Discrete Choice Modeling
Panel Data
Exercise 5

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 $

Part I. 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 $

4. (Continued) 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 $

6. An ordered choice model. Another interesting variable is HLTHSAT, the health satisfaction variable. It is coded 0,1,2,3,4. Fit an ordered probit model Use a pooled ordered probit model, and use a cluster correction for the standard errors. We're also interested in seeing the partial effects.

NLOGIT Tip: **ORDERED ; … ; Marginals ; Full ; Cluster = … $**

Part II. 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.

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

Where Φ is the standard normal cdf.   The inverse Mills ratio based on this model is $\lambda = \phi(\boldsymbol{\beta'x})/\Phi(\boldsymbol{\beta'x})$ 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.

Part III. Bootstrapping

A. 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 $

B. 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

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

The restricted model, under the null hypothesis that $\boldsymbol{\delta} = \mathbf{0}$ 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. 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.