

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

**Lab Session 6**

**Part I. Combining Revealed and Stated Preference Data**

This short exercise 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 can be built into the model by using a nested logit specification. The specification below embodies some 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.

**Data for this application are in sprp.lpj**

The data set is a survey sample of 2,688 trips in Sydney, Australia, 2 or 4 choices per situation. The sample consists of 672 individuals, 9408 observations in total. The choice situations are a revealed choice case in which the choices are as follows:

Revealed choice experiment:  
Revealed: Drive, ShortRail, Bus, Train

The revealed choice is followed by one or two hypothetical choice situations in which the individual chooses from among 4 of six experimental choices:

Hypothetical choice experiment:  
Drive, ShortRail, Bus, Train, LightRail, ExpressBus

Data Editor					
101/900 Vars: 11111 Rows: 9408 Cl: Cell: 0					
	ID	CITY	SPRP	SPEXP	ALTIJ
1 »	1000	1	1	0	1
2 »	1000	1	1	0	4
3 »	1000	1	2	1	5
4 »	1000	1	2	1	6
5 »	1000	1	2	1	8
6 »	1000	1	2	1	10
7 »	1000	1	2	2	5
8 »	1000	1	2	2	6
9 »	1000	1	2	2	9
10 »	1000	1	2	2	10
11 »	1000	1	2	3	5
12 »	1000	1	2	3	6
13 »	1000	1	2	3	7
14 »	1000	1	2	3	8
15 »	1001	1	1	0	1

**Each person makes four choices from a choice set that includes either two or four alternatives.**

**The first choice is the RP between two of the RP alternatives**

**The second-fourth are the SP among four of the six SP alternatives.**

**There are ten alternatives in total.**

The choice attributes in the model are

Cost –Fuel or fare  
Transit time  
Parking cost  
Access and Egress time

The revealed preference are a choice based sample. The sample market shares for the RP choices differ systematically from the known shares, .592,.208,.089,.111. The role of these in the models is seen in the statement of the list of the choices in the model commands below.

```
?
? 1. Nested logit to reveal scaling difference between RP and SP choices
?
? Sample is 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
;tree=Commute [ rp (RPDA,RPRS,RPBS,RPTN),
spda(SPDA), sprs(SPRS), spbs(SPBS),
sptn(SPTN), splr(SPLR), spbw(SPBW)]
;ivset: (rp)=[1.0] ;rul ;maxit=50
;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*fueled + tmrs*time + cavda*carav /
U(SPRS) = srsasc + invc*fueled + tmrs*time/
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 $

?
? 2. Using only Revealed Preference Data. Simple MNL
?
NLOGIT ; if[sprp = 1] ? Using only RP data
; lhs=chosen,cset,altij ; choices=RPDA,RPRS,RPBS,RPTN
; model:
U(RPDA) = rdasc + fl*fcost + tm*autotime/
U(RPRS) = rrsasc + fl*fcost + tm*autotime/
U(RPBS) = rbsasc + ptc*mptrfare + mt*mptrtime/
U(RPTN) = ptc*mptrfare + mt*mptrtime$

?
? 3. Using only Stated Preference Data. Simple MNL
?
SAMPLE ; all$
NLOGIT ; if[sprp = 2] ? Using only SP data
; lhs=chosen,cset,alt ; choices=SPDA,SPRS,SPBS,SPTN,SPLR,SPBW
; crosstab
; model:
U(SPDA) = dasc + cst*fueled + tmcar*time + prk*parking
+ pincda*pincome +cavda*carav/
U(SPRS) = rsasc+cst*fueled + tmcar*time + prk*parking/
U(SPBS) = bsasc+cst*fared + tmpt*time + act*acctime + egt*eggttime/
U(SPTN) = tnasc+cst*fared + tmpt*time + act*acctime + egt*eggttime/
U(SPLR) = lrasc+cst*fared + tmpt*time + act*acctime + egt*eggttime/
U(SPBW) = cst*fared + tmpt*time + act*acctime + egt*eggttime$
```

```

?
? 4. Using all data. Nested logit reveals scaling
?
SAMPLE ; All$
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
        ; tree=mode[rp(RPDA,RPRS,RPBS,RPTN),spda(SPDA),
          sprs(SPRS),spbs(SPBS),sptn(SPTN),splr(SPLR),spbw(SPBW)]
        ; ivset: (rp)=[1.0] ; rul
        ; maxit = 50
        ; model:
U(RPDA) = rdasc + invc*fcost + tmrs*autotime
          + pinc*pincome + CAVDA*CARAV/
U(RPRS) = rrsasc + invc*fcost + tmrs*autotime/
U(RPBS) = rbsasc + invc*mptrfare + mtpt*mptrtime/
U(RPTN) = cstrs*mptrfare + mtpt*mptrtime/
U(SPDA) = sdasc + invc*fueled + tmrs*time+cavda*carav
          + pinc*pincome/
U(SPRS) = srsasc + invc*fueled + tmrs*time/
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$

```

```

?
? 5. Using all data. Random parameters model connects choice
? situations.
?
SAMPLE ; All$
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
        ; rpl ; pds=4 ; halton ; pts=25 ; fcn=invc(n)
; model:
U(RPDA) = rdasc + invc*fcost + tmrs*autotime
          + pinc*pincome + CAVDA*CARAV/
U(RPRS) = rrsasc + invc*fcost + tmrs*autotime/
U(RPBS) = rbsasc + invc*mptrfare + mtpt*mptrtime/
U(RPTN) = cstrs*mptrfare + mtpt*mptrtime/
U(SPDA) = sdasc + invc*fueled + tmrs*time+cavda*carav
          + pinc*pincome/
U(SPRS) = srsasc + invc*fueled + tmrs*time/
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$

```

## Part II. Student Project: A Discrete Choice Model for Health Care and Moral Hazard

This exercise will suggest some open ended investigations for you to carry out on your own, to design and use a discrete choice model.

This set of exercises uses the health care data contained in `healthcare.lpj`.

A. We are interested in the variable `HSAT`, which contains the answer to a general question about the individuals satisfaction with their health. This is an ordered outcome coded 0, 1, ..., 10.

1. Using the entire sample, formulate an appropriate model for health satisfaction, and estimate the parameters of your model. Include `; MARGINAL EFFECTS` in your model command so you can obtain a listing of the partial effects of the variables in your model. You might want a more elaborate analysis of the partial effects, such as given by

```
PARTIALS ; effects: ..the variable & other variable = low(change)high  
; Outcome = one of the specific values that HSAT takes $
```

or

```
PARTIALS ; effects: .. the variable | other variable = v1,v2,... ; plot(ci)  
; Outcome = one of the specific values that HSAT takes $
```

(The `; Plot`, which can be used in either command, will produce a plot of the partial effect with confidence limits.) Now, using what we know to be an inappropriate estimator, fit your model by OLS and compare the coefficients you get (a) to the ordered choice model coefficients and (b) to the partial effects.

2. The sample contains roughly equal numbers of men and women. The variable `FEMALE` is a dummy variable which equals 1 for women and 0 for men. You can use this variable to deal with the samples of men and women separately. Test the hypothesis that the same model that you specified in part 1 applies to both men and women, versus the alternative hypothesis that different equations should be fit for the two sexes. Use a likelihood ratio test to carry out the test.
3. Your professor is uncertain whether income and education are significant determinants of health satisfaction. Test the hypothesis, separately for men and women. What do you find?

B. We are going to analyze the individual's choice of whether to obtain public insurance (`PUBLIC`). (We will motivate this in the next exercise.) This is a binary choice.

1. Formulate an appropriate model and analyze this choice. What variables should appear in the model. Use both probit and logit models and compare your results. Does it matter?
2. As in part A. we are interested in whether the model differs for men and women. Fit the model separately for men and women and test whether the two groups can be described by the same model. Use a likelihood ratio test.
3. A middle ground in part 2 might be to fit the same model for both men and women, but allow certain variables to impact the outcome differently. Examine this approach using

the education variable. Instead of having a single constant in your model, include the two dummy variables FEMALE and MALE = 1 – FEMALE. (You must CREATE MALE.) Now, instead of including EDUC in the equation, include the two variables MALE\*EDUC and FEMALE\*EDUC. Now, fit your binary choice model using the entire sample and with these four variables in the model. Request the partial effects. Do you find a noticeable difference between the marginal effects for the two education variables?

- C. The study from which this data set is taken was concerned with the variables DocVis (number of visits to the doctor) and HospVis (number of hospital visits). During our discussion, we have also used created variables Doctor = 1(DocVis > 0) and Hospital = 1(HospVis > 0). The authors of the paper were interested in the question of moral hazard, that is, whether the presence of insurance significantly determined usage of the health care system. They focused on the variable ADDON, which is an indicator of whether the person has private insurance. For purpose of this exercise, we will examine variable PUBLIC, which is an indicator of whether the individual has public insurance. (One must have public insurance to obtain the Addon insurance.) For this exercise, you are to examine the issue of moral hazard in the context of a count model (Poisson or negative binomial) for DocVis and a binary choice model (probit or logit) for Doctor. The question is whether the insurance variable is a significant determinant of the dependent variable. Formulate a model to test the hypothesis. What else should be in the equation? Carry out the test using both count data models and the binary choice models. Carry out the analysis for men and women separately. What do you find?
- D. As we discussed in class, these are panel data. (The group count variable is called TI.) The hypothesis tests you carried out in part B are influenced by the effect of the clustering of the observations in the panel. Devise an appropriate panel data treatment (cluster, robust inference), fixed effects, or random effects, and repeat the analysis. Do your results change?