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

Department of Economics

Econometric Analysis of Panel Data

Spring 2008 – Tuesday, Thursday: 1:00 – 2:20

Professor William Greene

Midterm Examination

This examination has four parts. Weights applied to the four parts will be 15, 15, 30 and 40. This is an open book exam. You may use any source of information that you have with you. You may not phone or text message or email or Bluetooth (is that a verb?) to "a friend," however.

Part I. Fixed and Random Effects

Define the two basic approaches to modeling unobserved, time invariant effects in panel data. What are the different assumptions that are made in the two settings? What is the benefit of the fixed effects assumption? What is the cost? Same for the random effects specification. Now, consider the possibility that the unobserved effects are not time invariant.? How does your answer change?

Two approaches are fixed effects and random effects. In the "effects model,"
$y_{it} = x_{it}'\beta + c_i + \varepsilon_{it}$, x_{it} is exogenous with respect to ε_{it}.
FE: c_i may be correlated with x_{it}.
Benefits: General approach,
Robust – estimator of β is consistent even if RE is the right model.
Cost: Many parameters, inefficient if RE is correct.
Precludes time invariant variables.
RE: c_i is uncorrelated with x_{it}
Benefits: Tight parameterization – only one new parameter
Efficient estimation – use GLS
Allows time invariant parameters
Cost Unreasonable orthogonality assumption
Inconsistent if RE is the right model.
Random parameters case. Replace the model statement with $y_{it} = x_{it}'\beta_i + \varepsilon_{it}$, $\beta_i = \beta + w_i$.
Case 1: w_i may be correlated with x_{it}. This is the counterpart to FE. In this case, it is necessary to fit the equations one at a time. Requires that there be enough observations to do so, so $T \geq K$. The efficient estimator is equation by equation OLS. Same benefits (robustness) and costs (inefficiency) as FE
Case 2; w_i is uncorrelated with x_{it}. This RP model can be fit
An efficient estimator will be the matrix weighted FGLS estimator. (Swamy et al.) This would be a two step estimator, just like FGLS for the RE model.
This model can also be fit by simulation – we mentioned this briefly in class, and
will return to it later this semester.

If the unobserved heterogeneity is time varying, then taking deviations from means will not remove it from the model. Returning to the model specification, we now have

$Y_{it} = \beta'x_{it} + c_{it} + \varepsilon_{it}$

If c_{it} is uncorrelated with x_{it} then it can be simply added to the disturbance in the model, and the model becomes a simple linear regression that can be fit by OLS. This is the RE case. In the FE case in which c_{it} is correlated with x_{it} we have a classic left out variable problem, and there is no way to proceed.

Part II. Minimum Distance Estimation

Munnell's 1990 study of public capital productivity was based on output, capital and labor data for 48 states (not Alaska and Hawaii), and 17 years. Variables are y_{it} = log of gross state product and
$\mathbf{x}_{it} = (\log K_{it}, \log L_{it})$ where K, and L are capital and labor. The model I propose is

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, i = 1,\ldots,48, t = 1970,\ldots,1986.$$

where $E[\varepsilon_{it}|\mathbf{x}_{it}] = 0$ for all i and t.

$$E[\varepsilon_{it}\varepsilon_{js}] = \sigma_{ij} \text{ if } t = s \text{ and } 0 \text{ if } t \neq s.$$

(I.e., states are correlated because of common macroeconomic conditions, but there is no correlation across time.) I propose to fit this model by the following strategy:

1. Estimate the equation separately for each state.
2. Use a minimum distance estimator to reconcile the 48 competing estimators of β

1. Does this procedure produce 48 sets of consistent estimators of the parameters of the model? Explain.

The estimators are unbiased, since this is a classical regression model. Consistency would hang on the usual assumptions about the data, plus an assumption that T was increasing. In the usual panel data case, we would assume that T is fixed, in which case, the answer would be no. With 17 years of data, and potentially more years, increasing T might make some sense. In principle, consistency of a panel data hangs on n increasing, but increasing numbers of years is certainly more plausible than increasing the number of states. So, consistency in this context is hardly assured.

2. Assuming that σ_{ij} equals zero when $i \neq j$. (i.e., no correlation across states, but different variances), show how to compute the minimum distance estimator of β.

If there is no correlation of the disturbances across states, we would have estimated α_i as efficiently as possible, since we used the data for each state to estimate the state specific α_i. But, we have 48 estimates of β. We can use a minimum distance estimator to reconcile the 48 of them by minimizing with respect to β

$$\sum_{i=1}^{48} (b_i - \beta)' V_i (b_i - \beta)$$

As long as the weighting matrix for the MDE is positive definite, any one will produce a consistent estimator. The most efficient estimator weights the components by the inverse of their respective covariance matrices. Thus, we would use for V_i the inverse of $s_i^2(X_i'X_i)^{-1}$. This produces a weighted average of the b_is, $b = \Sigma_i A_i b_i$ in which the weighting matrix is the diagonal of

$$A_i = [\Sigma_i \{s_i^2(X_i'X_i)^{-1}\}^{-1}]^{-1} \{s_i^2(X_i'X_i)^{-1}\}^{-1}$$

3. Show that a seemingly unrelated regression estimator can be used to estimate α_i and β efficiently if I do not make the assumption that $\sigma_{ij} = 0$ when $i \neq j$.

You can write the model in the model in the form of a seemingly unrelated regression model by just stacking the equations, but keeping a separate constant term for each state. It turns out that this produces precisely the MDE.

4. The first assumption made above is that the disturbance in each period is uncorrelated with the regressors in that period. Suppose the assumption is strengthened to $E[\varepsilon_{it}|\mathbf{x}_{is}] = 0$ for all *i*,*t* and *s*. That is, the disturbance in each period for each state is uncorrelated with the regressors in every period for that state. Does this weaken the claim of efficiency made in part 3? Explain. Is there an alternative estimator available that is more efficient than the one in 3?

By assuming only exogeneity, we have assumed that the disturbance ε_{it} in each period is uncorrelated with the x_{it} in that period. This produces the moment equations

$$\Sigma_t \, x_{it} \, \varepsilon_{it} = 0, \text{ for } i = 1,\ldots, 48.$$

Solving this for the one estimator of β produces the estimator in part 2. But, if we have strict exogeneity, then we have many additional equations. For example, this would imply the moment equations:

$$(1/n)\Sigma_i \, x_{i1} \, \varepsilon_{i2} = 0$$

The estimator implied by this moment condition is is the "IV" estimator

$$b_{12} = (X_2'X_1)^{-1}X_2'y_1$$

where X_2 is the 48 observations on x_{it} for period 1, and X_2 and y_1 are defined likewise. If you start listing these out, wou will see that this actually produces hundreds of additional estimators of β that can be added to the computation in 2. (One might wonder if this is really such a good idea.)

Part III. Dynamic Model

Consider the dynamic, linear, cross country, *random effects* regression model

$$y_{it} = \alpha + \beta x_{it} + \delta z_{it} + \gamma y_{i,t-1} + u_i + \varepsilon_{it},\ t = 1,...,4 \text{ (and } y_{i,0} \text{ is observed data).}$$

in which i is a country and t is a year; y_{it} is national income per capita, z_{it} is domestic investment and x_{it} is a measure of national labor input. You have 30 countries and 4 years of data.

1. Show that the pooled ordinary least squares estimator is inconsistent.

The variable $y_{i,t-1}$ must be correlated with u_i which appears in the disturbance.

2. Show how the Hausman and Taylor approach can be used to obtain consistent estimators of $(\alpha,\beta,\delta,\gamma)$.

Clearly it is an example of the H&T framework, though there are no time invariant variables in the model save for the constant In the H&T framework, x_{i1} is (x,z) and x_{i2} is (y_{-1}). z_{i1} is (1) and z_{i2} is null. Just counting variables, we find $K_1 = 2$ is certainly greater than $L_2 = 0$, so the H&T approach does appear to be available.

3. Let $w_{it} = (y_{it} - \alpha - \beta x_{it} - \delta z_{it} - \gamma y_{i,t-1})$. Consider the set of instruments $\mathbf{f}_{it} = (1, x_{it}, z_{it}, x_{i,t-1}, z_{i,t-1})$. Let **F** be a 120×4 matrix of instrumental variables, and **X** be the 120×4 matrix of data in the model. Does the simple strategy of pooling the panel and simply using two stage least squares with **F** as the set of instruments produce a consistent estimator of the parameters? Explain.

One would expect that the lagged values of x_{it} and z_{it} would, indeed, be valid instrumental variables in this model, so with no further assumptions that would make them endogenous, yes, 2sls would work in this case.

4. Suppose the model is modified to allow the coefficient on z_{it} to differ across countries.

$$y_{it} = \alpha + \beta x_{it} + \delta_i z_{it} + \gamma y_{i,t-1} + u_i + \varepsilon_{it},\ t = 1,...,4 \text{ (and } y_{i,0} \text{ is observed data).}$$

Can you propose a consistent estimator of the parameters of this model when δ_i varies across countries? Explain.

This is going to be difficult. A fixed effects approach won't work. In principle, a "mixed fixed" effects estimator can be constructed, by building an interaction term between z_i and a country dummy variable and having a separate term for each z_i. This leaves the lagged dependent variable in the model, so the problem of the endogeneity of $y_{i,t-1}$ remains. This estimator is not consistent. Likewise, a random parameters approach ($\delta_i = \delta + w_i$) might seem appealing, which turns the model into an RPM with a random constant and one random slope. But, that lagged y still remains. So, consistency in this case, without an instrumental variable approach or a maximum likelihood estimator is not going to be obtainable.

Part IV. Analysis of Panel Data

The following analysis is based on a panel of data on the U.S. airline industry (from back in the good old days before flying was less pleasant than root canal surgery). The original data are an unbalanced panel of observations on 25 airlines with numnber of observations ranging from 2 to 15. One of the airlines is missing some essential data (for our purposes), so it was dropped leaving 24 airlines. We also doctored the data a bit, converting a time varying variable, POINTS = the number of cities served, to NODES = the airline average of POINTS, which is time invariant. The variables in the data set that are used in the regressions below are as follows:

C_{it} = total costs
Q_{it} = total output in revenue passenger miles
Pj_{it} = prices of five inputs, M=Materials, L=Labor, P=Property, F=Fuel, E=Equipment
LF_{it} = load factor (average proportion of seats occupied on a flight)
$Stage_{it}$ = average stage length = average length of flights
$Nodes_i$ = number of nodes (airports) in the airlines' route map in the given year.

We begin with a basic loglinear model, in which a variable name preceded by "L" indicates logs:

$$LC_{it} = \beta_1 LPM + \beta_2 LPF + \beta_3 LPL + \beta_4 LPE + \beta_5 LPP + \theta LQ + \gamma_1 LF + \gamma_2 Lstage + \gamma_3 LNodes + c_i + \varepsilon_{it}$$

1. The assumption of linear homogeneity in prices is an essential part of the theory underlying the cost function. This would be $\Sigma_{k=1}^{5}\beta_k = 1$. How would you test the restriction of linear homogeneity in the input prices in the context of the pooled linear regression model? Do the results given below provide the statistics you need to carry out the test? If yes, show how to do it. If not, explain why not – i.e., what do you need that is not provided.

In order to test the hypothesis of the restriction, in principle, one could use a Wald test. But, the covariance matrix for the unrestricted model is not given, so this will not work. The alternative is to compare the restricted and unrestricted models using the fit measures. One would normally do this using an F statistic based on the two R^2s, which might appear to be

F(1,236) = [(.9960416-.9943532)/1] / [(1 - .9960416)/(246 – 10)] = 100.66249.

This seems highly significant. The problem is that the restricted regression is obtained by imposing the restriction. This is done by subtracting LLP from the other 4 log prices, and from the dependent variable. Note in the second regression, there are only 4 price variables, and the dependent variable which was LC becomes LCP. The test must be carried out using the sum of squared residuals, not the R^2s. This is

F(1,236) = [(1.606532-1.439515)/1] / [(1.606532)/(246 – 10)] = 25.534884.

Still highly significant. The hypothesis would be rejected.

2. Using the pooled least squares results, test the hypothesis that $\gamma_1 = \gamma_2 = \gamma_3 = 0$. Can you carry out this test using the fixed effects results? Explain? How would you carry out this test using the random effects results? (Note, the precise numbers are easy to manipulate on paper, but tedious actually to manipulate. Just show me what you would like to compute – you need not actually carry out the computation.)

For this test, the only way to proceed is the Wald test. The statistic would be

$W = (c1, c2, c3)' V^{-1} (c1, c2, c3)$

Where c1, c2, and c3 are the estimates in the pooled results

```
|LF      |     .64803***       .04247028    15.259   .0000  -1.1094900|
|LSTAGE  |    -.09635***       .02204232    -4.371   .0000   6.0411056|
|NODES   |     .00239***       .00027239     8.777   .0000   72.983740|
```

And the covariance matrix needed is the lower right 3×3 matrix in the picture shown under the regression results. (No need to do the actual calculation.)

3. Based on the results given, which model do you think the analyst should report as their best estimates, the pooled least squares results, the fixed effects results or the random effects results? Justify your answer with the statistical evidence.

This judgment has to be based on the model that does not include NODES, as this is time invariant. These are the last results given below. The reported results,

```
| Lagrange Multiplier Test vs. Model (3) =  425.17 |
| ( 1 df, prob value =  .000000)                   |
| (High values of LM favor FEM/REM over CR model.) |
| Fixed vs. Random Effects (Hausman)     =   34.44 |
| ( 7 df, prob value =  .000014)                   |
```

Include an LM statistic of 425.17, which is a chi squared with 1 degree of freedom. This is large, and we can confidently reject the "no effects model." The Hausman statistic given of 34.44 is given with a P-value of .000014, which suggests that we should reject the random effects model in favor of the fixed effects model.

4. The hypothesis of constant returns to scale is $\theta = 1$. Carry out a test of this hypothesis using the model that you chose in part 3.

The fixed effects results are

```
+--------+-------------+----------------+--------+--------+----------+
|Variable| Coefficient  | Standard Error |t-ratio |P[|T|>t]| Mean of X|
+--------+-------------+----------------+--------+--------+----------+
|LPMP    |     .51163***       .05845300     8.753   .0000   2.5223929|
|LPFP    |     .21330***       .01622798    13.144   .0000   2.1502564|
|LPLP    |     .12843**        .04951746     2.594   .0101   2.8079962|
|LPEP    |    -.02844          .03435212     -.828   .4085   1.8883002|
|LQ      |     .24755***       .03597221     6.882   .0000  -1.1728132|
|LF      |     .60151***       .03907522    15.394   .0000  -1.1094900|
|LSTAGE  |    -.15213***       .03313002    -4.592   .0000   6.0411056|
+--------+------------------------------------------------------------+
```

To test the hypothesis, we would refer the statistic (.24755 – 1)/.03597 to the standard normal table. The statistic is 20.91, which is large. We would reject the hypothesis.

5. Notice that in the first set of results, the sum of squared residuals for the fixed effects estimator is .5100564 . In the second set of results, where the time invariant variable NODES is removed from the regression, the sum of squared residuals given for the fixed effects regression is .5100564 again!!. Shouldn't the sum of squared residuals increase when a variables is removed to the regression? Can you explain this strange outcome?

Unfortunately, as noted in class during the exam, you did not actually have the results to observe this outcome. The result follows from the fact that when we fit the FE model without NODES, we get a sum of squares of .5100564. To try to add NODES to the model, we are adding a variable that is a linear combination of variables that are already in the model. This cannot improve the fit of the model, so it produces the same sum of squares.

6. Using the first set of regression results, test the hypothesis that all the constant terms in the fixed effects model are equal to each other.

The first set of regression results for the panel data treatment includes

```
+---------------------------------------------------------------------+
|            Test Statistics for the Classical Model                  |
+---------------------------------------------------------------------+
|       Model            Log-Likelihood  Sum of Squares    R-squared |
|(1)  Constant term only     -366.94408  .2845018183D+03    .0000000 |
|(2)  Group effects only      -58.93258  .2325641805D+02    .9182556 |
|(3)  X - variables only      251.76657  .1859991285D+01    .9934623 |
|(4)  X and group effects     386.41860  .6224075306D+00    .9978123 |
+---------------------------------------------------------------------+
|                        Hypothesis Tests                             |
|         Likelihood Ratio Test           F Tests                     |
|         Chi-squared   d.f.  Prob.       F    num. denom.   P value |
|(2) vs (1)   616.023     23  .00000  108.425    23     222   .00000 |
|(3) vs (1)  1237.421      7  .00000 5166.595     7     238   .00000 |
|(4) vs (1)  1506.725     30  .00000 3268.709    30     215   .00000 |
|(4) vs (2)   890.702      7  .00000 1116.933     7     215   .00000 |
|(4) vs (3)   269.304     23  .00000   18.587    23     215   .00000 |
+---------------------------------------------------------------------+
```

The question asks for the test of "Model 4" vs. "Model 3" above. The F statistic is given at the bottom of the table, 18.587, with a P value of .00000. We would reject the hypothesis that all the constant terms are the same.

7. In a cost function such as this, the assumption that the output variable is exogenous is sometimes justified by an appeal to the regulatory environment in which some regulatory body sets the prices for the firm and they must accept all demand that is forthcoming. The argument works for electricity or gas providers. It probably doesn't work for profit maximizing airlines. In general terms, how would you want to change your estimation strategy to deal with the possibility that the output variable is endogenous in the model.

If this were the case, I would look for a version of instrumental variables that accommodates fixed effects. In fact this is straightforward. We would need, first, to obtain the IV. In hand, an FE transformation (deviations from means) does solve the problem. The real problem is locating the valid instrumental variable.

8. The random effects model in the first results embodies an undesirable assumption of uncorrelatedness of c_i and the independent variables. The fixed effects model has many coefficients and is inefficient (possibly). The Mundlak approach represents a compromise of these two. Describe how to use Mundlak's estimator in this model.

Mundlak's approach would write the FE as

$\alpha_i = \delta' \text{x-bar}_i + u_i$

where x-bar is the group means of the time varying variables - and time invariant variables stay in the model. This turns the FE model into an RE model which contains the group means as additional variables.