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

Lab Session 1 Assignment
Basic Regression and Binary Choice Modeling

I. 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 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 = 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        ; 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 $
```

(Notice that the kernel estimator for log income seems to be to the right of the normal distribution. Why? The normal distribution has the same mean and variance as the log income for the full data set which spans 1984 to 1994 while the kernel estimator for log income is based only on the last year. This gives some evidence that 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 5th to 10th 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.

II. Binary Choice with Cross Sections

This exercise will involve estimating and analyzing binary choice models. We will analyze the panel probit, manufacturing innovation data. The data set is PanelProbit.lpj. The commands are already in the editor, so we just need to reset and open the new project. To reset the program, use Project → Reset… and clear the workspaces. There is no need to save the current project. Now, use File → Open Project … to open PanelProbit.lpj. These data are a panel. The data set appears as follows:

```
+-----------------------------------------------------------------------------+
|****************************************************************             |
|Panel probit data: Stacked                                                   |
|N = 1270, T = 5                                                              |
|EMPLP  = Employment                                                          |
|IM = Industry employment                                                     |
|IP = dependent variable, innovation, binary                                  |
|IMUM = imports share                                                         |
|FDIUM = FDI share                                                            |
|SP = relative size                                                           |
|PROD = productivity                                                          |
|SALES = sales                                                                |
|LOGSALES = log sales                                                         |
|RAWMTL, INVGOOD, CONSGOOD, FOOD = sector dummies                             |
|T = period, T1,T2,T3,T4,T5 = period dummy variables                          |
|FIRM = firm ID                                                               |
|Authors Model = (one,logsales,sp,imum,fdium,prod,rawmtl,invgood)             |
|****************************************************************             |
|Panel Probit Data -  Wide form                                               |
|For observations with T=1 (ignore the others)                                |
|IP84...IP88 = 5 years of IP                                                  |
|EMPLP84...EMPLP88                                                            |
|IM84...IM88                                                                  |
|IMUM84...IMUM88                                                              |
|FDIUM84...FDIUM88                                                            |
|PROD84...PROD88                                                              |
|SALES84...SALES88                                                            |
|LSALES84...LSALES88                                                          |
+-----------------------------------------------------------------------------+
```

1. Different Functional forms.

As we saw in class, the different distributions chosen for the binary choice model each imply a scaling of the coefficients. Superficially, it appears that the model results depend heavily on the distribution. But, this is illusory. The differences essentially disppear when we examine the partial effects rather than the raw coefficients. The following will illustrate this effect for three specific functional forms. The commands compute and assemble the results in tables that enable convenient viewing.

```
Probit      ; Lhs = IP ; Rhs = x ; Table = Probit  $
Partials    ; effects: x ; summary(table=ProbitME) ; means $
Logit       ; Lhs = IP ; Rhs = x ; Table = Logit   $
Partials    ; effects: x ; summary(table=LogitME)  ; means $
Arctan      ; Lhs = IP ; Rhs = x ; Table = Arctan    $
Partials    ; effects: x ; summary(table=ArctanME) ; means $
Maketable   ; Probit,Logit,Arctan$
Maketable   ; ProbitME,LogitME,ArctanME $
```

2. The Linear Probability Model

Some recent applications have used linear regression to fit a 'linear probability' rather than employ the usual probit model. What does least squares do in a binary choice setting. As might be expected from the previous exercise, the coefficients on obtains are very different. Are the results? The following compares the results of the linear probability model to those of a logit model, both in terms of the coefficients and the partial effects? The results suggest what is actually happening when one uses a linear probability model. The coefficients are approximating the partial effects (at the means of the data) of the appropriate nonlinear binary choice model.

```
Regress      ; Lhs = IP ; Rhs = x ; Table = LinearPM $
Partials     ; effects: x ; summary(table=linearME) $
Maketable    ; Logit,LinearPM $
Maketable    ; LogitME,LinearME $
```

3. A Robust Covariance Matrix.

It is now common to compute a 'robust' sandwich type of estimator when fitting a binary choice model. As we discussed in class, there is not much in the way of failures of the model assumption to which the MLE could be robust. Nonetheless, it might be of interest how much difference it makes. The robust estimator is $\mathbf{H}^{-1}(\mathbf{G}'\mathbf{G})\mathbf{H}^{-1}$, where $\mathbf{H}$ is the negative of the Hessian of the log likelihood and G is the $n\times K$ matrix of first derivatives, by observation, of the log densities. The following computes the conventional estimator, $\mathbf{H}^{-1}$ and the robust estimator. We then report the two sets of results side by side.

```
Probit       ; Lhs = IP ; Rhs = X ; Table = standard $
Probit       ; Lhs = IP  ; Rhs = X ;  RobustVC ; Table = Robust $
Probit       ; Lhs = IP  ; Rhs = X ; Cluster = Firm ; Table = Cluster $
Maketable    ; Standard, Robust $
```

With one notable exception, the so-called robust estimator doesn't matter much. But, the clustering seems to make a large difference. Again, this is to be expected.

4. Creating a Plot of Probabilities.

Once estimation is completed, there are a variety of useful post estimation computations that can be carried out with the estimated model. To begin, it is useful to display the predicted probabilities produced by the model. The following estimates a probit model for innovation, then simulates the probabilities over the range of logSales. The plot is generated by dividing the range into 20 parts from the sample minimum of logSales to the maximum. A listing of the probabilities averaged over the sample with all other variables taking their observed values is shown, followed by a plot with a confidence interval around the prediction.

```
Probit       ; if[t=1] ; Lhs = IP ; Rhs = one,IMUM,FDIUM,SP,logsales $
Calc         ; low = .5*Min(LogSales)
             ; high = 1.5*Max(LogSales)
             ; incrmnt = .05*(high-low) $
Simulate     ; scenario: logsales & logsales = Low(incrmnt)high
             ; plot(ci)
             ; title=Simulation of Innovation Probabilities vs. Log Sales$
```

5. Fit Measures

The binary choice models are not fit by least squares, and there is no R squared-like statistic to measure the correlation between the predictions of the model and the observed data. Many ad hoc measures have been proposed. The most widely known is McFadden's pseudo R squared, which as discussed in class, does not actually measure anything like the fit of the model to the data. We examined a number of others in class. The following fits a probit model and stores the predicted probabilities. It then computes predictions by the rule 'Predict y = 1 if fitted probability is greater than T*.' The routine lets you choose T*. The usual choice is T* = .5, however, for these data, the best choice – the choice that produces the most frequent match (zero or one) between actual and prediction – is less than .5. Some authors label this statistic the 'count R squared.' Try different values in the second execute command to find that value. Note that this strategy, in general, is not optimal because the MLE is not chosen to maximize the number of correct predictions. Manski's maximum score estimator does just that. The last command computes the Mscore estimator and displays the fit obtained.

```
Probit          ; Lhs = IP ; Rhs = X ; Summarize ; Prob = Pfit $
Proc = Fit(limit) $
Create          ; ipfit = Pfit > limit ; Correct = (IPfit = ip) $
Crosstab        ; Lhs = IP ; Rhs = IPFIT $
Calc            ; List ; CountRsq = Sum(correct) / N $
EndProc $
Exec            ; Proc = fit(.4) $
Calc            ; MeanIP = xbr(ip) $
Exec            ; Proc = fit(MeanIP) $
Exec            ; Proc = fit(.42)$  Try different values.  Which is best?
MScore          ; if[t=1] ; Lhs = IP ; Rhs = X $
```

6. Partial Effects for a Quadratic and for Interaction Terms

Marginal effects in the binary choice models are complicated functions of the parameters and the data. They are more so when the index function contains complex functions of the data. Suppose, for example,

$$P = \Phi(\boldsymbol{\beta}'\mathbf{x} + \alpha_0\text{logSales} + \alpha_1\text{logSales}^2).$$

The marginal effect of logSales, which is the effect on the probability of a one percent change in sales is

$$\partial P/\partial\text{logSales} = \phi(\boldsymbol{\beta}'\mathbf{x} + \alpha_0\text{logSales} + \alpha_1\text{logSales}^2) \times (\alpha_0 + 2\alpha_1\text{logSales})$$

Computing these properly is a longstanding, widely discussed issue in modern software. The problem, in general, is in obtaining the right single effect for logSales rather than separate effects for the two parts, neither of which give the right answer. Recent versions of Stata (with 'Margins') and NLOGIT (with PARTIALS and SIMULATE) have automated the computation of these types of effects. The following does several computations around this formulation. The probit model contains the indicated quadratic term in logSales. The first command computes the average partial effects for logSales and fdium. The second computes the average partial effect for logSales while varying fdium from .05 to 1.0 in steps of .05, and plots the results. This calculation is done using the delta method. The next Partials command does the same thing, but uses the method of Krinsky and Robb. Since K&R involves a large amount of computation, we have speeded it up by using only the observations with T = 1, which is the first of 5 years of the data. The Wald command shows how to compute the partial effects another way, by actually programming the function. In this example, Wald is more complicated than necessary. In other applications, it might be preferred.

```
Namelist      ; X = One,IMUM,FDIUM,SP $
Probit        ; Lhs = ip ; rhs = x,logsales,logsales^2 $
Partials      ; effects: logsales / fdium $
Partials      ; effects: logsales & fdium = .05(.05)1 ; plot(ci) $
Partials      ; if[t = 1] ; effects: logsales & fdium = .05(.05)1 ; plot(ci)
              ; k&r ; pts = 50 $ Add this to the command to use K&R
? Use Wald instruction for Delta method or K&R. (Add ;K&R)
Namelist      ; FullX = x,logsales,logsales^2
Wald          ; Start  = b ; Var = Varb
              ; Labels = beta0,beta1,beta2,beta3,a0,a1
              ; Fn1 = ME_logS   = n01(beta0'fullX)*(a0+2*a1*Logsales)
              ; Fn2 = ME_fdium = n01(beta0'fullX)*beta2
              ; Average $
```

Computing partial effects for models with interaction terms presents the same challenges as nonlinearities, but yet more complex. The model below is

$$P = \Phi(\beta_1 + \beta_2\, imum + \beta_3\, fdium + \beta_4\, sp + \beta_5\, sp\times imum + \beta_6\, sp\times fdium$$
$$+ \beta_7\, logSales + \beta_8\, logSales^2).$$

It contains the same nonlinearities as the previous model, plus the interaction terms in sp with the other two variables. The following estimates the probit model then computes partial effects for sp, evaluated at the sample range of values. All terms are accounted for.

```
Namelist      ; x=one,imum,fdium,sp,sp*imum,sp*fdium,logsales,logsales^2 $
Probit        ; Lhs = ip ; rhs = x $
Partials      ; effects: sp & sp = .05(.05)1 ; plot (ci) $
```

7. A Group of Dummy Variables for a Set of Categories

The data set also includes a set of sector dummy variables for four sectors. It might be interesting to examine the different results for the four sectors. The Namelist instruction defines the data matrix Sector which contains all four dummy variables. One of them must be dropped in estimation. The probit command contains ‘sector.’ – note the ending dot. This instructs NLOGIT to drop the last category. We will want all four categories for the next instruction. In the results of the probit estimation the coefficients on the first three dummy variables relate to the change in the predicted probability related to the omitted category, in this case, food. We might be interested in different transitions. For example, an interestng margin might be a comparison of the raw materials sector to the consumer goods sector. The Partials command requests a transition matrix that computes these transition probabilities.

```
? Group of dummy variables
Namelist      ; Sector = rawmtl,invgood,consgood,food $
Probit        ; Lhs = ip ; Rhs = x,sector. $
Partials      ; Effects : sector ; transition $
```

8. Testing for Structural Change.

A common test is for homogeneity of the parameter vector across different groups. For example, in our application here, it might be interesting to test whether underlying structural of the model has changed over the five year period of the data. Consider the structure

$$P_{it} = F(\boldsymbol{\beta}_t'\mathbf{x}_{it}),\ i = 1,\ldots,1270,\ t = 1,\ldots,5\ (1993 \text{ to } 1997)$$

which allows for different coefficient vectors in each year. We are interested in testing the hypothesis

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \boldsymbol{\beta}_4 = \boldsymbol{\beta}_5$$
$$H_1: \text{ not } H_0.$$

In a linear regression context, this would be a 'Chow' test and would be tested with an F test. Since this is not a linear regression model, we can't use the F test here. The easiest way to do this test is with a likelihood ratio test. The strategy is to fit the restricted model (pool the 5 years of data) and the unrestricted model (estimate the model separately for each year), and compare the log likelihoods. The log likelihood for the unrestricted model is the sum of the five years. Here is how you can automate this computation. The last part of the last CALC displays the 95% critical value from the chi-squared table. There are two ways to proceed. The first set of commands builds the procedure from first principles. The second uses a built in procedure. Carry out the test. What do you conclude? Should the null hypothesis be rejected? Repeat the test using a logit model instead of a probit model. Does the conclusion change? Try the exercise again while adding the sector dummy variables to the model. To do these, it is only necessary to change the model name from PROBIT to LOGIT, or the NAMELIST command by adding variables to it.

```
Namelist     ; x=one,imum,fdium,sp,sp*imum,sp*fdium,logsales,logsales^2 $
Probit       ; Lhs = IP ; Rhs = X ; quietly $  (Suppress the model results)
Calc         ; Logl0 = Logl ; Logl1 = 0 ; i = 0 $
Procedure
Probit       ; If[t = i] ; Quietly ; Lhs = IP ; Rhs = X $ (Suppress model results)
Calc         ; Logl1 = Logl1 + Logl $
EndProc $
Execute      ; i = 1,5 $ (This suppresses the individual year results.)
Calc         ; List ; Chisq = 2*(Logl1 - Logl0) ; Df = 4*Col(X) ; Ctb(.95,df) $

? Use internal automated procedure
Probit       ; for[ (test) t ] ; lhs=ip;rhs=x;quietly$
```

9. Hypothesis Tests:

This exercise will illustrate the three methods of carrying out hypothesis tests. Two tests are carried out. All of the procedures save for the last carry out the test of whether the sector dummy variables should be included in the index function in the probit model. In the last test, The model is

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i$$
$$\varepsilon \sim N[0,\sigma_i^2],\ \ \sigma_i = \exp(\boldsymbol{\gamma}'\mathbf{z}_i).$$
$$y_i = 1(y_i^* > 0]$$

and the test of whether $\gamma = 0$ is carried out using an LM test. The (small) advantage of the LM test is that it is not actually necessary to estimate the model to carry out the test as the statistic is based on the restricted, homoscedastic model.

```
Namelist        ; X       = One,IMUM,FDIUM,LogSales $
Namelist        ; Sectors = RawMtl,InvGood$
? We include Sectors in the model then test the hypothesis that the
? two coefficients are zero.
Probit          ; if[t=5] ; Lhs = IP ; Rhs = X $
Calc            ; Logl0 = LogL $
? Built in command tests using chi squared.
Probit          ; if[t=5] ; Lhs = IP ; Rhs = X,Sectors
                ; Parameters ; test: sectors $
? Likelihood ratio test.
Calc            ; Logl1 = LogL ; List ; LRstat = 2*(logL1 - LogL0) $
? Wald test using matrix algebra
Calc            ; List ; Ctb(.95,2) $
Calc            ; KX = Col(X) ; K1 = KX + 1 ; Kc = Col(Sectors); K = KX + KC$
Matrix          ; c = B(K1:K) ; vc = Varb(K1:K , K1:K) $
Matrix          ; List ; Waldstat = c'<vc>c  $
? Wald test using the Wald command that automates the matrix commands.
Wald            ; start = b ; Var = Varb
                ; labels=Kx_d,Kc_c ; fn1 = c1 - 0 ; fn2 = c2 - 0 $
? Lagrange multiplier test for omitted variables
Probit          ; Lhs = IP ; Rhs = X ; Quietly $
Probit          ; Lhs = IP ; Rhs = X,Sectors
                ; Start = b,0,0 ; LMTest $
? Lagrange multiplier test for heteroscedasticity
Probit          ; if[t=5] ; Lhs = IP ; Rhs = x ; Quietly $
Probit          ; if[t=5] ; Lhs = IP ; Rhs = x ; Het
                ; Hfn = Sectors ; Start = b,0,0 ; LMTest $
```

10. Simulation:

Using the binary choice model simulator, examine how a 1.1 fold increase in LOGSALES which corresponds to a roughly 10% increase in sales would affect the probability of innovation. The BinaryChoice command carries out a simulated change in every observation, and shows what would happen to the predicted sample responses. The Simulate command displays the average predicted probabilities over a range of values of logSales.

```
Probit          ; Lhs = IP ; Rhs=one,logsales,imum,fdium $
BinaryChoice    ; Lhs = IP ; Rhs = one,logsales,imum,fdium
                ; model=probit ; start=b ; scenario: logsales * = 1.1 $
Simulate        ; scenario: & logsales = 5(1)15 ; plot (ci) $$
```