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

Department of Economics

Econometric Analysis of Panel Data

Professor William Greene Phone: 212.998.0876
Office: KMC 7-78 Home page:www.stern.nyu.edu/~wgreene
Office Hours: TR, 3:00 - 5:00 Email: wgreene@stern.nyu.edu
URL for course web page:
www.stern.nyu.edu/~wgreene/Econometrics/PanelDataEconometrics.htm

Final Examination: Spring 2007

This is a 'take home' examination. Today is Tuesday, May 1, 2007. Your answers are due on Friday, May 11, 2007. You may use any resources you wish – textbooks, computer, the web, etc. – but please work alone and submit only your own answers to the questions.

Part I. (Continuing a (now) tradition) The Hausman and Taylor Estimator

Write out a full statement of the procedure that Hausman and Taylor devised for estimation of the parameters in a panel data model in which some independent variables are correlated with the time invariant part of the disturbance in a random effects model. Now, show how the Arellano/Bond/Bover estimator uses the Hausman and Taylor result.

Part II. Panel Data Regressions

Munnell (1990) analyzed the productivity of public capital at the state level using a Cobb-Douglas production function. We will use the data from that study to estimate a three level log linear regression model

$$\ln GSP_{it} = \alpha_i + \beta_1 \ln p\_cap_{it} + \beta_2 \ln hwy_{it} + \beta_3 \ln water_{it} + \beta_4 \ln priv\_cap_{it} + \beta_5 \ln util_{it} + \beta_6 \ln emp_{it} + \beta_7 unemp_{it} + \varepsilon_{it}$$

where the variables in the model are

GSP = gross state product,
p_cap = public capital,
hwy = highway capital,
water = water utility capital,
priv_cap = private capital,
util = utility capital,
emp = employment (labor),
unemp = unemployment rate.

The data set consists of a balanced panel of 48 states and 17 years (1970-1986). There are 816 observations in total, arranged by state, then year within the state. The data are posted on the home page for the course as a LIMDEP project file,

http://www.stern.nyu.edu/~wgreene/Econometrics/Munnell-Productivity.lpj

and in three other formats, .txt which is a simple text file, 816 lines of data plus a line of names at the top of the file, .xls which you can read directly into Excel or any other program directly, and .lim which is a *LIMDEP* command file that is the same as the .txt file except it contains a READ command. Do this exercise with *LIMDEP* (or *NLOGIT*), or any other software you wish to use.

a. Fit the "pooled" model and report your results
b. Fit a random effects model and a fixed effects model. Use your model results to decide which is the preferable model. If you find that neither panel data model is preferred to the pooled model, show how you reached that conclusion. As part of the analysis, test the hypothesis that there are no "state effects."
c. Assuming that there are "latent individual (state) effects," the asymptotic covariance matrix that is computed for the pooled estimator, $s^2(\mathbf{X'X})^{-1}$, is inappropriate. What estimator can be computed for the covariance matrix of the pooled estimator that will give appropriate standard errors?
d. The hypothesis of constant returns in the production of *GSP* would be that

$$H_0: \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 = 1$$

Test this hypothesis in the context of the model in a. and in the context of your preferred model in part b. Do you reach the same conclusion in both cases?

Part III. Instrumental Variable and GMM Estimation

For the setting in part 2, suppose we specify the following random effects model,

$$\ln GSP_{it} = \alpha + \beta_1 \ln p\_cap_{it} + \beta_2 \ln hwy_{it} + \beta_3 \ln water_{it} + \beta_4 \ln priv\_cap_{it} + \beta_5 \ln util_{it} + \beta_6 \ln emp_{it} + \beta_7 unemp_{it} + \gamma \ln GSP_{i,t-1} + \varepsilon_{it} + \underline{u}_i$$

a. Is the OLS estimator a consistent estimator of the parameters of this model? Explain.
b. Estimate the parameters of the model using Anderson and Hsiao's instrumental variable estimator. Explain the procedure you are using.
c. Describe a GMM estimator for this model.
d. Suppose the model is respecified as

$$\ln GSP_{it} = \alpha + \beta_1 \ln p\_cap_{it} + \beta_2 \ln hwy_{it} + \beta_3 \ln water_{it} + \beta_4 \ln priv\_cap_{it} + \beta_5 \ln util_{it} + \beta_6 \ln emp_{it} + \beta_7 unemp_{it} + \gamma_i \ln GSP_{i,t-1} + \varepsilon_{it}$$

Note that there is now no "state" effect, but the coefficient on the lagged dependent variable is state specific. Describe the different approaches to estimation of this model, and what problems arise with these different methods.

Part IV. Binary Choice Models

The German data on health care utilization on the course website

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

as well as .xls and .txt formats, contain several variables related to health utilization, satisfaction and insurance. For the moment, we will focus on the individual health assessment, HSAT, which is coded 0 to 10 in the raw data. I want to define the new variable

HEALTHY = 1 if HSAT > 6
0 otherwise

I am interested in a binary choice model for HEALTHY, with independent variables

$\mathbf{z}_{it}$ = *one,age,educ,female,hhninc,hhkids,married*

(a) If you fit a pooled **logit** model, there is the possibility that you might be ignoring unobserved heterogeneity (effects). Wooldridge argues that when one fits a probit model while ignoring unobserved heterogeneity, the raw coefficient estimator (MLE) is inconsistent, but the quantity of interest, the "Average Partial Effects" might well be estimated appropriately. Explain in detail what he has in mind here. Do you think that his result carries over to the logit model as the alternative to the probit model?
(b) The health care utilization data are an unbalanced panel, with number of observations per person ranging from 1 to 7. Suppose we were to estimate a "fixed effects" logit model by "brute force," just by including the 7,293 dummy variables needed to create the empirical model. What would the properties of the resulting estimator likely be? What is "the incidental parameters problem?" (Hint, this is more a thinking question than something you can derive.)
(c) How would I proceed to use Chamberlain's estimator to obtain a consistent slope estimator for the fixed effects logit model.
(d) Describe in detail how to fit a random effects logit model using quadrature and using simulation for the part of the computations where they would be necessary, under the assumption that the effects are uncorrelated with the other included exogenous variables.
(e) Using the random effects logit model that you described in part (d), describe how you would test the hypothesis that the same probit model applies to men and women. (Well, not quite the same. Notice that the gender variable appears in $\mathbf{z}_{it}$. So, answer this question with reference to a model that contains a reduced $\mathbf{z}_{it}$ that does not include *female*.

Part V. Sample Selection

Suppose, instead of the logit model in Part IV, we begin with a probit model involving the same variables,

$$HEALTHY = 1\ (\mathbf{z}_{it}'\boldsymbol{\alpha} + u_{it}) > 0$$

. We also consider an (admittedly, not very well specified) model for the number of Hospital visits,

$$HOSPVIS_{it} = \beta_1 + \beta_2\, AGE_{it} + \beta_3\, EDUC_{it} + \beta_4\, PUBLIC_{it} + \varepsilon_{it}$$

Suppose we now seek to examine the behavior of HEALTHY people. (We thus propose to ignore the unhealthy ones.) We consider the case in which the latent heterogeneity variables (u_{it} and ε_{it}) that enter the two equations are correlated (bivariate normal).

(a) What problem would arise if we simply use linear regression to estimate the parameters of the second equation?
(b) Does the problem go away if we simply change the regression to

$$HOSPVIS_{it} = \beta_1 + \beta_2\, AGE_{it} + \beta_3\, EDUC_{it} + \beta_4\, PUBLIC_{it} + \gamma HEALTHY_{it} + \varepsilon_{it}$$

and use the full data set to run our linear regression for the HOSPVIS equation?
(c) For the moment, ignore the discrete nature of *HOSPVIS*, and think of it as the dependent variable in a linear regression. Show how analysis of these data for the observations for which *HEALTHY* = 1 might translate into a familiar "sample selection" model. Describe in detail the computations one would do to fit such a model.
(d) Use the selection model that you described in (c) to analyze this variable, treating it as continuous. Fit a sample selection model, using several regressors that you select from the data set. (Note, we are ignoring the panel nature of the data set.) For those of you not using *Stata* or *LIMDEP* to do the computations, standard textbooks such as Greene describe the computations needed to compute the appropriate asymptotic covariance matrix. If you are able to do the computations, describe what you computed. If not, describe the computations you would program if you could.

Part VI. A Loglinear Model

The lognormal model is often used to analyze nonnegative random variables, such as time until failure of electric or electronic components. However, the lognormal has a long, thick tail that is sometimes viewed as putting too much mass in the extreme regions of the distribution. An alternative model is the inverse Gaussian distribution,

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi y_i^3}} \exp\left[-\frac{1}{2\sigma^2} \frac{(y_i - \lambda_i)^2}{\lambda_i^2 y_i} \right], y_i > 0, \sigma > 0, \lambda_i > 0.$$

To introduce observed heterogeneity into the model, we now specify $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$. The parameters to be estimated are the vector of slopes, $\boldsymbol{\beta}$ and the scale parameter, σ.

(a) Write out the log likelihood for this model and the first order conditions for estimation of $\boldsymbol{\beta}$ and σ. {Hint: $\partial \ln f(y_i)/\partial\boldsymbol{\beta} = [\partial \ln f(y_i)/\partial\lambda_i](\partial\lambda_i/\partial\boldsymbol{\beta})$ and $\partial\lambda_i/\partial\boldsymbol{\beta} = \lambda_i\mathbf{x}_i$.}

(b) Show how to use Newton's method to solve the likelihood equations.

(c) How would you test the null hypothesis H_0: σ=1?

(d) How would you estimate the asymptotic covariance matrix for the maximum likelihood estimators?

(e) For this random variable, $E[y_i|\mathbf{x}_i] = \lambda_i$ and $\text{Var}[y_i|\mathbf{x}_i] = \sigma^2\lambda_i^3$ Derive the partial effects for the model and describe how to compute them and how to estimate standard errors.

(f) Suppose we prefer to use a Bayesian (MCMC) approach to estimation of the parameters. We propose the following priors for

$p(\boldsymbol{\beta})$ ~ normal with mean $\boldsymbol{\beta}^0$ and covariance matrix $\boldsymbol{\Sigma}^0$
$p(\sigma)$ ~ exponential with parameter σ^0

Describe how the Bayesian estimation would proceed using a Gibbs sampler. (This does not call for a detailed derivation. Just describe in moderately detailed terms how Bayesian estimation would proceed, what the computations would be and what is "estimated.")