Discrete Choice Modeling

Assignment 1

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. \( N = 27326 \), Groupsizes range from 1 to 7, 2293 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 / 1000
- **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; otherwise = 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 ; Yearxxxx = 1 $  
```

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

```
INCLUDE ; New ; Year1991 = 1 $  
```

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.

1. Among other variables that will appear in your model, you will include INCOME. Obtain some descriptive measures for income (mean, standard deviation, histogram, kernel density estimator). Describe the income variable.

2. We are going to be interested in gender differences in choices, so FEMALE should also appear in your model. 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.

4. As a side issue, 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 education. Just include EDUC*EDUC in your ;Rhs list.) Test the hypothesis that education (and its square, jointly) is not a significant determinant of INCOME.

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.

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.