Lecture 11: Panel Data

Thomas Chadefaux

Contents

Panel data denotes data that in which the same units are surveyed two or more times. Panel data offers a number of advantages, including a large number of observations (nT) and a possible solution to the problem of omitted variable bias.

1 Motivation: The Problem of Unoberserved Heterogeneity.

Suppose that we are interested in explaining house prices in various cities, and we think that that effect might be explained by crime in that particular city. I.e.,

$$HP_{it} = \beta_0 + \beta_1 crime_{it} + \varepsilon$$

But clearly, there are also other factors that might explain house prices, and hence this model is misspecified. These other factors include:

- City-specific factors. For example, San Francisco is more expensive than Detroit, partly because San Francisco is on the Ocean whereas Detroit is landlocked. Other factors might include race, demographics, education, etc.
- Time-dependent factors, which do not vary across city. For example, perhaps we are in the middle of a housing crisis, or there has been a huge boom in construction.

To take into account these factors, we could specify our model as:

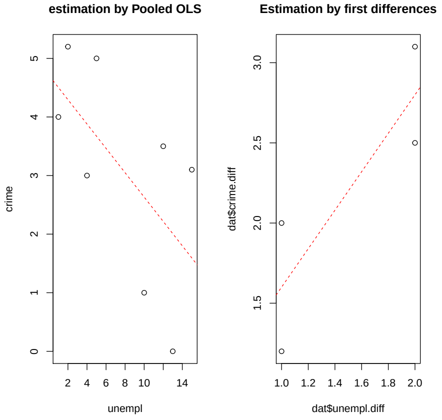
$$HP_{it} = \underbrace{\beta_0}_{\text{constant}} + \beta_1 crime_{it} + \underbrace{\mathbf{v_t}}_{\text{time-dependent term}} + \underbrace{\mathbf{c_i}}_{\text{city-dependent term}} + \underbrace{\varepsilon_{it}}_{\text{regular error term}}$$

More generally, in panel data individuals are observed at different points in time. Panel data are most useful when we suspect that the outcome variable depends on explanatory variables which are not observable but correlated with the observed IVs. If these omitted variables are constant over time, then panel data estimators allow to consistently estimate the effect of observed IVs.

2 The Problem with (pooled) OLS

Remember that for OLS to be consistent (i.e., unbiased in large sample, i.e. (loosely) $E[b] = \beta$), we need that $cov(\epsilon_{it}, x_{it}) = 0$. But now our error term includes not only ε_{it}, but also the city-specific error term, c_i. So the requirement now is that $cov(c_i + \epsilon_{it}, x_{it}) = 0$.

The problem is that there is a good chance that c_i and x_{it} are correlated. In the housing example, for example, demographics are probably correlated with crime, but since demographics enters c_i, we have a problem (for example, as age goes up, crime rate might go down; opposite for education, etc.). Note that this is yet another example of omitted variable bias. I.e., OLS will be biased and inconsistent.



In the context of cross-section regression, there is no easy way to deal with this unobserved heterogeneity. If c_i is uncorrelated with any of the regressors, then we can use OLS and the coefficients will remain unbiased. But if they are correlated, then we face the problem of endogeneity. So either we control for it directly or indirectly by finding a proxy, or we rely on instrumental variables. But adequate instrumental variables are difficult to find.

When we have observations over time, however, we will be able to use past and future information as controls. But we need a different estimator than OLS.

3 Estimators

3.1 Fixed Effects

- **First Differences** Instead of looking at levels, let's look at differences:

  $$HP_{it} - HP_{it-1} = \Delta HP_{it} = \beta_1 \Delta crime_{it} + \varepsilon$$

  Note that by taking the first difference, we have removed the unobserved heterogeneity c_i. That is intuitive, since we assumed that c_i did not vary over time.

  A small difficulty is that we must have variance in x_{it} across both time and city, or this term will disappear. Again, it makes sense given that taking first differences removes time-invariant covariates. This might be a problem if you are interested in, say, the effect of gender on income.

  In addition, there are some costs to the first differences estimator.

  - First, crime rate might vary across time and cities, but that variance across time might be small. That means your standard errors will be high.
  - Second, n degrees of freedom are lost because we removed the first observation for each of the n individuals.

- **Demeaning** With fixed effects, the idea is to subtract the mean values of the variables for each given individual. Let $\overline{x_i} = \frac{1}{T} x_{it}$

  $$\begin{aligned} HP_{it} - \overline{HP}_{it} &= \beta_1(crime_{it} - \overline{crime}_{it}) + (c_i - \overline{c}_i) + (\varepsilon - \overline{\varepsilon}) && (1) \\ \Rightarrow HP^*_{it} &= \beta_1 crime^*_{it} + \varepsilon^*, && (2) \end{aligned}$$

  where x^* denotes a time-demeaned variable. Practically, we are removing the average value for each individual, to get the time-demeaned value. Again, we have removed c_i, and hence there is no longer any covariance between our error term and the Xs (provided of course that $Cov(X_t, \varepsilon_{it}) = 0$.)

- **Dummy variable estimator.** include dummies for each city (minus one)

  $$HP_{it} = \beta_0 + \beta_1 crime_{it} + u_1\gamma_1 + u_2\gamma_2 + \ldots + u_{c-1}\gamma_{c-1} + \varepsilon$$

  . Here we are explicitly taking into account unobserved heterogeneity. Note that this method is exactly equivalent to the time-demeaning method, and so are their standard errors.

pros vs demeaning: explicit

cons vs demeaning: if the number of cities is large, then it becomes unwieldy, and demeaning makes more sense. In any case, it does not matter whether you use dummy variable estimation or demeaning. They yield the exact same results.

3.2 First differences vs. fixed effects/LSDV

How to choose between the two? If there are only two time periods, the two are equivalent. But when $t \geq 3$, then they are not.

If serially uncorrelated errors, then FE is better than FD. Why? Because with FD, we have $Cov(\Delta\varepsilon_{it}, \Delta\varepsilon_{it-1}) = Cov(\varepsilon_{it} - \varepsilon_{it-1}, \varepsilon_{it-1} - \varepsilon_{it-2}) \neq 0$.

But suppose now that ε is serially correlated, for example $\varepsilon_{it} = \varepsilon_{it-1} + u_{it}$, where u_{it} is some white noise, then $\Delta\varepsilon_{it} = u_{it}$, and hence FD is more efficient than FE because it is remove the serial correlation.

3.3 Random Effects

- **Motivation**
  - If $cov(x_{it}, c_i) \neq 0$, use FE or FD. But if not, then there is a more efficient estimator.
  - if $cov(x_{it}, c_i) = 0$, then use random effects. When might this be a reasonable assumption? perhaps we think that we have controlled for all relevant factors. Or perhaps the effect c_i is very small.

- **Estimation** It is tempting to use OLS, since we no longer have the issue of endogeneity. Indeed, pooled OLS is consistent. In fact, so are FE and FD, i.e., $b \rightarrow^p \beta$. But FE and FD have a problem: FD is wasting one period, and FE is estimating too many parameters. But pooled OLS also has problems. To see this, note that instead of ε_{it}, we have $u_{it} = c_i + \varepsilon_{it}$. But now look at the covariance between error terms over time:

$$E[u_{it}u_{is}] = E((c_i + \varepsilon_{it})(c_i + \varepsilon_{is})) \quad = E(c_i^2 + \varepsilon_{it}\varepsilon_{is} + c_i\varepsilon_{it} + c_i\varepsilon_{is}) = var[c_i] \tag{3}$$

So even if we assume 0 covariance between c_i and $\varepsilon_i t$ and between ε_{it} and ε_{is}, there is still covariance between c_i and c_i, which is in fact equal to the variance of c_i. But that means that our errors are serially correlated. How do we correct for serially correlated errors? We use FGLS, which in this context is called Random Effects (RE). Because we are correcting for serial correlation, the random effects estimator will be more efficient than both pooled and FE/FD.

How does the random effect estimator actually work?

$$HP_{it} - \lambda\overline{HP}_{it} = \beta_0(1-\lambda) + \beta_1(crime_{it} - \lambda\overline{crime}_{it}) + u_i - \lambda\overline{u}_{it},$$

where $u_{it} = c_i + \varepsilon_{it}$. At this point we do not know λ, but we can already see what would happen if $\lambda = 0$. Then we are back to OLS. On the other hand, if $\lambda = 1$, then we have the FE estimator. typically, however, $0 < \lambda < 1$.

But what is λ? We define

$$\lambda = 1 - \left(\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_c^2} \right)^{1/2}$$

σ_ε^2 is just the regular variance of our error term, and σ_c^2 is the variance of the unobserved heterogeneity. Suppose first that there is no unobserved heterogeneity. Then σ_c^2 is 0, and therefore $\lambda = 0$. In this case, RE = OLS, which makes sense since in the absence of unobserved heterogeneiety, there is no reason to use FE/RE in the first place!

If, on the other hand, $\sigma_c^2 = \infty$, then $\lambda = 1$, which means that RE=FE. This makes sense, because if the effect of c_i is large in the first place, we should be using FE.[1]

- **The Breusch-Pagan LM test** The goal here is to determine whether RE are justified. The basic idea is to test whether $\sigma_c^2 = 0$, which can be computed in R using plmtest (see below).

4 Fixed vs random effect

4.1 Pros and cons of random effects

Pros of random effects:

- S.E.$(b_{RE}) <$ S.E.(b_{FE}), because far fewer parameters to estimate.
- Allows us to estimate the effect of time-constant variables

Cons of random effects:

- Assumption that $cov(c_i, x_{it}) = 0$ is almost always violated, leading to endogeneity. Indeed, what we would need to have $cov(c_i, x_{it}) = 0$ is to assume that we have controlled for everything that matters, such that the remainng, unobserved variance c_i would be very small. But if the assumption does not hold, then b_{RE} is inconsistent.
- The effect c_i cannot be estimated

Of course, the choice between FE and RE will depend on the situation, and in fact there is a test that can help you make a decision: the Hausman test. More specifically, that test is testing whether $cov(c_i, x_{it}) = 0$

4.2 Hausman test for fixed vs random effects

Random effects assumes that $cov(c_i, x_{it}) = 0$. If that is the case, then both RE and FE are consistent, but RE is more efficient. BUT if the assumption is not true, then RE is inconsistent.

[1] Note that the θ that is reported in some program is

$$\theta = 1 - \left(\frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_c^2}} \right)$$

The Hausman statistic is defined as:

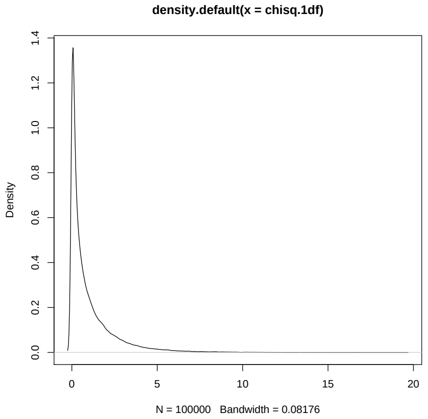
$$W = \frac{(b_{FE} - b_{RE})^2}{Var(b_{FE} - Var_{RE})}$$

Under the null hypothesis, this test is distributed chi squared with one d.f. What is the null hypothesis here? It is that $cov(c_i, x_{it}) = 0$.

But what is the intuition for the test? If the assumption that $cov(c_i, x_{it}) = 0$ is true, then both RE and FE are consistent, and hence their estimates will be the same. So the numerator will tend to 0. At the same time, the variance of FE will be larger, such that the denominator will be large. This mean that W will tend to be small. Given that we are dealing with a chi square distribution with 1 df, a small value is actually likely, and hence we cannot reject the null hypothesis.

Suppose now, on the contrary, that $cov(c_i, x_{it}) \neq 0$. Then the numerator is large, because $b_{RE} \neq b_{FE}$, because only b_{FE} is consistent under this assumption. Therefore W will be large, which is an unlikely value given the chi squared distribution with 1 df.

```r
chisq.1df <- rchisq(100000,1)
plot(density(chisq.1df))
```



captionHausman test: Chi square distribution with 1 df

5 Doing it all in R

Fixed effects in R: 3 equivalent ways of doing it. First, load the data:

```r
library(foreign)
Panel <- read.dta('http://dss.princeton.edu/training/Panel101.dta')
Panel$y <- Panel$y / 10000000
```

5.1 FE, Method 1– Do it manually by demeaning variables

```r
mean.y.by.country <- aggregate(data.frame(meany=Panel$y),
                               by=list(country=Panel$country), FUN=mean)
mean.x1.by.country <- aggregate(data.frame(meanx1=Panel$x1),
                                by=list(country=Panel$country), FUN=mean)
Panel <- merge(Panel, mean.y.by.country)
Panel <- merge(Panel, mean.x1.by.country)

y.demeaned <- Panel$y - Panel$meany
x1 <- Panel$x1 - Panel$meanx1
ols1 <- lm(y ~ x1 , data=Panel)
fe1 <- lm(y.demeaned ~ x1- 1)
```

5.2 FE, Method 2– Do it manually by adding dummy variables

```
fe2 <- lm(y ~ x1 + as.factor(country)- 1, data=Panel)
```

5.3 FE, Method 3– Use R’s plm package

```
library(plm)

## Loading required package:  Formula

fe3 <- plm(y ~ x1, data=Panel, index=c("country", "year"), model="within")
fixef(fe3)

##          A          B          C          D          E          F
##   88.05424 -105.78584 -172.28108  316.28269  -60.26220  201.07318
##          G
##  -98.47175
```

5.4 First differences

```
Panel$ylag <- unlist(tapply(Panel$y, Panel$country, function(x)c(rep(NA,1),x[1:(length(x)-1)])))
Panel$x1lag <- unlist(tapply(Panel$x1, Panel$country, function(x)c(rep(NA,1),x[1:(length(x)-1)])))
Panel$dy <- Panel$y - Panel$ylag
Panel$dx1 <- Panel$x1 - Panel$x1lag
fd1 <- lm(dy ~ dx1- 1, data=Panel)
```

5.5 Random Effects

```
re1 <- plm(y ~ x1, data = Panel, index = c('country', 'year'), model  = 'random')
stargazer(ols1, fe1, fe2, fe3, fd1, re1,
           column.labels=c('OLS', 'FE demeaned', 'FE w/ LSDV', 'FE using plm', 'FD', 'RE'),
          omit.stat=c("f", "ser"))
```

5.6 Run a Hausman test

```
phtest(fe3, re1)

##
##  Hausman Test
##
## data:  y ~ x1
## chisq = 3.674, df = 1, p-value = 0.05527
## alternative hypothesis: one model is inconsistent
```

	Dependent variable:					
	y	y.demeaned	y		dy	y
	OLS	*OLS*	*OLS*	*panel linear*	*OLS*	*panel linear*
	OLS	FE demeaned	FE w/ LSDV	FE using plm	FD	RE
	(1)	(2)	(3)	(4)	(5)	(6)
x1	49.499	247.562**	247.562**	247.562**		124.700
	(77.886)	(104.904)	(110.668)	(110.668)		(90.215)
as.factor(country)A			88.054			
			(96.181)			
as.factor(country)B			−105.786			
			(105.107)			
as.factor(country)C			−172.281			
			(163.151)			
as.factor(country)D			316.283***			
			(90.946)			
as.factor(country)E			−60.262			
			(106.429)			
as.factor(country)F			201.073*			
			(112.281)			
as.factor(country)G			−98.472			
			(149.272)			
dx1					235.517*	
					(119.285)	
Constant	152.432**					103.701
	(62.107)					(79.063)
Observations	70	70	70	70	63	70
R^2	0.006	0.075	0.440	0.075	0.059	0.027
Adjusted R^2	−0.009	0.061	0.368	0.066	0.044	0.027
Note:				*p<0.1; **p<0.05; ***p<0.01		

This suggest that we cannot reject the null hypothesis, i.e. that $cov(x_{it}, c_i) \neq 0$, and hence we should use RE. But it is barely rejected, so the answer is not so clear.

5.7 Are Time fixed effects needed?

Run an F test:

```
fe2 <-  lm(y ~ x1 + as.factor(country)- 1, data=Panel)
fe2b <- lm(y ~ x1 + as.factor(country) + as.factor(year) - 1, data=Panel)
anova(fe2, fe2b)

## Analysis of Variance Table
##
## Model 1: y ~ x1 + as.factor(country) - 1
## Model 2: y ~ x1 + as.factor(country) + as.factor(year) - 1
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1     62 4845371
## 2     53 4020073  9    825298 1.209 0.3094
```

The $pr(> F)$ is > 0.05, so we fail to reject the null that the coefficients for all years are jointly equal to zero, therefore no time fixed effects are needed here. We can also test whether country fixed effects were justified in the first place:

```
ols <-  lm(y ~ x1 , data=Panel)
fe2 <- lm(y ~ x1 + as.factor(country) - 1, data=Panel)
anova(ols, fe2)

## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + as.factor(country) - 1
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     68 6235911
## 2     62 4845371  6   1390540 2.9655 0.01307 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, fixed effects are justified.

5.8 Are Random Effects Justified?

Compare pooled OLS to RE using a Breusch-Pagan test. The test essentially tests whether there are significant differences across units.

```
pols <-  plm(y ~ x1, data=Panel, index=c("country", "year"), model="pooling")
plmtest(pols, type=c("bp"))
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan)
##
## data:  y ~ x1
## chisq = 2.6692, df = 1, p-value = 0.1023
## alternative hypothesis: significant effects
```

Here we failed to reject the null and conclude that random effects is not appropriate.

6 Standard Error Corrections for Panel Data

6.1 A Recap

Do not forget to correct for potentially incorrect standard errors...

- Huber-White SE: diagonal elements of $E[\varepsilon\varepsilon']$ differ. In R:
- Clustered SE: Allows the disturbances within each cluster to be correlated with each other but requires that the disturbances from different clusters be uncorrelated. $Var(\varepsilon)$ is block diagonal. To cluster by country in R, for example, suppose you have a column "country". Then run:
- Newey-West: correct for heteroskedasticity and autocorr. Similar to clustered SE, but instead of specifying a cluster variable, you specify a maximum order of correlation. In R:
- Panel-corrected SE (PCSE), Beck and Katz APSR 1995: The difference here is that we are interested in correlation across groups. It is assumed that there is no autocorrelation within group, or that it has been removed. Suppose there are 3 observations per unit. Then the error variance matrix would look like this:

$$Var(e) = \Omega = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \sigma_{12} & 0 & 0 & & \sigma_{1N} & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 & \sigma_{12} & 0 & & 0 & \sigma_{1N} & 0 \\ 0 & 0 & \sigma_1^2 & 0 & 0 & \sigma_{12} & & 0 & 0 & \sigma_{1N} \\ \sigma_{12} & 0 & 0 & \sigma_2^2 & 0 & 0 & & \sigma_{2N} & 0 & 0 \\ 0 & \sigma_{12} & 0 & 0 & \sigma_2^2 & 0 & & 0 & \sigma_{2N} & 0 \\ 0 & 0 & \sigma_{12} & 0 & 0 & \sigma_2^2 & \cdots & 0 & 0 & \sigma_{2N} \\ & & \vdots & & & \vdots & \ddots & & & \\ \sigma_{1N} & 0 & 0 & \sigma_{2N} & 0 & 0 & & \sigma_N^2 & 0 & 0 \\ 0 & \sigma_{1N} & 0 & 0 & \sigma_{2N} & 0 & & 0 & \sigma_N^2 & 0 \\ 0 & 0 & \sigma_{1N} & 0 & 0 & \sigma_{2N} & & 0 & 0 & \sigma_N^2 \end{bmatrix}$$

Figure 1: PCSE

6.2 How to do it in R

```r
library(pcse)
library(lmtest)
data(agl)
# OLS Estimation for a model of growth in OECD countries
lm1 <- lm(growth ~ lagg1 + opengdp + openex + openimp + central +
            leftc + inter + as.factor(year), data=agl)
summary(lm1)

# Estimate Huber-White SE:
library(sandwich)
coeftest(lm1, vcov. = vcovHC)

# Estimate clustered SE:
library(multiwayvcov)
lm1$clustered.se <- cluster.vcov(lm1, agl$country)
coeftest(lm1, lm1$clustered.se)

# Estimate Newey-West SE:
library(sandwich)
coeftest(lm1, vcov. = NeweyWest)
#or for the same idea as Newey-West but with Andrews weights (don't worry about it):
coeftest(lm1, vcov. = vcovHAC)

# Estimate Panel-Corrected SE:
library(pcse)
lm1.pcse <- pcse(lm1, groupN=agl$country, groupT=agl$year)
coeftest(lm1, vcov. = lm1.pcse$vcov)
```