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

Department of Economics

Econometric Analysis of Panel Data

Spring 2006 – Tuesday, Thursday: 1:00 – 2:20

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

This is a 'take home' examination. Today is Thursday, April 27, 2006. Your answers are due on Friday, May 12, 2006. 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. 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 (A&B - somehow Bover often manages to disappear from the references to this body of work.) uses the Hausman and Taylor result.

Part II. Panel Data Regressions

Using the Spanish dairy farm data on the course web site, fit pooled OLS, fixed effects and random effects linear regressions models. The central equation for the model is a translog production function

$$y_{it} = \alpha_i + \Sigma_k \beta_k x_{kit} + \Sigma_m \Sigma_{n\leq m} \gamma_{mn} x_{m,it} x_{n,it} + \varepsilon_{it}$$

(y_{it}, x_1, x_2, x_3, x_4 are already the logs of the output and four inputs.) Test the hypothesis of "no effects" vs. some effects. Then, use the data to decide which is the appropriate estimator for these data, random effects or fixed effects. Note that the data also include six year dummy variables For the second half of this exercise, extend your model to include a time effect (there are 6 years of data, drop one of them). Interpret the coefficients on the year dummies, report your results, and test the hypothesis that there is no year effect in the model.

Part III. GMM Estimation

Baltagi's world gasoline market data on the course website contain data on

LGASPCAR = log of consumption per car
LINCOMEP = log of per capita income
LRPMG = log of real price of gasoline
LCARPCAP = log of per capita number of cars

It is tempting to use our conventional estimators to analyze the consumption of gasoline as a function of income, price, and the number of cars. However, one could reasonably argue that while per capita income is exogenous in the equation, price surely is not, and the number of cars (per capita) is at least dubious. Suppose we wish to use an instrumental variable estimator rather than just least squares to fit the equation. I propose to use a time trend, t = 1,...,19 and lagged values of income as instrumental variables. Show how to fit the equation. Now, suppose I wish to improve efficiency by using two lags of these variables rather than just one. Does this improve the efficiency of the IV estimator? Explain. Finally, suggest how to proceed if the model specification is changed to include a lagged value of the dependent variable. What estimator would you use in this case. There are several possibilities. Consider more than one in your answer.

Part IV. Sample Selection

The data on health care utilization on the course website contains two discrete variables related to health insurance, PUBLIC which indicates whether the individual has public insurance and HOSPVIS which is a count of hospital visits.

We first consider just the PUBLIC variable.

(a) If we fit a pooled probit model (after deciding what should be in the model), there is the possibility that we 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.
(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" probit model. What would the properties of the resulting estimator likely be? What is "the incidental parameters problem?" Given what you know about incidental parameters, do you think that Wooldridge's analysis that you described in part (a) applies? (Hint, this is more a thinking question than something you can derive.)
(c) Describe in detail how to fit a random effects probit 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. Now, describe how to proceed under the assumption that the unobserved heterogeneity may be correlated with the other exogenous variables.

Now. we consider the HOSPVIS variable

(d) For the moment, ignore the discrete nature of HOSPVIS, and think of it as the dependent variable in a linear regression. We propose to fit a regression model of HOSPVIS on variables such as age, income and eduction. But, we propose to examine this variable using only the data on individuals who have public insurance. Show how analysis of these data for the observations for which PUBLIC = 1 might translate into a familiar "sample selection" model. Describe in detail the computations one would do to fit such a model – continuing (admitttedly erroneously) to treat the model for HOSPVIS as a linear regression.

(e) Use the selection model that you described in (d) to analyze HOSPVIS, 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 V. A Loglinear Model

Consider the geometric model for an integer variable that is the number of failures until the first success. Suppose θ is the success probability on any trial. Then,

$$\text{Prob}[\text{Y} = \text{y}] = \theta\,(1\text{-}\theta)^{\text{y}}.$$

Now, suppose we wish to turn this into a more interesting regression model. Suppose

$$\lambda = \exp(\boldsymbol{\beta}'\mathbf{x}),\ \text{and}\ \theta = 1\,/\,(1 + \lambda).$$

(a) Show that the resulting model, now with appropriate observation subscripts, is

$$\text{Prob}[\text{Y} = \text{y}_\text{i}] = \lambda_i^{y_i}(1+\lambda_i)^{-(1+y_i)}$$

(b) Write out the log likelihood for this model and the first order conditions for estimation of $\boldsymbol{\beta}$.
(c) It turns out that $\text{E}[\text{y}_\text{i}|\text{x}_\text{i}] = \lambda_\text{i}$. Use this result to devise a GMM estimator of $\boldsymbol{\beta}$. Will this differ from the MLE? Explain.
(d) Given the result in (c), nonlinear least squares might be an alternative estimator. Show how to estimate $\boldsymbol{\beta}$ using nonlinear least squares. Will this differ from the MLE? Explain.
(e) Which of the three estimators would you prefer? Explain.
(f) Show how to use Newton's method to compute the maximum likelihood estimator. How does this differ from the method of scoring? Explain.
(g) Just to think about how it would be done... Suppose we prefer to use a Bayesian approach to estimation of the parameters. We propose the following prior:

$\text{p}(\beta)$ ~ normal with mean 0 and covariance matrix A

Describe how the Bayesian estimation would proceed. (This does not call for a detailed derivation. Just describe in fairly general terms how Bayesian estimation would proceed, what the computations would be and what is "estimated.") (Note, by the way, it does not require a Gibbs sampler, either.)
(h) Suppose you have a panel of data on (y,x), and you are interested in extending the model framework to accommodate unobserved heterogeneity across individuals. What tools are available to you? How would you do it?
(i) (optional) I see two ways to show that $\text{E}[\text{y}_\text{i}|\text{x}_\text{i}] = \lambda_\text{i}$. The first is to derive the expected value the hard way by summing probabilities times outcomes. A second is to manipulate the result that the expectation of the gradient of the log likelihood is zero. Use the second of these to show the result. (A third is to use the moment generating function or characteristic function. Much to complicated if all you want is the mean...)