Exercise 1

Part I. Binary Choice Modeling

A. Fitting a Model with a Cross Section

This exercise uses the health care data contained in healthcare.lpj. The variables in the file are listed below.

```
Data from the Journal of Applied Econometrics Archive. This is an unbalanced
panel.N = 27326, Group sizes range from 1 to 7, 7293 groups.

id          person - identification number
female      female = 1; male = 0
year        calendar year of the observation
age         age in years
agesq       age squared
hsat        health satisfaction, coded 0 (low) - 10 (high)
handdum     handicapped = 1; otherwise = 0
handper     degree of handicap in percent (0 - 100)
income      household nominal monthly net income in German marks / 100000
hhkids      children under age 16 in the household = 1; otherwise = 0
educ        years of schooling
married     married = 1; otherwise = 0
haupts      highest schooling degree is Hauptschul degree = 1; otherwise = 0
reals       highest schooling degree is Realschul degree = 1; otherwise = 0
fachhs      highest schooling degree is Polytechnical degree = 1; otherwise = 0
abitur      highest schooling degree is Abitur = 1; otherwise = 0
univ        highest schooling degree is university degree = 1; otherwise = 0
working     employed = 1; otherwise = 0
bluec       blue collar employee = 1; otherwise = 0
whitec      white collar employee = 1; otherwise = 0
self        self employed = 1; otherwise = 0
beamt       civil servant = 1; otherwise = 0
docvis      number of doctor visits in last three months
hospvis     number of hospital visits in last calendar year
public      insured in public health insurance = 1; otherwise = 0
addon       insured by add-on insurance = 1; otherswise = 0
doctor      1 if number of doctor visits > 0
hospital    1 if number of hospital visits > 0
healthy     1 if hsat > 6, 0 otherwise
Year1984    dummy variable for year=1984
Year1985    dummy variable for year=1985
Year1986    dummy variable for year=1986
Year1987    dummy variable for year=1987
Year1988    dummy variable for year=1988
Year1991    dummy variable for year=1991
Year1994    dummy variable for year=1994
group       sequential identifier for groups, based on ID
ti          number of observations for the group, repeated
```

We are going to analyze the individual's choice of whether to obtain public insurance(PUBLIC). This is a binary choice, so your analysis will be done in this modeling framework. For this exercise, we will be using cross section methods, You will do your analysis using only one of the years of data.

Preliminaries. Set the sample to use only one year of data.

INCLUDE ; New ; Year = xxxx $ (one of 1984, 1985, 1986, 1987, 1988, 1991, 1994)

where you choose the year. For example, if you want to analyze the 1991 data, use

INCLUDE ; New ; Year = 1991 $

Keep this setting in place for the exercise. The command stream you create will be independent of the year, so if you want to analyze a different year, you need only reissue this command with the different year, then reuse the analysis commands.

NLOGIT Tip: To fit a model using only a specific year, in fact it is not necessary to reset the subsample. A command of the form
PROBIT ; If [Year = 1991] ; Lhs = … etc. $
Does the same thing, though the sample remains set at the full sample.

1. Among other variables that will appear in your model, you should include INCOME. Obtain some descriptive measures for income (mean, standard deviation, histogram, kernel density estimator). Describe the income variable. Or, you might think it a better idea to use the log of income, LOGINC. You can get some more details about the variable with

   QUANTILES ; Rhs = Income or Loginc $

2. We are are going to be interested in gender differences in choices, so FEMALE should also appear in your model. Use **DSTAT** to describe this variable.

3. What other variables will you include in your equation? Choose a set of other variables to include in your equation? To keep it manageable, choose only 4 or 5 variables. You can define the set of variables conveniently with

   NAMELIST ; xp = the list of variables $ (Include ONE as a variable.)

4. As a side isssue, you are interested in interrelationships among your variables. In particular, do the data contain evidence that INCOME is explained by other variables in the data set? Use a linear regression to explain INCOME. Include in your model both EDUC and EDUC*EDUC. (You need not compute the square of eduction. Just include EDUC*EDUC in your ;Rhs list.) Test the hypothesis that education (and its square, jointly) is not a significant determinant of INCOME.

   REGRESS ; Lhs = Income ; Rhs = one,educ,educ*educ,age,female $

5. Fit both probit probit and logit models using your specification in 3. compare your results. Does the functional form matter?

NLOGIT Tip: To get a convenient comparison, you can use

PROBIT ; … (your specification) ; Table = probit $
LOGIT ; … (your specification) ; Table = logit $
MAKETABLE ; probit, logit $

Choose one of the model forms, probit or logit, and continue the analysis below using that model. (As we discussed in class, it is not important which one is chosen.)

6. We are interested in whether the model differs for men and women. Fit your probit or logit model separately for men and women and test whether the two groups can be described by the same model. Use a likelihood ratio test.

NLOGIT Tip: To use a subsample,
LOGIT ; If [female = 1] ; Lhs = … (your specification) $
CALC ; LoglF = logL $
Similarly for male (female = 0), then
LOGIT ; … (full sample) $
CALC ; LogLMF = logL $
Then carry out the test.
You can also subsample, using **LOGIT ; If [female = 1 & year = 1991] ; … $**

NLOGIT Tip: This test can be automated with

Model … ; For [(test) Female = *,0,1] ; Lhs = … etc. $

7. Using the pooled model (the last one you fit in part 6), now obtain the partial effects for your variables.

NLOGIT Tip: **PARTIALS ; Effects: variable / variable / … ; Summary $**
Note, if your model has an interaction term in it, or a nonlinearity such as EDUC*EDUC, you do not include the interaction term or nonlinearity in the list of variables in PARTIALS – only include the original variables.

8. Fit a probit or logit model that includes an interaction term between FEMALE and EDUC. That is, along with your other variables, include FEMALE*EDUC in the ;Rhs list. Is the interaction statistically significant. Compute the partial effects two ways.

Model ; … (your specification) ; MarginalEffects $

Then, after the model (probit or logit)

PARTIALS ; Effects ; female / educ ; Summary $

Note the difference between the two sets of results. The first set are incorrect. The second set are correct. (;MarginalEffects does not pick up the interaction term correctly. PARTIALS does.)

B. The Delta Method

The delta method is used to compute standard errors for nonlinear (or linear) functions of asymptotically normally distributed estimators. Here is an example. We begin with a probit model.

$$\text{Prob}(y=1|x) = \Phi(\boldsymbol{\beta}'\mathbf{x})$$

Where Φ is the standard normal cdf. The inverse Mills ratio based on this model is $\lambda = \phi(\boldsymbol{\beta}'\mathbf{x})/\Phi(\boldsymbol{\beta}'\mathbf{x})$ where ϕ is the standard normal density. You will first fit the probit model for the full sample, then compute the inverse Mills ratio using the subsample with FEMALE=1. You will use the delta method to compute a standard error. The computation is done two ways, first by computing the function at the means of the data, second by computing the function for each individual, then averaging the functions. Do the results differ by the two methods?

NLOGIT Tip: You can use the following template:

```
NAMELIST ; xp = … your specification $
PROBIT ; lhs = doctor ; rhs=xp $
WALD ; if[female=1]
     ; parameters = b ; covariance = varb ; labels = kreg_b
     ; fn1 = n01(b1'xp)/phi(b1'xp) $
```

The **WALD** command does the computation at the means of the data. Note, in the command, kreg is the number of variables in the previous model command. In the labels definition, kreg_b defines the list as b1,b2,b3,… The construction b1'xp computes the index function using the x vector and the parameter vector starting with b1.
Add **; Average** to the WALD command to compute the average function value instead.
Use **; K&R** to request the Krinsky and Robb Method.

Another interesting function from the normal distribution is the variance of the truncated normal, which is

$$\sigma^* = 1 - \lambda(\lambda + \boldsymbol{\beta}'\mathbf{x})$$

You can analyze this function by adding

```
; fn2 = 1 - fn1 *(fn1 + b1'xp) $
```

to your WALD command. Try it.

C. Bootstrapping

C.1. Nonlinear Function

Bootstrapping is a method generally used to estimate the standard errors for an estimator. We can also use it as an alternative to the delta method. Use bootstrapping to estimate the standard error of the sample average IMR computed in Part II.

NLOGIT Tip: You can use the following template for this exercise. You must define the namelist, XP. Use the definition you provided earlier.

```
PROCEDURE  $
PROBIT ;  quietly ; lhs=doctor ; rhs = xp $
CREATE ;  imr = n01(b'xp)/phi(b'xp) $
CALC   ;  meanimr = female'imr/sum(female) $
ENDPROC $
EXEC ; n=100 ; bootstrap=meanimr ; histogram $
```

C.2. Test Statistic

Bootstrapping is often used to explore the distributions of test statistics. We'll try that here. The probit model with heteroscedasticity would be

$$\text{Prob}(y=1|\mathbf{x},\mathbf{z}) = \Phi\left[\frac{\beta'\mathbf{x}}{\exp(\delta'\mathbf{z})}\right]$$

The restricted model, under the null hypothesis that $\delta = \mathbf{0}$ is the original probit model. We will use our data on doctor visits to examine the LM statistic for testing this hypothesis. We start by simulating data that exactly obey the assumptions of the model. Note, this exercise uses the **XP** namelist that you defined earlier.

```
PROBIT; lhs = doctor ; rhs = xp $ (Obtains the 'true' coefficients.)
CREATE ; ysim = (b'xp + rnn(0,1))> 0 $ (Simulates the homoscedastic data)
```

In the simulated data, the true coefficients are the MLE probit estimates. There is no heteroscedasticity. NLOGIT will compute an LM statistic for a hypothesis if you provide the restricted estimates as starting values and specify MAXIT=0. We'll test the hypothesis that the data are heteroscedastic depending on gender (FEMALE)

```
PROBIT ; Lhs = ysim ; Rhs = xp $ (this computes the restricted estimates)
PROBIT ; Lhs = ysim ; Rhs = xp ; Het ; Hfn = female ; start = b,0 ; Maxit = 0 $
```

This will report the LM statistic. What value did you get? What is the critical value for the test? We'll now explore the distribution of the statistic under the true null hypothesis of homoscedasticity

```
PROCEDURE $
PROBIT      ;  quietly ; lhs=ysim ; rhs = xp $
PROBIT      ;  quietly ; lhs=ysim ; rhs=xp
            ;  het ; hfn=female;start=b,0;maxit=0$
ENDPROC $
EXECUTE     ; n=100 ; bootstrap=lmstat $
HISTOGRAM;rhs=bootstrp$
```

It will be interesting to see if the real data we are using display evidence of heteroscedasticity. You can do the test just by changing ysim to doctor in the two pairs of PROBIT commands in the discussion above. What do you find?

C.3. Estimated Parameter Vector.

The most common use of bootstrapping is to compute variances and covariance matrices for estimators. We'll do that for a vector of partial effects as scaled coefficients, based on a logit model. Here is the template you can use, once again based on your specification of the model in your **XP** namelist.

```
PROCEDURE $
LOGIT ; quiet ; Lhs = healthy ; Rhs = xp ; Prob = p $
CREATE ; scale = p*(1-p) $
CALC ; avgscale = xbr(scale) $
MATRIX ; ape = avgscale * b $
ENDPROC $
EXECUTE ; n = 50 ; bootstrap = ape $
```

You can compare your results to the results using the delta method by

```
LOGIT ; quiet ; Lhs = healthy ; Rhs = xp ; Marginal  $
```

Note that the comparison will become more favorable if you increase the number of bootstrap replications.

Part II. Panel Data

We continue our analysis of the healthcare data. For these exercises, we will use the smaller subset of the full data set, HealthData.lpj

In this exercise, we will be estimating and analyzing panel data models.

Preliminaries: You must declare the panel data set before fitting the models. After loading the project, use

```
SETPANEL ; Group = id ; Pds = ti $
```

A. Binary Choice and Ordered Choice Model Estimates

1. The first variable of interest is DOCTOR, a dummy variable that equals 1 if the number of doctor visits is greater than zero, and zero if not. Describe this variable. Is the sample relatively balanced, or highly unbalanced?

2. Begin the analysis by fitting pooled probit and logit models. Use at least 3 of the independent variables in the data set including FEMALE as one of them. Use the definition

```
NAMELIST ; XP = … your list of variables … (do not include HEALTHY) $
```

We will use your definition of xp in several exercises below. You'll be able to explore variations in the results just by changing this definition.

3. Since the data are a panel, your pooled estimator is ignoring the correlation across the observations in the households. Before fitting the appropriate panel data model, compute the pooled probit model with robust, cluster corrected standard errors. Compare the results to what you obtained in part 2.

NLOGIT Tip: Use the same **PROBIT** command you used in part 2, but add
; Cluster = id

4. At this point, we will look at fixed and random effects estimators. We start with fixed effects. Fixed effects models require within group (time) variation of the independent variables. Choose three variables, and define a namelist,

   NAMELIST ; xfe = your 3 variables $ (for example, age, income, hhkids)

   A familiar choice for the fixed effects model is the LOGIT specification. There are two approaches, the 'conditional' estimator (Chamberlain's) and the unconditional (Greene, brute force). It is well known that the second of these is biased due to the incidental parameters problem. The familiar 100% bias applies when T = 2. The average group size in our panel is closer to 4, so the bias should be smaller. Let's find out. Compute the unconditional and conditional estimators and compare the results.

   NLOGIT Tip: **LOGIT ; Lhs = doctor ; Rhs = xfe ; Panel ; Table = logit_c $**
   LOGIT ; Lhs = doctor ; Rhs = xfe ; Panel ; FEM ; Table = logit_u $
   MAKETABLE ; logit_c,logit_u $

   The conditional estimator of the logit model eliminates the fixed effects. The unconditional estimator computes the constant terms (when it can) along with the slopes. Examine the estimated constant terms for your model.

   NLOGIT Tip: **LOGIT ; Lhs = … ; Rhs = … ; Panel ; FEM ; Parameters $**
   Notice in the reported output above the coefficients it is indicated that the panel contains 550 individuals, but 307 are skipped because of inestimable ai. These are groups in which y_{it} is always 1 or always 0. Look in the project window in the Matrices folder. You will find a matrix named APLHAFE. Double click this matrix to display it. The values -1.d20 and +1.d20 are fillers for the groups for which ai could not be computed. (The **;Parameters** in your command requests this matrix.)

5. Compute the coefficients of the random effects probit model using your specification of **XP**. Note, there are two ways to do the estimation, the Butler and Moffitt method using quadrature, and maximum simulated likelihood. (Your model must contain a constant term.)

   NLOGIT Tip: **PROBIT ; Lhs = … ; Rhs = one,… ; Panel ; Random $ (B&M)**
   PROBIT ; Lhs = … ; Rhs = one,… ; Panel
   ; RPM ; Fcn = one(n) ; Halton ; Pts = 50 $