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

Lab Session 1 Assignment
Basic Regression

This first assignment will help you get started using NLOGIT with some familiar estimation and analysis computations. This assignment is based on the German health care data discussed in class. They are an unbalanced panel of 7,293 households observed in 1 to 7 years from 1984 to 1994 (with a couple gaps).

Load the health care panel data: project file healthcare.lpj

For most of our present purposes, however, we will treat them as a cross section of 27,326 observations. Since these are panel data, we define them as a panel now – later it will be convenient to move back and forth between panel and pooled data treatments.

```
SETPANEL    ; group = id ; pds = ti $ ti = the group size
CREATE      ; t = ndx(id,1) $           t  = the within group index, 1,2,...,Ti
CREATE      ; yr = map(year) $ year - 1983 ; if(yr=8)yr=6 ; if(yr=11)yr=7 $
```

1. Descriptive Statistics

First, let's take a look at the data. Use

```
DSTAT       ; rhs = * $  (The * means all variables in the active data set.)
```

to get some descriptive statistics for the variables in the data set. The analysis below will focus on the income variable. We are going to use log(income) as the dependent variable in the regressions. The following describes this variable and examines whether it appears to be normally distributed by comparing it to a random sample of draws from the normal distribution with the same mean and standard deviation. The kernel estimator uses only the 1994 data.

```
CREATE      ; loginc=log(income)
CALC        ; if[year = 1994] incbar=xbr(loginc);sdv=sdvinc(loginc)$
CREATE      ; normal=rnn(incbar,sdvinc)$
KERNEL      ; if[year = 1994] ; rhs=loginc,normal
            ; title=Kernel Estimator for Log of Income ; grid $
```

(Notice that the kernel estimator for log income seems to be to the right of the normal distribution. Perhaps incomes in Germany were growing in this period. We can learn a bit more about the income in the data by using boxplots. The following produces boxplots for the incomes of the female household heads by year. Depending on how they are draws, boxplots can be distorted by extreme observations in any data set. In the following, we take a simple strategy, and just restrict attention to a range that includes most of the data. The set of plots reveals both the skewed nature of the distribution and the upward trend.

```
BOXPLOT        ; if [Female=1 & Income < 2]
               ; rhs = Income
               ; str = year
               ; labels = 1984,1985,1986,1987,1988,1991,1994
               ; title = Boxplots of Income for Females by Year in GSOEP Data$
```

I am also interested in the education variable. In the original data, education is coded in part years, so a histogram is not very pretty. I will look at the full years of education by converting **educ** to integers.

```
HISTOGRAM      ; if [ year = 1994] ; rhs = educ $
CREATE         ; yrseduc = int(educ) $
HISTOGRAM      ; if[year = 1994] ; rhs = yrseduc ; title=Full Years of Education $
```

The distinctive lump at 18 years in the figure probably shows that the data are censored – it appears that education above 18 years is coded as 18. A histogram for a continuous variable will only look good if the data really are continuous. It is better to use a kernel estimator. You can try typing the command

```
KERNEL         ;if[year=1994];rhs=educ$
```

somewhere in the editing window and submitting it to see the effect. Note the mode near 18 years.

2. Linear Regression and Testing Hypotheses

For this exercise, we will pool the data and not explicitly use the panel aspect. For convenience, we define a couple of namelists with

```
NAMELIST       ;  demogrfc = age, female, married $
NAMELIST       ;  years = year1984, year1985,year1986,year1987,year1988,year1991 $
```

(We have omitted year1994, so year1994 is the base year.)

To start, we are going to do some linear regression modeling using the variable income as the dependent variable. We will fit a simple least squares regression with

```
REGRESS        ; lhs = loginc ; rhs = one, demogrfc, years $
CALC           ; rsq0 = rsqrd $
```

(An alternative way to handle a categorical variable is to use the internal procedures. We have a variable YR which indexes the years. We can use

```
REGRESS        ; lhs = loginc ; rhs = one, demogrfc, #yr $
```

Does education help to explain the variation in income? Add educ to the regression and test the hypothesis that the coefficient on education equals zero using an F test.

```
REGRESS        ; lhs = loginc ; rhs = one, demogrfc, years, educ $
CALC           ; rsq1 = rsqrd $
CALC           ; list ; fstat = ((rsq1 – rsq0)/1) / ((1-rsq1)/(n-kreg) $
CALC           ; list ; Ftb(.95,1,(n-kreg))$
```

What did you find? Note, the results contain two statistics for carrying out this test, the F statistic and a t statistic reported with the regression results. What are the results? The last instruction retrieves the critical value from the F table in case we do not remember it.

There are a variety of ways to test hypotheses. The program will compute a Wald (chi squared) statistic for you as part of the command. In the regression above, add **;Test:educ=0** to the command and resubmit it. Now, test the joint hypothesis that neither gender nor education are significand in the model. Use **;Test:educ=0,female=0.** This arrangement can also be used to set up constraints and test individual hypotheses.

```
REGRESS     ; lhs = loginc ; rhs = one,demogrfc,years ; cls: married = 0 $
```

Test the hypothesis that the three coefficients in **demogrfc** all equal zero. What do you find? There is a yet easier way to do this:

```
REGRESS     ; lhs = income ; rhs = one,demogrfc,years; cluster=id ; test : demogrfc$
```

We also want to test for the presence of 'time' effects in the regression model. There are several ways to do this: (Easiest) There is a built in function. Recall that years is the set of dummy variables, collected in a namelist. We can do the following Wald test using our robust covariance matrix:

```
REGRESS     ; lhs = loginc ; rhs = one,demogrfc,years ; test: years $
```

(More transparent) We can use matrix algebra. The **REGRESS** command provides the coefficients (matrix B) and covariance matrix (VARB) for us to use in matrix algebra and other commands. For example, in the regression command, the years variables are the 5^{th} to 10^{th} variables. We can use

```
MATRIX      ; by=b(5:10) ; vy=varb(5:10,5:10) $
MATRIX      ; list ; wld = by'<vy>by $
```

3. Partial Effects.

Consider the nonlinear regression model

$$\text{Loginc} = \beta_1 + \beta_2\text{Age} + \beta_3\text{Educ} + \beta_4\text{Female} + \beta_5\text{Age*Educ} + \beta_6\text{Age}^2 + \beta_7\text{Age*Female} + \beta_8\text{Educ*Female} + \varepsilon$$

What are the partial effects of Age and Educ on Loginc? Differentiating, we get

$$\partial\text{Loginc}/\partial\text{Age} = \beta_2 + \beta_5\text{Educ} + 2\beta_6\text{Age} + \beta_7\text{Female}$$

$$\partial\text{Loginc}/\partial\text{Educ} = \beta_3 + \beta_5\text{Age} + \beta_8\text{Female}.$$

What is the male – female income differential?

$$(\text{Loginc}|\text{Female=1}) – (\text{Loginc}|\text{Female=0}) = \beta_4 + \beta_7\text{Age} + \beta_8\text{Educ}.$$

How can you compute these and obtain standard errors for them? There are built in functions. First fit the regression with the interaction terms made explicit.

```
REGRESS       ; Lhs = loginc ; rhs = one,age, educ, female, age*educ,
                         age^2, age*female, educ*female $
```

(a) Effect of age computed for education fixed at 12,14,16,18,20, and averaged over sample observations.

```
PARTIAL       ; effects: age | educ = 12,14,16,18,20 $
```

(b) Effect of education computed for ages of 25, 28, 31, …, 64. Plot of the values with confidence intervals.

```
PARTIAL       ; effects: educ & age = 25(3)64 ; plot(ci) $
```

(c) Effect for female, for three levels of education, age 25 to 64 at each education level. Plots of three sets of values. We compute the partial effects and the predictions of the regression.

```
PARTIAL       ; effects:   female | educ = 12,16,20 & age = 25(5)64 ; plot $
SIMULATE      ; scenario: female | educ = 12,16,20 & age = 25(5)64 ; plot $
```

(d) When the model contains a set of categories, such as levels of education, say coded with 4 dummy variables: LTHS (less than high school), HS (high school), COLL (college) or GRAD (postgraduate), the partial effects for each dummy variable compute the effect relative to the base category. It might be interesting to compute the other partial effects. For example, suppose that LTHS is the base. We might compute the impact on income of achieving some college education. The following shows how to compute such a 'transition matrix.' In the regression, the educ variable is replaced by the group of variables, in the primary effect and in the interactions.

```
? Examine threshold effects of education
    CREATE        ; LTHS   = YrsEduc < 12
                  ; HS     = YrsEduc = 12
                  ; COLL   = (yrseduc > 12)*(yrseduc<=16)
                  ; GRAD   = yrseduc > 16 $
    NAMELIST      ; degree  = LTHS,HS,COLL,GRAD $
    REGRESS       ; lhs = income
    ? Note dot after degree. Drops last category when it is expanded.
                  ; rhs = one,age,degree., female, degree.*age,
                        age^2, age*female,  degree.*female $
    Partials  ; effects: degree ; transition $
```

```
-----------------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=INCOME   Mean                 =          .35214
             Standard deviation   =          .17686
----------   No. of observations  =           27326  DegFreedom   Mean square
Regression   Sum of Squares       =         86.6414          13       6.66473
Residual     Sum of Squares       =         768.040       27312        .02812
Total        Sum of Squares       =         854.682       27325        .03128
----------   Standard error of e  =          .16769  Root MSE          .16765
Fit          R-squared            =          .10137  R-bar squared     .10095
Model test   F[ 13, 27312]        =       237.00187  Prob F > F*       .00000
--------+--------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
  INCOME|  Coefficient       Error       z     |z|>Z*         Interval
--------+--------------------------------------------------------------------
Constant|    -.23222***      .02362    -9.83  .0000     -.27851    -.18593
     AGE|     .02883***      .00086    33.44  .0000      .02714     .03052
    LTHS|     .08552***      .01835     4.66  .0000      .04956     .12148
      HS|     .10676***      .02444     4.37  .0000      .05885     .15466
    COLL|     .02968         .02126     1.40  .1627     -.01198     .07135
  FEMALE|     .01132         .01130     1.00  .3166     -.01083     .03346
LTHS*AGE|    -.00562***      .00042   -13.30  .0000     -.00645    -.00479
  HS*AGE|    -.00458***      .00056    -8.19  .0000     -.00568    -.00349
COLL*AGE|    -.00233***      .00050    -4.67  .0000     -.00330    -.00135
 AGE^2.0|    -.00026***   .8780D-05   -29.80  .0000     -.00028    -.00024
        |Interaction AGE*FEMALE
Intrct05|    -.00089***      .00018    -4.84  .0000     -.00126    -.00053
        |Interaction LTHS*FEMALE
Intrct06|     .02469***      .00878     2.81  .0049      .00748     .04190
        |Interaction HS*FEMALE
Intrct07|     .02505**       .01177     2.13  .0334      .00197     .04812
        |Interaction COLL*FEMALE
Intrct08|     .00910         .01089      .84  .4032     -.01224     .03044
--------+--------------------------------------------------------------------
Note: nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
-----------------------------------------------------------------------------
---------------------------------------------------------------------
Partial Effects  Analysis for Linear Regression Function
---------------------------------------------------------------------
Effects of switches between categories in DEGREE   (dummy variables)
Results are computed by average over sample observations
LTHS    = .7731   HS      = .0632   COLL    = .0950   GRAD    = .0687
---------------------------------------------------------------------
df/dDEGREE         Partial    Standard
From --> To        Effect      Error     |t|  95% Confidence Interval
---------------------------------------------------------------------
LTHS     LTHS        .00000     .00000     .00      .00000      .00000
LTHS     HS          .06668     .00442   15.10      .05802      .07533
LTHS     COLL        .08023     .00386   20.81      .07268      .08779
LTHS     GRAD        .14743     .00444   33.23      .13873      .15613
LTHS     (Other)     .09679     .00261   37.09      .09167      .10190
HS       LTHS       -.06668     .00442   15.10     -.07533     -.05802
HS       HS          .00000     .00000     .00      .00000      .00000
HS       COLL        .01356     .00563    2.41      .00252      .02459
HS       GRAD        .08075     .00604   13.38      .06892      .09258
HS       (Other)    -.04773     .00439   10.86     -.05635     -.03912
COLL     LTHS       -.08023     .00386   20.81     -.08779     -.07268
COLL     HS         -.01356     .00563    2.41     -.02459     -.00252
COLL     COLL        .00000     .00000     .00      .00000      .00000
COLL     GRAD        .06719     .00564   11.90      .05613      .07826
COLL     (Other)    -.06439     .00383   16.79     -.07191     -.05688
GRAD     LTHS       -.14743     .00444   33.23     -.15613     -.13873
GRAD     HS         -.08075     .00604   13.38     -.09258     -.06892
GRAD     COLL       -.06719     .00564   11.90     -.07826     -.05613
GRAD     GRAD        .00000     .00000     .00      .00000      .00000
GRAD     (Other)    -.13471     .00441   30.52     -.14337     -.12606
---------------------------------------------------------------------
* (Other) = conditional share weighted average of all switch effects
---------------------------------------------------------------------
```

```
-----------------------------------------------------------------
Partial Effects Transition Matrix for DEGREE
There are  4 categories (sample %)
  01=LTHS     (77.31) 02=HS       ( 6.32) 03=COLL     ( 9.50)
  04=GRAD     ( 6.87)
Entry = effect on outcome of switch from row category to column
Switch to Other is unspecified switch out of row category
--------+--------------------------------------------------------
        |     01     02     03     04  Other
--------+--------------------------------------------------------
    LTHS|   .000   .067   .080   .147   .097
      HS|  -.067   .000   .014   .081  -.048
    COLL|  -.080  -.014   .000   .067  -.064
    GRAD|  -.147  -.081  -.067   .000  -.135
--------+--------------------------------------------------------
```

4. Panel Data

We will examine panel data later in the course. We'll take a brief look at some of the operations here. The GSOEP data are a panel. There is probably correlation across observations, which may mean that although least squares is consistent, the standard errors need correcting. Do we see 'cluster effects' in the standard errors? We consider two approaches. In the first, we correct the OLS standard errors for the correlation across observations in a group. In the second, we use the fixed and random effects approaches to fit the model.

```
SAMPLE ;all $
REGRESS       ; lhs = loginc ; rhs = one,demogrfc,years ; Table = OLS $
REGRESS       ; lhs = loginc ; rhs = one,demogrfc,years ; cluster=id ; Table = Cluster $
MAKETABLE     ; OLS,Cluster ; Standard errors$
REGRESS       ; lhs = loginc ; rhs = one,demogrfc,years ; panel $
```

Note, it is only necessary to add ;Panel to the command to request the estimators. At the very beginning of this exercise, we used **SETPANEL** to declare the form of the panel. It is also possible to request just fixed effects with **;Fixed** or random effecs with **;Random**. Notice the treatment of a time invariant variable in the fixed effects model.