Problem set 2: Panel Data Models

1. You have a sample of N individuals for T years. Suppose you estimate by OLS the annual income equation:

$$y_{it} = \alpha_0 + \alpha_1 ed_i + \alpha_2 age_{it} + \alpha_3 (ed_i \times age_{it}) + \gamma y_{it-1} + u_{it}$$

where ed_i represents the years of education of the ith individual, age_{it} represents the age of the individual i in period t and u_{it} represents all unobservables.

   (a) Suppose you estimate γ as 0.82 with the standard error of 0.12. State a set of sufficient assumptions for the consistency of the OLS estimator in this context.

   (b) Describe an alternative estimation technique that you could use to evaluate the validity of some of your assumptions. Justify your choice and explain carefully the conditions under which your alternative estimator is consistent.

2. Consider the model with a single regressor x_{it}

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}$$

where α_i represents an unobserved effect fixed over time and u_{it} is a homoskedastic error term which is independent over time (t) and individuals (i). There are N randomly sampled individuals, each observed for $T = 4$ time periods. Assume that $E(u_{it} \mid X) = 0$ for all i and that $E(u_{it}u_{is} \mid X, \text{ any } t \text{ and } s : t \neq s) = 0$, where X represents the $NT \times 1$ data matrix.

   (a) Derive the covariance matrix of the Within Groups estimator and for the random effects estimator.

   (b) Explain how you could test the assumption that $E(\alpha_i \mid x_{it}) = 0$

3. You wish to study the effects of unionisation on wages using a panel of N individuals and T time periods. You wish to allow for the following phenomena: a) unionised firms select the higher ability workers and b) workers with bad productivity shocks join the union sector.

   (a) Set up a suitable model and explain how these phenomena are reflected in your specification.

   (b) Explain how you would estimate this model and present the estimator. Carefully state any assumptions you make.

4. Suppose you decide to estimate the single β parameter in

$$y_{it} = x_{it}\beta + f_i + u_{it}$$

by OLS on the first differences model when x_{it} is strictly exogenous and there are $T > 2$ time periods of data available for N individuals. Assume f_i is unobserved and $\text{var}(\Delta x_{it}) > 0$ where $\Delta x_{it} = x_{it} - x_{it-1}$.

   (a) Show that this estimator is consistent.

   (b) Derive its variance assuming that u_{it} is serially uncorrelated and homoskedastic.

   (c) Compare its variance to that of the within groups estimator for β. (Hint: one of the difficulties arises from the fact that $\triangle\ u_{it}$ is an $MA(1)$ process. Hence, there is a special form of serial correlation.)

5. Suppose you wish to estimate a dynamic model of the form

$$\begin{aligned} y_{it} &= \beta x_{it} + f_i + u_{it} \\ u_{it} &= \rho u_{it-1} + e_{it} \end{aligned}$$

where f_i is an unobserved fixed effect and the unobservables e_{it} are independent and identically distributed over time. The single regressor x_{it} may be correlated with f, is uncorrelated with e_{it} but is not strictly exogenous.

   (a) Derive a consistent estimator for β. State carefully any assumptions you might have to make and also the minimum number of observations required for estimation.

   (b) What is the covariance matrix of your estimator?

   (c) Suggest a way of testing the hypothesis that $\rho = 0$ and describe a consistent estimator for β under the hypothesis that $\rho = 0$. State carefully any assumptions you might have to make and also the minimum number of observations required for estimation.

   (d) Would your estimation strategy change if there was no fixed effect when:

      i. $\rho = 0$?

      ii. $\rho \neq 0$?

6. You have a panel data set which contains repeated observations on log real annual earnings (`lny`) for a number of individuals. For each individual you also observe the age (`age`) and an

education indicator (`educ`). This takes the values of 1 to 4 with 1 being the lowest education group. Finally `year` is an indicator of time and `newid` is a personal identification code. The data is stored in STATA format and sorted by individual and year. It is named incpanel.dta.

(a) Estimate by OLS a dynamic earnings equation using as explanatory variables lagged income, education, age, age squared and time dummies.

Notice that dummies for a discrete variable x can be constructed as follows:

```
ta(x), gen(xd)
```

(b) Using the `egen` function in STATA construct individual means of the data and using these perform a within groups transformation on income.

(c) Estimate the model using within groups. Will you include education? Will you include the time dummies? Will you include age and age squared?

(d) Explain what the coefficients on the time dummies mean.

(e) Create the first differences of income. To create the first differences of a variable x in a panel follow these steps:

```
sort newid year
by newid:  gen dx=x-x[_n-1]
```

(f) Estimate the model in first differences using OLS and then IV. Will you include time dummies, age and age squared. What instruments did you use? Compare the results.

(g) Comment on the validity of the standard errors that the package provides in each case.

7. Use Cornwell.dta

(a) Estimate a random effects and a fixed effects models relating the logarithm of crime rate to `lprbarr`, `lprbconv`, `lprbpris`, `lavgsen`, and `lpolpc`.

(b) Compute the regression-based version of the Hausman test comparing RE and FE.

(c) Add the wage variables (in logarithmic form), and test for joint significance after estimation by fixed effects.

(d) Estimate the equation by first differences, and comment on any notable changes. Do the standard errors change much between fixed effects and first differences?