Applied Econometrics
Lecture 11

Nathaniel Higgins

ERS and JHU

15 November 2010

Outline

Multinomial models

Unordered	Ordered
Multinomial logit	Ordered logit
Multinomial probit	Ordered probit
Conditional logit	
~~Nested logit~~	

Framework

- Outcomes for the i^{th} chooser (the i^{th} individual) are denoted y_i
- An outcome is a choice among several alternatives or options
- There are m alternatives
- The probability that the j^{th} alternative is chosen by the i^{th} chooser is given by $Pr(y_i = j) = F(x_i, \beta)$
- Note that C&T use the notation $F(x_i, \theta)$ rather than $F(x_i, \beta)$
- We want to estimate β using maximum likelihood
- We want to obtain $\widehat{\beta}^{MLE}$

Framework

- Outcomes for the i^{th} chooser (the i^{th} individual) are denoted y_i
- An outcome is a choice among several alternatives or options
- There are m alternatives
- The probability that the j^{th} alternative is chosen by the i^{th} chooser is given by $Pr(y_i = j) = F(x_i, \beta)$
- Note that C&T use the notation $F(x_i, \theta)$ rather than $F(x_i, \beta)$
- We want to estimate β using maximum likelihood
- We want to obtain $\widehat{\beta}^{MLE}$

Framework

- Outcomes for the i^{th} chooser (the i^{th} individual) are denoted y_i
- An outcome is a choice among several alternatives or options
- There are m alternatives
- The probability that the j^{th} alternative is chosen by the i^{th} chooser is given by $Pr(y_i = j) = F(x_i, \beta)$
- Note that C&T use the notation $F(x_i, \theta)$ rather than $F(x_i, \beta)$
- We want to estimate β using maximum likelihood
- We want to obtain $\widehat{\beta}^{MLE}$

Framework

- Outcomes for the i^{th} chooser (the i^{th} individual) are denoted y_i
- An outcome is a choice among several alternatives or options
- There are m alternatives
- The probability that the j^{th} alternative is chosen by the i^{th} chooser is given by $Pr(y_i = j) = F(x_i, \beta)$
- Note that C&T use the notation $F(x_i, \theta)$ rather than $F(x_i, \beta)$
- We want to estimate β using maximum likelihood
- We want to obtain $\widehat{\beta}^{MLE}$

Framework

- Outcomes for the i^{th} chooser (the i^{th} individual) are denoted y_i
- An outcome is a choice among several alternatives or options
- There are m alternatives
- The probability that the j^{th} alternative is chosen by the i^{th} chooser is given by $Pr(y_i = j) = F(x_i, \beta)$
- Note that C&T use the notation $F(x_i, \theta)$ rather than $F(x_i, \beta)$
- We want to estimate β using maximum likelihood
- We want to obtain $\widehat{\beta}^{MLE}$

Multinomial choice

- We want a framework that enables us to specify all the choices in a single model, rather than as a bunch of individual models
- Why bother with the complication? Why not just write m individual probit/logit models?

$$y = \{car, not\ car\}$$
$$y = \{metro, not\ metro\}$$
$$y = \{walk, not\ walk\}$$

Multinomial choice

- We want a framework that enables us to specify all the choices in a single model, rather than as a bunch of individual models
- Why bother with the complication? Why not just write *m* individual probit/logit models?

$$
\begin{aligned}
y &= \{car, not\ car\} \\
y &= \{metro, not\ metro\} \\
y &= \{walk, not\ walk\}
\end{aligned}
$$

Multinomial choice

- We want a framework that enables us to specify all the choices in a single model, rather than as a bunch of individual models
- Why bother with the complication? Why not just write m individual probit/logit models?

$$y = \{car, not\ car\}$$
$$y = \{metro, not\ metro\}$$
$$y = \{walk, not\ walk\}$$

Multinomial choice

- When individuals face *m* choices simultaneously and make a single choice, it makes intuitive sense to build a model that reflects this fact
- The single model enforces that the probability of choice sum to 1
- Allows us to evaluate marginal effects as they influence *all* the possible choices
- An increase in an independent variable *x* cannot increase the probability of every choice — some alternatives become more likely and so other alternatives must become *less* likely

- If an individual faces a choice of m alternatives, we think that individual i will choose alternative j if the j^{th} alternative is *better* than all other alternatives
- In economics-world, a choice is "better" if it results in a higher utility
- So ... an individual selects alternative j if alternative j yields more utility than any other available alternative

$$y_i = j \text{ if } U_{ij} > U_{ik} \; \forall k \neq j$$
$$\Rightarrow Pr(y_i = j) = Pr(U_{ij} > U_{ik}) \; \forall k \neq j$$

- If an individual faces a choice of m alternatives, we think that individual i will choose alternative j if the j^{th} alternative is *better* than all other alternatives
- In economics-world, a choice is “better” if it results in a higher utility
- So . . . an individual selects alternative j if alternative j yields more utility than any other available alternative

$$y_i = j \text{ if } U_{ij} > U_{ik} \; \forall k \neq j$$
$$\Rightarrow Pr(y_i = j) = Pr(U_{ij} > U_{ik}) \; \forall k \neq j$$

Multinomial choice
Logic of the model

- If an individual faces a choice of m alternatives, we think that individual i will choose alternative j if the j^{th} alternative is *better* than all other alternatives
- In economics-world, a choice is "better" if it results in a higher utility
- So ... an individual selects alternative j if alternative j yields more utility than any other available alternative

$$
\begin{aligned}
y_i = j \text{ if } U_{ij} > U_{ik} \; \forall k \neq j \\
\Rightarrow Pr(y_i = j) = Pr(U_{ij} > U_{ik}) \; \forall k \neq j
\end{aligned}
$$

- We want to express utility as a function of data that we observe
- Denote the "observable" utility that individual i gets from selecting alternative j by V_{ij}
- Assume that utility (like everything else) is made up of observable and unobservable parts

$$U_{ij} = V_{ij} + \epsilon_{ij} \tag{1}$$

- We model observable utility V_{ij} as a function of independent variables and parameters

$$V_{ij} = x_{ij}\beta + z_i\gamma_j \tag{2}$$

- Put (1) and (2) together and we get:

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij} \tag{3}$$

- We want to express utility as a function of data that we observe
- Denote the "observable" utility that individual i gets from selecting alternative j by V_{ij}
- Assume that utility (like everything else) is made up of observable and unobservable parts

$$U_{ij} = V_{ij} + \epsilon_{ij} \tag{1}$$

- We model observable utility V_{ij} as a function of independent variables and parameters

$$V_{ij} = x_{ij}\beta + z_i\gamma_j \tag{2}$$

- Put (1) and (2) together and we get:

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij} \tag{3}$$

- We want to express utility as a function of data that we observe
- Denote the "observable" utility that individual i gets from selecting alternative j by V_{ij}
- Assume that utility (like everything else) is made up of observable and unobservable parts

$$U_{ij} = V_{ij} + \epsilon_{ij} \tag{1}$$

- We model observable utility V_{ij} as a function of independent variables and parameters

$$V_{ij} = x_{ij}\beta + z_i\gamma_j \tag{2}$$

- Put (1) and (2) together and we get:

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij} \tag{3}$$

- We want to express utility as a function of data that we observe
- Denote the "observable" utility that individual i gets from selecting alternative j by V_{ij}
- Assume that utility (like everything else) is made up of observable and unobservable parts

$$U_{ij} = V_{ij} + \epsilon_{ij} \tag{1}$$

- We model observable utility V_{ij} as a function of independent variables and parameters

$$V_{ij} = x_{ij}\beta + z_i\gamma_j \tag{2}$$

- Put (1) and (2) together and we get:

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij} \tag{3}$$

- We want to express utility as a function of data that we observe
- Denote the "observable" utility that individual i gets from selecting alternative j by V_{ij}
- Assume that utility (like everything else) is made up of observable and unobservable parts

$$U_{ij} = V_{ij} + \epsilon_{ij} \quad (1)$$

- We model observable utility V_{ij} as a function of independent variables and parameters

$$V_{ij} = x_{ij}\beta + z_i\gamma_j \quad (2)$$

- Put (1) and (2) together and we get:

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij} \quad (3)$$

Multinomial choice
Variable types

- OK, by now you are really used to seeing dependent variables modeled as a function of some independent variables x_{ij} and some parameters β
- So what's with this new z_i and γ_j stuff?

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- In the multinomial choice world, we think that it is the *interaction* of characteristics of individual i and alternative j that determine whether individual i chooses alternative j
- Let's look at x_{ij}, β, z_i, and γ_j one at a time

Multinomial choice
Variable types

- OK, by now you are really used to seeing dependent variables modeled as a function of some independent variables x_{ij} and some parameters β
- So what's with this new z_i and γ_j stuff?

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- In the multinomial choice world, we think that it is the *interaction* of characteristics of individual i and alternative j that determine whether individual i chooses alternative j
- Let's look at x_{ij}, β, z_i, and γ_j one at a time

- OK, by now you are really used to seeing dependent variables modeled as a function of some independent variables x_{ij} and some parameters β
- So what's with this new z_i and γ_j stuff?

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- In the multinomial choice world, we think that it is the *interaction* of characteristics of individual i and alternative j that determine whether individual i chooses alternative j
- Let's look at x_{ij}, β, z_i, and γ_j one at a time

- OK, by now you are really used to seeing dependent variables modeled as a function of some independent variables x_{ij} and some parameters β
- So what's with this new z_i and γ_j stuff?

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- In the multinomial choice world, we think that it is the *interaction* of characteristics of individual i and alternative j that determine whether individual i chooses alternative j
- Let's look at x_{ij}, β, z_i, and γ_j one at a time

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- x_{ij} are those independent variables that vary over choices (they are j-specific) and *possibly* vary across individuals
- β does not vary over individuals or choices
- z_i are those independent variables that vary over individuals (*choosers*) but not over choices (that's why z doesn't get a j subscript)
- γ_j is a set of coefficients (like β) — but the γ_j coefficients vary over choices

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- x_{ij} are those independent variables that vary over choices (they are j-specific) and *possibly* vary across individuals
- β does not vary over individuals or choices
- z_i are those independent variables that vary over individuals (*choosers*) but not over choices (that's why z doesn't get a j subscript)
- γ_j is a set of coefficients (like β) — but the γ_j coefficients vary over choices

Multinomial choice
Variable types

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- x_{ij} are those independent variables that vary over choices (they are j-specific) and *possibly* vary across individuals
- β does not vary over individuals or choices
- z_i are those independent variables that vary over individuals (*choosers*) but not over choices (that's why z doesn't get a j subscript)
- γ_j is a set of coefficients (like β) — but the γ_j coefficients vary over choices

Multinomial choice
Variable types

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- x_{ij} are those independent variables that vary over choices (they are j-specific) and *possibly* vary across individuals
- β does not vary over individuals or choices
- z_i are those independent variables that vary over individuals (*choosers*) but not over choices (that's why z doesn't get a j subscript)
- γ_j is a set of coefficients (like β) — but the γ_j coefficients vary over choices

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- x_{ij} are those independent variables that vary over choices (they are j-specific) and *possibly* vary across individuals
- β does not vary over individuals or choices
- z_i are those independent variables that vary over individuals (*choosers*) but not over choices (that's why z doesn't get a j subscript)
- γ_j is a set of coefficients (like β) — but the γ_j coefficients vary over choices

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- This model of utility allows us to model the effect of several different types of variables on the utility that individual i gets from selecting choice j:
  - Variables that vary over individuals (but not over choices)
  - Variables that vary over choices (but not individuals)
  - Variables that vary over *both* individuals and choices (more rare)

- Taking stock:
- We have a model of utility (3) that expresses utility as a function of observable and unobservable stuff

$$U_{ij} = x_{ij}\beta + z_i\gamma_j + \epsilon_{ij}$$

- And we think that individual i selects choice j if the utility of choice j is greater than all other options:

$$U_{ij} > U_{ik} \ \forall k \neq j$$

- This leads us to model the probability of individual i selecting choice j as the probability of U_{ij} being greater than all other U_{ik}

$$Pr(U_{ij} > U_{ik}) \; \forall k \neq j$$

- Notice that this is the same as the probability that the difference in utilities is less than 0

$$Pr(U_{ij} > U_{ik}) \; \forall k \neq j$$
$$Pr(U_{ik} - U_{ij} < 0) \; \forall k \neq j$$

$$Pr(U_{ik} - U_{ij} < 0)\ \forall k \neq j$$

- Expand this expression by inserting the definition of U

$$Pr(V_{ik} + \epsilon_{ik} - V_{ij} - \epsilon_{ij} < 0)\ \forall k \neq j$$

- Rearrange this just a little bit to get

$$Pr(\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik})\ \forall k \neq j$$

- This is the key equation that makes our model operational

$$Pr(U_{ik} - U_{ij} < 0)\ \forall k \neq j$$

- Expand this expression by inserting the definition of U

$$Pr(V_{ik} + \epsilon_{ik} - V_{ij} - \epsilon_{ij} < 0)\ \forall k \neq j$$

- Rearrange this just a little bit to get

$$Pr(\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik})\ \forall k \neq j$$

- This is the key equation that makes our model operational

- Depending on the distribution of ϵ, we get different models
- Notice that it is the *difference* between two sets of random variables ϵ_{ij} and ϵ_{ik} that we want to know the distribution of
- If we assume that $\epsilon_{ik} - \epsilon_{ij}$ is distributed normally, then we work with probit models
- If we assume that $\epsilon_{ik} - \epsilon_{ij}$ is distributed logistic, then we work with logit models
- (For those of you who want to know: the difference between two normal random variables is itself a normal random variable; the difference between two *extreme value* random variables is a logistic random variable. This means that we can make assumptions about the ϵs themselves which lead us to model the *differences* between ϵs in a way that we find convenient.)

Multinomial logit

- Because the logit model is a bit easier to work with we will start with the multinomial logit model rather than the multinomial probit model
- We use the multinomial logit model when we are interested in the effect of independent variables that are choice-invariant on the probability of selecting choice j
- To focus on these individual-specific variables, we look at observable utility only as a function of these variables — that is, we eliminate variable that are alternative-specific:

$$\begin{aligned} U_{ij} &= x_{ij}\beta + z_i\gamma_j + \epsilon_{ij} \\ &= \cancel{x_{ij}\beta} + z_i\gamma_j + \epsilon_{ij} \\ U_{ij} &= z_i\gamma_j + \epsilon_{ij} \end{aligned}$$

Multinomial logit

- Because the logit model is a bit easier to work with we will start with the multinomial logit model rather than the multinomial probit model
- We use the multinomial logit model when we are interested in the effect of independent variables that are choice-invariant on the probability of selecting choice j
- To focus on these individual-specific variables, we look at observable utility only as a function of these variables — that is, we eliminate variable that are alternative-specific:

$$
\begin{aligned}
U_{ij} &= x_{ij}\beta + z_i\gamma_j + \epsilon_{ij} \\
&= \cancel{x_{ij}\beta} + z_i\gamma_j + \epsilon_{ij} \\
U_{ij} &= z_i\gamma_j + \epsilon_{ij}
\end{aligned}
$$

Multinomial logit

- So we have:

$$Pr(y_i = j) =$$
$$= Pr(\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik})$$

$$= F(V_{ij} - V_{ik})$$
$$= F(z_i\gamma_j - z_i\gamma_k)$$

- Factor out the z_i and we get:

$$Pr(y_i = j) =$$
$$= F(z_i(\gamma_j - \gamma_k))$$

- where F is the logistic distribution function with mean 0 and scale parameter 1 (the same distribution we used for the regular logit model)

Multinomial logit

- So we have:

$$
\begin{aligned}
Pr(y_i = j) = & \\
&= Pr(\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik}) \\
&= F(V_{ij} - V_{ik}) \\
&= F(z_i\gamma_j - z_i\gamma_k)
\end{aligned}
$$

- Factor out the z_i and we get:

$$
\begin{aligned}
Pr(y_i = j) = & \\
&= F(z_i\left(\gamma_j - \gamma_k\right))
\end{aligned}
$$

- where F is the logistic distribution function with mean 0 and scale parameter 1 (the same distribution we used for the regular logit model)

Multinomial logit

- So we have:

$$
\begin{aligned}
Pr(y_i = j) = & \\
& = Pr(\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik}) \\
& = F(V_{ij} - V_{ik}) \\
& = F(z_i\gamma_j - z_i\gamma_k)
\end{aligned}
$$

- Factor out the z_i and we get:

$$
\begin{aligned}
Pr(y_i = j) = & \\
& = F(z_i\left(\gamma_j - \gamma_k\right))
\end{aligned}
$$

- where F is the logistic distribution function with mean 0 and scale parameter 1 (the same distribution we used for the regular logit model)

Multinomial logit

- This leads us to the multinomial logit model:

$$Pr(y_i = j) = \frac{exp(z_i\gamma_j)}{exp(z_i\gamma_1) + \ldots + exp(z_i\gamma_m)}$$

- or:

$$Pr(y_i = j) = \frac{exp(z_i\gamma_j)}{\sum_k exp(z_i\gamma_k)}$$

- As you may recall, we have *m* different alternatives that individual *i* is choosing between
- Since the probabilities $Pr(y_i = 1), \ldots, Pr(y_i = m)$ sum to 1, we only need to identify $m - 1$ of the γ parameters in order to determine all the probabilities (think of the red and black balls: we only needed to know how many red balls there were to then know how many black balls there were)

Multinomial logit

- This leads us to the multinomial logit model:

$$Pr(y_i = j) = \frac{exp(z_i\gamma_j)}{exp(z_i\gamma_1) + \ldots + exp(z_i\gamma_m)}$$

- or:

$$Pr(y_i = j) = \frac{exp(z_i\gamma_j)}{\sum_k exp(z_i\gamma_k)}$$

- As you may recall, we have m different alternatives that individual i is choosing between
- Since the probabilities $Pr(y_i = 1), \ldots, Pr(y_i = m)$ sum to 1, we only need to identify $m - 1$ of the γ parameters in order to determine all the probabilities (think of the red and black balls: we only needed to know how many red balls there were to then know how many black balls there were)

Multinomial logit

- Since we only need to identify $m - 1$ of the γ_js, a convenient normalization is $\gamma_1 = 0$

$$Pr(y_i = j) = \frac{exp(z_i\gamma_j)}{\sum_k exp(z_i\gamma_k)}$$

- Insert $\gamma_1 = 0$ into the equation above and we get:

$$\begin{aligned} Pr(y_i = j) &= \frac{exp(z_i\gamma_j)}{exp(0) + exp(z_i\gamma_2) + \ldots + exp(z_i\gamma_m)} \\ &= \frac{exp(z_i\gamma_j)}{1 + exp(z_i\gamma_2) + \ldots + exp(z_i\gamma_m)} \end{aligned}$$

Multinomial logit

- Since we only need to identify $m - 1$ of the γ_js, a convenient normalization is $\gamma_1 = 0$

$$Pr(y_i = j) = \frac{exp(z_i\gamma_j)}{\sum_k exp(z_i\gamma_k)}$$

- Insert $\gamma_1 = 0$ into the equation above and we get:

$$\begin{aligned} Pr(y_i = j) &= \frac{exp(z_i\gamma_j)}{exp(0) + exp(z_i\gamma_2) + \ldots + exp(z_i\gamma_m)} \\ &= \frac{exp(z_i\gamma_j)}{1 + exp(z_i\gamma_2) + \ldots + exp(z_i\gamma_m)} \end{aligned}$$

Multinomial logit

- Put it all back together and you get something that looks very similar to a regular logit model

$$Pr(y_i = j) = \frac{exp(z_i\gamma_j)}{1 + \sum_{k=2}^{m} exp(z_i\gamma_k)}$$

- This is essentially what you see on page 498 of C&T, but C&T haven't put the normalization into the formula yet
- We use this definition of the probability of each individual i selecting choice j as the basis for our model
- Anybody remember what we do with these probabilities?

Multinomial logit

- Put it all back together and you get something that looks very similar to a regular logit model

$$Pr(y_i = j) = \frac{exp(z_i\gamma_j)}{1 + \sum_{k=2}^{m} exp(z_i\gamma_k)}$$

- This is essentially what you see on page 498 of C&T, but C&T haven't put the normalization into the formula yet
- We use this definition of the probability of each individual i selecting choice j as the basis for our model
- Anybody remember what we do with these probabilities?

Framework

- We want to estimate a multinomial model by maximum likelihood
- To do this we set up a likelihood function
- We do this just like we have set up likelihood functions before
- Remember how we set up the likelihood function with the red ball / black ball example . . .

Maximum likelihood

- Four balls in a hat
- Pull balls out of the hat N times (generate N pieces of data)
- We get data: $(y_1, y_2, \ldots, y_N)$
- Suppose (just to have a concrete example) that the data looked like this:

$$(red, red, red, black, red, black, black, red, red, red)$$

- $N = 10$ (10 draws)
- Number of red balls drawn $= 7$
- Number of black balls drawn $= 3$
- We want to know how many red balls are in the hat (the number of red balls is equal to β, the parameter to be estimated)

Maximum likelihood

- If there had been one red ball and three black balls in the hat:
- The probability of observing the data we observed would be:

$$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) \quad (4)$$

- Or, more compactly:

$$\left(\frac{1}{4}\right)^7 \times \left(\frac{3}{4}\right)^3$$

- This is the probability of the first draw being a red ball, the second draw being a red ball, etc., so that we get *exactly* the draws in the exact order that we got them in (4) above
- (Note that this is not quite the same thing as the probability of getting 7 red balls and 3 black balls)

Maximum likelihood

- If, instead, if there had been two red balls and two black balls in the hat:
- The probability of observing

$$(red, red, red, black, red, black, black, red, red, red)$$

would be:

$$(\frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4})$$

or

$$\left(\frac{2}{4}\right)^7 \times \left(\frac{2}{4}\right)^3$$

Maximum likelihood

- If there were three red balls and one black ball in the hat then the probability of getting the data we observed would be:

$$\left(\frac{3}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}\right)$$

or

$$\left(\frac{3}{4}\right)^7 \times \left(\frac{1}{4}\right)^3$$

Maximum likelihood

- Compare the probability of observing

$$(\textit{red}, \textit{red}, \textit{red}, \textit{black}, \textit{red}, \textit{black}, \textit{black}, \textit{red}, \textit{red}, \textit{red})$$

- If there had been one red ball in the hat:

$$\left(\frac{1}{4}\right)^7 \times \left(\frac{3}{4}\right)^3 = 0.0000257492$$

- If there had been two red balls in the hat:

$$\left(\frac{2}{4}\right)^7 \times \left(\frac{2}{4}\right)^3 = 0.0009765625$$

- Or if there had been three red balls in the hat:

$$\left(\frac{3}{4}\right)^7 \times \left(\frac{1}{4}\right)^3 = 0.002085686$$

Maximum likelihood

- If there had only been $\beta = 1$ red ball in the hat, the chance of observing the data that we did is about 0.003%
- If there had been $\beta = 2$ red balls in the hat, the chance of observing the data that we did is approximately 0.1%, i.e. a little better
- Finally, if there had been $\beta = 3$ red balls in the hat, the chance of observing the data that we did is about 0.2%, twice as likely as if there had been two red balls!

Maximum likelihood

- What is our *maximum likelihood* estimate of β?
- Of the choices 1, 2, or 3 red balls, which one is most likely to have produced the data we observed?
- We pick the choice that results in the highest likelihood

$$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) = 0.0000257492$$

$$(\frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}) = 0.0009765625$$

$$(\frac{3}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{4}, \frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}) = 0.002085686$$

- We choose $\widehat{\beta}^{MLE} = 3$

- OK. Leave the balls-in-a-hat example behind.
- Speaking generally:
- The maximum likelihood concept is very "portable" — the basic concept works the same in all cases
- The question we seek to answer is always the same: What βs make the data that we observed the most likely?
- We select the parameters β that are most likely to have generated the data we observe $(y_1, y_2, \ldots, y_N)$ and call those βs our estimates $\widehat{\beta}^{MLE}$
- To do this we only need to be able to form the likelihood function

- The likelihood function always takes the same basic form
- To form the likelihood we need to be able to express the probability of each "event," or piece of data y_i occurring
- We need to be able to write down the probability that each observation turns out as it does, that: $Y_i = y_i$
- We write this probability in terms of our data and call it $p(y_i|x_i, \beta)$
  - This is what we did in the balls-in-a-hat example when we wrote down the probability that a given draw $y_i = 1$ (red ball) $= \frac{\beta}{4}$ and the probability that a given draw $y_i = 0$ (black ball) $= 1 - \frac{\beta}{4}$
  - Those two things were expressions of $p(y_i = 1|x_i, \beta)$ and $p(y_i = 0|x_i, \beta)$, respectively

Maximum likelihood
How the method is used

- The likelihood function always takes the same basic form
- To form the likelihood we need to be able to express the probability of each "event," or piece of data y_i occurring
- We need to be able to write down the probability that each observation turns out as it does, that: $Y_i = y_i$
- We write this probability in terms of our data and call it $p(y_i|x_i, \beta)$
  - This is what we did in the balls-in-a-hat example when we wrote down the probability that a given draw $y_i = 1$ (red ball) $= \frac{\beta}{4}$ and the probability that a given draw $y_i = 0$ (black ball) $= 1 - \frac{\beta}{4}$
  - Those two things were expressions of $p(y_i = 1|x_i, \beta)$ and $p(y_i = 0|x_i, \beta)$, respectively

- The likelihood function always takes the same basic form
- To form the likelihood we need to be able to express the probability of each "event," or piece of data y_i occurring
- We need to be able to write down the probability that each observation turns out as it does, that: $Y_i = y_i$
- We write this probability in terms of our data and call it $p(y_i|x_i, \beta)$
  - This is what we did in the balls-in-a-hat example when we wrote down the probability that a given draw $y_i = 1$ (red ball) $= \frac{\beta}{4}$ and the probability that a given draw $y_i = 0$ (black ball) $= 1 - \frac{\beta}{4}$
  - Those two things were expressions of $p(y_i = 1|x_i, \beta)$ and $p(y_i = 0|x_i, \beta)$, respectively

- We write down each and every likelihood function in exactly the same way:

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(y_i | x_i, \beta)$$

- How to write the likelihood function for a multinomial model?
- How do we write down the probability of each occurrence in a multinomial model?

$$L(y_1, \ldots, y_N) = \prod_{i=1}^{N} p(y_i | x_i, \beta)$$

- Start with $p(y_i | x_i, \beta)$ — what are the probabilities that we'd like to express?

Multinomial logit

- We would like to express the probability of each choice being made
- If there are three alternative choices 1, 2, and 3, then we write the likelihood function:

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

$$\prod_{i \in 1} p(y_i = 1 | z_i \gamma_1) \prod_{i \in 2} p(y_i = 2 | z_i \gamma_2) \prod_{i \in 3} p(y_i = 3 | z_i \gamma_3)$$

- If there are m alternative choices, then we write the likelihood function:

$$L(y_1, \ldots, y_N) = \prod_{j=1}^{m} \prod_{i \in j} Pr(y_i = j)$$

Multinomial logit

- We would like to express the probability of each choice being made
- If there are three alternative choices 1, 2, and 3, then we write the likelihood function:

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

$$\prod_{i \in 1} p(y_i = 1 | z_i \gamma_1) \prod_{i \in 2} p(y_i = 2 | z_i \gamma_2) \prod_{i \in 3} p(y_i = 3 | z_i \gamma_3)$$

- If there are m alternative choices, then we write the likelihood function:

$$L(y_1, \ldots, y_N) = \prod_{j=1}^{m} \prod_{i \in j} Pr(y_i = j)$$

Multinomial logit

- We would like to express the probability of each choice being made
- If there are three alternative choices 1, 2, and 3, then we write the likelihood function:

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

$$\prod_{i \in 1} p(y_i = 1 | z_i \gamma_1) \prod_{i \in 2} p(y_i = 2 | z_i \gamma_2) \prod_{i \in 3} p(y_i = 3 | z_i \gamma_3)$$

- If there are m alternative choices, then we write the likelihood function:

$$L(y_1, \ldots, y_N) = \prod_{j=1}^{m} \prod_{i \in j} Pr(y_i = j)$$

Multinomial logit

- When we express the probability that $y_i = j$ in the way we just derived, we can write the likelihood function:

$$L(y_1, \ldots, y_N) = \prod_{j=1}^{m} \prod_{i \in j} \frac{exp(z_i \gamma_j)}{1 + \sum_{k=2}^{m} exp(z_i \gamma_k)}$$

- We estimate the model by selecting the γ_js that maximize the likelihood function
- We find a different set of γs for each of the $m - 1$ alternative choices (remember we have a *base* case) that make the likelihood function above as big as it can be

Multinomial logit

- When we express the probability that $y_i = j$ in the way we just derived, we can write the likelihood function:

$$L(y_1, \ldots, y_N) = \prod_{j=1}^{m} \prod_{i \in j} \frac{exp(z_i \gamma_j)}{1 + \sum_{k=2}^{m} exp(z_i \gamma_k)}$$

- We estimate the model by selecting the γ_js that maximize the likelihood function
- We find a different set of γs for each of the $m - 1$ alternative choices (remember we have a *base* case) that make the likelihood function above as big as it can be

Multinomial logit

- When we express the probability that $y_i = j$ in the way we just derived, we can write the likelihood function:

$$L(y_1, \ldots, y_N) = \prod_{j=1}^{m} \prod_{i \in j} \frac{exp(z_i \gamma_j)}{1 + \sum_{k=2}^{m} exp(z_i \gamma_k)}$$

- We estimate the model by selecting the γ_js that maximize the likelihood function
- We find a different set of γs for each of the $m - 1$ alternative choices (remember we have a *base* case) that make the likelihood function above as big as it can be

Multinomial logit

- We find these estimates, $\widehat{\gamma_2}, \ldots, \widehat{\gamma_m}$ by optimizing the likelihood function
- Luckily, Stata does the optimization for us so we don't have to worry about how to find the correct γ_js
- We use the `mlogit` command to estimate this model

Stata code

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

Multinomial probit

- The multinomial probit model works almost exactly like the multinomial logit model
- The practical difference is that the model is harder to compute (because the normal distribution is harder to work with than the logistic distribution)
- This doesn't mean much to us if we have a relatively modestly-sized dset and a relatively small number of choice alternatives (m is small)
- If, however, m is large and/or N is very large, the multinomial probit model may be extremely time-consuming to compute

Multinomial probit

- The multinomial probit model works almost exactly like the multinomial logit model
- The practical difference is that the model is harder to compute (because the normal distribution is harder to work with than the logistic distribution)
- This doesn't mean much to us if we have a relatively modestly-sized dset and a relatively small number of choice alternatives (m is small)
- If, however, m is large and/or N is very large, the multinomial probit model may be extremely time-consuming to compute

Multinomial probit

- So why compute `mprobit` at all if it's so hard?
- There is an advantage to the multinomial probit model:
- The multinomial logit model forces us to make the assumption of independence of irrelevant alternatives (IIA). The multinomial probit model does *not* force us to make this assumption
- IIA says that adding an extra alternative does nothing to change the relative probability of any other two choices
- Might be OK in some circumstances — not such a good assumption when some choices are close substitutes

Multinomial probit

- So why compute `mprobit` at all if it's so hard?
- There is an advantage to the multinomial probit model:
- The multinomial logit model forces us to make the assumption of independence of irrelevant alternatives (IIA). The multinomial probit model does *not* force us to make this assumption
- IIA says that adding an extra alternative does nothing to change the relative probability of any other two choices
- Might be OK in some circumstances — not such a good assumption when some choices are close substitutes

Multinomial probit

- So why compute `mprobit` at all if it's so hard?
- There is an advantage to the multinomial probit model:
- The multinomial logit model forces us to make the assumption of independence of irrelevant alternatives (IIA). The multinomial probit model does *not* force us to make this assumption
- IIA says that adding an extra alternative does nothing to change the relative probability of any other two choices
- Might be OK in some circumstances — not such a good assumption when some choices are close substitutes

- Done with multinomial logit/probit for now
- On to conditional logit

Conditional logit

- Rewind way back to when we made the decision to eliminate all but the individual-specific (C&T call them case-specific) regressors:

$$\begin{aligned} U_{ij} &= x_{ij}\beta + z_i\gamma_j + \epsilon_{ij} \\ &= \cancel{x_{ij}\beta} + z_i\gamma_j + \epsilon_{ij} \\ U_{ij} &= z_i\gamma_j + \epsilon_{ij} \end{aligned}$$

- What if we wanted to identify coefficients on alternative-specific variables?
- We can do this by essentially doing exactly the opposite of what we did last time

$$\begin{aligned} U_{ij} &= x_{ij}\beta + z_i\gamma_j + \epsilon_{ij} \\ &= x_{ij}\beta + \cancel{z_i\gamma_j} + \epsilon_{ij} \\ U_{ij} &= x_{ij}\beta + \epsilon_{ij} \end{aligned}$$

Conditional logit

- By essentially the exact same reasoning that lead us to model last time, we get

$$Pr(y_i = j) = \frac{exp(x_{ij}\beta)}{\sum_{k=1}^{m} exp(x_{ik}\beta)}$$

- The only real difference here is that we don't use the same normalization because now there is only one set of βs (we don't have a different set of coefficients for each choice anymore, since we are identifying parameters that vary over choices)
- What's left to do? Just have to form the likelihood function and estimate

Conditional logit

- By essentially the exact same reasoning that lead us to model last time, we get

$$Pr(y_i = j) = \frac{exp(x_{ij}\beta)}{\sum_{k=1}^{m} exp(x_{ik}\beta)}$$

- The only real difference here is that we don't use the same normalization because now there is only one set of βs (we don't have a different set of coefficients for each choice anymore, since we are identifying parameters that vary over choices)
- What's left to do? Just have to form the likelihood function and estimate

Conditional logit

- The likelihood function is written in much the same way we did before, but with the new expression for the $Pr(y_i = j)$:

$$L(y_1, \ldots, y_N) = \prod_{j=1}^{m} \prod_{i \in j} \frac{exp(x_{ij}\beta)}{\sum_{k=1}^{m} exp(x_{ik}\beta)}$$

- We estimate the model by selecting the βs that maximize the likelihood function

Conditional logit

- We find these estimates, $\widehat{\beta}$, by optimizing the likelihood function
- Luckily, Stata does the optimization for us so we don't have to worry about how to find the correct $\widehat{\beta}$s
- We use the `clogit` command to estimate this model

Stata code

```
clogit y x1 x2 ..., group(groupid)
```

- But WAIT — conditional logit models can actually be run using a different command in Stata that is actually easier to use: `asclogit`

Conditional logit

- We find these estimates, $\widehat{\beta}$, by optimizing the likelihood function
- Luckily, Stata does the optimization for us so we don't have to worry about how to find the correct $\widehat{\beta}$s
- We use the `clogit` command to estimate this model

Stata code

```
clogit y x1 x2 ..., group(groupid)
```

- But WAIT — conditional logit models can actually be run using a different command in Stata that is actually easier to use: `asclogit`

Conditional logit

- `asclogit` *is* a conditional logit model
- It just so happens that Stata has some more flexible commands that make learning `asclogit` more useful than learning `clogit`

Stata code

```
asclogit y x1 x2 ..., case(idvar)///
alternatives(altvar)
```

Conditional logit

- Let's try it so you learn how to use `asclogit` — it's easier to learn this by doing
- We'll start by using a dset that is already in the correct format

Stata code

```
use choice.dta
* Explore the data a bit first
browse
```

- What do you see? How is the data organized?

Conditional logit

- The data is organized in *long* form
- Rather than having a single observation per individual, we have many
- Each observation represents a *possible* choice an individual could have made (an alternative)
- Notice that the `id` variable helps us realize which observations belong to which individual
- We have two types of variables: ones that do not vary within an individual and ones that do
- Which are which? Make a list.

Conditional logit

- The data is organized in *long* form
- Rather than having a single observation per individual, we have many
- Each observation represents a *possible* choice an individual could have made (an alternative)
- Notice that the `id` variable helps us realize which observations belong to which individual
- We have two types of variables: ones that do not vary within an individual and ones that do
- Which are which? Make a list.

Conditional logit

- The data is organized in *long* form
- Rather than having a single observation per individual, we have many
- Each observation represents a *possible* choice an individual could have made (an alternative)
- Notice that the `id` variable helps us realize which observations belong to which individual
- We have two types of variables: ones that do not vary within an individual and ones that do
- Which are which? Make a list.

Conditional logit

- The data is organized in *long* form
- Rather than having a single observation per individual, we have many
- Each observation represents a *possible* choice an individual could have made (an alternative)
- Notice that the `id` variable helps us realize which observations belong to which individual
- We have two types of variables: ones that do not vary within an individual and ones that do
- Which are which? Make a list.

Conditional logit

- Every individual has three choices of cars to purchase: American, Japanese, or a European car
- `income` varies across individuals, but does not vary across choices. Neither do `sex` or `size`
- `car` simply identifies the type of car that is an alternative
- `choice` indicates the car selected, so it obviously varies over alternatives (only one is selected)
- `dealer` also varies across alternatives
- So what is `dealer`? `dealer` represents the number of dealerships in the "neighborhood" (or car-searching area) of the individual

Conditional logit

- Let's run a conditional logit model to determine whether the number of dealers in a neighborhood influences the probability that a particular type of car will be purchased. Don't bother controlling for anything else (yet)

Stata code

```
asclogit choice dealer, case(id)///
alternatives(car)
```

- What do you get? Notice anything?

Conditional logit

- You get one coefficient on `dealer` and one each for the estimated constant on the two non-base alternatives
- Any idea what's going on?
- What the single coefficient is telling us is simple: the effect of having more dealerships in an individual's neighborhood is positive
- But, as before, we need to calculate marginal effects using a postestimation command, since coefficients are not equal to marginal effects in *any of these models*

Stata code

```
* Calc MEs
estat mfx
```

- Interpretation?

Conditional logit

- You get one coefficient on `dealer` and one each for the estimated constant on the two non-base alternatives
- Any idea what's going on?
- What the single coefficient is telling us is simple: the effect of having more dealerships in an individual's neighborhood is positive
- But, as before, we need to calculate marginal effects using a postestimation command, since coefficients are not equal to marginal effects in *any of these models*

Stata code

```
* Calc MEs
estat mfx
```

- Interpretation?

Conditional logit

- You get one coefficient on `dealer` and one each for the estimated constant on the two non-base alternatives
- Any idea what's going on?
- What the single coefficient is telling us is simple: the effect of having more dealerships in an individual's neighborhood is positive
- But, as before, we need to calculate marginal effects using a postestimation command, since coefficients are not equal to marginal effects in *any of these models*

Stata code

```
* Calc MEs
estat mfx
```

- Interpretation?

Conditional logit

- Suppose you didn't have `asclogit`. What to do?
- Create dummy variables for each category of `car`: `tab car, gen(cartype)`
- Now run a conditional logit model using `clogit` instead of `asclogit`
- Recall what we typed before: `asclogit choice dealer, case(id) alternatives(car)`
- Instead, this time type: `clogit choice dealer cartype*, group(id)`
- (Note: `margins` must be used to get marginal effects when using `clogit`)

- What some people refer to as a *mixed* model is what we use when we want to identify individual *and* alternative-specific variables in a single model
- In Stata, we use `asclogit` to run what many people refer to as mixed models (or fixed-effects logits)
- Try this with the same data: try to identify coefficients on `dealer` (as before) and this time add coefficients on `sex` and `income` to the model:

```
asclogit choice dealer, case(id)
alternatives(car) casevars(sex income)
```

- Try `estat mfx` to examine the marginal effects

Ordered data

- Finally, we need to discuss one more type of model that you might run in to: ordered models
- The predominant models of this type are the ordered probit and ordered logit models

Ordered logit/probit

- Grades
- Credit ratings (AAA, BBB, etc.)
- This model is actually pretty intuitive
- Once again $y*$ is an unobserved variable
- Let's play out the grades example:

$$
\begin{array}{lll}
y = Fail & \text{if} & y* = X\beta + \epsilon \leq \gamma_1 \\
y = D & \text{if} & \gamma_1 < y* = X\beta + \epsilon \leq \gamma_2 \\
y = C & \text{if} & \gamma_2 < y* = X\beta + \epsilon \leq \gamma_3 \\
y = B & \text{if} & \gamma_3 < y* = X\beta + \epsilon \leq \gamma_4 \\
y = A & \text{if} & \gamma_4 < y* = X\beta + \epsilon \leq \gamma_5
\end{array}
$$

Ordered logit/probit

- You can already see how this is going
- Let's just take one example (all the rest are analogous)

$$Pr(y = C) = \gamma_2 < y* = X\beta + \epsilon \leq \gamma_3$$

$$Pr(y = C) = \gamma_2 - X\beta < \epsilon \leq \gamma_3 - X\beta$$

- We get a condition for every possible grade
- If we assume that ϵ is distributed normally we can form a likelihood function
- All we need is an expression in terms of the distribution function (like always)

Ordered logit/probit

- Look at just the F outcome first

$$Pr(y = Fail) = Pr(X\beta + \epsilon \leq \gamma_1) = Pr(\epsilon \leq \gamma_1 - X\beta) = F(\gamma_1 - X\beta)$$

$$\begin{aligned}
Pr(y = Fail) &= F(\gamma_1 - X\beta) \\
Pr(y = D) &= F(\gamma_2 - X\beta) - F(\gamma_1 - X\beta) \\
Pr(y = C) &= F(\gamma_3 - X\beta) - F(\gamma_2 - X\beta) \\
Pr(y = B) &= F(\gamma_4 - X\beta) - F(\gamma_3 - X\beta) \\
Pr(y = C) &= F(\gamma_5 - X\beta) - F(\gamma_4 - X\beta)
\end{aligned}$$

Ordered logit/probit

- Look at just the F outcome first

$$Pr(y = Fail) = Pr(X\beta + \epsilon \leq \gamma_1) = Pr(\epsilon \leq \gamma_1 - X\beta) = F(\gamma_1 - X\beta)$$

$$\begin{aligned}
Pr(y = Fail) &= F(\gamma_1 - X\beta) \\
Pr(y = D) &= F(\gamma_2 - X\beta) - F(\gamma_1 - X\beta) \\
Pr(y = C) &= F(\gamma_3 - X\beta) - F(\gamma_2 - X\beta) \\
Pr(y = B) &= F(\gamma_4 - X\beta) - F(\gamma_3 - X\beta) \\
Pr(y = C) &= F(\gamma_5 - X\beta) - F(\gamma_4 - X\beta)
\end{aligned}$$

Ordered logit/probit

- In Stata we use `oprobit` or `ologit`

Stata code

```
oprobit y x1 x2 ...
ologit y x1 x2 ...
```

- Practical differences are almost nil, but `oprobit` is more popular (for reasons that escape me)
- The likelihood functions are not appreciably different in their difficultly to solve (because there is no “base case”)

Ordered probit

- Since ordered probit is a little bit more popular, let's try some examples
- First we need some data

Stata code

```
use ordwarm2.dta
```

- Each subject in the dset was asked to evaluate the following statement: "A working mother can establish just as warm and secure of a relationship with her child as a mother who does not work."
- The response is recoded in a variable called `warm`. It has four levels: 1 = Strongly Disagree (SD), 2 = Disagree (D), 3 = Agree (A) and 4 = Strongly Agree (SA).
- Other variables in the dset include `age`, `ed`, `male`, etc.

Ordered probit

- Try to run a model where the response of the individual is expressed as a function of the survey year (in case there were cultural changes over time), and the gender, race, age, education, and occupation of the respondent

Stata code

```
oprobit warm yr89 male white age ed prst
```

Ordered probit

- Try to run a model where the response of the individual is expressed as a function of the survey year (in case there were cultural changes over time), and the gender, race, age, education, and occupation of the respondent

Stata code

```
oprobit warm yr89 male white age ed prst
```

Ordered probit

- To get the marginal effects you can use `margins` as with the other commands we have talked about today
- Just for some variety, however, try `prchange`
- Most of you have already installed `prchange` from a previous assignment
- For those of you who haven't, you can install it now

Homework

- Read section 15.9
- Do exercise 4 on p. 533