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

Lab Session 1 Assignment
Basic Regression Modeling

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 farms observed in 1 to 7 years. For present purposes, however, we will treat them as a cross section of 27,326 observations.

Load the project file healthcare.lpj

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

```
DSTAT ; rhs = * $
```

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.

```
CREATE ; loginc=log(income)
CALC ; xbar=xbr(loginc);sdv=sdv(loginc)$
CREATE ; normal=rnn(xbar,sdv)$$
KERNEL ;rhs=loginc,normal$
```

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.)

2. 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 $
```

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) $
```

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?

3. Least absolute deviations (LAD) is an alternative to least squares. It works well in small samples, but provides little or no benefit in large samples. This exercise will demonstrate two aspects of estimation using LAD. (1) In a given sample, the LAD estimator is unique. But, the standard errors must be computed using bootstrapping. You must specify the number of bootstrap replications. Use **;NBT=the number**, for example **;NBT = 50**. We will use LAD on a subsample of the data. (If you try to use all 27,326 observations, the program will give an error message.)

```
SAMPLE ; 1 - 100 $  (use the first 100 observations)
REGRESS ; lhs = loginc ; rhs = one,demogrfc $
REGRESS ; lhs = loginc ; rhs = one,demogrfc ; alg=LAD ; nbt=25 $
```

Do the results differ much from OLS? (2) Since the LAD estimator of the standard errors is based on bootstrapping, a random number generator is used to generate the replications. Unfortunately, this means that if you run the regression again, you will get a different answer.

```
REGRESS ; lhs = loginc ; rhs = one,demogrfc ; alg=LAD ; nbt=25 $
```

Try it. Just submit the second **REGRESS** command again, and compare the two sets of results. The solution to this problem is to reset the seed of the random number generator to a specific value (any specific value) before using the bootstrap estimator. Use

```
CALC ; ran ( the value you choose) $ (use a 5 digit odd number)
```

Now compute the regression. Finally, execute the **CALC** command again and redo the regression.

4. These 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? Try this

```
SAMPLE ;all $
REGRESS ; lhs = loginc ; rhs = one,demogrfc,years ; cluster=id $
```

5. Hypothesis tests. We want to test for the presence of 'time' effects in the regression model. There are several ways to do this

a. (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 ; cluster=id  ; test: years = 0 $
```

b. (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 5th to 10th variables. We can use

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

c. 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$
```

d. This arrangement can also be used to set up constraints and test individual hypotheses.

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

6. Partial Effects. Consider the nonlinear regression model

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

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 falues with confidence intervals.

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

(3) Effect for female, for three levels of education, age 25 to 64 at each education level. Plots of three sets of values.

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

7. Finally, we will examine the residuals from a regression are nearly normally distributed. (This test is often applied to vectors of least squares residuals.) We can use a chi-squared sort of statistic to 'test' for nonormality. The test is based on the third and 4th moments of the variable – they should be 0 and 3, respectively. The test statistic is

$$C = N \times [\,(m_3/s^3)^2 / 6 \quad + \quad (m_4/s^4 - 3)^2/24]$$

where N is the sample size, su is the standard deviation of the residuals and m_3 and m_4 are the third and fourth sample moments. (We can't use the name '**s**' because like '**B**' and '**VARB**' it is a program reserved name.) After obtaining the residuals as above, you can compute the parts with the following, which is the Bowman and Shenton test – when applied to regression residuals, it is usually attributed to Bera and Jarque.

```
 REGRESS ; Lhs = loginc ; Rhs = one,demogrfc,years ; Res=e ; quietly $
CREATE ; v2 = e^2 ; v3 = v2*e ; v4 = v3*e $
CALC      ; sv = sqr (xbr (v2)) ;  m3 = xbr(v3) ; m4 = xbr (v4) $
CALC      ; List ; Chisq = N *( ( m3/sv^3)^2 / 6  +  ((m4/sv^4) – 3)^2 /24 )  $
```

An alternative approach sometimes used is the Kolmogorov-Smirnov test. The K-S test compares the empirical cdf of the data to that of the normal distribution with the same mean and variance.

```
CALC    ; List ; kst(e) $
```

Both tests strongly reject normality. Maybe a look at the data will help to explain the finding.

```
KERNEL ; Rhs = e $
```

It looks rather ‘normal.’ It is the long tail at the left that is the culprit.

8. These data are an interesting panel. We can explore the familiar approaches to modeling panel data. We start with the conventional estimators. (SETPANEL identifies the configuration of the data so the program can handle balanced or unbalanced panels conveniently.)

```
SAMPLE ; all $
SETPANEL ; group = id ; pds = ti $
REGRESS  ; Lhs = income ; Rhs = one,age,educ,female,female*age
        ; panel ; Fixed effects $
REGRESS  ; Lhs = income ; Rhs = one,age,educ,female,female*age
        ; panel ; Random effects $
```

9. It’s a minor point for our purposes, but there is an intriguing alternative method of computing the fixed effects estimator. A linear regression of the dependent variable, the levels of the time varying variables and the group means of those variables produces the desired coefficients. We note, however, that this does not produce the right standard errors – it uses the wrong degrees of freedom and the sum of squared residuals is much smaller than estimated by LSDV.

```
NAMELIST ; x=one,age,educ,hsat$
CREATE   ; ageb  = group mean(age,pds=ti)$
CREATE   ; educb = group mean(educ,pds=ti)$
CREATE   ; hsatb = group mean(hsat,pds=ti)$
REGR     ; lhs = income ; rhs = age,educ,hsat ; Panel ; fixed effects $
REGR     ; lhs = income ; rhs = age,educ,hsat,ageb,educb,hsatb $
```

10. A hierarchical model. The fixed and random effects models are an interesting place to begin the modeling of heterogeneity. A next step is a random parameters, or hierarchical model. (They are not quite synonymous.) Here, we fit a model of the form

$$\begin{aligned} \text{Income} &= \beta_1 + \beta_2\text{Age} + \beta_{3i}\,\text{Educ} + \beta_4\text{Hsat} + \beta_5\text{Working} + \varepsilon_{it}. \\ \beta_{3i} &= \gamma_1 + \gamma_2\text{Female}_i + w_i. \end{aligned}$$

So, the coefficient on Education is randomly distributed around a mean that shifts depending on whether the respondent is male or female. The model is estimated by maximum simulated likelihood.

```
REGRESS  ; Lhs = income ; Rhs = one,age,educ,hsat,working
         ; Panel ; RPM=female ; Pts=10 ; Halton
         ; Fcn = educ(n) ; Parameters $
 CREATE   ; b_educ = beta_i(_stratum)$
 CREATE   ; t=prd(id)$
 KERNEL   ; if[t=1] ; rhs=b_educ ; group=female$
```

You might try different specifications of this RP model.