EC402 - Problem Set 3

Konrad Burchardi

11th of February 2009

Introduction

Today we will

- briefly talk about the **Conditional Expectation Function** and
- lengthily talk about **Fixed Effects**: How do we calculate them, what do we estimate when including them, how does this help?

(i) Let us just define $\epsilon_i \equiv y_i - E[y_i|X_i]$.[1] Then by construction

$$E[\epsilon_i|X] = E[y_i - E[y_i|X_i]|X_i] = E[y_i|X_i] - E[y_i|X_i] = 0 \quad (1)$$

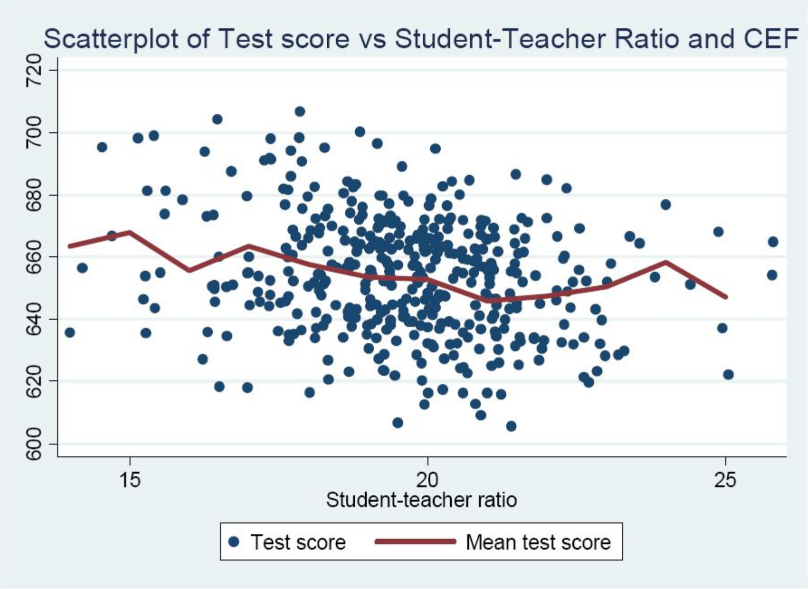
(ii) Let $h(X_i)$ be any function of X_i, then

$$E[h(X_i)\epsilon_i] = E[E[h(X_i) \cdot \epsilon_i|X_i]] = E[h(X_i) \cdot E[\epsilon_i|X_i]] = 0 \quad (2)$$

[1] To understand this, just think of the joint density function of y_i and X_i. What we are doing is that at any X_i we take the expectations over all possible y_is that come with this X_i and then define y_i as this expectation plus some error - all at a given X_i, so at one cut through the joint density function.

1. Conditional Expectation Function

A. CEF is a valuable summary of bivariate relationship in data



Source: Steve Pischke's Ec486 notes.

Any linear approximation of the CEF will be of the form Xb. We search for the best linear approximation minimizing the mean-squared-error, so we want to solve

$$min_b E[(E[y|X] - Xb)^2] \tag{3}$$

It turns out that the solution to

$$min_b E[(y - Xb)^2] \tag{4}$$

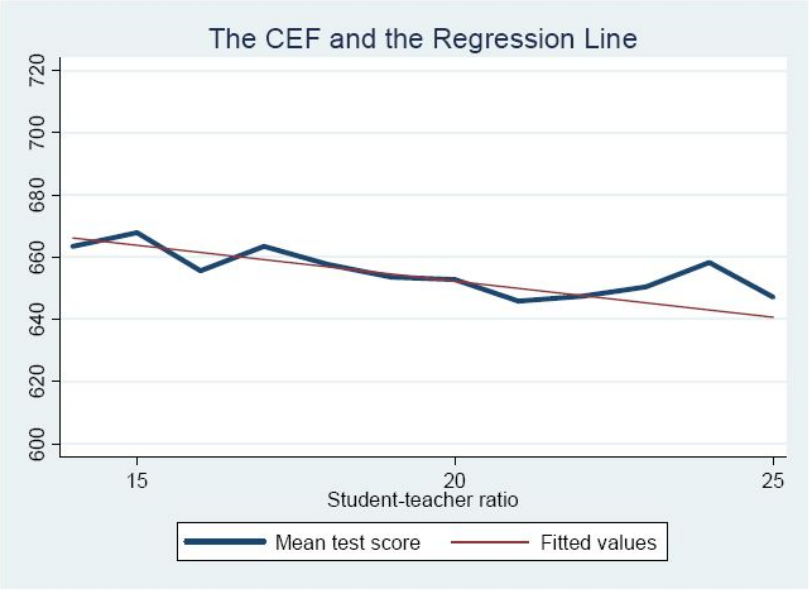
will be the same as to the first problem. But the solution to the second problem is OLS. You can see this since

$$\begin{aligned}
E(y - Xb)^2 &= E(y - E[y|X] + E[y|X] - Xb)^2 && (5)\\
&= E[(y - E[y|X]) + (E[y|X] - Xb)]^2 && (6)\\
&= E(y - E[y|X])^2 + E(E[y|X] - Xb)^2 && (7)\\
&\quad +2E[(y - E[y|X]) \cdot (E[y|X] - Xb)] && (8)
\end{aligned}$$

The first bit does not depend on b and hence does not affect the minimization, and the last bit is just $E[\epsilon \cdot h(X)]$, where $h(X)$ is a function of X. But this we showed is 0.

1. Conditional Expectation Function

B. $\hat{\beta}_{OLS}$ is best linear approximation to CEF



Source: Steve Pischke's Ec486 notes.

2. Repeated Cross Sectional Data

A. Are the time fixed effects correlated with the Xs?

Whether the δ_t and X are correlated or not has implications for $\hat{\beta}_{RE}$ and $\hat{\beta}_{FE}$:

Random Effects If the δ_t and Xs are uncorrelated, then Random Effects will be consistent and efficient (it is essentially GLS). But if δ_t and X are correlated, then it will be inconsistent.

Fixed Effects This is not the case for a Fixed Effects estimation, since this essentially gets away with δ_t. It is not efficient if δ_t and X are uncorrelated, but in both cases it will be consistent.

To **check whether the δ_t are independent of X** we can

- Either just look the the coefficients. If they are by some measure 'far' apart, e.g. their confidence intervals do not overlap, the reason might be that $\hat{\beta}_{RE}$ is inconsistent and hence far off.

- A more formal test would be the Wu-Hausman test. The idea is that if under some H_0 two estimators are consistent ($\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$) and one efficient ($\hat{\beta}_{RE}$), but under H_a the latter is not consistent anymore, we can make use of the fact that under H_0 are on average 0 and distributed as

$$(\hat{\beta}_{RE} - \hat{\beta}_{FE})'[Var(\hat{\beta}_{RE}) - Var(\hat{\beta}_{FE})]^{-1}(\hat{\beta}_{RE} - \hat{\beta}_{FE}) \sim^a \chi^2(k-1) \quad (9)$$

  If this quantity is very large (so the different of $\hat{\beta}_{RE}$ and $\hat{\beta}_{FE}$ is big) we reject the null. In our case H_0: 'δ_t and X are independent'

Think of some data where you are worried about **omitted variables**. For example

y_{it} being a variable capturing whether individual i broke his arm in week t and
x_{it} being a variable capturing for how many hours he worked.

Wanting to find the causal relation between the two, you should be worried about omitted variables. For example if there was snow in week t this might drive up the incidence of broken arms and drive down the number of hours worked - *which would give us a correlation between y and x even though there is no causal relation*.

Now, to the extend that the omitted variable effects all the same in week t, we can take care of this by just including time-fixed-effects. These will capture any omitted effect **which is the same for all** in week t. So we would run

$$y_{it} = X_{it}\beta + \delta_t + \epsilon_{it} \tag{10}$$

For this to work we obviously need that

- *the omitted variable indeed takes the same value for all individuals i in time t and*
- *that the omitted variable indeed has the same effect for all individuals i.*

If you want to test whether the effect of X on y is the same in all time periods, you can **just allow them to be different and then test the linear hypothesis that they are the same**.

So you would want to run the regression

$$y_{it} = X_{it}\beta_t + \delta_t + \epsilon_{it} \tag{11}$$

This will give you the same coefficient estimates as running t separate regressions (since you have for each time period a separate intercept and slope coefficient). However, running them jointly allows you to test the H_0: $\beta_1 = \beta_2 = ... = \beta_k$ in the simple fashion we are used to.

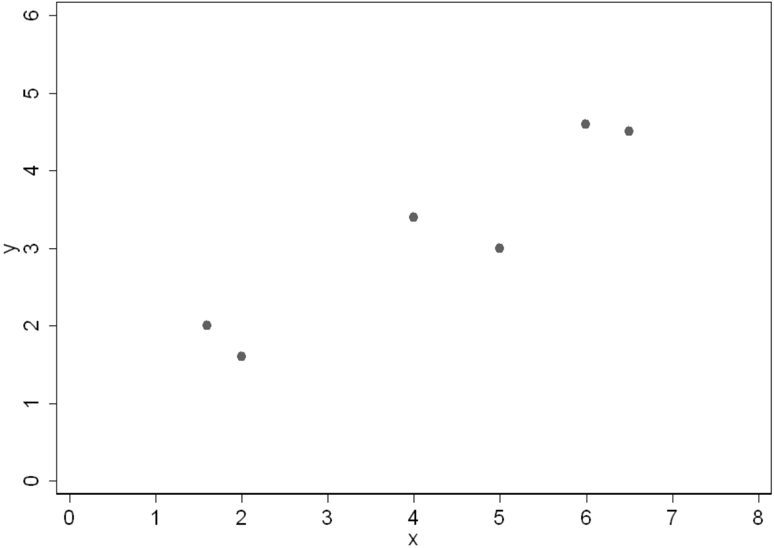
3. Using Fixed Effects - Example

A. Returns to schooling among identical twins

Ashenfelter and Rouse (1998) try to **identify the causal effect of schooling,** x_{if} **on wages outcomes,** y_{if}. This question has gotten much/most/almost all attention in the Labour literature - and inherently difficult because of omitted variable problems, e.g. students from a rich family might get both a long education and high wages, but they get the latter because of their parents' network and not their good education.

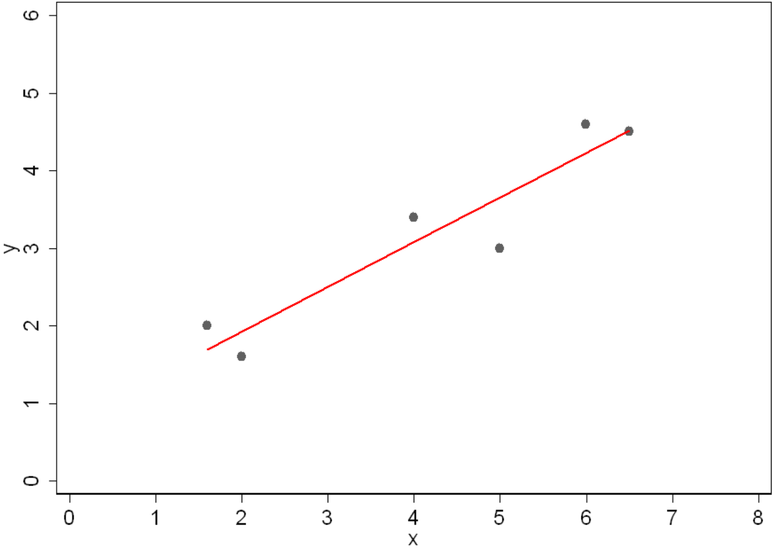
3. Using Fixed Effects - Example

A. **Suppose we do not include fixed effects.**



3. Using Fixed Effects - Example

A. **Suppose we do not include fixed effects.**

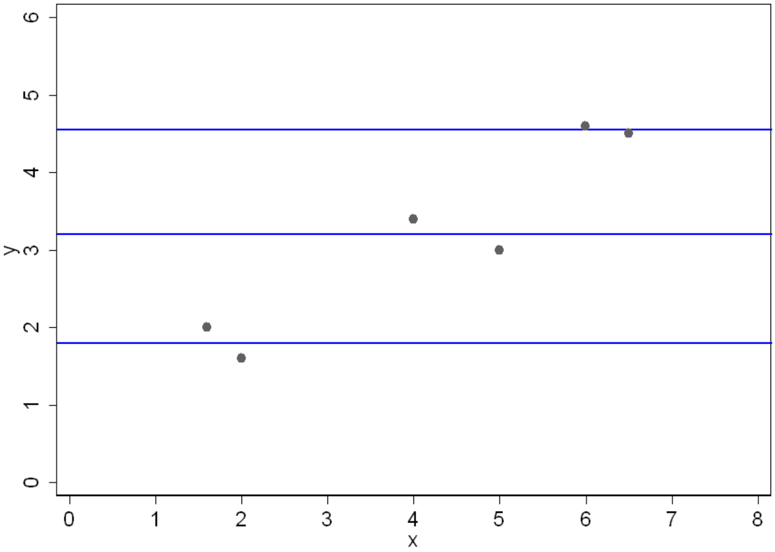


3. Using Fixed Effects - Example

A. Returns to schooling among identical twins

Their study tries to tackle this problem by **comparing how the wages differ for identical twins who received different amount of schooling**. They implement this with **family fixed effects**.

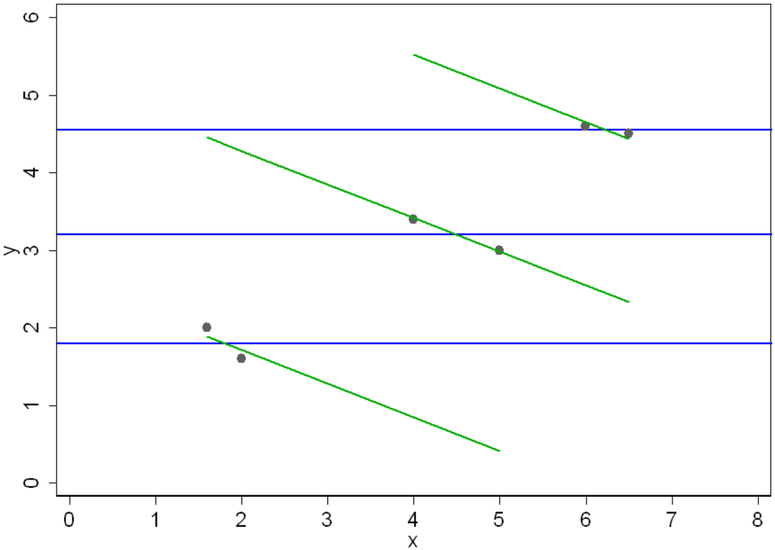
- How might this take care of the omitted variable 'ability'?
- Of which type of omitted variables can we take care of with family fixed effects?
- Of which type of omitted variables do family fixed effects not take care of?
- Of which unobserved factors can you think which drive the correlation between earnings and schooling among identical twins? Are they taken care of by including family fixed effects?



0
1
2
3
4
5
6
y
0
1
2
3
4
5
6
7
8
x

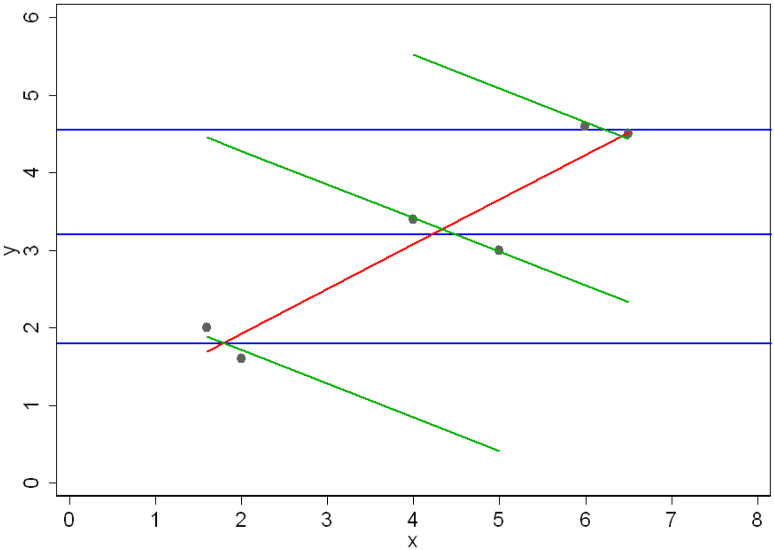
3. Using Fixed Effects - Example

A. Exploiting 'within' variation



3. Using Fixed Effects - Example

A. Compare to what OLS would have given!



3. Using Fixed Effects - Example

A. Returns to schooling among identical twins

We saw, that once we have included fixed effects, our estimator will explain differences in the outcome variable **within groups** (i.e. families) with the variation in the explanatory variable **within these groups**. That excludes many potential omitted variables.

But we still need to ensure that **within one group** the differences in the explanatory variable are indeed **as good as random**. So, in our example, the reasons why identical twins get different amounts of schooling should have nothing to do with their earnings! Or: We try to show that for our purposes the differences in schooling among identical twins are as good as random.

3. Using Fixed Effects - Example

A. Returns to schooling among identical twins

Potential Problem: There are **many things which might drive one of the twins to get more education than the other** and drive the earnings as well. We cannot look at all of them. But if we at least find that some of the potential candidates have no correlation with how many years of schooling one or the other twin gets, this gives us some comfort. That is what the authors do - and the question says they find no correlation.

If we *believe/assume* that **all other factors which we could not check are sufficiently correlated with those which we checked**, then it is unlikely that our estimates are biased by any of these omitted variables.

3. Using Fixed Effects - Example

C. Including just group level variables

Suppose we had not including fixed effects, but just controlled for the average level of education in the family. (Will this fit the data as well as fixed effects?) Then the effect of the own education goes down. This might have two reasons:

- There might be something else which drives both average education and earning up in the family, e.g. parents social network.
- [Or this might as well capture the fact that there are spill-over effects of education on the siblings.]

In any case it shows that the **initially estimated relation between an individual's education and his earnings was overestimated in (1)**.

We can estimate straightforwardly the OLS estimator for the following FE model

$$y_{ij} = \beta s_{ij} + \delta_j f_j + \epsilon_{ij}$$

Alternative we can calculate $\hat{\beta}_{OLS}$ using **partitioned regression results** as

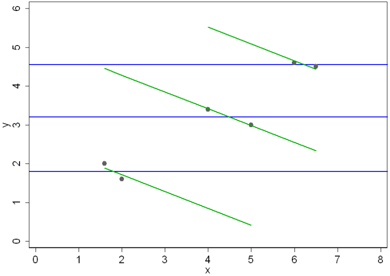
$$\begin{aligned} \hat{\beta}_{OLS} &= [s' M_F s]^{-1} s' M_F y \\ &= [(M_F s)'(M_F s)]^{-1} (M_F s)' M_F y \end{aligned}$$

where $M_F = I - F(F'F)^{-1}F'$ and F is the matrix of family dummy variables.

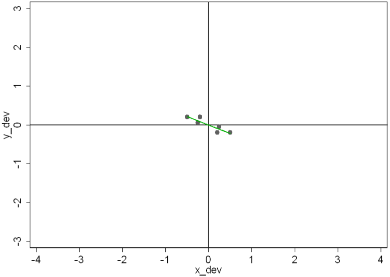
You can easily see why this is called the **difference estimator**. Just show that $M_F x$ gives a vector of deviations of x_{ij} from the family mean of x. (We have done something similar with seasonal fixed effects in Vassilis' part.)

3. Using Fixed Effects - Example

D. Exploiting ‘within’ variation



Fixed Effects estimation



Within estimation

In columns (3) and (5) the authors **instrument** for explanatory variable (years of schooling). Among other things this takes care of potential **measurement error**.

The fact that the coefficient estimates indeed increase relative to the respective non-instrumented counterparts, i.e. (2) and (4), might be explained by the fact that previously the coefficient estimates suffered from the **attenuation (i.e. downward) bias** with is cause by measurement error.[2]

[2] Note that this has pretty much nothing to do with the fixed effects issues.