mfx` or `margins` to carry out post estimation evaluation of marginal effects with these models

Multinomial logit vs probit

- There's little difference for lots of models
- But unlike in the binary case, there *are* important differences

Multinomial logit vs probit

- There's little difference for lots of models
- But unlike in the binary case, there *are* important differences
- Logit is easier than probit to calculate
- In some cases this doesn't matter
- When the dset is big, however, `mprobit` may be difficult to calculate (at least take a long time)
- AND if the number of alternatives is large, `mprobitt` might fail entirely (complicated integral calculations)

Multinomial logit vs probit

- There's little difference for lots of models
- But unlike in the binary case, there *are* important differences
- Logit is easier than probit to calculate
- In some cases this doesn't matter
- When the dset is big, however, `mprobit` may be difficult to calculate (at least take a long time)
- AND if the number of alternatives is large, `mprobitt` might fail entirely (complicated integral calculations)
- But of course there are advantages to `mprobit` otherwise we wouldn't ever bother with it
- `mlogit` requires an assumption that `mprobit` does not

Multinomial logit vs probit

- Independence of irrelevant alternatives (IIA)

Multinomial logit vs probit

- Independence of irrelevant alternatives (IIA)
- Adding an additional alternative D does not change the relative probability of choosing A over C
- This might be a fine assumption in lots of cases
- But in others it can be trouble

Multinomial logit vs probit

- Independence of irrelevant alternatives (IIA)
- Adding an additional alternative *D* does not change the relative probability of choosing *A* over *C*
- This might be a fine assumption in lots of cases
- But in others it can be trouble
- IF *D* and *A* are very close substitutes, we wouldn't think this made much sense
- You would ordinarily think that adding a new alternative that is very similar to the old alternative *A* would (approximately) halve the probability of choosing *A*
- `mlogit` cannot accommodate this sometimes reasonable assumption
- In this case, use `mprobit`

Multinomial logit/probit

- All the variables you include must be individual-specific!
- Yuck!
- This means that you cannot estimate coefficients of the impact of choices that very only by the alternative
- Things like the price of taking the metro or the price of driving, etc. cannot be estimated by the `mlogit` / `mprobit` commands!

Conditional logit/probit

- Use the conditional logit command `clogit` to estimate models where variables such as price will be important
- In the basic `clogit` model ONLY alternative-specific variables are identified (exactly the opposite problem as before)
- To identify both at the same time you interact a dummy variable with each individual-specific alternative
- A shortcut way to do this is given by the `asclogit` command

Homework

- Read: 15.1, 15.2, 15.3, 15.4, 15.5, 15.7 (except 15.7.3)
- Do: Exercise 2 at the end of the chapter (p. 533)
- A few tips appear on the next slide –>

Homework
Tips for understanding the reading

- Variables that do not vary across alternatives (remember that “alternatives” are the choices indicated by the dependent variable, e.g. metro vs. car vs. walk in the transportation example) are referred to in C &T as “case-specific.”
- If it makes it easier for you to understand, you can substitute the phrase “individual-specific” for “case-specific” when you are doing the reading
- See next slide –>

Homework
Next time

- To prepare for next class have a look at section 15.9