NYU STERN
NEW YORK UNIVERSITY LEONARD N. STERN SCHOOL OF BUSINESS

Department of Economics

Econometric Analysis of Panel Data

Assignment 7

This assignment uses the German health care data that we have used in class. You can download the data set from the course website in the form of an Excel spreadsheet, at

http://people.stern.nyu.edu/wgreene/Econometrics/healthcare.xls and .csv

You should be able to import one or the other of these files into Stata, SAS, etc. If you are using LIMDEP or NLOGIT, you can download the file in project format that you can load directly, at

http://people.stern.nyu.edu/~wgreene/Econometrics/healthcare.lpj

Part I. Random Effects Poisson Model

The Poisson model for a panel of data may be formulated

$$\text{Prob}[Y_{it} = y_{it}] = \exp(-\lambda_{it})\lambda_{it}^{y_{it}} / y_{it}!$$

Note, it's usually convenient to write the factorial as $\Gamma(y_{it}+1)$ where Γ is the gamma function. We'll use the usual loglinear specification $\lambda_{it} = \exp(\boldsymbol{\beta}'\mathbf{x}_{it})$. Consider now a random effects model,

$$\lambda_{it} = \exp(\boldsymbol{\beta}'\mathbf{x}_{it} + w_i)$$

Rather than using a normal distribution, we will suppose (as has been done historically) that w_i is distributed as 'log-gamma.' That is, $\exp(w_i)$ has a gamma density with mean 1. Denote $\exp(w_i)$ as u_i. Note that

$$\lambda_{it} = \exp(w_i)\exp(\beta' x_{it}) = u_i\,\phi_{it}$$

We will assume u_i has mean 1 – this is the same as assuming the mean of u_i in more familiar cases equals zero. We then have

$$u_i \sim \text{Gamma}(\theta,\theta)$$

(That is gamma with parameters θ and θ, so the mean is $\theta/\theta = 1$.) Thus,

$$f(u_i) = [\theta^{\theta}/\Gamma(\theta)]u_i^{\theta-1}\exp(-\theta u_i),\ u_i \geq 0$$

With this in place, then

$$P(y_{it}|u_i) = \exp(- u_i \phi_{it})(u_i \phi_{it})^{yit} / y_{it}!$$

What is the marginal distribution of y_{it}? You will obtain this by integrating u_i out of the joint distribution of y_{it} and u_i, which is

$$P(y_{it},u_i) = P(y_{it} | u_i) f(u_i)$$

You can do this using gamma integrals, fairly easily. The purpose for choosing the log-gamma variable to begin with is to have a conjugate marginal distribution for u_i. This sets up the convenient gamma integrals used to get rid of u_i. (HINT: This problem is solved in full for the cross section case in your text. The derivation here involves only a trivial change in some subscripts.)

The purpose for finding the marginal distbribution of y_i is to set up the density to use in the likelihood function. Suppose w_i were assumed to be normally distributed, rather than log-gamma. How would your approach to this problem have to change?

Part II. Panel Data Estimation

The variables in the German health care data set are

id	person - identification number
female	female = 1; male = 0
year	calendar year of the observation
age	age in years
hsat	health satisfaction, coded 0 (low) - 10 (high)
handdum	handicapped = 1; otherwise = 0
handper	degree of handicap in percent (0 - 100)
hhninc	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; otherswise = 0
numobs	= ni = number of observations on this individual, 1 ... 7
doctor	= 1 if docvis > 0, 0 otherwise
newhsat	same as hsat with some obvious coding errors corrected.

1. We begin with a conventional linear model. The variable hhninc is household income.

a. Specify a linear regression model for hhninc. (I.e., choose an appropriate set of independent variables for your model. Compute and report the coefficients of the linear regression.

b. Compute and report fixed and random effects models for hhninc. Using standard statistical procedures, determine which is the preferred model given your specification.

c. There are 7 years of data in the data set. The variable YEAR takes the values 1984, 1985, 1986, 1987, 1988, 1991, 1994. Create 6 of the 7 dummy variables you need to fit a two way fixed effects model then add the time dummies to your regression. Are the time effects significant, collectively? (HINT: Your income equation probably contains AGE. AGE is perfectly collinear with the family dummies and the time dummies, since, for example, $AGE_{i85} = AGE_{i,84} + T_{1985}$ and likewise for the other years. So, at least for this part of the exercise, you will have to take AGE out of your equation.)

2. The variable DOCTOR in the German health care data is a binary outcome that indicates whether or not the individual visited a doctor in the survey year.

a. Drawing on the list of variables in the data set, formulate a binary choice model for DOCTOR. (I.e., choose a list of appropriate independent variables). Then, fit simple probit and logit models, ignoring the panel nature of the data. (You may restrict the sample if you desire. For example, it might be convenient to use only observations with ti = 7, to create a balanced panel.) Compare the two sets of results.

b. How would you go about fitting an 'effects' model for this variable? What are the issues in doing so? If you are using Stata, LIMDEP or SAS, you can fit a random and/or fixed effects model. Do so, and report your findings. How do your results change, compared to the 'pooled' estimator you computed in part a.

III. An Effects Model

The following is from Wooldridge (problem 15.5, page 511.) Consider the probit model $p(y=1|\mathbf{z},q) = \Phi(z_1\delta + \gamma z_2 q)$ where q is independent of z_2 and q is distributed N(0,1). z_2 is observed but q is not.

(a) Find the partial effect of z_2 on the response probability, namely $\partial P(y=1|\mathbf{z},q)/\partial z_2$.

(b) Show that $P(y=1|\mathbf{z}) = \Phi[z_1\delta / (1 + \gamma^2 z_2^2)^{1/2}]$. (Hint: $y^* = z_1\delta + (\varepsilon + \gamma z_2 q)$, y = 1 if $y^* > 0$.)

(c) Define $\rho = \gamma^2$. How would you test $H_0{:}\rho = 0$.

(d) If you believe that $\rho > 0$, how would you estimate δ and ρ?

(e) (My own addition), supposing that z_1 varies through time, $z_{1,it}$ while z_2 is time invariant, $z_{2,i}$, how would you handle this estimation problem assuming you were given a panel of data on $(y_{it}, z_{1,it}, z_{2i})$.

IV. Another Effects Model

(Also from Wooldridge, problem 15.18.) Consider Chamberlain's random effects probit model,

$$\text{Prob}(y_{it} = 1) = \Phi(\boldsymbol{\beta}'\mathbf{x}_i + u_i),\ \text{Prob}(y_{it} = 0) = 1 - \text{Prob}(y_{it} = 1),$$

where $u_i \mid \mathbf{x}_i \sim N[\mu + \boldsymbol{\delta}'\bar{\mathbf{x}}_i,\ \sigma_u^2 \exp(\boldsymbol{\lambda}'\bar{\mathbf{x}}_i)]$

(so, u_i has conditional mean and variance that both depend on the group mean of the **x**'s.) This extends the random effects model to heteroscedasticity.

a. Find $P(y_{it} = 1| \mathbf{x}_{it}, a_i)$ where $a_i = u_i - E[u_i|x_i]$.

b. Derive the log likelihood function for estimation of the parameters in this model.

c. After you have estimated the parameters of the model, how would you estimate the marginal effects in this model?