Barcelona School of Economics
Albrecht Glitz

Problem Set 2: Panel Data

Please submit a document with your answers, including Stata or R output and programs, to our TA Janik Deutscher via the Google Classroom by ***Thursday, 29 January, 18:00*** *at the very latest. You can work in groups of up to four.*

1. You have a sample of N individuals for T years. Suppose you estimate by Pooled 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 a standard error of 0.05. State a set of sufficient assumptions for the consistency of the Pooled OLS estimator in this context.

b) Describe an alternative estimation technique and general procedure 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 following model:

$$y_{it} = \alpha + x_{it}\beta + z_{it}\gamma + f_i + u_{it} \qquad \text{for i=1,...,N and t=1,...,T,}$$

where x_{it} and z_{it} are scalars, f_i is a permanent unobserved effect and the error term u_{it} is homoscedastic and serially uncorrelated. Furthermore,

assume that x_{it} is strictly exogenous:

$$E(u_{it}|x_{i1}, ..., x_{iT}, f_i) = 0. \quad (1)$$

a) Suppose the only thing we can safely assume is that the orthogonality condition (1) holds and you take first differences to estimate the model. For different assumptions regarding the exogeneity of z_{it} and the relationship between z_{it} and f_i, state the properties of your estimator (consistency and efficiency) when OLS is used to estimate β and γ in the first differences model?

b) Now suppose T=5 and the additional assumption holds:

$$E(u_{it}|z_{i1}, ..., z_{it-1}, f_i) = 0.$$

i. How would you now estimate β and γ efficiently?
ii. Derive the variance of your estimator.

c) Suppose $z_{it} = y_{it-1}$ and you know that $\gamma = 1$. How would you estimate the model? Provide the estimator and its variance matrix.

3. Download the SOEP practise data set *soep_lebensz_en.dta*, available on the course website. As in Problem Set 1, construct the binary variable *has_kids* that indicates if a person at time t has any children. For each individual, keep only the first two time periods. In the following regressions, include as control variables only education and an individual's current standardized health status, and use non-clustered, non-robust standard errors for simplicity.

a) Estimate the effect of the child indicator on standardized life satisfaction in a first-difference model. Next, estimate the effect with a fixed effects regression. Do you expect the estimated coefficients to differ? How would you interpret the estimated coefficients that you obtain? (Hint:

Make sure that you explicitly exclude all variables that are differenced-out in the FD model.)

b) Start again with the full sample. This time keep all time periods. Re-estimate the FE and FD specifications of the previous question. Do the estimated coefficients differ? Why? Now assume that the assumptions for consistency of the FE and FD estimators hold in your model. In theory, when is the FD estimator efficient, when the FE estimator? In your context, which estimator would you expect to be more efficient? Why? Which estimate, $\hat{\beta}^{FD}$ or $\hat{\beta}^{FE}$, has the higher standard error? How could you make your standard errors more robust to deviations from the assumed structure of the idiosyncratic error terms?

c) Next, we build a dynamic model of life satisfaction. Intuitively, would you expect current life satisfaction and past life satisfaction to be positively or negatively related? Why? Now estimate a fixed effects model that includes, besides the first lag of life satisfaction, an individual's education and current standardized health status as well as an indicator for having children as additional control variables. Which sign has the coefficient of lagged life satisfaction? Is this what you expected? Do you think the coefficient is unbiased? Why? Why not?

d) Now, estimate your dynamic panel data model using the Arellano-Bond Estimator (Hint: use the Stata command `xtabond` or the R package `pdynmc`). Include one lag of the dependent variable and use at most 2 lags as instruments. Under which assumptions is your estimate of the coefficient on the lagged dependent variable unbiased? Test for serial correlation in the error terms u_{it}. What do you conclude? If unbiased, would you expect the Arellano-Bond estimate to be more positive or more negative relative to the FE estimate from the previous question? Why? Which sign has the coefficient of lagged life satisfaction now? Interpret the magnitudes of your estimated coefficients?