

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

**Lab Session 2**  
**Binary Choice Modeling with Panel Data**

This assignment will extend the models of binary choice and ordered choice to panel data frameworks. These exercises will use the health care data, `healthcare.lpj`  
 Since these are a panel data set, we begin by identifying it as one

```
SAMPLE ; All $  

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

**1. Cluster correction**

We start the exercise by noting that in fitting a model using a panel data set such as this one, conventional estimators such as the binary logit model may be consistent (that remains to be seen), but either way, the conventionally computed standard errors are likely to be biased downward. The 'cluster correction' is the usual device for dealing with this issue. We'll use a probit model to investigate. It seems evident in the table that the correction actually does correct something. This is suggestive of the presence of common effects that are inducing correlation across observations.

```
Sample ; All $  

Namelist ; X = One,age,educ,income,married,hhkids$  

Probit ; Lhs = public ; Rhs = X ; Table = Pooled $  

Matrix ; Var0 = Varb $ (Uncorrected covariance matrix)  

Probit ; Lhs = public ; Rhs = X ; Cluster = id ; Table = Cluster $  

Maketable ; Pooled,Cluster ; StandardErrors $  

?(Corrected covariance matrix)  

Matrix ; sepanel = diag(Varb) ; sepanel = sqrt(sepanel)$  

? PCTDIFF is the percentage difference between the estimated standard errors  

Matrix ; se0 = diag(Var0) ; se0 = sqrt(se0) ; Diff = Vecd(sepanel) - Vecd(se0)  

Matrix ; List ; PctDiff = 100*<se0>*Diff$
```

Variable	POOLED		CLUSTER	
	Estimate	Std.err	Estimate	Std.err
Constant	3.63082	.073	3.63082	.132
AGE	.00115	.001	.00115	.002
EDUC	-.17189	.004	-.17189	.008
INCOME	-.98023	.056	-.98023	.083
MARRIED	-.02779	.029	-.02779	.047
HHKIDS	-.06936	.025	-.06936	.040
.....	.....	.....	.....	.....
Log-L	-8320.24		-8320.24	
Log-L(0)	-9711.25		-9711.25	

## 2. Pooled and Random Effects Models

In the random effects linear regression setting, the least squares estimator is consistent, though the conventional asymptotic covariance matrix,  $s^2(\mathbf{X}'\mathbf{X})^{-1}$ , is inappropriate. There is a misperception in some parts that the same is true for nonlinear models such as the probit model. In fact, when the model is nonlinear, even a random effect induces a bias in the conventional estimator, usually toward zero. Let's find out.

```
Probit      ; Lhs = public ; Rhs = x ; RandomEffects ; Panel ; Table=REM $
MakeTable  ; Cluster,REM ; Standard errors $
```

Variable	CLUSTER		REM	
	Estimate	Std.err	Estimate	Std.err
Constant	3.63082	.132	12.3972	.438
AGE	.00115	.002	-.00509	.005
EDUC	-.17189	.008	-.59850	.025
INCOME	-.98023	.083	-1.58040	.157
MARRIED	-.02779	.047	.05148	.084
HHKIDS	-.06936	.040	-.11486	.079
Rho			.92380	.005
Log-L	-8320.24		-4860.56	
Log-L(0)	-9711.25			
LogLNoRE			-8320.24	

The displayed results seem to be consistent with the observation.

## 3. Conditional and Unconditional Logit Fixed Effects Estimation.

For the binary logit model, the Chamberlain form of the fixed effects estimator is consistent while the unconditional (brute force) fixed effects estimator is inconsistent. (This is the incidental parameters problem that arises when  $T$  is small. In our unbalanced panel here, the largest group size is 7, and most groups have less than that. Thus,  $T$  is small here.) Fit the logit model by the two approaches, and compare the results. Are they very different? To see if we can't highlight the effect, let's look at the standard case, with  $T = 2$ . How different are the results now? Remember, in the  $T=2$  case,  $\text{plim } \mathbf{b}_{MLE} = 2\boldsymbol{\beta}$  while  $\text{plim } \mathbf{b}_C = \boldsymbol{\beta}$ . Do the results seem to bear this out? Note, going forward. There is no conditional fixed effects estimator for the probit model

```
SAMPLE      ; All $
? Full Sample, T = 1 - 7
LOGIT       ; Lhs = Doctor ; Rhs = income,educ ; panel
            ; Table = Cond$ (Conditional)
LOGIT       ; Lhs = Doctor ; Rhs = income,educ ; panel ; Fixed
            ; Table = uncond $ (Unconditional)
MAKETABLE   ; Cond,Uncond $
? Two period
LOGIT       ; if[Ti=2]; Lhs = Doctor ; Rhs = income,educ ; panel
            ; Table = Cond$ (Conditional)
LOGIT       ; if[Ti=2]; Lhs = Doctor ; Rhs = income,educ ; panel; Fixed
            ; Table = uncond $ (Unconditional)
MAKETABLE   ; Cond,Uncond $
```

#### 4. Test for Fixed Effects vs. No Effects in the Logit Model

In order to test for the need for fixed effects in the logit model, we can't use the likelihood ratio test because the unrestricted estimator is inconsistent. We can use the Hausman test, instead in the conditional logit model. (This was developed in full in Cicchetti's 1986 application of this model.) This uses the chi-squared statistic

$$H = (\mathbf{b}_C - \mathbf{b}_R)' [\mathbf{V}_C - \mathbf{V}_R]^{-1} (\mathbf{b}_C - \mathbf{b}_R)$$

where 'C' refers to the Chamberlain, conditional estimator and 'R' refers to the 'restricted' estimator which has only a single constant term. Note that  $\mathbf{b}_R$  is the subvector of the restricted estimator that strips off the overall constant term – it keeps only the slope coefficients. Using the model suggested in the commands below, carry out the test. What is the result? Do you reject the hypothesis? (What is the null hypothesis?) Note, it is not guaranteed that the difference matrix in the statistic is positive definite. To find out if it is, we will look at the characteristic roots. They must all be positive. Are they?

```
Sample      ; All $
Logit       ; Lhs = Doctor ; Rhs = hhninc,educ,hhkids ; panel $
Matrix      ; bfe = B ; Vfe = VARB $
Logit       ; Lhs = Doctor ; Rhs = hhninc,educ,hhkids,one $
Matrix      ; db = bfe - b(1:3) ; dV = Vfe - Varb(1:3,1:3) $
Matr;list   ; root(dv)$
Matrix      ; List ; Hausman = db'<dv>db $
Calc        ; list ; ctb(.95,row(db))$
```

#### 5. Testing for Random Effects

Testing for random effects in the probit model is straightforward using the likelihood ratio or Wald test. The LM test can also be used, but it has a peculiar feature that requires some special calculations. (See

[http://web-docs.stern.nyu.edu/old\\_web/economics/docs/workingpapers/2012/Greene-LMTestsRandomEffects\\_Sept2012.pdf](http://web-docs.stern.nyu.edu/old_web/economics/docs/workingpapers/2012/Greene-LMTestsRandomEffects_Sept2012.pdf))

The LM test has an advantage over the other two in that it can be computed without fitting the random effects model. The following computes the three test statistics. The first Probit command fits the pooled model, but informs the program that it is using a panel to compute the LM statistic. The LM statistic appears in the initial results above the coefficient estimates. The second Probit command fits the full RE model then we obtain the likelihood ratio and Wald statistics for the tests.

```
Probit      ; Lhs = Doctor ; Rhs = one,income,educ,hhkids ; Panel ; LMTest $
Calc        ; Logl0 = logl $
Probit      ; Lhs = Doctor ; Rhs = one,income,educ,hhkids ; par
            ; Panel ; Random ; Hpt=8 $
Calc        ; Logl1 = logl ; list ; lrtest = 2*(logl1 - logl0) $
Calc        ; k1=kreg+1;list ; Waldtest = b(k1)^2/varb(k1,k1) $
```

## 6. Mundlak's Approach – Correlated Random Effects

The fixed effects model holds that

$$y_{it}^* = \alpha_i + \beta'x_{it} + \varepsilon_{it}$$

and  $E[\alpha_i|X_i] = g(X_i)$ ; that is, the effects are correlated with the other exogenous variables,  $X_i$ . Mundlak's suggestion is that the correlation can be accommodated through the group means (using the time varying variables),

$$\alpha_i = \alpha + \gamma'\bar{x}_i + u_i$$

where  $u_i$  is uncorrelated with  $X_i$ . Inserting the second equation into the first, we obtain what amounts to a random effects model;

$$y_{it}^* = \alpha + \beta'x_{it} + \gamma'\bar{x}_i + u_i + \varepsilon_{it}$$

This suggests a parsimonious treatment of the common effects in the model, and as well, a test of the presence of fixed effects. If the formulation above is (approximately) appropriate, then a test of the null hypothesis that  $\gamma = 0$  would be a test of the null hypothesis of the random effects against the alternative of the fixed effects model. The following uses the Mundlak formulation to test for fixed effects. The commands involve some new features, so there is some annotation. The integration needed to compute random effects estimates is time consuming. We use Hermite integration. The default setting is 64 points. To speed this up (by a factor of 8), we set the number of points to 8.

```
Namelist      ; xm  = income,educ,hhkids $
? This creates new variables that have nothing in them yet. These are placeholders.
Namelist      ; (new); xmbar = incbar,educbar,kidsbar $
? This command computes group means of the variables in the namelist.
Create        ; xmbar=GroupMean (xm, Pds=ti) $
? This is the basic random effects computation.
Probit        ; lhs = Doctor ; Rhs = one,xm ; panel ; Random ; hpt=8 $
Calc          ; logl0 = logl ; rho0=rho $
Matrix        ; beta0 = b $
? This computes the LM statistic at the starting values, which are zero for the group
? means.
Probit        ; lhs = Doctor ; Rhs = one,xm,xmbar ; Random ; panel ; hpt=8
               ; start=beta0,0,0,0,rho0 ; lmtest $
? This command requests the Wald statistic for the test.
Probit        ; lhs = Doctor ; Rhs = one,xm,xmbar ; Random ; panel ; hpt=8
               ; test: xmbar $
Calc          ; logl1 = logl ; List ; LRstat = 2*(logl1 - logl0) $
```

## 7. Estimating the Random effects model as a Random Parameters Model.

The random effects model is a random parameters model with, in this case, only a random constant. (We will examine more elaborate specifications later in the course.) The RP formulation would be

$$y_{it}^* = (\alpha + \sigma w_i) + \beta' x_{it} + \varepsilon_{it}.$$

Thus far, we have fit the REM using the Butler and Moffitt quadrature based method. We will now fit the model as an RP model using maximum simulated likelihood. We will then compare the results for the two methods. As you saw in parts 5 and 6, the Butler and Moffitt estimator reports RHO in the output – this equals the squared correlation between observations in a group. The simulation estimator reports SIGMA, the standard deviation of the common individual effect,  $w_i$ . To compare the two estimates of  $\rho$ , you must compute  $\rho^* = \sigma^2 / (1 + \sigma^2)$  from the random parameters estimates. What do you find? Are the estimates of the other slopes nearly the same?

```
Namelist      ; X = income,educ,hhkids,one $
Probit        ; if[Ti = 7] ; lhs = hospital ; Rhs = x ; panel ; random effects ; Hpt = 8 $
Calc         ; RhoRE = Rho $
Probit        ; if[Ti = 7] ; lhs = hospital ; Rhs = x ; panel ; RPM ; Fcn=One(n) ; Pts = 20 $
Calc         ; K1 = Col(X) + 1 $
Calc         ; SigmaRP = B(K1) ; List ; rhoRE ; RhoRP = SigmaRP^2/(1+SigmaRP^2) $
```

## 8. Fixed and Random Effects.

The disadvantage of the random effects estimator is that it requires an assumption that the individual effects are uncorrelated with the included variables. If that assumption is not met, the estimator is inconsistent. The fixed effects estimator is inconsistent when T is not large. Thus, both estimators have problems. Chamberlain's conditional estimator provides a way to estimate the logit fixed effects model consistently. One other complication is that the conditional estimator is only available for the logit model. We can fit a random effects logit model, but it is a bit peculiar in that the underlying utility function has a normally distributed random term and a logistically distributed common effect. The mixture of the two seems a bit peculiar. Nonetheless, we will use that model for this experiment.

```
Namelist      ; x = income,educ,hhkids $
Logit         ; if[Ti = 7] ; lhs = doctor ; Rhs = x,one
              ; random ; panel ; hpt=8 ; Table = REM $
Logit         ; if[Ti = 7] ; lhs = hospital ; Rhs = x ; panel ; Table = FEMCond$
Logit         ; if[Ti = 7] ; lhs = hospital ; Rhs = x ; panel ; FEM ; Table = FEM $
Maketable     ; REM,FEM,FEMCond $
```

## 9. Dynamic Probit Model

The following computes the parameters of the dynamic probit model that we discussed in class. To do the estimation, we will take advantage of what we found in part 7, and use maximum simulated likelihood rather than quadrature to fit the model. This approach builds on the Mundlak model in part 7. Note the computation of the matrix of group means could be done a bit more compactly.

```
Sample      ; All $
Setpanel    ; group = id ; pds = ti $
Namelist    ; xit = age,income,hhkids,hsat,married $
Create      ; agebar = group mean (age, pds=ti) $
Create      ; incbar = group mean (income, pds=ti) $
Create      ; kidsbar = group mean (hhkids, pds=ti) $
Create      ; hsatbar = group mean (hsat, pds=ti) $
Create      ; marrbar = group mean (married, pds=ti) $
Namelist    ; xbar = agebar,incbar,kidsbar,hsatbar,marrbar $
Create      ; t = ndx(id,1) $
Create      ; yi0 = Group obs1(public,pds=ti) $
Create      ; ylag = public[-1] $
Probit      ; if[t > 1] ; Lhs = public ; Rhs = one,xit,xbar,yi0,ylag
            ; Maxit = 10 ; Panel ; RPM ; Fcn=one(n) ; Pts=25 ; Halton $
```