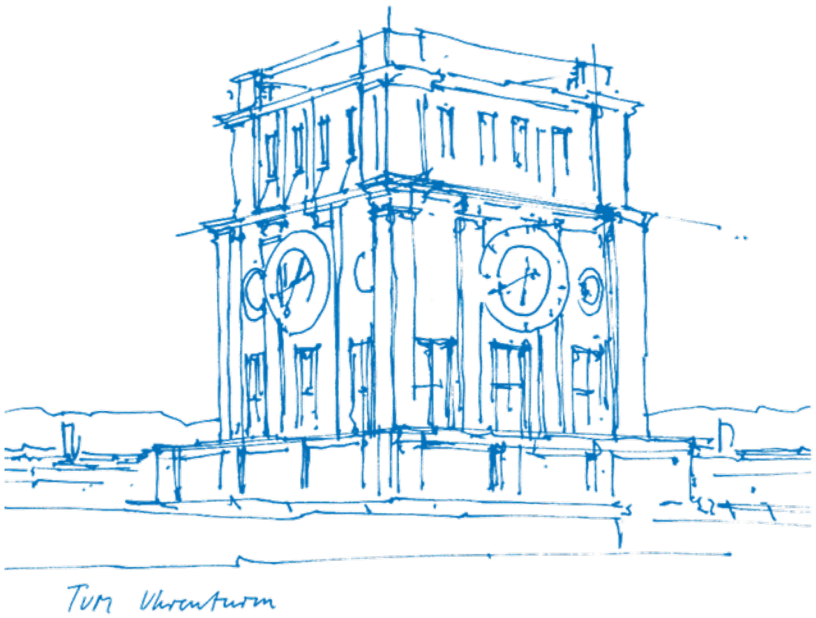
Applied Econometrics
TUM School of Management
Technical University of Munich

Empirical Research Methods - Lecture 7

Prof. Dr. Helmut Farbmacher
Technical University of Munich
TUM School of Management

Munich, May 31, 2024



Omitted Variables and Panel Data Estimation

Omitted Variables

Omitted Variables

Example: Private returns to education

Suppose the **population model** is (for ease of exposition w/o a constant)

$$y_i = \beta x_i + \gamma q_i + u_i \,.$$

where y is wage, x is education measured in years of schooling and q is a general measure of ability (unobservable for the researcher). It may capture factors like social skills, ambitions, etc.

We assume that our dataset has been generated by this process. We are also willing to assume that $E(x_i u_i) = 0$. However, while x is observed, q is not.

Omitted Variables

Because data on q is missing, the best we can do is to **estimate** the following model

$$y_i = \alpha x_i + \varepsilon_i$$

where ε_i is a composite error term.

What is the probability limit of the OLS estimator for α?

Can we say something about the potential direction of the bias?

Omitted Variables

Wage is determined by: $y_i = \beta x_i + \gamma q_i + u_i$

We instead estimate: $y_i = \alpha x_i + \varepsilon_i$

Auxiliary regression: $q_i = \delta x_i + w_i$

Then,

$$\begin{aligned} y_i &= \beta x_i + \gamma(\delta x_i + w_i) + u_i \\ &= \underbrace{(\beta + \gamma\delta)}_{\alpha} x_i + \underbrace{(\gamma w_i + u_i)}_{\varepsilon_i} \end{aligned}$$

If y is regressed on x alone, α will be the estimated slope of x.

Omitted Variables

$$y_i = \underbrace{(\beta + \gamma\delta)}_{\alpha} x_i + \underbrace{(\gamma w_i + u_i)}_{\varepsilon_i}$$

We expect that $\alpha > \beta$ (Why?)

Intuition:

- Say, education has on average a positive effect on wages ($\beta > 0$).
- People with higher general abilities might also be more successful in school on average ($\delta > 0$; **o.v. bias condition I**).
- And presumably, employers are also willing to pay on average higher wages to people with higher general abilities ($\gamma > 0$; **o.v. bias condition II**).
- General abilities is hence an omitted variable/confounding effect.

The return to education will be **overestimated** because $0 < \beta < \beta + \underbrace{\gamma\delta}_{>0}$.

Omitted Variables

We have to include q in our regression to avoid falsely attributing the explanatory power of q to x.

Example: The parameter of education is biased if we do not control for other dimensions of ability.

The parameter of x will be estimated consistently **only** when either $\delta = 0$ or $\gamma = 0$.

Otherwise, it will suffer from omitted variable bias. Generally, we have

	$\delta > 0$	$\delta < 0$
$\gamma > 0$	positive bias	negative bias
$\gamma < 0$	negative bias	positive bias

Omitted Variables

We can only estimate the true private returns to education if we keep all other omitted variables constant, we often say “we control for these other variables”.

To do so, of course, means we have to **know and observe** these relevant economic variables when we use OLS.

There are several strategies in econometrics to identify causal effects even if **unknown or unobserved** omitted variables exist.

In the next lecture we are going to discuss two simple yet powerful strategies, namely randomization and differences-in-differences estimation.

Omitted Variables

Moreover, Panel data estimators or instrumental variables estimators may also help to get rid of **unknown or unobserved** omitted variables.

Panel data estimators help us to get rid of **time-constant** omitted variables. We will discuss panel data estimation now.

Instrumental variables estimation may help us even with **time-varying** omitted variables. We will briefly discuss this method at the end of this course.

Panel Data Estimation

Panel Data Estimation

Consider a linear bivariate regression model for cross-sectional data

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Under OLS.1 $[\mathrm{E}(x_i \varepsilon_i) = 0]$ & OLS.2 $[\mathrm{E}(x_i^2) \neq 0]$, we can estimate β_1 consistently.

Assumption OLS.1 is often **problematic** when we analyze **observational data**.
For example,

1. Private returns to education (y_i: wage, x_i: education, ε_i: ability)
2. Effect of firm entry on price
   (y_i: price, x_i: # of competitors, ε_i: unobserved demand factors)

Under certain conditions we can deal with this issue using panel data.

Panel Data Estimation

Panel data is characterized by **repeated observations**.
i denotes the cross-sectional units of our analysis (e.g. individuals or firms), which we observe over several time periods t. Each observation is thus indexed by two subscripts ($i = 1, \ldots, N$ & $t = 1, \ldots, T$).

Basic **panel data model**

$$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$$

where y_{it} and y_{is} for $s \neq t$ represent the same individual's outcome across two time periods t and s.

A panel dataset is called "balanced" if T is the same for each individual in our dataset and it is called "unbalanced" if the number of time periods vary, i.e. there is a T_i .

The panel data estimators we are going to discuss work for balanced and unbalanced panels (as long as $T_i > 1$ for all cross-sectional units).

Panel Data Estimation

Balanced panel data

id	year	sales	salary
1	2019	310	2400
1	2020	150	2400
1	2021	180	2500
2	2019	790	3500
2	2020	1460	5000
2	2021	1520	6000
3	2019	120	1800
3	2020	150	2000
3	2021	180	2300

Unbalanced panel data

id	year	sales	salary
1	2019	310	2400
1	2021	180	2500
2	2019	790	3500
2	2020	1460	5000
2	2021	1520	6000
3	2020	150	2000
3	2021	180	2300

Panel Data Estimation

Example: Effect of Firm Entry on Prices:

- p_{it} is the **average price** of a haircut in market i in year t
- x_{it} is the **number of salons** in market i in year t

We are interested in the effect of competition on the price

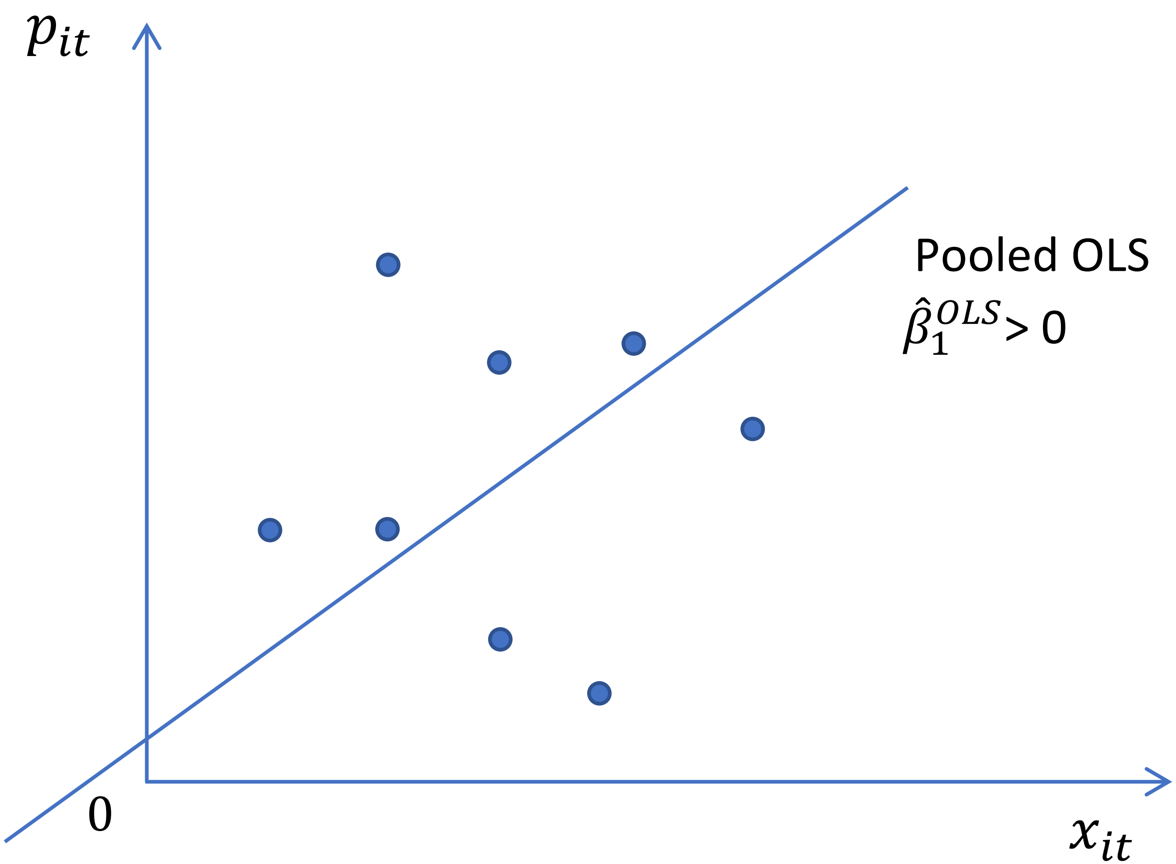
$$p_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$$

Economic intuition: the more competitors, the lower the price ($\beta_1 < 0$).

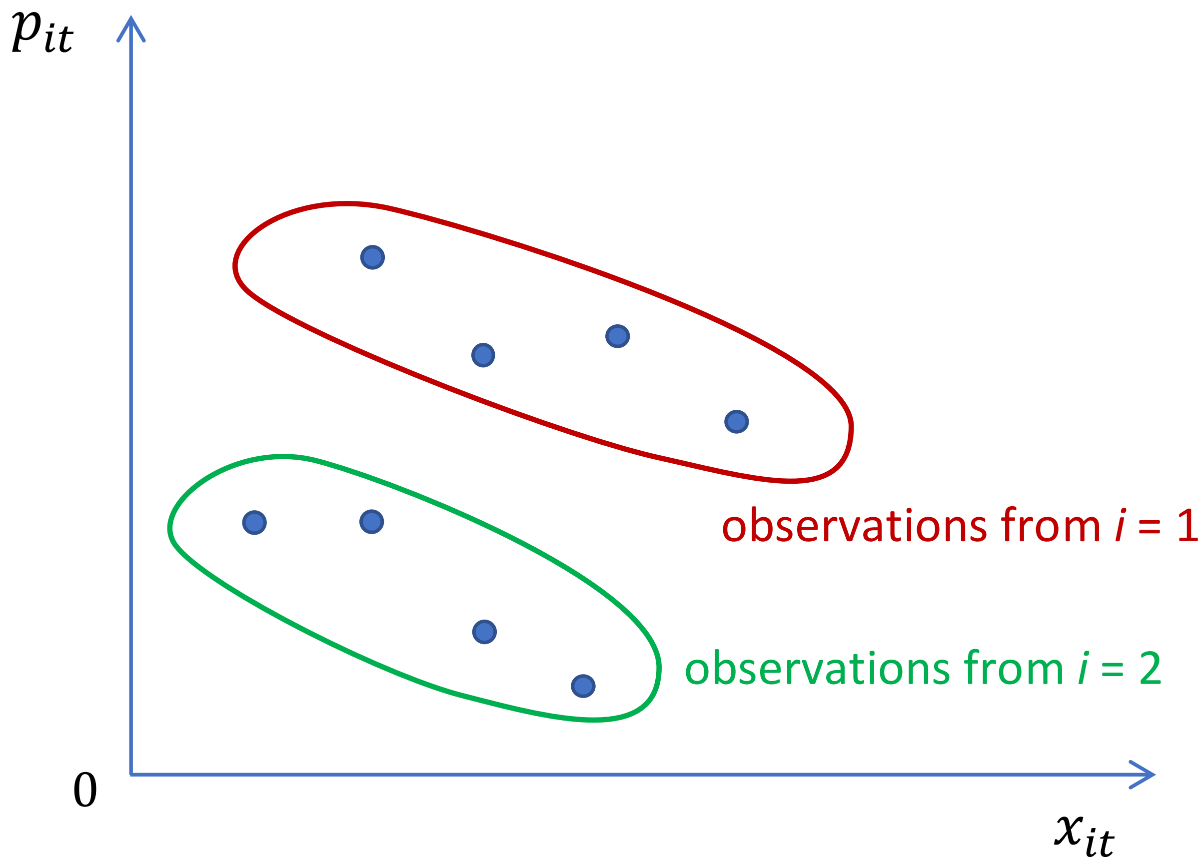
ε_{it} captures every relevant variable that we do not use in our regression (say, because we do not observe it in our data).

Does pooled OLS of p_{it} on x_{it} capture the effect of competition on price only? - **Unlikely!**

Panel Data Estimation



Panel Data Estimation



Panel Data Estimation

Example for omitted variables: Markets with larger demand are generally more attractive (consider a salon in Schwabing compared to a salon in a small village).

Many factors affect the attractiveness of a market (e.g., differences in the population composition or availability of a bus stop).

These factors are captured by the error term ε_{it} if we cannot control for them.

OLS.1, $\left[\mathrm{E}(x_{it}\,\varepsilon_{it}) = 0\right]$, is crucial for consistency of the OLS estimates.

An economic argument that would invalidate OLS.1 could be the following...

Panel Data Estimation

Consider a variable q that measures the attractiveness of a market.
q is an omitted variable if

1. salons in more attractive markets can charge higher prices, i.e. $\gamma > 0$ and
2. salons are more likely to open in more attractive markets, i.e. $\mathrm{Cov}(x, q) > 0$.

Now, in a panel data model we can distinguish **two types of omitted variables**:

- Omitted variables that are time-varying (collected in u_{it}).
- Omitted variables that are time-constant (collected in c_i).

That is, we consider a model with a composite error term $\varepsilon_{it} = c_i + u_{it}$.

Depending on the time horizon of our data, we may be willing to consider the population composition or the availability of a bus stop as time-constant omitted variables.

Panel Data Estimation

$$p_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it} \quad \text{with} \quad \varepsilon_{it} = c_i + u_{it}$$

By the linearity of the expectation operator, we know

$$\mathrm{E}(x_{it}\,\varepsilon_{it}) = \mathrm{E}(x_{it}\,c_i) + \mathrm{E}(x_{it}\,u_{it})$$

Therefore, for OLS.1 to hold we need both

$$\mathrm{E}(x_{it}\,c_i) = 0 \quad \text{and} \quad E(x_{it}\,u_{it}) = 0$$

Now, the fixed effects (FE) estimator allows to relax OLS.1. If we have a panel dataset and use FE estimation, we can consistently estimate β_1 if

$$E(x_{it}\,u_{it}) = 0$$

while $\mathrm{E}(x_{it}\,c_i)$ can be unequal (or equal) to zero.

Panel Data Estimation

Implementation of the FE estimator:

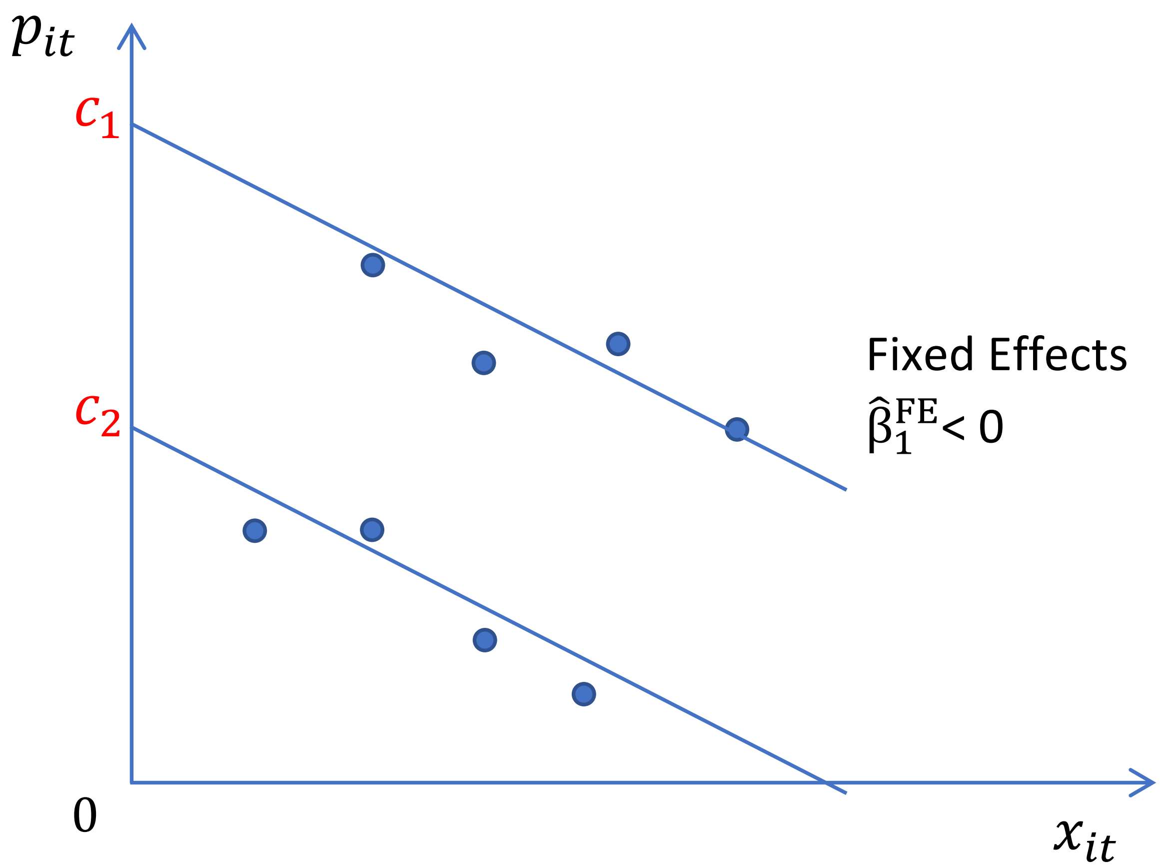
1. Construct a dummy variable for each market i
2. Run OLS of p_{it} on x_{it}, **including** (all but one) market dummies

FE estimator allows for each cross-sectional unit (here market) to have its own constant.

Ceteris-paribus intuition: Holding markets fixed, we estimate the effect of x_{it} on p_{it}.

We only use the **variation within** the cross-sectional units to estimate β_1.

Panel Data Estimation



Panel Data Estimation

Note that the parameters of the market dummies reflect **everything** that

1. affects the price of a haircut and
2. does not vary over time (i.e. time-invariant variables)

That is c_i captures **all time-invariant characteristics** of market i. Not only the unobserved characteristics but also the observed ones.

With FE estimation we can, therefore, not estimate the effect of a time-constant **observed** variable x_i (e.g., the gender wage gap could not be estimated with FE).

Reason: We still use OLS to estimate FE regressions and, therefore, we also need OLS.2 $\left[\mathrm{E}(x_{it}^2) \neq 0\right]$ to hold.

In a FE estimation OLS.2 means that we need within-subject variation in x over time.

Recommended reading

For next week please read chapter:

4.0 Potential Outcomes Causal Model
4.1 Physical Randomization

`https://mixtape.scunning.com/index.html`

Contact

Helmut Farbmacher
office.econometrics@mgt.tum.de