Econometrics II

Lecture 6: Panel Data Part I

Måns Söderbom*

18 April 2011

*Department of Economics, University of Gothenburg. Email: mans.soderbom@economics.gu.se. Web: www.economics.gu.se/soderbom, www.soderbom.net.

1. Introduction

Recall that the main theme of the Angrist-Pischke book concerns how to estimate causal effects with non-experimental data. With such data, we must usually address the problem posed by selection, or omitted variables. These are often referred to as confounding factors.

If confounding factors are observed in the data, the solution is simple: just control for these in the regressions - see discussion in AP Chapter 3.

If important confounding factors are unobserved, we may be able to estimate the parameters of interest consistently using instrumental variables - see AP Chapter 4.

Now we will turn to an alternative setting in which we can control for certain types of unobserved variables without using instrumental variables. This setting is one in which **panel data** are available.

Panel data-sets follow a random sample of individuals (or firms, households, etc.) over time.

The big advantage of working with panel data is that we will be able to control for **individual-specific, time-invariant, unobserved heterogeneity**, the presence of which could lead to bias in standard estimators like OLS. We can also estimate dynamic equations.

In this lecture I will first discuss what the data need to look like, for the econometrics that then follow to be relevant. I then discuss standard panel data estimators. This is followed by a brief discussion of model selection. Finally, I discuss problems posed by endogeneity in the context of panel data models.

Chapter 5 in Angrist & Pischke is quite brief, and casts the panel data estimators in a treatment framework. I will discuss this in my next lecture.

Useful references for this lecture:

- Greene 9.1-9.5.
- *Optional*: Wooldridge (2002) "Cross Section and Panel Data": Chapters 7.8; 10; 11. (Personally I think this exposition is much better than what you'll find in Greene and AP). Some of my notes below draw on Wooldridge's presentation.

2. Combined Time Series and Cross Section Data

Panel data combine a time series dimension with a cross section dimension, in such a way that there are data on N individuals (or firms, countries...), followed over T time periods. Not all data-sets that combine a time series dimension with a cross section dimension are panel data-sets, however. It is important to distinguish **panel data** from **repeated cross-sections**.

2.1. A Panel Data-Set

- Panel data contains information on the **same** cross section units - e.g. individuals, countries or firms - over time. The structure of a panel data set is as follows:

id	year	yr92	yr93	yr94	x1	x2
1	1992	1	0	0	8	1
1	1993	0	1	0	12	1
1	1994	0	0	1	10	1
2	1992	1	0	0	7	0
2	1993	0	1	0	5	0
2	1994	0	0	1	3	0
(...)	(...)	(...)	(...)	(...)	(...)	(...)

where id is the variable identifying the individual that we follow over time; yr92, yr93 and yr94 are time dummies, constructed from the year variable; x1 is an example of a **time varying variable** and x2 is an example of a **time invariant variable**.

- In microeconomic data, N (the number of individuals, firms...) is typically large, while T is small.
- In aggregate data, longer T is more common.
- The econometric theory discussed by Greene in Chapter 9 and AP Chapter 5 assumes that N is 'large' while T is 'small'.
- In the opposite case, say $N = 5$ countries and $T = 40$ years, the topic becomes multiple time series.

- Throughout these lectures, I will focus mostly on the large N, small T case. I leave it to Joakim Westerlund to cover the case where T is large.

- If the time periods for which we have data are the same for all N individuals, e.g. $t = 1, 2, ...T$, then we have a **balanced panel.** In practice, it is common that the length of the time series and/or the time periods differs across individuals. In such a case the panel is **unbalanced.**

- Analyzing unbalanced panel data typically raises few additional issues compared with analysis of balanced data. However if the panel is unbalanced for reasons that are not entirely random (e.g. because firms with relatively low levels of productivity have relatively high exit rates), then we may need to take this into account when estimating the model. This can be done by means of a sample selection model (more on this estimator in Lecture 10). We abstract from this particular problem here.

- **Repeated cross sections** are **not** the same as panel data. Repeated cross sections are obtained by sampling from the same population at different points in time. The identity of the individuals (or firms, households etc.) is **not** recorded, and there is **no attempt to follow individuals over time**. This is the key reason why pooled cross sections are different from panel data. Had the id variable in the example above not been available, we would have referred to this as a pooled repeated cross-section data-set.

2.2. New Opportunities

When we have a dataset with both a time series and a cross-section dimension, this opens up new opportunities in our research. For example:

- Larger sample size than single cross-section, and so you should be able to obtain **more precise estimates** (i.e. lower standard errors).

- You can now ask how certain effects evolve over time (e.g. time trend in dependent variable; or changes in the coefficients of the model).

- Panel data enable you to solve an **omitted variables problem**.

- Panel data also enable you to estimate dynamic equations (e.g. specifications with lagged dependent variables on the right-hand side).

3. Using Panel Data To Address an Endogeneity Problem

- Arguably the main advantage of panel data is that such data can be used to solve an **omitted variables problem**. Suppose our model is

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + (\alpha_i + u_{it}),$$

  $t = 1, 2, ..., T$, where we observe y_{it} and $\boldsymbol{x}_{it}$, and α_i, u_{it} are not observed. Our goal is to estimate the parameter $\boldsymbol{\beta}$. As usual, $\boldsymbol{x}_{it}$ is a $1 \times K$ vector of regressors, and $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters to be estimated.

- Throughout this lecture I will assume that the residual u_{it}, which varies both over time and across individuals, is serially uncorrelated.

- Our problem is that we do not observe α_i, which is constant over time for each individual (hence no t subscript) but varies across individuals. Hence if we estimate the model in levels using OLS then α_i will go into the error term: $v_{it}^{OLS} = \alpha_i + u_{it}$.

- What would be the consequence of α_i going into the error term?

- If α_i is **uncorrelated** with $\boldsymbol{x}_{it}$, then α_i is just another unobserved factor making up the residual. It is true that OLS will not be BLUE, because the error term v_{it}^{OLS} is serially correlated:

$$corr\left(v_{it}^{OLS}, v_{i,t-s}^{OLS}\right) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2}$$

  for $s = 1, 2, ...$ (this calculation assumes that u_{it} is non-autocorrelated; more on this below). This

suggests some feasible generalized least squares estimator could be preferable (this is indeed the case; see below). Notice that OLS would be consistent, however, and the only substantive problem with relying on OLS for this model is that the standard formula for calculating the standard errors are wrong. This problem is straightforward to solve, e.g. by clustering the standard errors on the individuals.

- But if α_i is **correlated** with $\boldsymbol{x}_{it}$, then putting α_i in the error term can cause serious problems. This, of course, is an omitted variables problem, so we can use some of familiar results to understand the nature of the problem. For the single-regressor model,hence

$$p\lim \hat{\beta}^{OLS} = \beta + \frac{cov\left(x_{it}, \alpha_i\right)}{\sigma_x^2},$$

  which shows that the OLS estimator is inconsistent unless $cov\left(x_{it}, \alpha_i\right) = 0$. If x_{it} is positively correlated with the unobserved effect, then there is an upward bias. If the correlation is negative, we get a negative bias.

- Can you think about applications for which a specification like the following

$$y_{it} = \beta \boldsymbol{x}_{it} + \left(\alpha_i + u_{it}\right),$$

  would be appropriate? How about:

  - Individual earnings
  - Household expenditures
  - Firm investment
  - Country income per capita.

What factors can reasonably be represented by α_i? Can these be assumed uncorrelated with x_{it}?

3.1. Model 1: The Fixed Effects ("Within") Estimator

- Model:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + (\alpha_i + u_{it}), \qquad t = 1, 2, ...T; i = 1, 2, ..., N, \tag{3.1}$$

  where I have put $\alpha_i + u_{it}$ within parentheses to emphasize that these terms are unobserved.

- Assumptions about unobserved terms:

  - Assumption 1.1: α_i freely correlated with $\boldsymbol{x}_{it}$
  - Assumption 1.2: $E(\boldsymbol{x}_{it}u_{is}) = \mathbf{0}$ for $s = 1, 2, ..., T$ (strict exogeneity)

- We have seen that if α_i is correlated with the variables in the $\boldsymbol{x}_{it}$ vector, there will be an endogeneity problem which would bias the OLS estimates. Under assumptions 1.1 and 1.2, we can use the **Fixed Effects** (FE) or the **First Differenced** (FD) estimator to obtain consistent estimates of $\boldsymbol{\beta}$ allowing α_i to be freely correlated with $\boldsymbol{x}_{it}$.

- Note that strict exogeneity rules out feedback from past u_{is} shocks to current $\boldsymbol{x}_{it}$. One implication of this is that FE and FD will not yield consistent estimates if $\boldsymbol{x}_{it}$ contains lagged dependent variables $(y_{i,t-1}, y_{i,t-2}, ...)$.

- If the assumption that $E(\boldsymbol{x}_{it}u_{is}) = \mathbf{0}$ for $s = 1, 2, ..., T$, does **not** hold, we may be able to use instruments to get consistent estimates. This will be discussed later.

To see how the FE estimator solves the endogeneity problem that would contaminate the OLS estimates, begin by taking the average of (3.1) for **each individual** - this gives

$$\bar{y}_i = \bar{\boldsymbol{x}}_i\boldsymbol{\beta} + (\alpha_i + \bar{u}_i), \qquad i = 1, 2, ..., N, \tag{3.2}$$

where $\bar{y}_i = \left(\sum_{t=1}^{T} y_{it}\right)/T$, and so on.[1] Now subtract (3.2) from (3.1):

$$\begin{aligned} y_{it} - \bar{y}_i &= (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)\boldsymbol{\beta} + (\alpha_i - \alpha_i + u_{it} - \bar{u}_i), \\ y_{it} - \bar{y}_i &= (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)\boldsymbol{\beta} + (u_{it} - \bar{u}_i), \end{aligned}$$

which we write as

$$\ddot{y}_{it} = \ddot{\boldsymbol{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}, \qquad t = 1, 2, ...T; i = 1, 2, ..., N, \tag{3.3}$$

where $\ddot{y}_{it}$ is the **time-demeaned data** (and similarly for $\ddot{\boldsymbol{x}}_{it}$ and $\ddot{u}_{it}$).

This transformation of the original equation, known as the ***within transformation****, has eliminated* α_i *from the equation.*

- Hence, we can estimate $\boldsymbol{\beta}$ consistently by using OLS on (3.3). This is called the **within estimator** or the **Fixed Effects estimator.**
- You now see why this estimator requires strict exogeneity: the equation residual in (3.3) contains all residuals $u_{i1}, u_{i2}, ..., u_{iT}$ (since these enter $\ddot{u}_{it}$) whereas the vector of transformed explanatory variables contains all values of the explanatory variables $x_{i1}, x_{i2}, ..., x_{iT}$ (since these enter $\bar{\boldsymbol{x}}_i$). Hence we need $E(\boldsymbol{x}_{it}u_{is}) = \mathbf{0}$ for $s = 1, 2, ..., T$, or there will be endogeneity bias if we estimate (3.3) using OLS.
- In Stata, we obtain FE estimates from the 'xtreg' command if we use the option 'fe', e.g.
- xtreg yvar xvar, i(firm) fe
- Rather than time demeaning the data, couldn't we just estimate (3.1) by including one dummy variable for each individual (or country, firm...)? Indeed we could, and it turns out that this is exactly the same estimator as the within estimator (can you prove this?). If your N is large, so that you have a large number of dummy variables, this may not be a very practical approach however.

[1] Without loss of generality, the exposition here assumes that T is constant across individuals, i.e. that the panel is balanced.

3.2. Model 2: The First Differencing Estimator

- Model:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + (\alpha_i + u_{it}), \qquad t = 1, 2, ...T; i = 1, 2, ..., N, \tag{3.4}$$

  where α_i, u_{it} are unobserved.

- Assumptions about unobserved terms:

  - Assumption 2.1: α_i freely correlated with $\boldsymbol{x}_{it}$
  - Assumption 2.2: $E(\boldsymbol{x}_{it}u_{is}) = \mathbf{0}$ for $s = t, t-1$. This is a weaker form of strict exogeneity than what is required for FE, in the sense that $E(\boldsymbol{x}_{it}u_{i,t-2}) = \mathbf{0}$, for example, is not required). Thus, if there is feedback from u_{it} to $\boldsymbol{x}_{it}$ that takes more than two periods, FD will be consistent whereas FE will not (hence weaker form of strict exogeneity).

- Starting from the model in (3.4), but rather than time-demeaning the data (which gives the FE estimator), we now difference the data:

$$\begin{aligned} y_{it} - y_{i,t-1} &= (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})\boldsymbol{\beta} + (\alpha_i - \alpha_i + u_{it} - u_{i,t-1}), \\ \Delta y_{it} &= \Delta \boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta u_{it}. \end{aligned} \tag{3.5}$$

  Clearly this removes the individual fixed effect, and so we can obtain consistent estimates of $\boldsymbol{\beta}$ by estimating the equation in first differences by OLS.

- You now see why this estimator requires strict exogeneity: the equation residual in (3.5) contains the residuals u_{it} and $u_{i,t-1}$ whereas the vector of transformed explanatory variables contains $x_{it}, x_{i,t-1}$. Hence we need $E(\boldsymbol{x}_{it}u_{is}) = \mathbf{0}$ for $s = t, t-1$, or there will be endogeneity bias if we estimate (3.5) using OLS.

FE or FD?

- So FE and FD are two alternative ways of removing the fixed effect. Which method should we use?

- First of all, when $T = 2$ (i.e. we have only two time periods), FE and FD are exactly equivalent and so in this case it does not matter which one we use (try to prove this).

- But when $T \geq 3$, FE and FD are not the same. Under the null hypothesis that the model is correctly specified, FE and FD will differ only because of sampling error whenever $T \geq 3$. Hence, if FE and FD are significantly different - so that the differences in the estimates cannot be attributed to sampling error - we should worry about the validity of the strict exogeneity assumption.

- If u_{it} is a random walk ($u_{it} = u_{i,t-1} + \xi_{it}$), then Δu_{it} is serially uncorrelated and so the FD estimator will be more efficient than the FE estimator.

- Conversely, under "classical assumptions", i.e. $u_{it} \sim iid\left(0, \sigma_u^2\right)$, the FE estimator will be more efficient than the FD estimator (as in this case the FD residual Δu_{it} will exhibit negative serial correlation).

3.3. Model 3: The Pooled OLS Estimator

- Model:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + (\alpha_i + u_{it}), \qquad t = 1, 2, ...T; i = 1, 2, ..., N \tag{3.6}$$

  where α_i, u_{it} are unobserved.

- Assumptions about unobserved terms:

  - Assumption 3.1: α_i is **uncorrelated** with $\boldsymbol{x}_{it}$: $E(\boldsymbol{x}_{it}\alpha_i) = \mathbf{0}$

  - Assumption 3.2: $E(\boldsymbol{x}_{it}u_{it}) = \mathbf{0}$ ($\boldsymbol{x}_{it}$ predetermined)

- Note that A3.1 is **stronger** than A1.1 and A2.1 whereas A3.2 is **weaker** than A1.1 and A2.1. Clearly under these assumptions, $v_{it}^{OLS} = (\alpha_i + u_{it})$ will be uncorrelated with $\boldsymbol{x}_{it}$, implying we can estimate $\boldsymbol{\beta}$ consistently using OLS. In this context we refer to this as the **Pooled OLS (POLS) estimator**.

- To do inference based on the conventional OLS estimator of the covariance matrix, we need to assume homoskedasticity and no serial correlation in the data. Both of these assumptions can be restrictive, especially the latter one. As a rule of thumb, it is a good idea to obtain an estimate of the covariance matrix that is **robust** to heteroskedasticity **and** autocorrelation, using the following 'sandwich' formula:

$$V^R\left(\hat{\boldsymbol{\beta}}^{POLS}\right) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i' \boldsymbol{X}_i\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{X}_i\right)^{-1},$$

where $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT})$ and $\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, \hat{u}_{i2}, ..., \hat{u}_{iT})$. In Stata, we get this by using the option 'cluster', e.g.

regress yvar xvar, cluster(firm)

3.4. Model 4: The Random Effects Estimator

- Model:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + (\alpha_i + u_{it}), \quad t = 1, 2, ...T; i = 1, 2, ..., N.$$

- Assumptions about unobserved terms:
  - Assumption 4.1: α_i uncorrelated with $\boldsymbol{x}_{it}$: $E(\boldsymbol{x}_{it}\alpha_i) = \mathbf{0}$
  - Assumption 4.2: $E(\boldsymbol{x}_{it}u_{is}) = \mathbf{0}$ for $s = 1, 2, ..., T$ (strict exogeneity)

- Note that this combines the strongest assumption underlying FE/FD estimation (strict exogeneity) with the strongest assumption underlying POLS estimation (no correlation between time invariant part of residual and the explanatory variables). Why we need these assumptions will be clear below.

- So clearly Models 1-3 above will all give consistent estimates under A4.1-2.

- Consider using POLS in this case. It is straightforward to show that POLS is inefficient since the

residual $v_{it}^{OLS} = (\alpha_i + u_{it})$ is serially correlated. To see this, note that

$$\begin{aligned} E\left(v_{it}^{OLS}, v_{i,t-s}^{OLS}\right) &= E[(\alpha_i + u_{it})(\alpha_i + u_{i,t-s})] \\ &= E\left(\alpha_i^2 + \alpha_i u_{it} + \alpha_i u_{i,t-s} + u_{it}u_{i,t-s}\right) \\ &= E\left(\alpha_i^2\right) \\ &= \sigma_\alpha^2, \end{aligned}$$

and so

$$\begin{aligned} corr\left(v_{it}^{OLS}, v_{i,t-s}^{OLS}\right) &= \frac{E\left(v_{it}^{OLS}, v_{i,t-s}^{OLS}\right)}{\sqrt{\sigma_{v_t}^2 \sigma_{v_{t-s}}^2}} \\ corr\left(v_{it}^{OLS}, v_{i,t-s}^{OLS}\right) &= \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2}, \end{aligned} \tag{3.7}$$

for $s = 1, 2, ...$, since $\sigma_{v_t}^2 = \sigma_{v_{t-s}}^2 = \sigma_\alpha^2 + \sigma_u^2$. (this calculation assumes that u_{it} is non-autocorrelated, which I have already assumed). If we are concerned with **efficiency**, we may want to consider a GLS estimator that takes this serial correlation into account. Also note that if σ_α^2 is high relative to σ_u^2 the serial correlation in the residual will be high. As a result the conventional estimator of the covariance matrix for the OLS estimator will not be correct (but we've seen how we can fix this problem for POLS by 'clustering').

- The **Random Effects** (RE) estimator is a GLS estimator that takes (3.7) into account. This works as follows.

The RE Transformation

- Using GLS involves transforming the original equation, so that the transformed equation fulfils the assumptions underlying the classical linear regression model.

- Panel data model

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \alpha_i + u_{it}.$$

- Define

$$\lambda = 1 - \left(\frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2} \right)^{1/2}.$$

- Multiply λ by the individual average of the original equation:

$$\lambda \bar{y}_i = \lambda \bar{\boldsymbol{x}}_i \boldsymbol{\beta} + \lambda \bar{v}_i^{RE}.$$

- Subtract this expression from the original equation:

$$y_{it} - \lambda \bar{y}_i = (\boldsymbol{x}_{it} - \lambda \bar{\boldsymbol{x}}_i)\, \boldsymbol{\beta} + \left(v_{it}^{RE} - \lambda \bar{v}_i^{RE} \right).$$

  Using OLS on this, the transformed equation, gives the random effects GLS estimator.

- This estimator is efficient, because

$$v_{it}^{RE} - \lambda \bar{v}_i^{RE}$$

  is now serially uncorrelated (see Appendix for a proof).

- The parameter λ is not known and so it has to be estimated first. This involves estimating σ_u^2 and σ_α^2. There are various ways of doing this. The simplest, perhaps, is to use POLS in the first stage to obtain estimates of the composite residual $\hat{v}_{it}$. Based on this, we can calculate σ_α^2 as the covariance between $\hat{v}_{it}$ and $\hat{v}_{i,t-1}$ (for instance), and

$$\hat{\sigma}_u^2 = \hat{\sigma}_v^2 - \hat{\sigma}_\alpha^2.$$

  We can then plug $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_u^2$ into the formula for λ, and then estimate the transformed equation.

- In Stata, we can obtain the RE GLS estimator by using the command 'xtreg', e.g.

xtreg yvar xvar, i(firm)

- If $\lambda = 0$, what do we get? What if $\lambda = 1$?

[EXAMPLE: Section 1 in appendix - production function estimates.]

4. Model Selection

We have discussed four estimators of the panel data model:

1. Fixed Effects
2. First Differences
3. Pooled OLS (i.e. OLS estimation of the levels equation)
4. Random Effects

Which one should we use? The following tests will provide some guidance:

- Testing for non-zero correlation between the unobserved effect and the regressor(s): FE versus RE.
- Testing for the presence of an unobserved effect: RE versus pooled OLS

Testing for non-zero correlation between the unobserved effect and the regressor(s). Reference: Greene, 9.5.4.

- An important consideration when choosing between a random effects and fixed effects approach is whether α_i is correlated with $\boldsymbol{x}_{it}$. To test the hypothesis that α_i is **uncorrelated** with $\boldsymbol{x}_{it}$, we can use a **Hausman test**.
- Recall that the Hausman test in general involves comparing one estimator which is consistent regardless of whether the null hypothesis is not true or not, to another estimator which is only consistent under the null hypothesis.
- In the present context, the FE estimator is consistent regardless of whether α_i is or isn't correlated with x_{it}, while the RE requires this correlation to be zero in order to be consistent. Strict exogeneity is assumed for both models.
- The null hypothesis is that **both models are consistent**, and a statistically significant difference is therefore interpreted as evidence against the RE model. If we cannot reject the null, we may decide to use the RE model in the analysis on the grounds that this model is efficient.

- The Hausman statistic is computed as

$$\begin{aligned} H &= \left(\hat{\beta}^{FE} - \hat{\beta}^{RE}\right)' \left[var\left(\hat{\beta}^{FE}\right) - var\left(\hat{\beta}^{RE}\right)\right]^{-1} \\ &\quad \times \left(\hat{\beta}^{FE} - \hat{\beta}^{RE}\right), \end{aligned}$$

  using matrix notation. Under the null hypothesis, this test statistic follows a chi-squared distribution with M degrees of freedom, where M is the number of time varying explanatory variables in the model.

- Notice that the Hausman test by default compares **all** the parameters in the model across the two estimators. Sometimes we are primarily interested in a single parameter, in which case we can use a t test that ignores the other parameters. If β_1 is the element in β that we wish to use in the test, then the Hausman t statistic is simply

$$t^H = \frac{\hat{\beta}_1^{FE} - \hat{\beta}_1^{RE}}{\left[se\left(\hat{\beta}_1^{FE}\right)^2 - se\left(\hat{\beta}_1^{RE}\right)^2\right]^{1/2}},$$

  where se denotes the standard error of the estimated coefficient. Notice that $t^H = \sqrt{H}$ if $M = 1$. The t^H statistic has a standard normal distribution i.e. absolute values of t^H in excess of 1.96 suggests the null hypothesis should be rejected.

- Note that it is sometimes the case that there are large differences between the FE and RE point estimates of the coefficients but, due to high standard errors, the Hausman statistic fails to reject. What should be done in this case? A typical response is to conclude that the RE assumptions hold and to focus on the RE estimates. In doing so, however, we may make a Type II error: failing to reject an assumption when it is false. We can't really know if this is the case or not. Some judgement is required here. If RE and FE estimates differ a lot but the Hausman test does not reject, then this is worth mentioning in the analysis.

Testing for the presence of an unobserved effect: The Breusch-Pagan test. Reference: Greene, 9.5.3.

- If the regressors are strictly exogenous and u_{it} is non-autocorrelated and homoskedastic, then POLS and RE will both be efficient if there are no unobserved effects, i.e. $\sigma_\alpha^2 = 0$. If $\sigma_\alpha^2 > 0$ then RE is efficient (provided, of course, that α_i is uncorrelated with the explanatory variables).

- The most common test of $H_0 : \sigma_\alpha^2 = 0$ is the Lagrange multiplier test due to **Breusch and Pagan** (1980). This test is calculated in Stata by means of the instruction 'xttest0', following RE estimation. The test statistic is based on the pooled OLS residuals, and is written

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_i \left(\sum_t \hat{v}_{it} \right)^2}{\sum_i \sum_t \hat{v}_{it}^2} - 1 \right]^2$$

  for a balanced panel, where $\hat{v}_{it}$ is the estimated pooled OLS residual. Under the null hypothesis, LM is distributed as chi-squared with one degree of freedom.

- I will not derive the BP test statistic, but the intuition is as follows:

  - Under the null hypothesis $\left(\sum_t \hat{v}_{it} \right)^2 = \sum_t \hat{v}_{it}^2$, in which case the term within [.] will be zero.

  - However, if $\sigma_\alpha^2 > 0$, then $\left(\sum_t \hat{v}_{it} \right)^2 > \sum_t \hat{v}_{it}^2$, hence $LM > 0$. Notice that for $T = 2$,

$$\left(\sum_t \hat{v}_{it} \right)^2 = \hat{v}_{i1}^2 + 2\hat{v}_{i1}\hat{v}_{i2} + \hat{v}_{i2}^2,$$

    and $2\hat{v}_{i1}\hat{v}_{i2}$ will be strictly positive if $\sigma_\alpha^2 > 0$.

[EXAMPLE: Appendix Section 1 - model specification tests.]

5. Extensions

Now I will discuss some further issues that may arise when estimating panel data models. Most of this material is covered by Greene (albeit not in Chapter 12) but it should be enough just to focus on the

notes below.

5.1. What to do if some of your explanatory variables are time invariant?

The model:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + (\alpha_i + u_{it})\,.$$

Suppose you do not believe that the time invariant term α_i is uncorrelated with all your explanatory variables. You are therefore considering FE or FD. However, suppose some or all of the variables in the $\boldsymbol{x}_{it}$ vector are time invariant, i.e. they do not change over time:

$$\boldsymbol{x}_{it} = \left[\begin{array}{cc} \boldsymbol{w}_{it} & \boldsymbol{z}_i \end{array}\right],$$

where $\boldsymbol{w}_{it}$ is a $1 \times P$ vector of time varying variables, and $\boldsymbol{z}_i$ is a $1 \times (K - P)$ vector of time invariant variables. The model is now rewritten as

$$y_{it} = \boldsymbol{w}_{it}\boldsymbol{\beta}_1 + \boldsymbol{z}_i\boldsymbol{\beta}_2 + (\alpha_i + u_{it})\,.$$

The problem is now obvious: FE and FD will eliminate $\boldsymbol{z}_i$ and so $\boldsymbol{\beta}_2$ is not directly identified by these methods.

Now suppose you believe that α_i is uncorrelated with $\boldsymbol{z}_i$ (but correlated with $\boldsymbol{w}_{it}$). In this case, rather than resorting the POLS or RE, a better approach will be a two-stage procedure:

1. Estimate

$$y_{it} = \boldsymbol{w}_{it}\boldsymbol{\beta}_1 + (\lambda_i + u_{it})$$

using the FE or FD estimator. Back out the estimated fixed effects (one for each individual):

$$\hat{\lambda}_i = y_{it} - \boldsymbol{w}_{it}\hat{\boldsymbol{\beta}}_1.$$

Theoretically, $\hat{\lambda}_i$ is an estimate of $\boldsymbol{z}_i\boldsymbol{\beta}_2 + \alpha_i$.

2. Run the following regression:

$$\hat{\lambda}_i = \boldsymbol{z}_i\boldsymbol{\beta}_2 + v_i.$$

Under the assumption that α_i is uncorrelated with $\boldsymbol{z}_i$, this is a consistent estimator of $\boldsymbol{\beta}_2$.

In a more general case where some of the variables in $\boldsymbol{z}_i$ are correlated with α_i, identification of all the parameters in the model may still be possible if some of the variables in $\boldsymbol{w}_{it}$ are uncorrelated with α_i. For details on the **Hausman-Taylor estimator**, see Chapter 12.8 in Greene (optional).

5.2. What if Strict Exogeneity Doesn't Hold?

The model, again:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + (\alpha_i + u_{it}).$$

Strict exogeneity:

$$E(\boldsymbol{x}_{it}u_{is}) = 0, \quad s = 1, 2, ..., t, ..., T$$

- If α_i is correlated with $\boldsymbol{x}_{it}$ then - as we have seen - OLS and RE will generally be inconsistent, and FE or FD may be used instead, provided strict exogeneity holds. The FD approach, for example, involves transforming the data and then estimating the model

$$y_{it} - y_{i,t-1} = (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})\boldsymbol{\beta} + (u_{it} - u_{i,t-1})$$

  using OLS. Naturally, the resulting estimate of $\boldsymbol{\beta}$ will only be consistent if $(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})$ is uncorrelated with the equation residual $(u_{it} - u_{i,t-1})$. This will hold under strict exogeneity.

- But: **strict exogeneity may not hold!** Examples:
  - A seemingly mild form of non-exogeneity in this context is when u_{it} is uncorrelated with x_{it}

but correlated with **future** values of the regressors, e.g.

$$x_{i,t+1} = \varphi u_{it} + e_{i,t+1}.$$

A production function in which labour demand in period $t+1$ responds to unobserved productivity shocks in period t, represented by u_{it}, is one example of this. In this case there is **sequential exogeneity -** but not strict exogeneity.

– A second form of non-exogeneity occurs when there is **contemporaneous correlation** between the regressor and the residual, caused by, for instance, omitted variables or measurement errors in the explanatory variable uncorrelated with the true value of the regressor. That is, $E(x_{it}u_{it}) \neq 0$.

- In both of these cases, FE, FD and RE estimates will typically be inconsistent.
- In other words, if strict exogeneity does not hold, the panel data approach will not solve all our problems. If so, we need to do more work.[2]

5.3. Sequential Exogeneity

A less restrictive assumption than strict exogeneity is **sequential exogeneity**, which we state as

$$E(\boldsymbol{x}_{it}u_{is}) = 0, \quad s = t, t+1, ..., T.$$

When this assumption holds we will say that the $\boldsymbol{x}_{it}$ are sequentially exogenous. The key difference compared to strict exogeneity is that

$$E(\boldsymbol{x}_{it}u_{is}) = 0, \quad s = 1, 2, ..., t-1$$

[2] In the discussion of non-exogeneity I focus exclusively on the Fixed Effects (FE) and First Difference (FD) models and not the Random Effects (RE) model. The reason is that the standard version of the latter model does not even allow for correlation between the regressors and the time invariant individual effect.

is **not** imposed under sequential exogeneity. Thus, a process like

$$x_{i,t+1} = \varphi u_{it} + e_{i,t+1},$$

where $\varphi \neq 0$ and e_{it} is exogenous and non-autocorrelated, is consistent with sequential exogeneity, but not with strict exogeneity .

What is the implication of sequential exogeneity for the FE model? Recall that the FE model is equivalent to running an OLS regression on data expressed in deviations from individual means:

$$\ddot{y}_{it} = \ddot{\boldsymbol{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it},$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$, $\ddot{\boldsymbol{x}}_{it} = \boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i$ and $\ddot{u}_{it} = u_{it} - \bar{u}_i$. Hence for the FE model to be consistent we require

$$E\left(\ddot{\boldsymbol{x}}_{it}\ddot{u}_{it}\right) = 0$$

as usual. For the single regressor model, we have

$$\begin{aligned} E\left(\ddot{x}_{it}\ddot{u}_{it}\right) &= E\left[\left(x_{it} - \bar{x}_i\right)\left(u_{it} - \bar{u}_i\right)\right] \\ E\left(\ddot{x}_{it}\ddot{u}_{it}\right) &= E\left(x_{it}u_{it}\right) - E\left(\bar{x}_i u_{it}\right) - E\left(x_{it}\bar{u}_i\right) + E\left(\bar{x}_i\bar{u}_i\right) \end{aligned}$$

where $E\left(x_{it}u_{it}\right) = 0$ by implication of sequential exogeneity, but $E\left(\bar{x}_i u_{it}\right)$, $E\left(x_{it}\bar{u}_i\right)$ and $E\left(\bar{x}_i\bar{u}_i\right)$ will be non-zero since strict exogeneity does not hold.

We can do a similar analysis of the FD model:

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta u_{it},$$

hence

$$\begin{aligned} E\left(\Delta x_{it}\Delta u_{it}\right) &= E\left[\left(x_{it}-x_{i,t-1}\right)\left(u_{it}-u_{i,t-1}\right)\right] \\ &= E\left(x_{it}u_{it}\right)+E\left(x_{i,t-1}u_{i,t-1}\right)-E\left(x_{it}u_{i,t-1}\right)-E\left(x_{i,t-1}u_{it}\right) \\ &= -E\left(x_{it}u_{i,t-1}\right) \end{aligned}$$

which is different from zero unless there is strict exogeneity.

Example: Introduction to Dynamic Panel Data Models Consider the simple **autoregressive model**:

$$y_{it}=\beta_1 y_{i,t-1}+\left(\alpha_i+\varepsilon_{it}\right), t=1,2,...,T, \tag{5.1}$$

where $0\leq\beta_1<1$ and ε_{it} is non-autocorrelated (i.e. the process is stationary). Some observations:

- Since α_i is part of the process that generates the explanatory variable $y_{i,t-1}$, we have $E\left(y_{i,t-1}\alpha_i\right)>0$. Hence POLS or RE will not work - you'd expect these estimators to produce an estimate of β_1 that is **upward biased**.
- Also, since $\varepsilon_{i,t-1}$ determines $y_{i,t-1}$, strict exogeneity does not hold. Hence FD and FE will not work.
- To see the latter result more clearly, consider estimating (5.1) in first differences:

$$y_{it}-y_{i,t-1}=\beta_1\left(y_{i,t-1}-y_{i,t-2}\right)+\left(\varepsilon_{it}-\varepsilon_{i,t-1}\right). \tag{5.2}$$

While the FD transformation has eliminated α_i, we see that the differenced residual $\left(\varepsilon_{it}-\varepsilon_{i,t-1}\right)$ generally will not be uncorrelated with $\left(y_{i,t-1}-y_{i,t-2}\right)$, since $\varepsilon_{i,t-1}$ impacts on $y_{i,t-1}$:

$$y_{i,t-1}=\beta_1 y_{i,t-2}+\alpha_i+\varepsilon_{i,t-1},$$

which follows from (5.1). We see that

$$\varepsilon_{i,t-1} \nearrow \Longrightarrow \left\{ \begin{array}{c} (\varepsilon_{it} - \varepsilon_{i,t-1}) \searrow \\ y_{i,t-1} \nearrow \Longrightarrow (y_{i,t-1} - y_{i,t-2}) \nearrow \end{array} \right\},$$

thus

$$E\left[(\varepsilon_{it} - \varepsilon_{i,t-1})(y_{i,t-1} - y_{i,t-2})\right] < 0,$$

and so we expect the bias β_1 in from estimating (5.1) by first differences to be **negative.** Hence, you'd expect this estimator to produce an estimate of β_1 that is **downward biased**

A similar result can be derived for the FE model.

[EXAMPLE - Section 2 in the appendix: AR(1) models for log sales.]

5.4. Contemporaneous Correlation between the Regressor and the Residual

Consider the model

$$y_{it} = \beta_1 z_{it} + \beta_2 x_{it} + \alpha_i + u_{it},$$

where x_{it} is correlated with the individual effect α_i and **contemporaneously** correlated with u_{it}:

$$\begin{aligned} E(\alpha_i x_{it}) &\neq 0, \\ E(u_{it} x_{it}) &\neq 0, \end{aligned} \tag{5.3}$$

while z_{it} is strictly exogenous, but possibly correlated with α_i:

$$\begin{aligned} E(\alpha_i z_{it}) &\neq 0, \\ E(u_{it} z_{it}) &= 0. \end{aligned}$$

The correlation between x_{it} and u_{it} can be due to, for example:

- Omitted variables

- Measurement errors

In general, FD, FE, POLS and RE will not give consistent estimates if there is non-zero contemporaneous correlation between the regressor(s) and the error term.

Suppose the above equation represents a production function. Then FE or FD estimation will recognize that α_i may be correlated with the input x_{it}, for instance, because firms with better management (which is time invariant) use more inputs. But it is also possible that u_{it}, which we can think of as an time varying demand shock, is correlated with the input: when demand is high, firms increase the level of the input and vice versa.

- **What To Do Then...?**

If strict exogeneity does not hold while at the same time the time invariant effect is correlated with the explanatory variables, none of the estimators considered in this lecture will be consistent. However, we may be able to obtain consistent parameter estimates by using instruments.

PhD Programme: Econometrics II

Department of Economics, University of Gothenburg
Appendix Lecture 6
Måns Söderbom

Econometric Analysis of Company-Level Panel Data

Reference:

Blundell, R.W. and Bond, S.R. (2000), 'GMM estimation with persistent panel data: an application to production functions', Econometric Reviews, 19, 321-340 (http://www.ifs.org.uk/publications.php?publication_id=2722)

1. *Estimating a simple production function*

In this section we consider the results of four basic panel data models – i.e. POLS, RE, FE, FD – for a simple Cobb-Douglas production function of the following form:

$$y_{it} = \beta_1 n + \beta_2 k + \gamma_t + (\alpha_i + \varepsilon_{it})$$

where y_{it} denotes log sales, n_{it} is log employment, k_{it} is log capital stock, γ_t is a time effect common to all firms, α_i is a firm specific time invariant effect, ε_{it} is a time varying residual, and *i,t* denote firm and year, respectively. The data is a balanced panel of 509 R&D-performing US manufacturing companies observed for 8 years, 1982-89. These data have been analyzed previously by Blundell and Bond (2000).

Table 1.1 OLS, levels

```
. reg y n k yr2-yr8, cluster(id)

Linear regression                                      Number of obs =    4072
                                                       F(  9,   508) = 2507.63
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9693
                                                       Root MSE      =  .35256

                                    (Std. Err. adjusted for 509 clusters in id)
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |   .5578836   .0308763    18.07   0.000     .4972227    .6185445
           k |   .4322828   .0274846    15.73   0.000     .3782853    .4862803
         yr2 |  -.0568626   .0083657    -6.80   0.000    -.0732982   -.0404269
         yr3 |   -.050041   .0110933    -4.51   0.000    -.0718355   -.0282465
         yr4 |  -.0875714   .0135255    -6.47   0.000    -.1141442   -.0609987
         yr5 |   -.092866    .016461    -5.64   0.000     -.125206   -.0605259
         yr6 |  -.0580931   .0174944    -3.32   0.001    -.0924634   -.0237228
         yr7 |  -.0211632   .0185846    -1.14   0.255    -.0576754    .015349
         yr8 |  -.0382923    .020265    -1.89   0.059    -.0781058    .0015213
       _cons |   3.046843   .0915369    33.29   0.000     2.867005    3.22668
------------------------------------------------------------------------------

. test n+k=1
```

```
 ( 1)  n + k = 1

       F(  1,   508) =    1.58
            Prob > F =    0.2095
```

Table 1.2 Random effects GLS, levels

```
. xtreg y n k yr2-yr8

Random-effects GLS regression                   Number of obs      =      4072
Group variable: id                              Number of groups   =       509

R-sq:  within  = 0.7352                         Obs per group: min =         8
       between = 0.9727                                        avg =       8.0
       overall = 0.9683                                        max =         8

Random effects u_i ~ Gaussian                   Wald chi2(9)       =  27784.44
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |   .6449503   .0128055    50.36   0.000     .6198519    .6700487
           k |   .3240523   .0111113    29.16   0.000     .3022746      .34583
         yr2 |  -.0479095   .0094174    -5.09   0.000    -.0663673   -.0294517
         yr3 |  -.0304206   .0095667    -3.18   0.001    -.0491709   -.0116702
         yr4 |  -.0562695   .0098609    -5.71   0.000    -.0755966   -.0369424
         yr5 |  -.0530942   .0101615    -5.23   0.000    -.0730105    -.033178
         yr6 |  -.0137367   .0103345    -1.33   0.184    -.0339921    .0065186
         yr7 |   .0262007   .0104546     2.51   0.012     .0057102    .0466913
         yr8 |    .018625   .0109099     1.71   0.088    -.0027581     .040008
       _cons |   3.450817   .0425702    81.06   0.000     3.367381    3.534253
-------------+----------------------------------------------------------------
     sigma_u |  .31878629
     sigma_e |  .14715329
         rho |  .82434862   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. estimates store re

. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        y[id,t] = Xb + u[id] + e[id,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                       y |   4.041093       2.010247
                       e |   .0216541       .1471533
                       u |   .1016247       .3187863

        Test:   Var(u) = 0
                                 chi2(1) =  9433.49
                              Prob > chi2 =     0.0000

. test n+k=1

 ( 1)  n + k = 1

           chi2(  1) =   25.04
         Prob > chi2 =    0.0000
```

Table 1.3 Fixed effects ("within")

```
. xtreg y n k yr2-yr8, fe

Fixed-effects (within) regression               Number of obs      =      4072
Group variable: id                              Number of groups   =       509

R-sq:  within  = 0.7379                         Obs per group: min =         8
       between = 0.9706                                        avg =       8.0
       overall = 0.9661                                        max =         8

                                                F(9,3554)          =   1111.47
corr(u_i, Xb)  = 0.5988                         Prob > F           =    0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |   .6544609   .0144048    45.43   0.000     .6262184    .6827034
           k |   .2329073    .013637    17.08   0.000     .2061702    .2596443
         yr2 |  -.0376406   .0093042    -4.05   0.000    -.0558828   -.0193985
         yr3 |  -.0076445   .0096071    -0.80   0.426    -.0264805    .0111914
         yr4 |  -.0234513   .0100955    -2.32   0.020    -.0432449   -.0036578
         yr5 |  -.0136103   .0105543    -1.29   0.197    -.0343034    .0070829
         yr6 |   .0314121   .0108748     2.89   0.004     .0100907    .0527335
         yr7 |   .0753576   .0111072     6.78   0.000     .0535805    .0971347
         yr8 |   .0764164   .0118166     6.47   0.000     .0532485    .0995844
       _cons |   3.863804   .0529288    73.00   0.000      3.76003    3.967578
-------------+----------------------------------------------------------------
     sigma_u |  .42922318
     sigma_e |  .14715329
         rho |  .89482518   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(508, 3554) =    38.90           Prob > F = 0.0000

. test n+k=1

 ( 1)  n + k = 1

       F(  1,  3554) =  121.32
            Prob > F =    0.0000

. estimates store fe

. hausman fe re

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       fe           re         Difference          S.E.
-------------+----------------------------------------------------------------
           n |    .6544609     .6449503        .0095106        .0065967
           k |    .2329073     .3240523        -.091145        .0079061
         yr2 |   -.0376406    -.0479095        .0102689               .
         yr3 |   -.0076445    -.0304206         .022776        .0008802
         yr4 |   -.0234513    -.0562695        .0328182        .0021636
         yr5 |   -.0136103    -.0530942         .039484        .0028525
         yr6 |    .0314121    -.0137367        .0451488        .0033849
         yr7 |    .0753576     .0262007        .0491569        .0037512
         yr8 |    .0764164      .018625        .0577915        .0045393
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(9) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      143.16
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

Table 1.4 First Differences

```
. reg dy dn dk yr3-yr8, cluster(id)

Linear regression                                      Number of obs =    3563
                                                       F(  8,   508) =  108.19
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4072
                                                       Root MSE      =  .14536

                                    (Std. Err. adjusted for 509 clusters in id)
------------------------------------------------------------------------------
             |               Robust
          dy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          dn |   .4759999    .029565    16.10   0.000     .4179151    .5340847
          dk |   .2242109   .0360815     6.21   0.000     .1533236    .2950983
         yr3 |   .0700838   .0101958     6.87   0.000     .0500527    .0901149
         yr4 |   .0147162   .0106677     1.38   0.168     -.006242    .0356745
         yr5 |   .0381541   .0107752     3.54   0.000     .0169846    .0593235
         yr6 |   .0796229   .0092489     8.61   0.000      .061452    .0977938
         yr7 |   .0774642   .0097661     7.93   0.000     .0582773    .0966512
         yr8 |   .0325291   .0100251     3.24   0.001     .0128332    .0522249
       _cons |  -.0289894   .0081968    -3.54   0.000    -.0450933   -.0128856
------------------------------------------------------------------------------

. test dn+dk=1

 ( 1)  dn + dk = 1

       F(  1,   508) =  105.39
            Prob > F =    0.0000
```

2. *Towards dynamics: Results for simple AR(1) specifications*

We now consider the results for a simple AR(1) specification for log sales:

$$y_{it} = \rho y_{i,t-1} + (\alpha_i + \varepsilon_{it})$$

Table 2.1 Pooled OLS

```
. reg y y_1 , cluster(id)

Linear regression                                      Number of obs =    3563
                                                       F(  1,   508) =       .
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9912
                                                       Root MSE      =  .18715

                                    (Std. Err. adjusted for 509 clusters in id)
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         y_1 |     .98832    .001953   506.06   0.000     .9844831    .9921569
       _cons |   .1150687    .013331     8.63   0.000     .0888781    .1412593
------------------------------------------------------------------------------
```

Table 2.2 Fixed Effects (within)

```
. xtreg y y_1 , fe

Fixed-effects (within) regression               Number of obs      =      3563
Group variable: id                              Number of groups   =       509

R-sq:  within  = 0.5879                         Obs per group: min =         7
       between = 0.9981                                        avg =       7.0
       overall = 0.9912                                        max =         7

                                                F(1,3053)          =   4355.96
corr(u_i, Xb)  = 0.9783                         Prob > F           =    0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         y_1 |   .7380082    .011182    66.00   0.000     .7160832    .7599332
       _cons |   1.573936   .0652301    24.13   0.000     1.446037    1.701835
-------------+----------------------------------------------------------------
     sigma_u |  .51579652
     sigma_e |  .16594538
         rho |  .90620103   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(508, 3053) =     2.91           Prob > F = 0.0000
```

Table 2.3 First differences

```
. ge dy_1=d.y_1
(1018 missing values generated)

. reg dy dy_1

      Source |       SS       df       MS              Number of obs =    3054
-------------+------------------------------           F(  1,  3052) =   91.69
       Model |  2.92890642     1  2.92890642            Prob > F      =  0.0000
    Residual |   97.491852  3052  .031943595           R-squared     =  0.0292
-------------+------------------------------           Adj R-squared =  0.0288
       Total |  100.420758  3053  .032892486           Root MSE      =  .17873

------------------------------------------------------------------------------
          dy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dy_1 |   .1608638   .0167995     9.58   0.000     .1279242    .1938034
       _cons |   .0438578   .0033411    13.13   0.000     .0373068    .0504089
------------------------------------------------------------------------------
```

So the estimate of ρ varies between 0.16 (FD) and 0.99 (POLS). What should we believe and why?