Discrete Choice Modeling William Greene Stern School of Business, New York University

Lab Session 4

Part 1. Conditional Logit and Nested Logit Models

This assignment will consist of some simple exercises with the multinomial logit model. The data for the exercises is

MultinomialChoice.lpj

This project file contains both the brand choices data used in the first set of exercises and the travel mode data used later. Altogether, there are 12,800 observations in the brand choices data. The travel mode data appear in the first 840 rows of the data area.

1. <u>Test for functional form of utility functions</u>. The discrete choice model we will use in this exercise is

U(brand) = β_1 Fashion + β_2 Quality + β_3 Price + β_4 ASC4 + ε_{brand} ,

for brand = brand1, brand2, brand3 and none. Fashion, Quality and Price are all zero for NONE, while ASCNONE is 1 for NONE and zero for the others. This is a convenient way to consider the 'none of the above' choice. We are interested in testing the hypothesis that the price enters the utility function quadratically, rather than linearly. Thus, we test for significance of an additional variable, $PriceSq = Price^2$. The commands below can carry out the test. What do you find?

```
? Test for significance of squared term
CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
    ; Rhs = Fash,Qual,Price,Asc4 $
CALC ; L0 = logl $
CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
    ; Rhs = Fash,Qual,Price,Pricesq,Asc4 $
CALC ; L1 = logl $
CALC ; list ; chisq = 2*(L1 - L0) $
```

2. <u>Structural change</u>. Are men's preferences the same as women's? We carry out the equivalent of a Chow test for structural change. What do you find. Is the null hypothesis: H_0 : $\beta_M = \beta_W$ for the vector of parameters in the model rejected or not?

SAMPLE ; All \$
CLOGIT ; For[Male = 0] ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
 ; Rhs = Fash,Qual,Price,Asc4 \$
CALC ; LogIF = LogL \$
MATRIX ; db = b ; dv = varb \$
CLOGIT ; For[Male = 1] ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
 ; Rhs = Fash,Qual,Price,Asc4 \$
CALC ; LogIM = LogL \$
MATRIX ; db = db - b ; dv = dv + varb \$
CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
 ; Rhs = Fash,Qual,Price,Asc4 \$
CALC ; LogIMF = LogL \$
CALC ; LogIMF = LogL \$

? Likelihood Ratio Test CALC ; List; Chisq = 2*(LogIM + LogLF - LOGLMF); Ctb(.95,8) \$? Wald test MATRIX; List; Wald = db'<dv>db \$? Monginal Efforts We estimate a marginal affect (of price)

3. <u>Marginal Effects</u>. We estimate a marginal effect (of price) in the MNL model. What are the estimates of the own and cross elasticities across the three brands? What is the evidence of the IIA assumption in these results?

```
    ? 3. Examine the marginal effect of price on brand choice
    SAMPLE ; All $
    CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
    ; Rhs = Fash,Qual,Price,Asc4
    ; Effects : Price (*) $
```

4. <u>Impact of a price change</u>. What would happen to the market shares of the three brands if the price of Brand 1 of shoes rose by 50%. What would happen to the market shares if the prices of all three brands rose by 50%

? ? 4. What would happen to the market shares of the three brands If the price of Brand 1 of shoes rose by 50%. What would ? ? happen to the market shares if the prices of all three ? brands rose by 50%? ? SAMPLE : All \$ CLOGIT : Lhs = Choice : Choices=Brand1.Brand2.Brand3.None ; Rhs = Fash,Qual,Price,Asc4 ; Effects : Price (*) \$ CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None ; Rhs = Fash,Qual,Price,Asc4 Simulation = * Scenario: Price (Brand1) = [*] 1.5 \$ CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None ; Rhs = Fash,Qual,Price,Asc4 Simulation = * ; Scenario: Price (Brand1, Brand2, Brand3) = [*] 1.5 \$

5. <u>**Testing for IIA**</u>. Is Brand3 an irrelevant alternative in the choice model? Given the way the data are constructed, one wouldn't think so. Here we investigate. Carry out the Hausman to test the IIA assumption using Brand 3 as the omitted alternative. What do you find?

```
?

? 5. Testing for IIA

?

SAMPLE ; All $

CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None

; Rhs = Fash,Qual,Price,Asc4 $

CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None

; Rhs = Fash,Qual,Price,Asc4 ; IAS = Brand3 $
```

6. **Functional form and marginal impact**. Do men pay more attention to fashioni than women? To investigate, we fit the choice model with a different coefficient on fashion for men and women. Then, simulate the model so as to see what happens when the variable which carries this effect into the model is zero'd out. What are the results? How do you interpret your findings?

```
?
?
?
6. Do men pay different attention to fashion than women?
? Is the difference statistically significant?
?
?
Create ; MaleFash = Male*Fash $
CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
; Rhs = Fash,Qual,Price,Asc4,MaleFash $
CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
; Rhs = Fash,Qual,Price,Asc4,MaleFash $
CLOGIT ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
; Rhs = Fash,Qual,Price,Asc4,MaleFash
; Simulation = * ; Scenario: MaleFash(*) = [*] 0 $
```

7. <u>Heteroscedastic extreme value model</u>. Fit the MNL model while allowing the variances to differ across the utily functions. First, fit the basic MNL model. Then, allow the variances to vary. Finally, allow the variances to vary across utililties and with age and sex. In each case, obtain the marginal effects with respect to price. Does the change in the model specification produce changes in the impacts?

```
?
?
? 7. Building heterogeneity into the model
?
NLOGIT (none) ; Lhs = Choice ; Choices = Brand1,Brand2,Brand3,None
; Rhs = Fash,Qual,Price,ASC4
; Effects: Price(Brand1,Brand2,Brand3) $
NLOGIT (het) ; Lhs = Choice ; Choices = Brand1,Brand2,Brand3,None
; Rhs = Fash,Qual,Price,ASC4
; Het ; Effects: Price(Brand1,Brand2,Brand3) $
NLOGIT (variance, het) ; Lhs = Choice ; Choices = Brand1,Brand2,Brand3,None
; Rhs = Fash,Qual,Price,ASC4
; Het ; Effects: Price(Brand1,Brand2,Brand3) $
NLOGIT (variance, het) ; Lhs = Choice ; Choices = Brand1,Brand2,Brand3,None
; Rhs = Fash,Qual,Price,ASC4
; Het ; Hfn = Male,age25,age39 ; Effects: Price(Brand1,Brand2,Brand3) $
```

8. <u>Constraints</u>. Test the hypothesis that the variances in the four utility functions are all equal. Since one of them is normalized to one, this is done by testing whether the first J-1 are equal. In NLOGIT's HEV model, the first set of values reported are $(\sigma_j/\sigma_J - 1)$, so the desired test can be carried out by testing the hypothesis that these three (J-1) coefficients are zero. Carry out the test using the brand choice data. What do you find?

?
?
8. Testing for homoscedasticity with a Wald test
?
NLOGIT (het) ; Lhs = Choice ; Choices = Brand1,Brand2,Brand3,None
 ; Rhs = Fash,Qual,Price,ASC4
 ; Het ; Par \$
MATRIX ; c = b(5:7) ; vc = Varb(5:7,5:7) ; List ; WaldStat = c'<vc>c \$

9. <u>Testing for variance heterogeneity</u>. Are age and sex significant determinants of the variances in the utility functions. Test the hypothesis that they are not using a likelihood ratio test. What do you find?

```
?
? 9. Testing for heterogeneity in the heteroscedasticity
?
NLOGIT (het) ; Lhs = Choice ; Choices = Brand1,Brand2,Brand3,None
    ; Rhs = Fash,Qual,Price,ASC4
    ; Het $
CALC ; LogLR = LogL $
NLOGIT (variance, het) ; Lhs = Choice ; Choices = Brand1,Brand2,Brand3,None
    ; Rhs = Fash,Qual,Price,ASC4
    ; Het ; Hfn = Male,age25,age39 $
CALC ; LogIU = LogL $
CALC ; LogIU = LogL $
CALC ; List ; LRTest = 2*(LogLU - LogLR) ; Ctb(.95,3) $
```

This part of the assignment will use the mode choice, conditional logit data. In what follows, be sure that you are only using the first 840 rows in the combined data set. The command to set this data set is

```
SAMPLE ; 1 – 840 $
```

You can see the number of observations in the current sample at the top of the project window, as shown below.



If this value is not 840 at any time, you can just issue the sample command to reset the sample.

1. Nested logit model. We begin with a simple nested logit model.

```
?
? 1. Basic nested logit
?
NLOGIT ; Lhs = Mode ; Choices=Air,Train,Bus,Car
; Rhs = One,GC,TTME,INVT,INVC
; Tree = Private(Air,Car),Public(Train,Bus)
; Show Tree $
CALC ; LOGLU = LOGL $
```

2. <u>**RU1 and RU2**</u>. These are different formulations of the model. They are not simple reparameterizations of the model, so they will not give identical results in a finite sample. Which is the appropriate to use is up to the analyst. There is no way to test the specification as a hypothesis.

```
?
? 2. Different normalization
?
NLOGIT ; Lhs = Mode ; Choices=Air,Train,Bus,Car
; Rhs = One,GC,TTME,INVT,INVC
; Tree = Private(Air,Car),Public(Train,Bus)
; RU2
; CrossTab$
```

3. <u>Constrained nested logit model</u>. Constraining the IV parameters to equal 1 returns the original multinomial logit model. Use this device to test the restriction. Note that this specification test is whether the MNL is appropriate, against the alternative of the nested logit model.

```
?
? 3. Constrain IV parameters to produce MNL model
?
NLOGIT ; Lhs = Mode ; Choices=Air,Train,Bus,Car
; Rhs = One,GC,TTME,INVT,INVC
; Tree = Private(Air,Car),Public(Train,Bus)
; IVSET:(Private,Public)=[1] $
CALC ; LOGLR = LOGL $
CALC ; List ; LRTEST = 2*(LOGLU - LOGLR) $
```

4. <u>Degenerate branch</u>. A branch that contains only one alternative is labeled 'degenerate' (for reasons lost to antiquity). The RU1 and RU2 normalizations produce different results for such models. Fit the two and examine the effect.

```
?
? 4. Degenerate branch. Two normalizations
?
NLOGIT ; Lhs = Mode ; Choices=Air,Train,Bus,Car
; Rhs = One,GC,TTME,INVT,INVC
; Tree = Fly(Air),Ground(Car,Train,Bus) $
NLOGIT ; Lhs = Mode ; Choices=Air,Train,Bus,Car
; Rhs = One,GC,TTME,INVT,INVC
; Tree = Fly(Air),Ground(Car,Train,Bus) ; RU2 $
```

5. <u>Alternative approaches to reveal scaling</u>. The nested logit model can be modified to act like the heteroscedastic extreme value buy making all branches contain one alternative. This will allow a different scale parameter in each branch. The HEV model is another way to do this. Are the results similar?

```
?
? 5. Use nested logit to reveal scaling.
?
NLOGIT ; Lhs = Mode ; Choices=Air,Train,Bus,Car
; Rhs = One,GC,TTME,INVT,INVC
; Tree = Fly(Air),Drive(Car),Rail(Train),Ride(Bus)
; IVSET: (Ride) = [1]
; Par $
NLOGIT ; Lhs = Mode ; Choices=Air,Train,Bus,Car
; Rhs = One,GC,TTME,INVT,INVC
; HET ; SDV = SA,ST,1.0,SC $
```

6. Generalized nested logit model. The GNL model is a fairly exotic formulation (not yet in the mainstream) of the nested logit model in which alternatives may appear in more than one branch. The model allocates a portion of the alternative to the various branches. We fit one here, and leave the interpretation of the resulting model to the analyst.

```
?
? 6. A Generalized Nested Logit Model
?
NLOGIT ; Lhs = Mode ; Choices=Air,Train,Bus,Car
; Rhs = One,GC,TTME,INVT,INVC
; Tree = Fast(Air,Car,Train),Public(Train,Bus)
; GNL $
7 HEV Model _ Fit on HEV model with these
```

7. <u>**HEV Model**</u>. Fit an HEV model with these data, allowing the variances of the utilities to differ across alternatives. Use a likelihood ratio test to test for equal variances. Examine the impact of the heteroscedasticity on the marginal effect of IN VEHICLE TIME (INVT).

dastic vs. Heteroscedastic Extreme Value
; Lhs = Mode
; Choices = Air,Train,Bus,Car
; Rhs = TTME,INVC,INVT,GC,One
; Effects: INVT(*) \$
; LR = LogL \$
; Lhs = Mode
; Choices = Air,Train,Bus,Car
; Rhs = TTME,INVC,INVT,GC,One
; Het
; Effects: INVT(*) \$
; LU = LogL \$
; List ; LRTEST = 2*(LU - LR); Ctb(.95,3) \$

This assignment involves a sampling of latent class models. Though there are, of course, many aspects of the underlying models, latent class modeling, itself, is fairly uncomplicated. That is, beyond the underlying models, latent class modeling involves a small number of straightforward principles. In this exercise, we will fit a handful of latent class models to different kinds of choice variables.

8. <u>Income effect on mode choice model model</u>. Does income affect the means in the utility functions of the mode choice model, or the variances? We will use a Vuong test to explore the question. The initial random utility model has

 $U_{ij} = \beta_1 TTME_{ij} + \beta_2 INVC_{ij} + \beta_3 INVT_{ij} + \beta_4 GC_{ij} + \alpha_j + \delta_j Income_i + \epsilon_{ij}$

where $Var[\varepsilon_{ij}] = \sigma^2$, the same for all utilities. The second form of the model is

$$U_{ij} = \beta_1 TTME_{ij} + \beta_2 INVC_{ij} + \beta_3 INVT_{ij} + \beta_4 GC_{ij} + \alpha_j + \epsilon_{ij}$$

where $Var[\epsilon_{ij}] = \sigma_j^2 \times exp(\gamma \text{ Income}_i)$. These models are not nested, so we cannot use a likelihood ratio test to test one against the other. We use a Vuong test, instead. We fit each model, then for each, we retrieve LogL_i, the contribution of each individual to the log likelihood. The Vuong statistic is computed by first obtaining

 $m_i \ = \ LogL_{i0} \ \text{-} \ LogL_{i1}$

where $LogL_{i0}$ and $LogL_{i1}$ are the contributions to the log likelihood for the null model and the alternative model, respectively. We then compute the Vuong statistic,

$$V = \frac{\sqrt{n} \ \overline{m}}{s_m}$$

CALC

The limiting distribution of the Voung statistic is standard normal. Large positive values (using 1.96 for 95% confidence) favor the null hypothesis, large negative values favor the alternative hypothesis. Note, in the calculations below, for a MNL model with J alternatives, NLOGIT stores the individual log likelihoods with the last alternative, in this case CAR. The CASC variable is a dummy variable which equals one for the CAR alternative, so it is a convenient device to restrict our sample to the observations we want for our computation. Carry out the test. What do you conclude?

? ? 8.Heterogeneous. Does Income affect the means or the variances? ? SAMPLE ; All \$ NLOGIT ; Lhs = Mode ; Choices = Air, Train, Bus, Car ; Rhs = TTME, INVC, INVT, GC; Rh2=One, Hinc \$ CREATE ; LOGLMean = Logl_Obs \$ NLOGIT ; Lhs = Mode ; Choices = Air,Train,Bus,Car ; Rhs = TTME, INVC, INVT, GC, one ; Het ; Hfn = HINC \$ CREATE ; LogIVar = LogI_Obs \$; V = LogIMean - LogLVar \$ CREATE REJECT ; CASC = 0 \$

; List ; Vuong = sqr(n) * xbr(v) / sdv(v) \$

Part 2. Latent Class and Random Parameter Models

This assignment will consist of some simple exercises with the multinomial logit model. The data for the exercises is

MultinomialChoice.lpj

This project file contains both the travel mode choices data used in the first set of exercises and the brand choices data used later. Altogether, there are 12,800 observations in the brand choices data. The travel mode data appear in the first 840 rows of the data area.

The first part of the assignment will use the mode choice, conditional logit data. In what follows, be sure that you are only using the first 840 rows in the combined data set. The command to set this data set is

SAMPLE ; 1 - 840 \$

You can see the number of observations in the current sample at the top of the project window, as shown below.

🔂 MultinomialChoice.lpj *		
Data: U; 44444 Rows: 840 Obs 🛛 🔫	_	
Date Date Date Date Date Date Date Date	E	
Data	1.	

If this value is not 840 at any time, you can just issue the sample command to reset the sample. When you wish to use the brand choices data in the second part of the assignment, use the command

SAMPLE ; All \$

and the sample setting will revert to the 12,800 observations for the brand choices data.

1. <u>Multinomial probit model</u>. Do the multinomial logit and multinomial probit models give similar results? You can't tell directly from the coefficient estimates because of scaling and normalization, so you have to rely on other indicators such as marginal effects. Fit a multinomial probit and a multinomial logit model, and compare the results. Note, estimation of the MNP model is <u>extremely</u> slow, so we have set it up with a very small number of replications and stopped the iterations at 25. This particular model would take 30-50 iterations to finish.

? ? 1. Multinomial Probit Model ? NLOGIT ; Lhs = Mode ; Choices = Air,Train,Bus,Car ; Rhs = TTME,INVC,INVT,GC; Rh2=One,Hinc ; Effects:GC(*) \$

```
NLOGIT ; Lhs = Mode ; MNP ; PTS = 10 ; Maxit = 25 ; Halton
; Choices = Air,Train,Bus,Car
; Rhs = TTME,INVC,INVT,GC; Rh2=One,Hinc
; Effects:GC(*) $
```

<u>The next set of computations is based on the brand choices data.</u> Be sure that the sample setting uses all the data. Use

SAMPLE ; All \$

to set the sample correctly.

Note that in these simulated data, the true underlying model really is a latent class data generating mechanishm, with three classes.

1. **Latent class model for brand choice**. First, fit a simple three class model with constant class probabilities. Then, fit the same model, but allow the class probabilities to very with age and sex. Finally, since we know that the true model is a three class model, we explore what happens when the model is over fit by fitting a four class model.

```
?
? (1) Basic 3 class model.
2
Nlogit ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
   ; Rhs = Fash,Qual,Price,ASC4
   ; LCM ; Pds = 8 ; Pts = 3 $
?
? (2) 3 class model. Class probabilities depend on covariates
Nlogit ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
    ; Rhs = Fash,Qual,Price,ASC4
    ; LCM=Male,Age25,Age39 ; Pds = 8 ; Pts = 3 $
?
? (3) Overspecified model. 4 class model. The true model
   underlying the data has three classes
?
Nlogit ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
    ; Rhs = Fash,Qual,Price,ASC4
    ; LCM ; Pds = 8 ; Pts = 4 $
```

3. **<u>Random parameters models</u>**. We fit two specifications of a random parameters model. We also test the null hypothesis that the parameters are nonrandom.

```
?
? (4) Random parameters model
?
Nlogit ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
    ; Rhs = Fash,Qual,Price,ASC4 $
CALC ; logl0 = logl $
Nlogit ; Lhs = Choice ; Choices=Brand1,Brand2,Brand3,None
    ; Rhs = Fash,Qual,Price,ASC4
    ; RPL ; Fcn= Fash(n),Price(n)
    ; Pds = 8 ; Pts = 25 $
CALC ; logl1 = logl $
CALC ; List ; chisq = 2*(logl1 - logl0) $
```

How many degrees of freedom are there for this test? Is the null hypothesis rejected?

2. <u>Error Components logit model</u>. Fit the simple brand choice model with the addition of a person specific random effect. Note that here, we will take advantage of the fact that this is a panel. The same person is observed 8 times in each choice situation, so we assume that the effect does not change from one choice setting to the next. To speed this up, for purpose of the exercise, we use only 10 points in the simulation estimator. After obtaining the estimates, interpret your estimated model.

```
?
? (5) An Error Components Logit model
?
ECLOGIT ; Lhs = Choice ; Choices = Brand1,Brand2,Brand3,None
; Rhs = Fash,Qual,Price,ASC4
; Pts = 10 ; Pds = 8
; ECM = (Brand1,Brand2,Brand3),(none) $
```

Part 3. Combining Revealed and Stated Preference Data

This short assignment consists of estimation of a model using a data set that combines stated and revealed preference data. The different scaling needed to accommodate the two parts of the data set is built into the model by using a nested logit specification. The specification below embodies many of the more advanced features of the conditional logit model, including the nesting with degenerate branches to reveal the scaling and choice based sampling in the revealed preference data. The data set is also complicated by having the choice sets vary across individuals, with each individual choosing from a possibly different subset of the master choice set.

To carry out the assignment, you need only load the sprp.lpj data set, then execute the command set below, which is replicated in LabAssignment-10.lim. Then, examine the estimates of the model components to see how the specification has (or has not) captured the important features of this data set.

/* ? ? Data for this application are in SPRP.LPJ ? ? Application of revealed/stated preference data ? Uses nested logit to handle scaling. ? Sample is also choice based, as shown by weights. ? Choice variable is CHOSEN ? Number of choices in choice set is CSET ? Specific choices from master set given by ALTIJ ? FCOST = fuel cost ? AUTOTIME = time spent commuting by car. ? Numerous other variables in the data set are not used here. ? */ NLOGIT ;lhs=chosen,cset,altij ;choices=RPDA,RPRS,RPBS,RPTN,SPDA,SPRS,SPBS,SPTN,SPLR,SPBW /.592,.208,.089,.111, 1.0, 1.0, 1.0, 1.0, 1.0,1.0,1.0 ;tree=Commute [rp (RPDA,RPRS,RPBS,RPTN), spda(SPDA), sprs(SPRS), spbs(SPBS), sptn(SPTN),

splr(SPLR),spbw(SPBW)] ;ivset: (rp)=[1.0] ;ru1 ;maxit=150

;model:

U(RPDA) = rdasc + invc*fcost+tmrs*autotime /

U(RPRS) = rrsasc + invc*fcost+tmrs*autotime /

U(RPBS) = rbsasc + invc*mptrfare+mtpt*mptrtime/

U(RPTN) = cstrs*mptrfare+mtpt*mptrtime/

U(SPDA) = sdasc + invc*fueld + tmrs*time+cavda*carav /

U(SPRS) = srsasc + invc*fueld + tmrs*time/? cavrs*carav/

U(SPBS) = invc*fared + mtpt*time +acegt*spacegtm/

U(SPTN) = stnasc + invc*fared + mtpt*time+acegt*spacegtm/

U(SPLR) = slrasc + invc*fared + mtpt*time+acegt*spacegtm/

U(SPBW) = sbwasc + invc*fared + mtpt*time+acegt*spacegtm\$