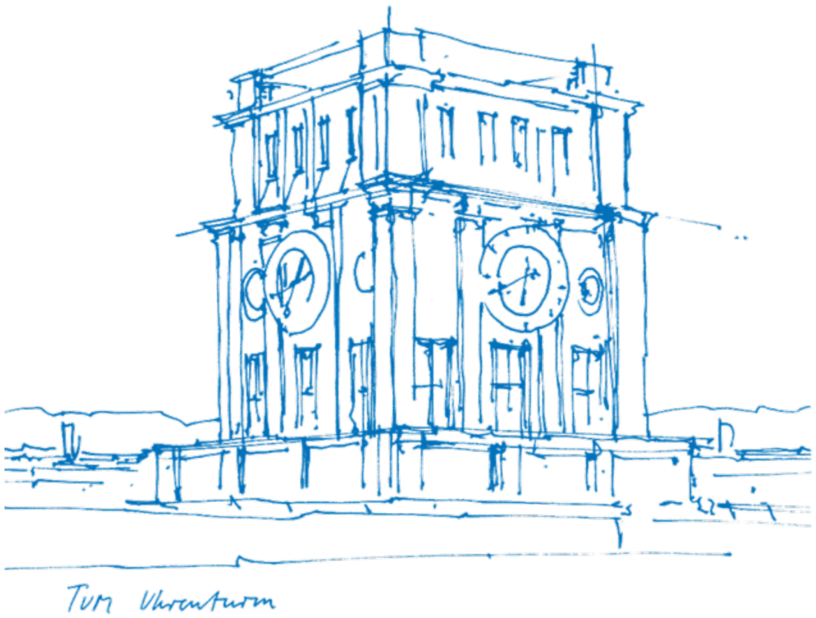
Empirical Research Methods - Lecture 8

Prof. Dr. Helmut Farbmacher
Technical University of Munich
TUM School of Management
Munich, April 9, 2024



Program Evaluation & Identification Strategies

Outline of today's lecture

Program Evaluation

Program Evaluation

In economics we are often interested in the **causal effect of a treatment** (e.g., a policy intervention) on an outcome variable. For example,...

- the causal effect of attending college on wages;
- the causal effect of a training program on the probability of finding a job;
- the causal effect of copayments on health care demand.

Standard econometric arguments: Is the binary treatment indicator exogenous?

Arguments in the evaluation literature look slightly different (although the underlying problem is the same).

Helps us to think about causality from a different perspective.

Program Evaluation

Potential Outcome Framework

Potential Outcome Framework

So far, we assumed that each individual outcome is characterized by a data vector $\{y_i, x_i, u_i\}_{i=1}^{n}$.

However, in the evaluation literature we think about two **potential outcomes**

$$\begin{aligned} Y_i(0) &: \text{potential outcome if individual } i \text{ not treated} \\ Y_i(1) &: \text{potential outcome if individual } i \text{ treated} \end{aligned}$$

Potential outcomes are defined for each individual BUT they will never be realized at the same time (since an individual i is either treated or not).

Define the **treatment status** as

$$D_i = \begin{cases} 0 & \text{if individual } i \text{ is not treated} \\ 1 & \text{if individual } i \text{ is treated} \end{cases}$$

Note: Capitalization of letters is standard in the program evaluation literature.

Potential Outcome Framework

The following **observational rule** gives us the **realized outcome** Y_i

$$Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$$

We only observe the realized outcome Y and the treatment status D in our dataset. That is, we observe

- the potential outcome $Y(0)$ for those being untreated $(D = 0)$ and
- the potential outcome $Y(1)$ for those being treated $(D = 1)$.

However, we do not oberve the **counterfactual** outcomes:

- $Y(1)$ is the counterfactual outcome for those being untreated $(D = 0)$ and
- $Y(0)$ is the counterfactual outcome for those being treated $(D = 1)$.

Potential Outcome Framework

Y_0	Y_1	D	Y
6.862733	7.819566	0	6.862733
8.364915	8.818250	0	8.364915
7.226931	7.904501	0	7.226931
8.649052	9.221686	0	8.649052
8.821402	8.924327	0	8.821402
6.136669	7.036494	1	7.036494
7.584316	7.830404	1	7.830404
8.677257	8.719317	1	8.719317
7.654305	7.982226	1	7.982226
7.369844	8.324348	1	8.324348

Potential Outcome Framework

Y_0	Y_1	D	Y
6.862733	NA	0	6.862733
8.364915	NA	0	8.364915
7.226931	NA	0	7.226931
8.649052	NA	0	8.649052
8.821402	NA	0	8.821402
NA	7.036494	1	7.036494
NA	7.830404	1	7.830404
NA	8.719317	1	8.719317
NA	7.982226	1	7.982226
NA	8.324348	1	8.324348

Potential Outcome Framework

D	Y
0	6.862733
0	8.364915
0	7.226931
0	8.649052
0	8.821402
1	7.036494
1	7.830404
1	8.719317
1	7.982226
1	8.324348

Program Evaluation

Treatment Effect

Treatment Effect

The **individual treatment effect** is the difference between the two potential outcomes:

$$Y_i(1) - Y_i(0)$$

We cannot observe this quantity since we either observe $Y_i(1)$ or $Y_i(0)$ but never both quantities for the same individual i at the same time.

However, we may be able to identify the **Average Treatment Effect**:

$$\mathrm{ATE} = \mathrm{E}\big(Y_i(1) - Y_i(0)\big)$$

The ATE is the expected effect of the treatment when we randomly draw an individual from the population.

Treatment Effect

Another population quantity of interest is the **Average Treatment Effect on the Treated**:

$$\mathrm{ATET} = \mathrm{E}\big(Y_i(1) - Y_i(0)|D_i = 1\big)$$

The ATET is the expected effect of the treatment when we randomly draw an individual from the sub-population that received the treatment.

Neither the ATE **nor** the ATET can be estimated without additional assumptions because we never observe $Y_i(0)$ and $Y_i(1)$ for the same individual i at the same time.

Treatment Effect

Of course, we can estimate the expected difference in the **realized outcome** between untreated and treated individuals, that is

$$\begin{aligned} \Delta &= \mathrm{E}(Y_i|D_i=1) - \mathrm{E}(Y_i|D_i=0) \\ &= \mathrm{E}\big(Y_i(1)|D_i=1\big) - \mathrm{E}\big(Y_i(0)|D_i=0\big) \end{aligned}$$

In many economic applications it is, however, unlikely that Δ equals the ATE.

Why?

Treatment Effect

Selection into treatment is a fundamental problem if we want to evaluate a program.

Example: Let $Y(0)$ and $Y(1)$ be an individual's wages without and with training, then selection into treatment could mean that individuals with higher education

- are more likely (or willing) to do the training and
- have higher values of $Y(0)$ anyway.

Trick to see this:

$$\begin{aligned}
\Delta &= \mathrm{E}\big(Y_i(1)|D_i = 1\big) - \mathrm{E}\big(Y_i(0)|D_i = 0\big) \overbrace{- \mathrm{E}\big(Y_i(0)|D_i = 1\big) + \mathrm{E}\big(Y_i(0)|D_i = 1\big)}^{\text{expanding by counterfactual term}} \\
&= \mathrm{E}\big(Y_i(1)|D_i = 1\big) - \mathrm{E}\big(Y_i(0)|D_i = 1\big) + \mathrm{E}\big(Y_i(0)|D_i = 1\big) - \mathrm{E}\big(Y_i(0)|D_i = 0\big) \\
&= \mathrm{E}\big(Y_i(1) - Y_i(0)|D_i = 1\big) + \mathrm{E}\big(Y_i(0)|D_i = 1\big) - \mathrm{E}\big(Y_i(0)|D_i = 0\big)
\end{aligned}$$

Treatment Effect

Selection into treatment is a fundamental problem if we want to evaluate a program.

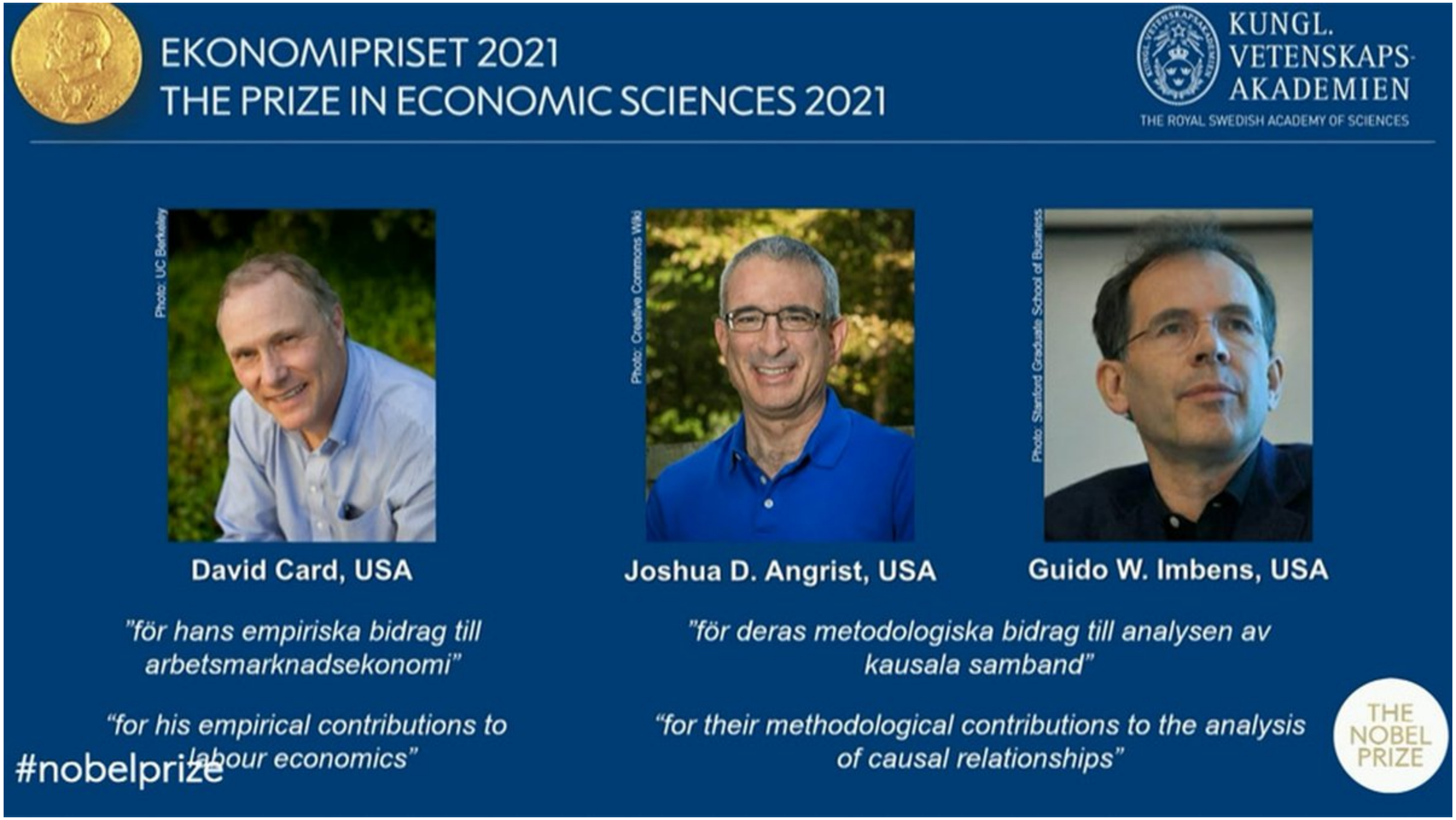
Example: Let $Y(0)$ and $Y(1)$ be an individual's wages without and with training, then selection into treatment could mean that individuals with higher education

- are more likely (or willing) to do the training and
- have higher values of $Y(0)$ anyway.

Trick to see this:

$$\begin{aligned}
\Delta &= \mathrm{E}\big(Y_i(1)|D_i=1\big) - \mathrm{E}\big(Y_i(0)|D_i=0\big) \overbrace{-\mathrm{E}\big(Y_i(0)|D_i=1\big) + \mathrm{E}\big(Y_i(0)|D_i=1\big)}^{\text{expanding by counterfactual term}} \\
&= \mathrm{E}\big(Y_i(1)|D_i=1\big) - \mathrm{E}\big(Y_i(0)|D_i=1\big) + \mathrm{E}\big(Y_i(0)|D_i=1\big) - \mathrm{E}\big(Y_i(0)|D_i=0\big) \\
&= \underbrace{\mathrm{E}\big(Y_i(1)-Y_i(0)|D_i=1\big)}_{\mathrm{ATET}} + \underbrace{\mathrm{E}\big(Y_i(0)|D_i=1\big) - \mathrm{E}\big(Y_i(0)|D_i=0\big)}_{\text{Selection Effect}}
\end{aligned}$$

Treatment Effect



Suppose we have access to observational data for a large number of individuals on their earnings, Y_i, and whether they have completed high school, $D_i = 1$, or not, $D_i = 0$. The observed average difference in earnings across the two groups, is an estimate of $\Delta = E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$. It is unlikely that Δ equals the causal effect of interest, however. To see this formally, subtract and add $E[Y_i(0)|D_i = 1]$, a counterfactual term, which yields:

$$\underbrace{\Delta}_{\text{Difference in means}} = \underbrace{E\big[\big(Y_i(1) - Y_i(0)\big)\big|D_i = 1\big]}_{\text{causal effect on high school completers}} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

The first term in this expression, $E\big[\big(Y_i(1) - Y_i(0)\big)\big|D_i = 1\big]$, is the causal effect of interest. In our high school example, it provides the answer to the question: What is the effect on wage earnings of completing high school, among those that did so? The second term, $E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$, represents selection bias. It measures how different the earnings would have been in the two populations — completers and non-completers — had they not completed high school. For reasons described above, high school completers would likely earn more than non-completers, even without high-school, implying that the selection effect is positive in this example. In such a case, the comparison of means, Δ, provides an upward biased estimate of the causal effect of high school education on earnings.

Identification Strategies

Identification Strategies

$Y_i(0)$: potential outcome if individual i is not treated

$Y_i(1)$: potential outcome if individual i is treated

Observational rule gives us the realized outcome:

$$Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$$

Realized outcome mixes two effects:

$$\begin{aligned}\Delta &= \mathrm{E}\big(Y_i(1)|D_i = 1\big) - \mathrm{E}\big(Y_i(0)|D_i = 0\big) \\ &= \underbrace{\mathrm{E}\big(Y_i(1) - Y_i(0)|D_i = 1\big)}_{\text{ATET}} + \underbrace{\mathrm{E}\big(Y_i(0)|D_i = 1\big) - \mathrm{E}\big(Y_i(0)|D_i = 0\big)}_{\text{Selection Effect}}\end{aligned}$$

Now, we are going to discuss several **strategies to disentangle both effects**.

Identification Strategies

Randomization

Randomization

Random assignment of treatment (say, in a real experiment) is the **gold standard** to learn treatment effects.

Randomized Controlled Trials (RCTs) are often used in economics nowadays.

Banerjee, Duflo and Kremer received the Nobel Prize in Economics 2019 more or less for their RCTs to alleviate global poverty.

Analyzing experimental data is (almost) always very easy. Why is this the case?

Randomization

Say, we want to estimate the ATET:

$$\begin{aligned} \text{ATET} &= \mathrm{E}\big(Y_i(1) - Y_i(0) | D_i = 1\big) \\ &= \mathrm{E}\big(Y_i(1) | D_i = 1\big) - \mathrm{E}\big(Y_i(0) | D_i = 1\big) \end{aligned}$$

where the second equality just uses the linearity of the expectation operator.

We can only estimate the difference in the realized outcome:

$$\begin{aligned} \Delta &= \mathrm{E}(Y_i | D_i = 1) - \mathrm{E}(Y_i | D_i = 0) \\ &= \mathrm{E}\big(Y_i(1) | D_i = 1\big) - \mathrm{E}\big(Y_i(0) | D_i = 0\big) \quad \text{[holds by definition of } Y_i\text{]} \end{aligned}$$

Randomization

Say, we want to estimate the ATET:

$$\begin{aligned} \text{ATET} &= \mathrm{E}\big(Y_i(1) - Y_i(0)|D_i = 1\big) \\ &= \mathrm{E}\big(Y_i(1)|D_i = 1\big) - \mathrm{E}\big(Y_i(0)|D_i = 1\big) \end{aligned}$$

where the second equality just uses the linearity of the expectation operator.

We can only estimate the difference in the realized outcome:

$$\begin{aligned} \Delta &= \mathrm{E}(Y_i|D_i = 1) - \mathrm{E}(Y_i|D_i = 0) \\ &= \mathrm{E}\big(Y_i(1)|D_i = 1\big) - \mathrm{E}\big(Y_i(0)|D_i = 0\big) \quad \text{[holds by definition of } Y_i\text{]} \end{aligned}$$

So far, ATET and Δ are <u>not</u> equal.

Let's see how randomization of treatment affects the program evaluation...

Randomization

Randomization of treatment implies that D is **independent** of any other variable. This means that D is also independent of the potential outcomes, $Y(0)$ and $Y(1)$.

We can use the independence of D from the potential outcomes in the following way:

$$\underbrace{\mathrm{E}\big(Y_i(0)|D_i=1\big)}_{\text{counterfactual}} = \mathrm{E}\big(Y_i(0)|D_i=0\big) = \underbrace{\mathrm{E}\big(Y_i|D_i=0\big)}_{\text{can be estimated}}$$

The first equality holds by independence of D and $Y(0)$, which follows from the randomization of the treatment. The second equality holds by the definition of Y.

That is under randomization we have

$$\begin{aligned}
\Delta &= \mathrm{E}\big(Y_i(1)|D_i=1\big) - \mathrm{E}\big(Y_i(0)|D_i=0\big) \\
&= \mathrm{E}\big(Y_i(1)|D_i=1\big) - \mathrm{E}\big(Y_i(0)|D_i=1\big) \\
&= \mathrm{ATET}
\end{aligned}$$

Randomization

Randomization enables us to estimate the ATET from the realized outcome ($\Delta = \mathrm{ATET}$). The following regression is an easy way to do so:

$$Y_i = \beta_0 + \beta_1 D_i + U_i$$

Because D is a (binary) dummy regressor, we get as marginal effect the mean difference between treated and nontreated outcomes, i.e.,

$$\beta_1 = \mathrm{E}(Y_i | D_i = 1) - \mathrm{E}(Y_i | D_i = 0) = \Delta$$

which under randomization equals the ATET.

An alternative way to think about consistency of $\widehat{\beta}_1$ is that $\mathrm{E}(D_i U_i) = 0$, which means that OLS.1 is fulfilled, because D is independent of U due to randomization.

What does OLS.2 mean here?

Randomization

Of course, in this setting we also know that $\mathrm{ATET} = \mathrm{ATE}$ because D is independent of any other random variable (due to randomization).

That is we also have

$$\mathrm{ATET} = \mathrm{E}\big(Y_i(1) - Y_i(0)|D_i = 1\big) = \mathrm{E}\big(Y_i(1) - Y_i(0)\big) = \mathrm{ATE}$$

The last result is a unique characteristic of randomization.

In a general setting (i.e. **no** randomization of treatment), there are <u>not</u> many reasons to assume that the average treatment effect in the entire population is equal to the average treatment effect in the treated population.

People with larger **individual** treatment effects have a higher incentive to participate.

Identification Strategies

Selection on (Un)Observables

Selection on (Un)Observables

Selection on observables and/or selection on unobservables is a huge problem in applied economics.

Suppose that $Y(1)$ and $Y(0)$ are an individual's wages with and without training and D is an indicator whether this person did or did not participate in the training.

If people can choose to participate, then it could be that individuals with higher education

1. are more likely (or more willing) to do the training
2. have already a higher value of $Y(0)$

In this case D and $Y(0)$ are not independent.

Selection on (Un)Observables

If selection only depends on **observable variables**, we can simply add these variables as additional exogenous regressors $\mathbf{x}$, i.e. we estimate

$$Y_i = \beta_0 + \beta_1 D_i + \mathbf{x}_i'\gamma + U_i$$

If selection also depends on **unobservable variables**, differences-in-differences or instrumental variables estimation could help to estimate the treatment effect.

Identification Strategies

Differences-in-Differences Estimation

Differences-in-Differences Estimation

Differences-in-Differences (DiD) estimation is a method that is often used to evaluate programs.

It is particularly well suited for the analysis of natural experiments. In a **natural experiment**, individuals are (as good as) randomly assigned to a treatment and a control group.

In principle DiD estimation also works if there is selection on (un)observables.

Data requirement: We can observe individuals of both groups for two periods (before and after the treatment). A repeated cross-section is sufficient (we do not need a real panel dataset).

Differences-in-Differences Estimation

Treatment group: Individuals that are affected by the treatment.
Control group: Individuals that are not affected by the treatment.

Note that both groups need to be unaffected by the treatment **before** the reform.
That's why DiD estimation is often used to evaluate policy changes.

Now, let's partition the data into four groups:

	Before	After
Control	$\bar{y}_{00}$	$\bar{y}_{01}$
Treatment	$\bar{y}_{10}$	$\bar{y}_{11}$

where $\bar{y}_{ta}$ denotes the mean outcome in the respective group.

Differences-in-Differences Estimation

From this table we can now already calculate the differences-in-differences:

	Before	After	Difference over time
Control	$\bar{y}_{00}$	$\bar{y}_{01}$	$\bar{y}_{01} - \bar{y}_{00}$
Treatment	$\bar{y}_{10}$	$\bar{y}_{11}$	$\bar{y}_{11} - \bar{y}_{10}$
Differences over groups	$\bar{y}_{10} - \bar{y}_{00}$	$\bar{y}_{11} - \bar{y}_{01}$	DiD

$$\mathrm{DiD}_v = \bar{y}_{11} - \bar{y}_{10} - [\bar{y}_{01} - \bar{y}_{00}]$$
$$\mathrm{DiD}_h = \bar{y}_{11} - \bar{y}_{01} - [\bar{y}_{10} - \bar{y}_{00}]$$

Of course, $\mathrm{DiD}_v = \mathrm{DiD}_h = \mathrm{DiD}$.

While this calculation gives us the point estimate, it does <u>not</u> provide an easy way to calculate standard errors. Therefore, it is better to employ a linear model representation of the DiD...

Differences-in-Differences Estimation

Linear model representation of DiD estimation:

$$y = \alpha + \beta \text{After} + \gamma \text{Treat} + \delta \text{After} \cdot \text{Treat} + u$$

with

$$\text{After} = \begin{cases} 1 & \text{if observations after the treatment} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\text{Treat} = \begin{cases} 1 & \text{if observations belongs to treatment group} \\ 0 & \text{otherwise} \end{cases}$$

Note that Treat $= 1$ also for observations **before** the treatment (if they would belong to the treated group after the treatment).

Differences-in-Differences Estimation

$$y = \alpha + \beta \text{After} + \gamma \text{Treat} + \delta \text{After} \cdot \text{Treat} + u$$

Interpretation of the parameters in this regression:

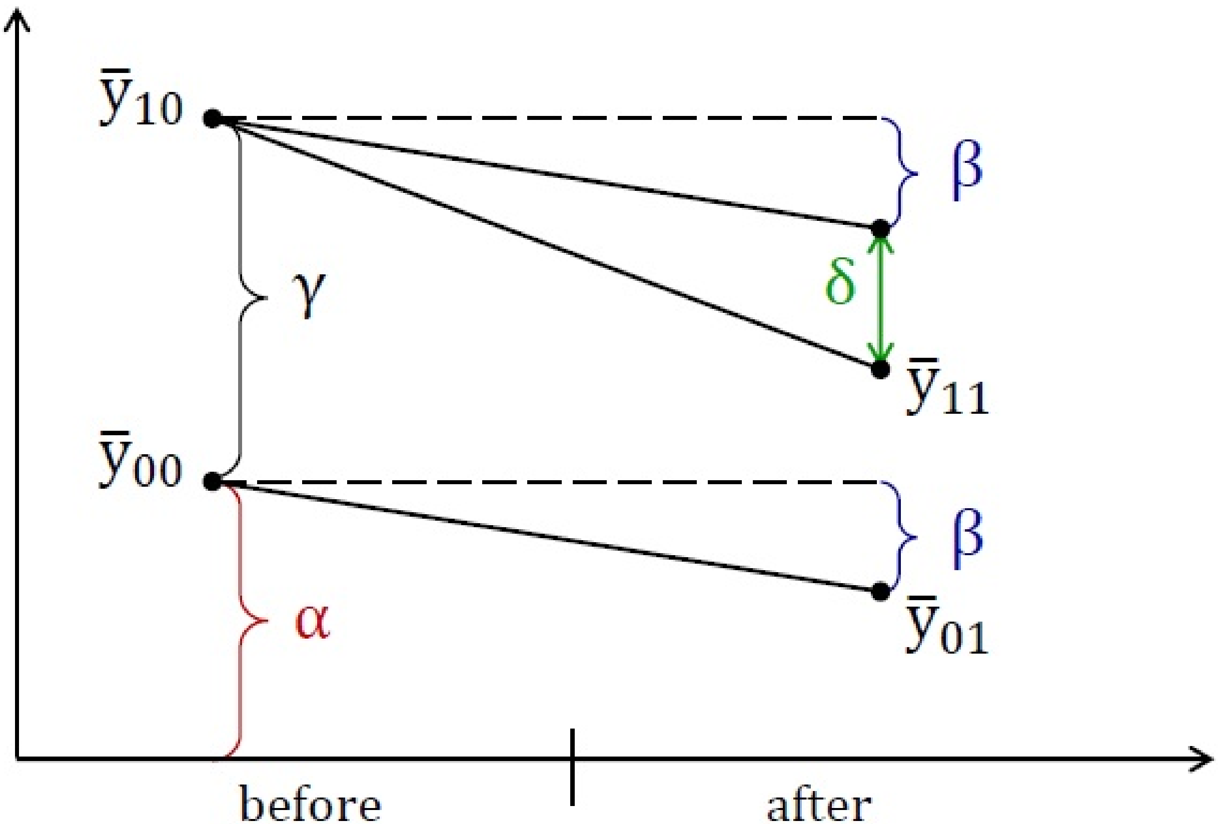
	Before	After	Difference over time
Control	α	$\alpha+\beta$	β
Treatment	$\alpha+\gamma$	$\alpha+\beta+\gamma+\delta$	$\beta+\delta$
Differences over groups	γ	$\gamma+\delta$	δ

That is, we can interpret $\widehat{\delta}$ as the DiD estimate and additionally get valid standard errors of this estimate.

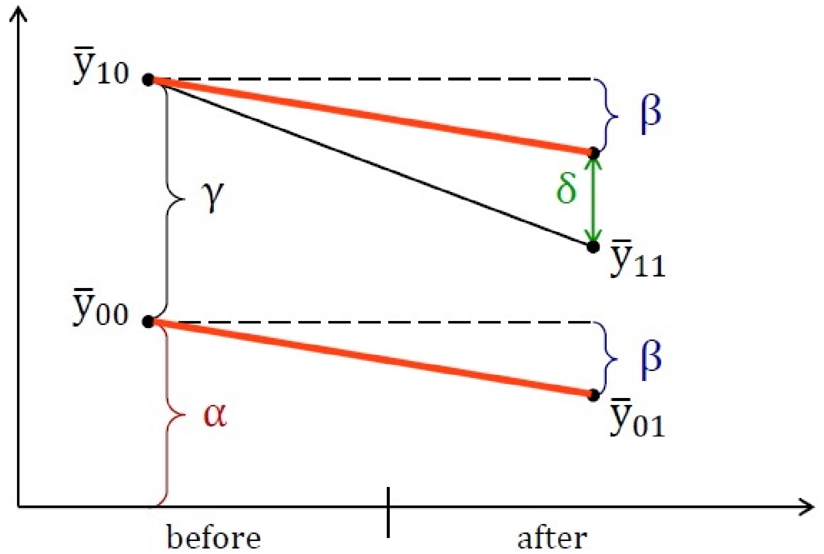
Moreover, we can easily include control variables in the linear model.

Differences-in-Differences Estimation

Graphical illustration



Differences-in-Differences Estimation



Crucial assumption: Treated observations would follow **the same trend** as control observations in the absence of the treatment (**common trend assumption**).

This is an identifying assumption, which in general cannot be tested with the data at hand.

Differences-in-Differences Estimation: Application

Example: Causal effect of minimum wage on employment.

Theoretical considerations: employment falls due to downward-sloping labour demand curve.

Card and Krueger (1994) "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania" *American Economic Review* Vol. 84 No.4

On April 1st, New Jersey (NJ) increased its minimum wage from $4.25 to $5.05 .

(Adjacent) Pennsylvania (PA) did not change its minimum wage.

Differences-in-Differences Estimation: Application

Data:

They surveyed 410 fast-food stores in NJ and PA before and after the minimum wage increase in NJ.

Fast-food industry: most employees are paid close to the minimum wage; looking at average income earners does not make sense.

Research Strategy:

Differences-in-differences estimation with NJ as treatment group and PA as control group.

Differences-in-Differences Estimation: Application

Figure: Part of Table 3 in Card and Krueger (1994)

Variable	Stores by state: PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	–2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	–0.14 (1.07)
3. Change in mean FTE employment	–2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Outcome: Full-time-equivalent (FTE) employment counts per store

Unexpected results: Employment actually rose in NJ relative to PA after the minimum wage increase.

Applied Econometrics

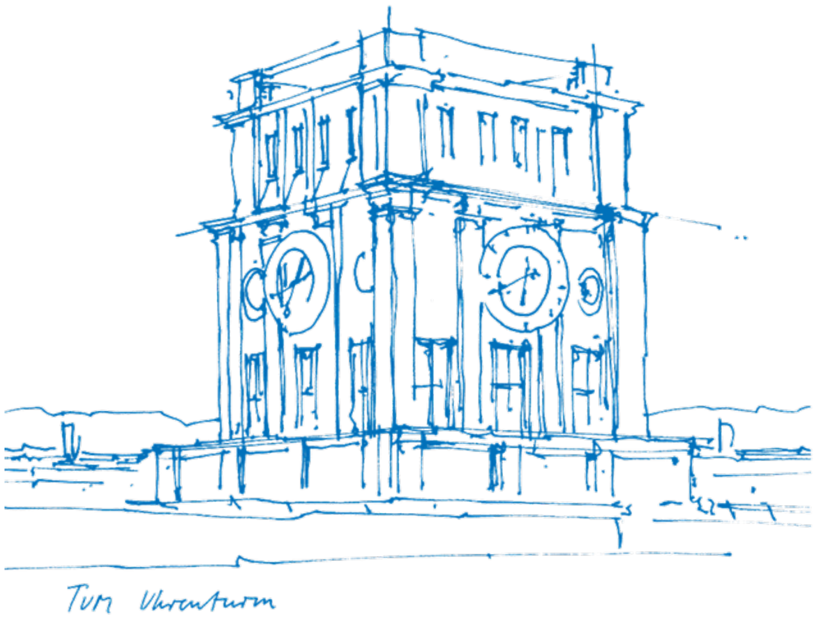
Contact

Helmut Farbmacher
office.econometrics@mgt.tum.de

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 $\left[\mathrm{E}(x_i \varepsilon_i) = 0\right]$ & OLS.2 $\left[\mathrm{E}(x_i^2) \neq 0\right]$, 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

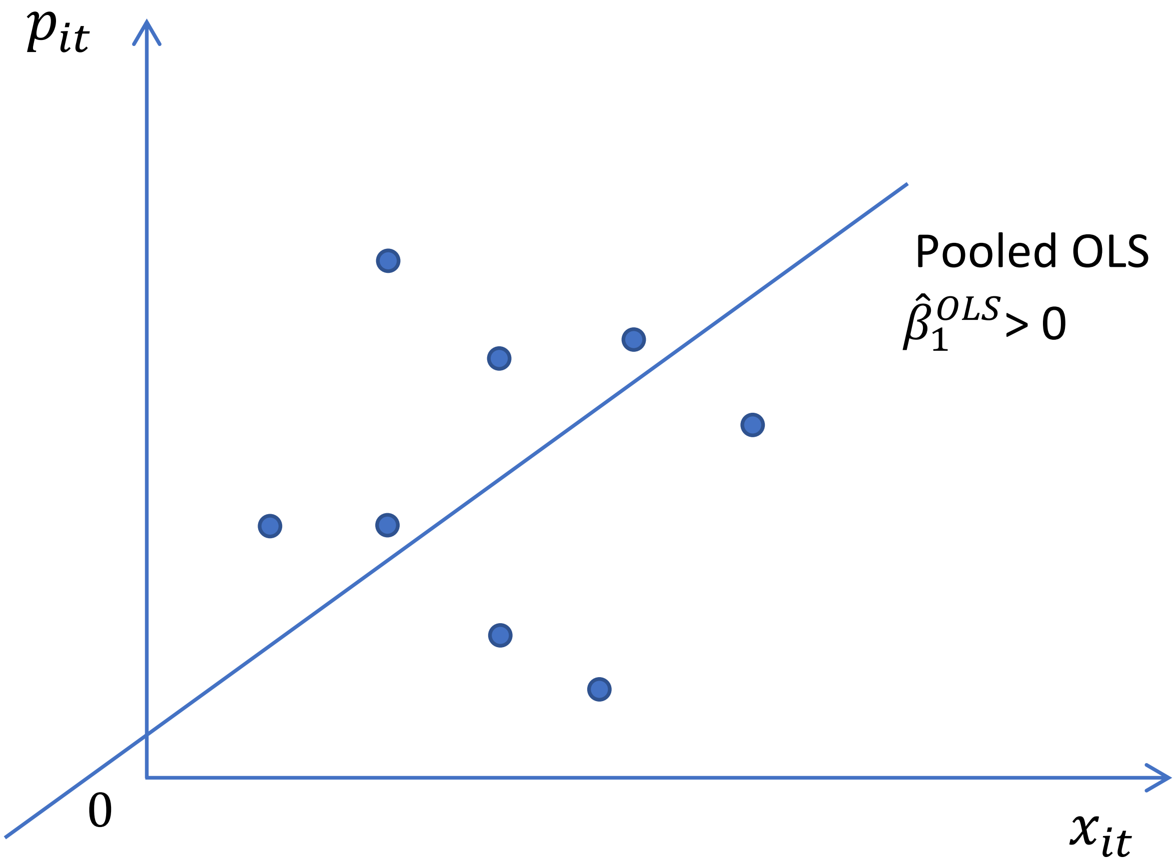
$$p_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$$

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

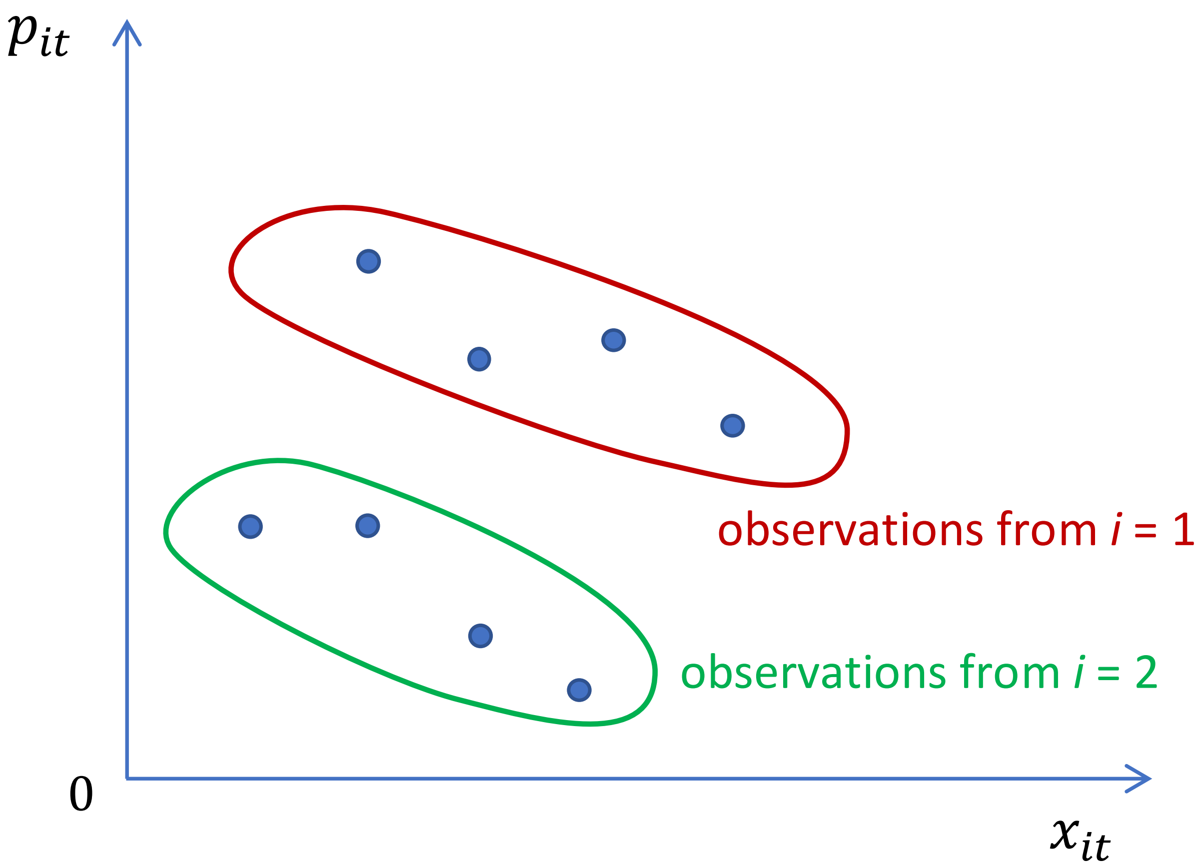
ε_{it} captures every relevant variable that we do not use in our regression (say, because we do not observe it in our data).

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

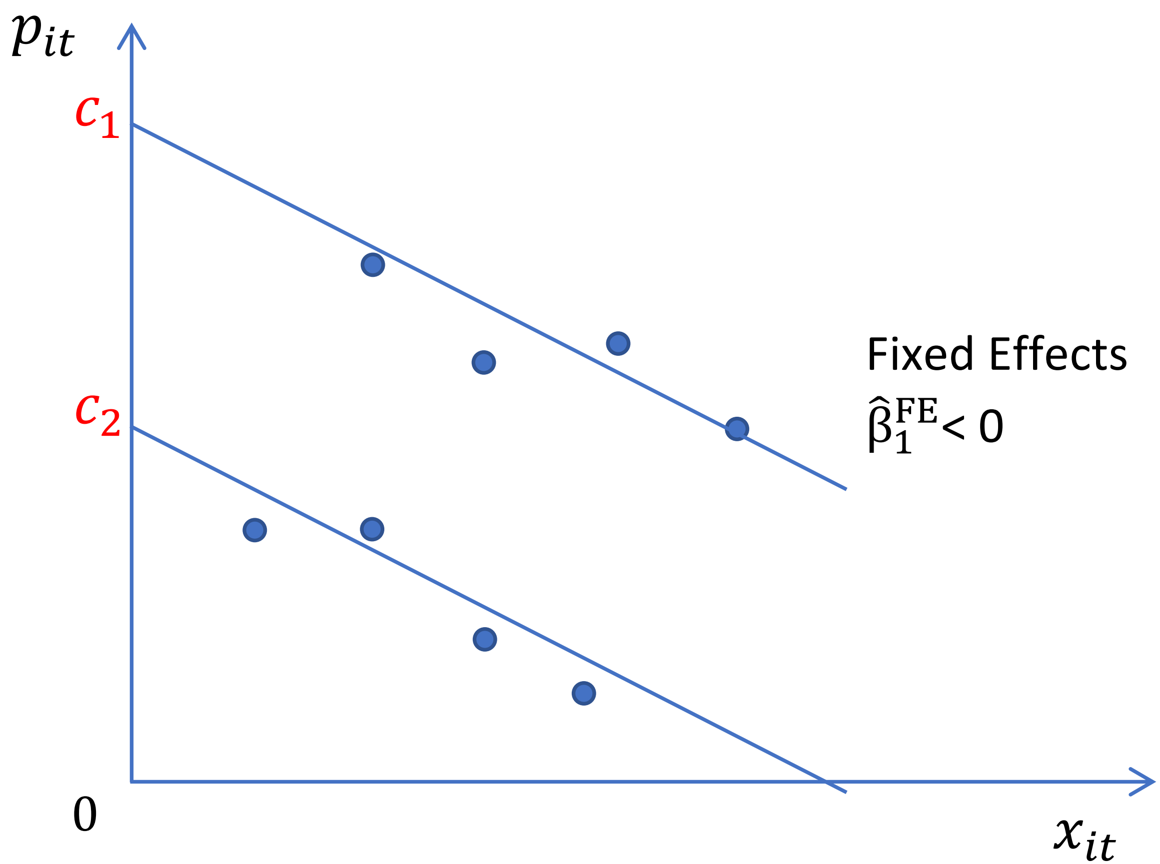
FE estimator allows for each cross-sectional unit (here market) to have its own constant.

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.

Applied Econometrics

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`

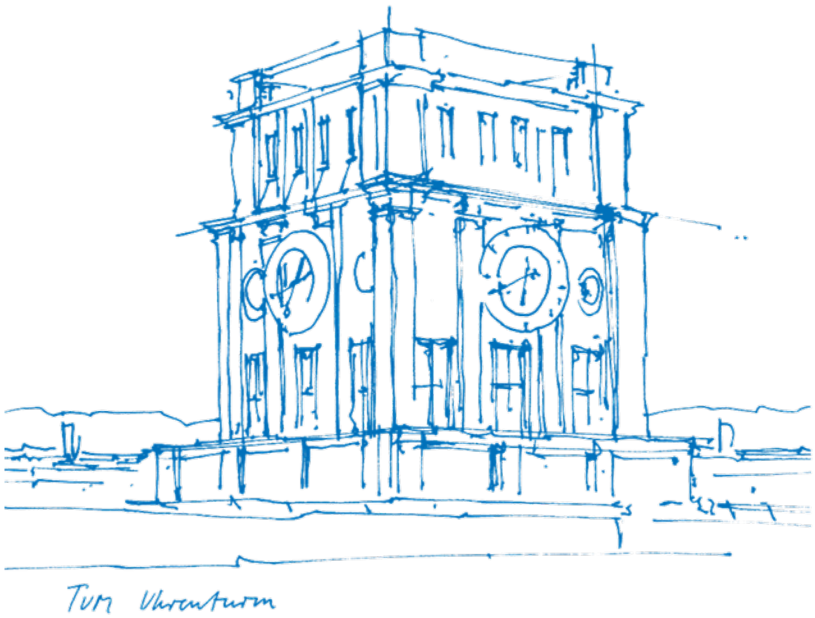
Applied Econometrics

Contact

Helmut Farbmacher
office.econometrics@mgt.tum.de

Empirical Research Methods - Lecture 6

Prof. Dr. Helmut Farbmacher
Technical University of Munich
TUM School of Management
Munich, May 27, 2024



Linear Regression (Part IV)

Outline of today's lecture

Interpretation of Parameters

- Logarithms in Regressions
- Polynomials
- Interactions
- Categorical Regressors (two categories)
- Categorical Regressors (multiple categories)

Hypothesis Testing

- Different ways to convey similar information
- Sidenote 1: Significance vs Relevance
- Sidenote 2: Multiple testing

Interpretation of Parameters

Logarithms in Regressions

Continuous Regressors (Recap)

$$y_i = \alpha + \beta x_i + u_i$$

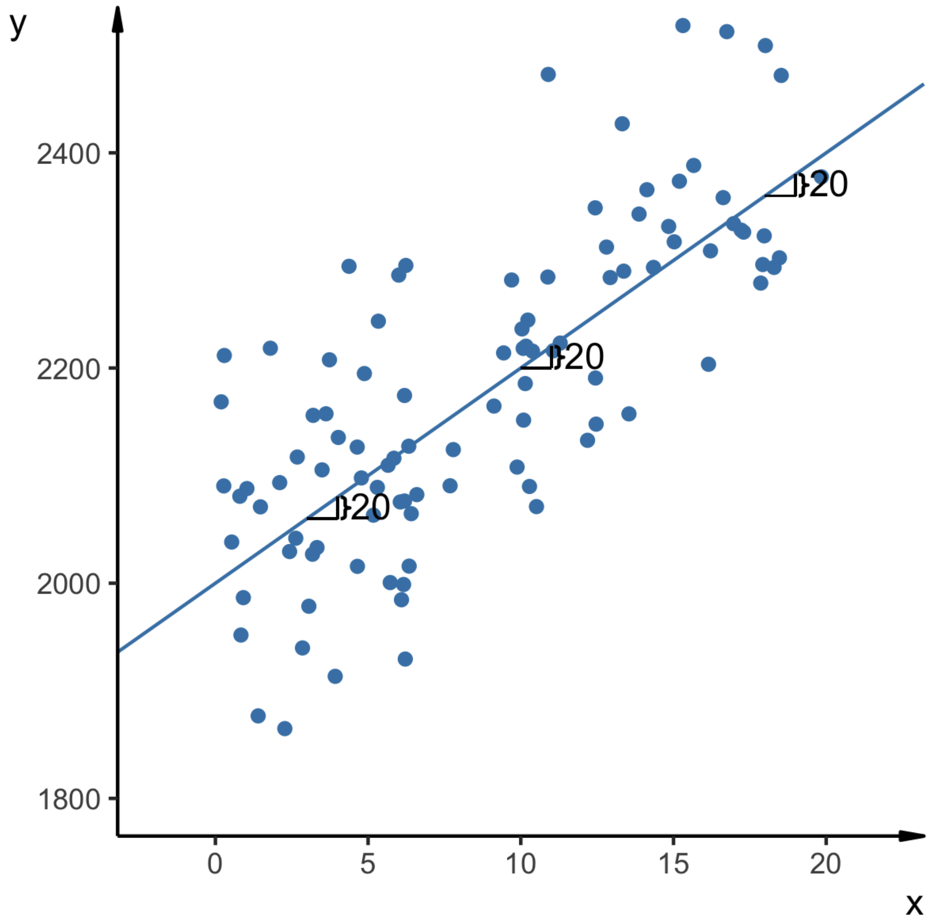
When a regressor x is continuously distributed, we define the **marginal effect** of a change in x as derivative

$$\frac{\partial \mathrm{E}(y|x)}{\partial x} = \beta$$

In the linear regression model the derivative is simply the slope parameter (β).

That is the marginal effect of x is **constant** in a linear model. Particularly, it does <u>not</u> depend on the level of y or x so far.

Continuous Regressors (Recap)



Logarithms in Regressions

We can also model nonlinear relations in a linear regression model. Note that the "linearity" in a linear regression is just referring to the parameters in the model.

That is, the following regressions are linear regression models,

$$\begin{aligned} y &= \alpha + \beta ln(x) + u \\ ln(y) &= \alpha + \beta x + u \\ ln(y) &= \alpha + \beta ln(x) + u \end{aligned}$$

although the natural logarithm $ln(\cdot)$ clearly is a nonlinear function.

The variables being log-transformed must take on positive values only.

Logarithms in Regressions

Linear-Log:	$y = \alpha + \beta ln(x) + u$	A 1% change in x is associated with a change in y of 0.01β.
Log-Linear:	$ln(y) = \alpha + \beta x + u$	A change in x by one unit (Δx = 1) is associated with a 100*β% change in y.
Log-Log:	$ln(y) = \alpha + \beta ln(x) + u$	A 1% change in x is associated with a β% change in y. β is thus the elasticity of y with respect to x.

Logarithms in Regressions

Data info:

The *ceosal2* dataset contains information on U.S. CEOs, including age, degree, salary, tenure, etc.from 1990.

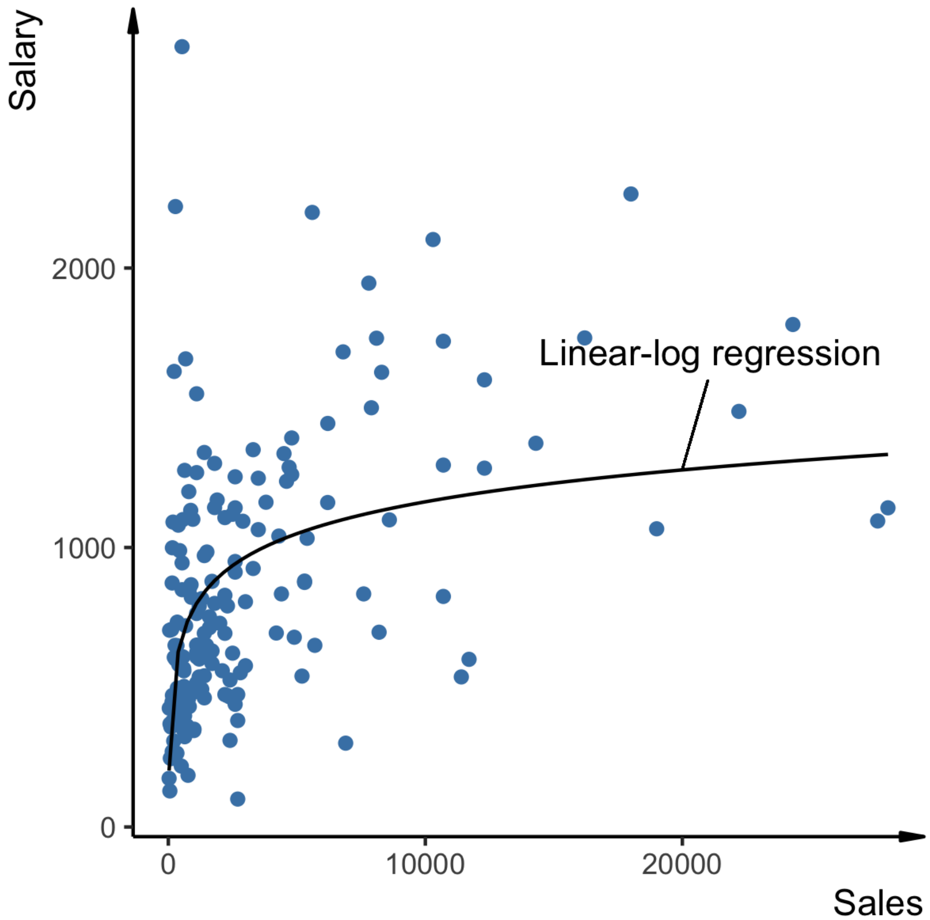
Variables:

salary	compensation of the CEO (in $1000)
sales	firm sales, millions

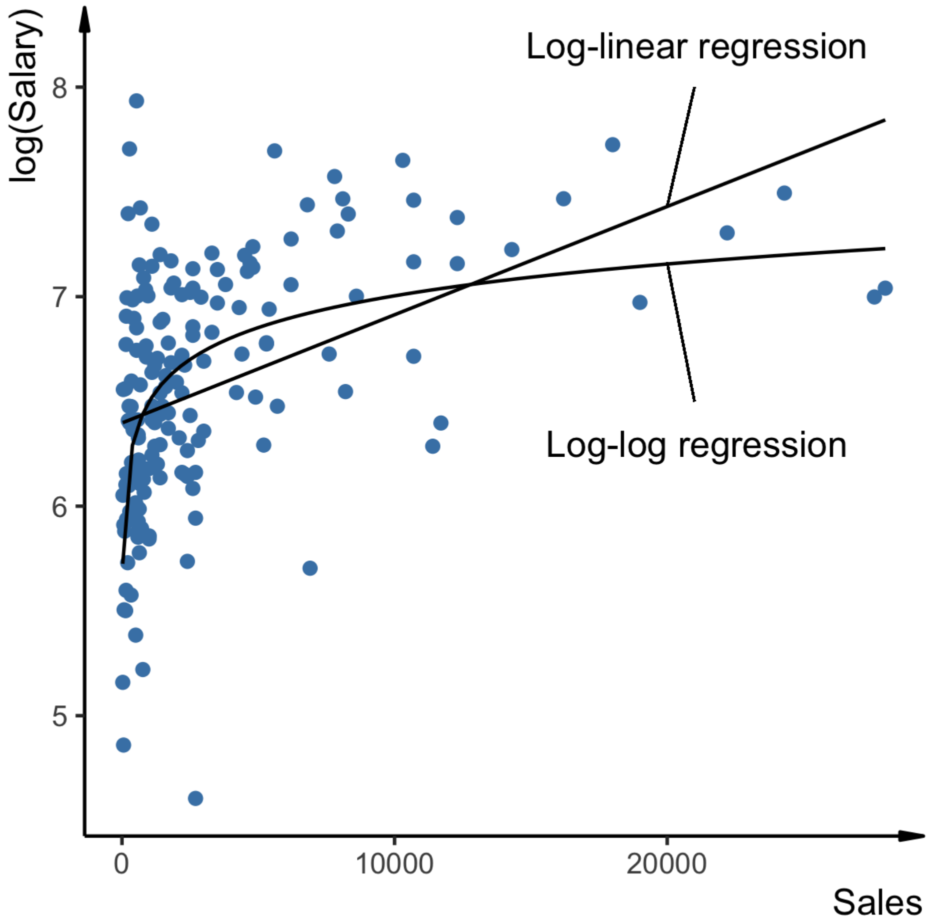
Data can be downloaded here:
`https://econpapers.repec.org/paper/bocbocins/ceosal2.htm`.

You can also load the data via the *wooldridge* package in *R*, use the function *data("ceosal2")* after installing the package.

Logarithms in Regressions



Logarithms in Regressions



Interpretation of Parameters

Polynomials

Polynomials

Example: quadratic effect of job experience

$$wage = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 exp^2 + u$$

Wage is explained by education and a quadratic function of experience.

The model has three explanatory variables:
education, experience and experience squared.

The marginal effect of experience is:

$$\frac{\partial \mathrm{E}(wage|\boldsymbol{x})}{\partial exp} = \beta_2 + 2\beta_3 exp$$

Polynomials

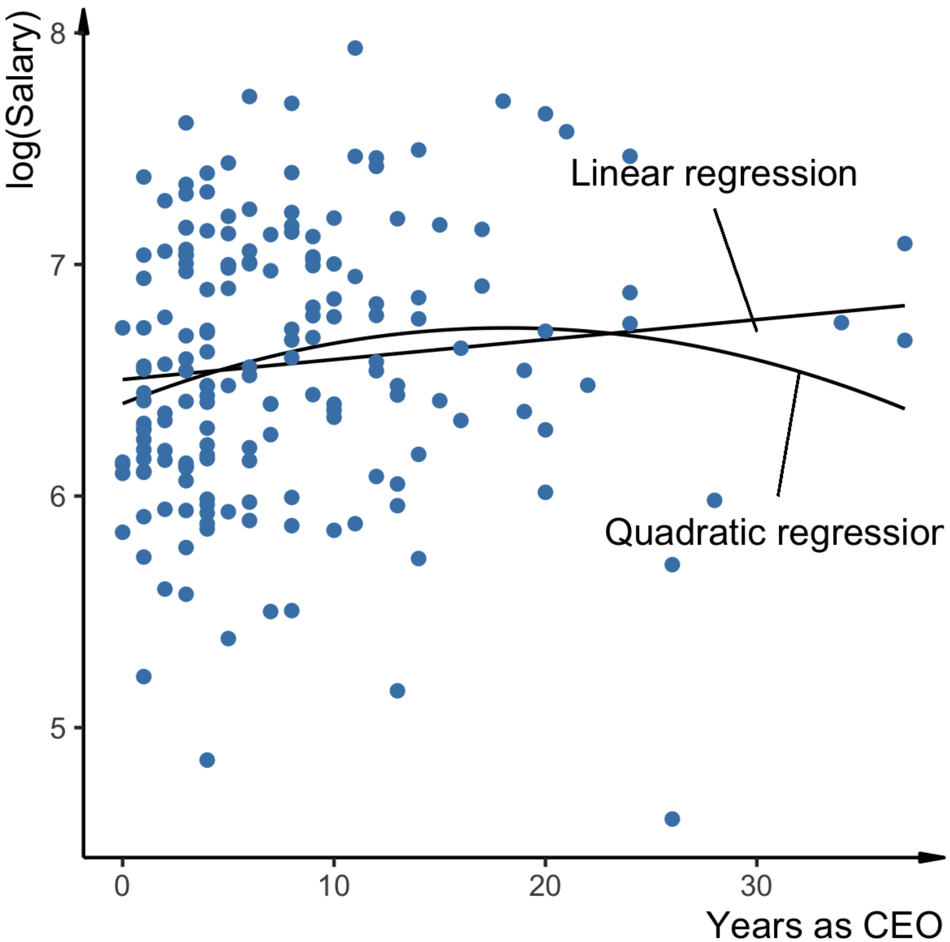
Example: CEO salary

$$log(salary) = \beta_0 + \beta_1 log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u$$

Model assumes a (ceteris paribus) constant elasticity relationship between CEO salary and the sale of the firm.

Model assumes a (ceteris paribus) quadratic relationship between log CEO salary and his or her tenure with the firm.

Polynomials



Interpretation of Parameters

Interactions

Binary & Continuous Regressors (Recap)

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + u$$

When a regressor x is **binary**, we get as marginal effect a group difference

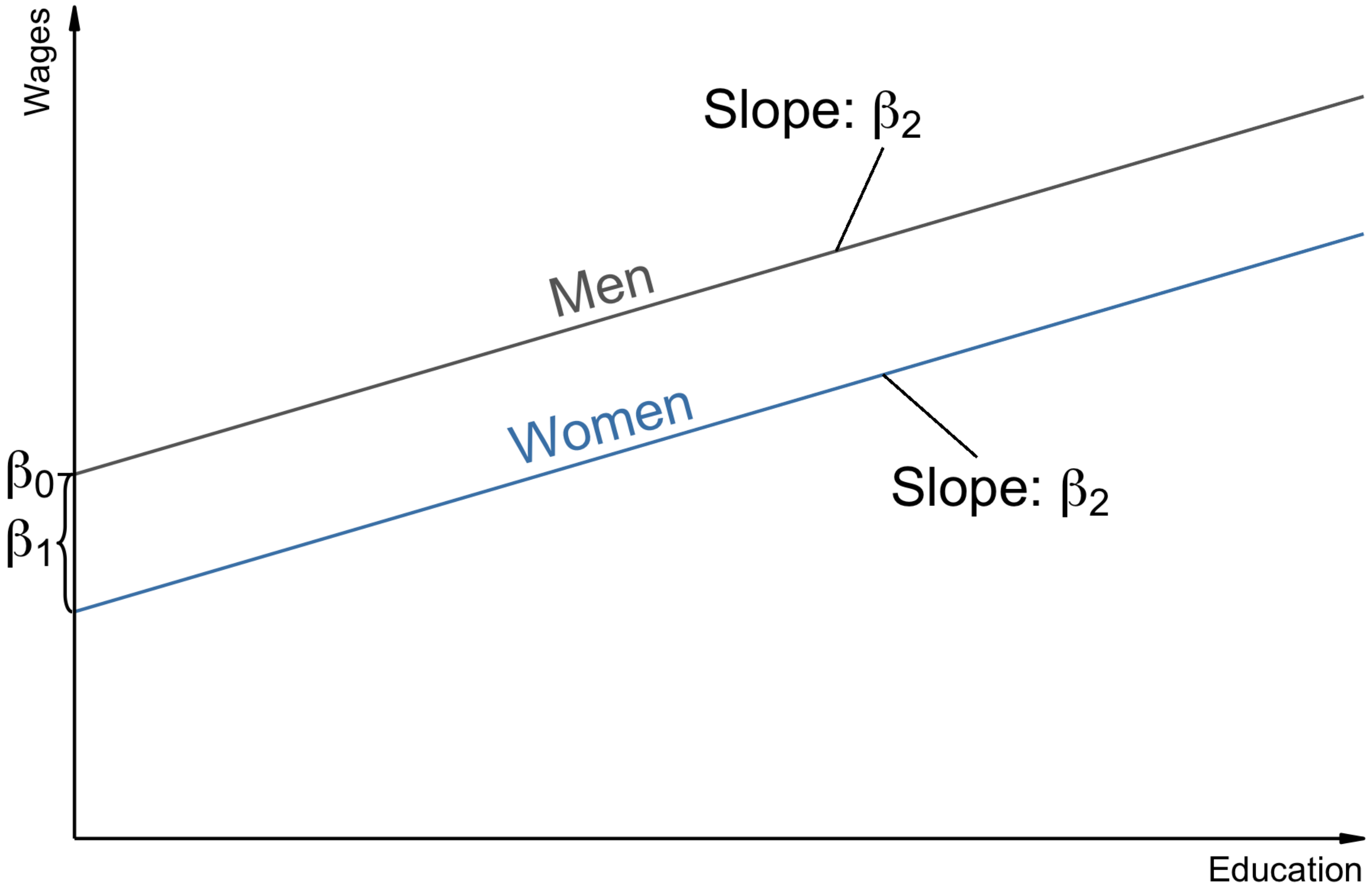
$$\beta_1 = \mathrm{E}[wage|educ, female = 1] - \mathrm{E}[wage|educ, female = 0]$$

$$\beta_2 = \frac{\partial \mathrm{E}[wage|educ, female]}{\partial educ}$$

hence, β_1 is an **intercept shifter** and β_2 is the slope parameter.

Interpretation: β_1 gives the difference in wages between women and men, holding education fixed.

Binary & Continuous Regressors (Recap)



Interactions

An important practical technique is the use of **interaction terms**.

Suppose we want to allow that the effect of schooling differs b/w women and men

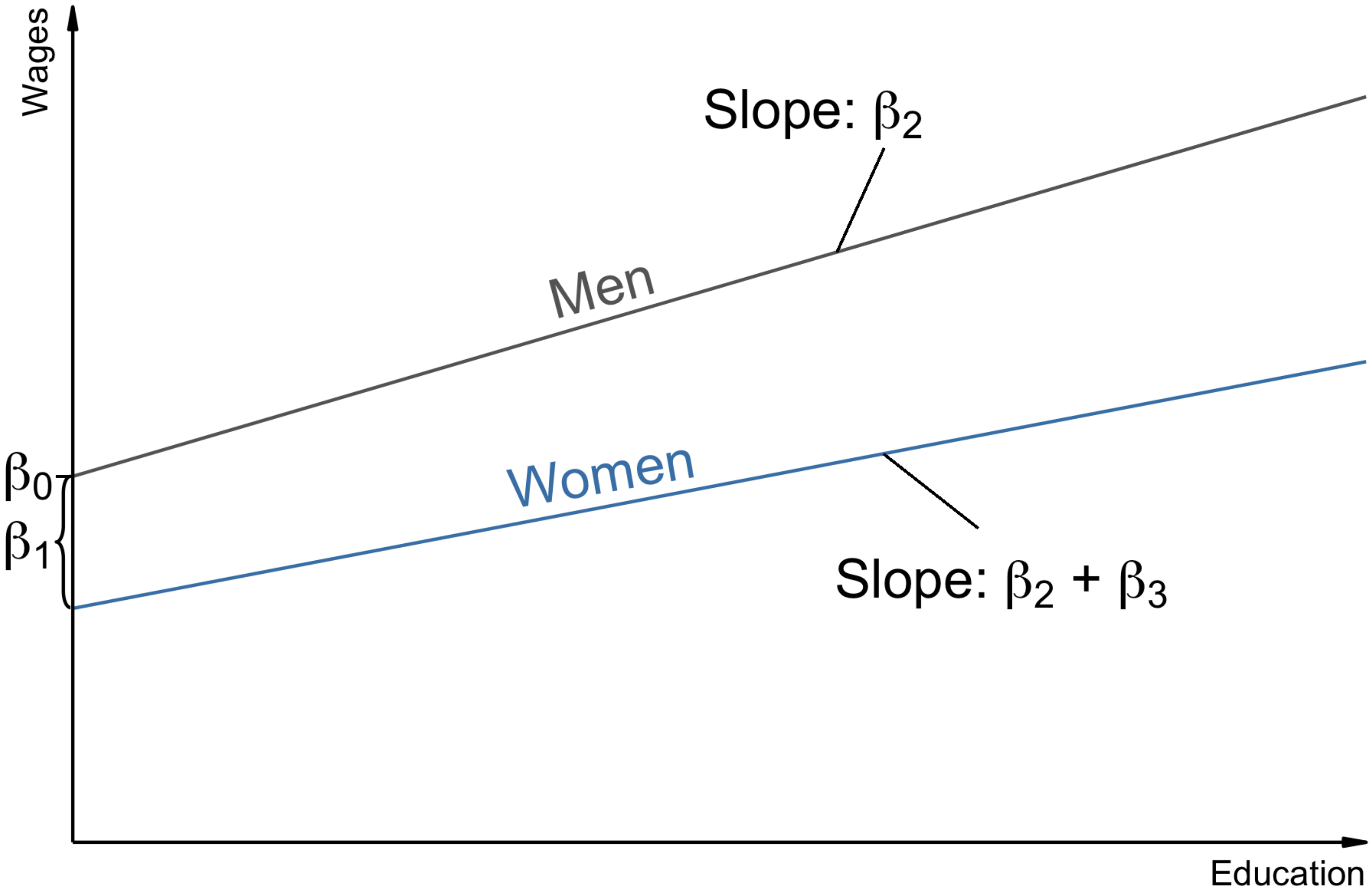
$$wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 female \times educ + u$$

The marginal effect of *educ* on *wage* now is

$$\frac{\partial \mathrm{E}(wage|\boldsymbol{x})}{\partial educ} = \beta_2 + \beta_3 female$$

It is straightforward to test for an interaction effect $H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$.

Interactions: Graphical Illustration



Interpretation of Parameters

Categorical Regressors (two categories)

Categorical Regressors (two categories)

Consider a categorical variable with mutually exclusive categories.

When using dummy variables, one category has to be omitted.

For example, $male + female = 1$, there is perfect collinearity between *male*, *female* and the *constant* (**dummy variable trap**).

We call the omitted category the **reference or baseline category**.

1. $wage = \beta_0 + \beta_1 educ + \beta_2 female + u \rightarrow$ Reference category are men ($female = 0$)

   or

2. $wage = \alpha_0 + \alpha_1 educ + \alpha_2 male + u \rightarrow$ Reference category are women ($male = 0$)

Categorical Regressors (two categories)

How does our dataset look like with a dummy variable? Say, we have a constant, educ (in years) and a dummy for female (3rd variable):

$$\begin{pmatrix} 1 & 9 & 1 \\ 1 & 12 & 1 \\ 1 & 8 & 0 \\ 1 & 15 & 1 \\ 1 & 15 & 0 \end{pmatrix}$$

Now, let's add a dummy variable for male (4th variable) in this dataset:

$$\begin{pmatrix} 1 & 9 & 1 & 0 \\ 1 & 12 & 1 & 0 \\ 1 & 8 & 0 & 1 \\ 1 & 15 & 1 & 0 \\ 1 & 15 & 0 & 1 \end{pmatrix}$$

$male + female = 1$, **perfect collinearity** between *male*, *female* and the *constant*.

Categorical Regressors (two categories)

Statistical software drops one of the dummies if you include *male* and *female*.

General rule: We always have to leave one of the categories out of the regression (**reference/baseline category**)

Does it matter which category we use as reference? Not really!

We can change the reference category

$$\begin{aligned} wage &= \beta_0 + \beta_1 educ + \beta_2 female + u \\ &= \beta_0 + \beta_1 educ + \beta_2 (1 - male) + u \\ &= (\beta_0 + \beta_2) + \beta_1 educ - \beta_2 male + u \\ &= \alpha_0 + \alpha_1 educ + \alpha_2 male + u \end{aligned}$$

where $\alpha_0 = \beta_0 + \beta_2$, $\alpha_1 = \beta_1$, and $\alpha_2 = -\beta_2$

The constant differs, the sign (of the dummy) flips, but the conclusion is the same.

Categorical Regressors (two categories)

Data info:

The Current Population Survey (*cps09mar*) contains information on employment, earnings, educational attainment, income etc. for 57.000 U.S. households (March 2009).

Variables:

earnings	total annual earnings
education	years of education (based on highest degree)
female	1 if female, 0 otherwise
male	1 if male, 0 otherwise

Data can be downloaded here: `https://www.ssc.wisc.edu/~bhansen/econometrics/`.

Categorical Regressors (two categories)

$$lm(formula = earnings \sim female + education, data = cps09mar)$$

```
MODEL INFO:
Observations: 50742
Dependent Variable: earnings
Type: OLS linear regression

MODEL FIT:
F(2,50739) = 5810.40, p = 0.00
R² = 0.19
Adj. R² = 0.19

Standard errors: OLS
-----------------------------------------------------
                         Est.      S.E.   t val.     p
----------------- ----------- --------- -------- ------
(Intercept)         -40230.70   1089.59   -36.92   0.00
female              -20770.84    423.35   -49.06   0.00
education             7480.62     76.27    98.08   0.00
-----------------------------------------------------
```

$\widehat{\beta}_2 = \ldots$

Categorical Regressors (two categories)

$$lm(formula = earnings \sim male + education, data = cps09mar)$$

```
MODEL INFO:
Observations: 50742
Dependent Variable: earnings
Type: OLS linear regression

MODEL FIT:
F(2,50739) = 5810.40, p = 0.00
R² = 0.19
Adj. R² = 0.19

Standard errors: OLS
------------------------------------------------------
                         Est.       S.E.   t val.     p
----------------- ----------- ---------- -------- ------
(Intercept)         -61001.55    1119.75   -54.48   0.00
male                 20770.84     423.35    49.06   0.00
education             7480.62      76.27    98.08   0.00
------------------------------------------------------
```

$\widehat{\alpha}_2 = \ldots$

Interpretation of Parameters

Categorical Regressors (multiple categories)

Categorical Regressors (multiple categories)

When a regressor x can take on **multiple discrete** values (e.g. occupation, industry, region, educational degrees), we get as marginal effects several group differences.

Let D represent a categorical variable assuming J **distinct** values.

We do not want to estimate

$$y = \beta_0 + \beta_1 D + u$$

because β_1 has no precise interpretation. All group differences would be set to β_1 . Simply recoding categories would give different estimates.

Solution: create $J-1$ dummy variables (omit one, J_{ref}, which becomes the reference category)

$$d_{ij} = \begin{cases} 1 & \text{if } D_i = j \\ 0 & \text{otherwise} \end{cases}$$

Categorical Regressors (multiple categories)

Example: highest education obtained; $degree \in \{$ BSc, MSc, PhD$\}$

$$degree_1 = \begin{cases} 1 & \text{if } degree = \text{BSc} \\ 0 & \text{otherwise} \end{cases}$$

$$degree_2 = \begin{cases} 1 & \text{if } degree = \text{MSc} \\ 0 & \text{otherwise} \end{cases}$$

$$degree_3 = \begin{cases} 1 & \text{if } degree = PhD \\ 0 & \text{otherwise} \end{cases}$$

Note that degrees are defined as "highest degree obtained", and so they are **mutually exclusive**.

Categorical Regressors (multiple categories)

i	D	d_1	d_2	d_3
1	category A	1	0	0
2	category B	0	1	0
3	category A	1	0	0
4	category C	0	0	1
5	category C	0	0	1
6	category B	0	1	0
7	category B	0	1	0
8	category A	1	0	0
9	category A	1	0	0
10	category C	0	0	1

Categorical Regressors (multiple categories)

We can then estimate the following model:

$$wage = \beta_0 + \beta_1 degree_2 + \beta_2 degree_3 + u$$

where "BSc" is the **baseline/reference category**.

The intercept β_0 is the mean outcome of those in the reference group (i.e. BSc)

β_1 is the difference in wages between Master ($degree = 2$) and Bachelor degree (baseline):

$$\beta_1 = \mathrm{E}(wage|MSc) - \mathrm{E}(wage|BSc)$$

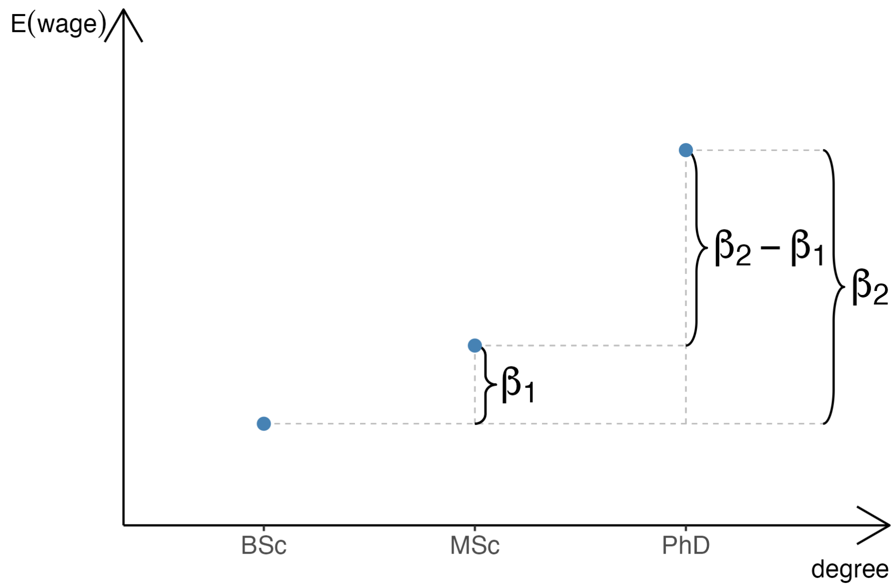
β_2 is the difference in wages between PhD ($degree = 3$) and Bachelor degree (baseline):

$$\beta_2 = \mathrm{E}(wage|PhD) - \mathrm{E}(wage|BSc)$$

Categorical Regressors (multiple categories)

If we want to learn the difference between two groups of a categorical regressor, we only need to contrast their coefficients. We get the difference b/w PhD ($degree = 3$) and Master ($degree = 2$) by comparing β_1 and β_2, that is

$$\begin{aligned}\beta_2 - \beta_1 &= \mathrm{E}(wage|PhD) - \mathrm{E}(wage|BSc) - [\mathrm{E}(wage|MSc) - \mathrm{E}(wage|BSc)] \\ &= \mathrm{E}(wage|PhD) - \mathrm{E}(wage|MSc)\end{aligned}$$



Categorical Regressors (multiple categories)

We can estimate

$$y = \beta_0 + \sum_{j \neq J_{ref}} \beta_j d_j + u$$

The interpretation of the regression coefficients is

$$\beta_j = \mathrm{E}(y|d=j) - \mathrm{E}(y|d=J_{ref})$$

and therefore **relative to the reference category** (J_{ref}).

The intercept β_0 is here the average outcome of those in the reference group.

If we want to compare group k with group m, we need to compare their coefficients:

$$\beta_k - \beta_m = \mathrm{E}(y|d=k) - \mathrm{E}(y|d=m)$$

Hypothesis Testing (in large samples)

Hypothesis Testing

We consider the following (bivariate) regression:

$$y_i = \alpha + \beta x_i + u_i \,.$$

Suppose we want to test $H_0 : \beta = \beta_{H_0}$ vs $H_1 : \beta \neq \beta_{H_0}$ for an hypothesized value (β_{H_0}).

For example, we want to test whether the private returns to education are equal to 1000 € ($\beta_{H_0} = 1000$) or whether the private returns to education are equal to 0 € ($\beta_{H_0} = 0$), that is education, x, has no effect at all on wages, y.

Keep in mind:

- β is the true (unknown) value in the population
- $\widehat{\beta}$ is the OLS estimate in our sample
- β_{H_0} is a (known) hypothesized value (often this value is simply 0)

Hypothesis Testing

We are looking for sufficient evidence against the null hypothesis.

Idea: Construct a test with a known distribution under the null hypothesis (H_0). Reject the null hypothesis if the value of the test statistic is too large.

Ingredients: $\widehat{\beta} \xrightarrow{d} N\left(\beta, \mathrm{Var}(\widehat{\beta})\right)$

The t-statistic is simply a transformation of $\widehat{\beta}$ such that it would be **standard** normally distributed under the null hypothesis (i.e. if $\beta = \beta_{H_0}$ is true)

$$t = \frac{\widehat{\beta} - \beta_{H_0}}{\sqrt{\mathrm{Var}_{\widehat{\beta}}}} \left[= \frac{\widehat{\beta} - \beta_{H_0}}{\mathrm{SE}}\right]$$

A large (in absolute values) t-statistic is evidence against the null hypothesis.

Hypothesis Testing

The t-statistic behaves asymptotically like a **standard** normal random variable if the null hypothesis is true (i.e. under H_0).

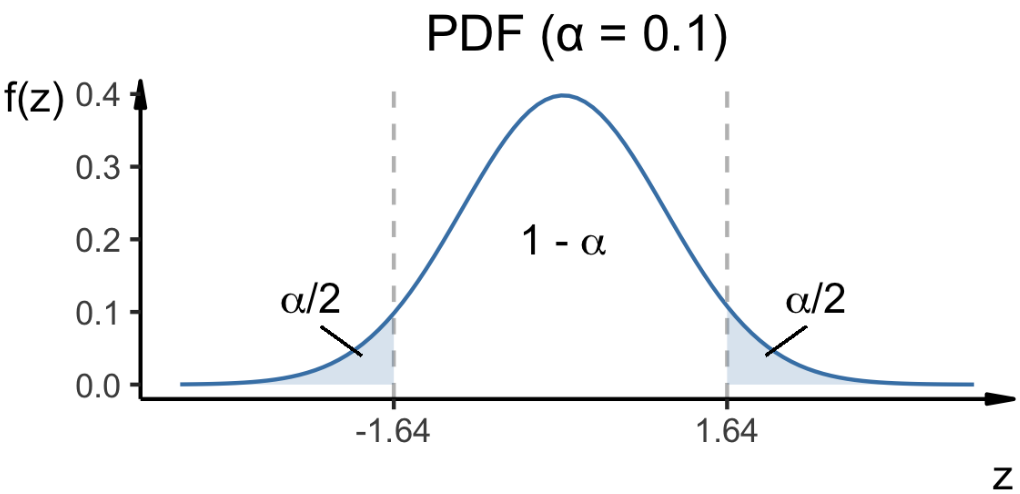
This is a convenient result because we then need to remember only the critical values of a **standard** normal distribution:

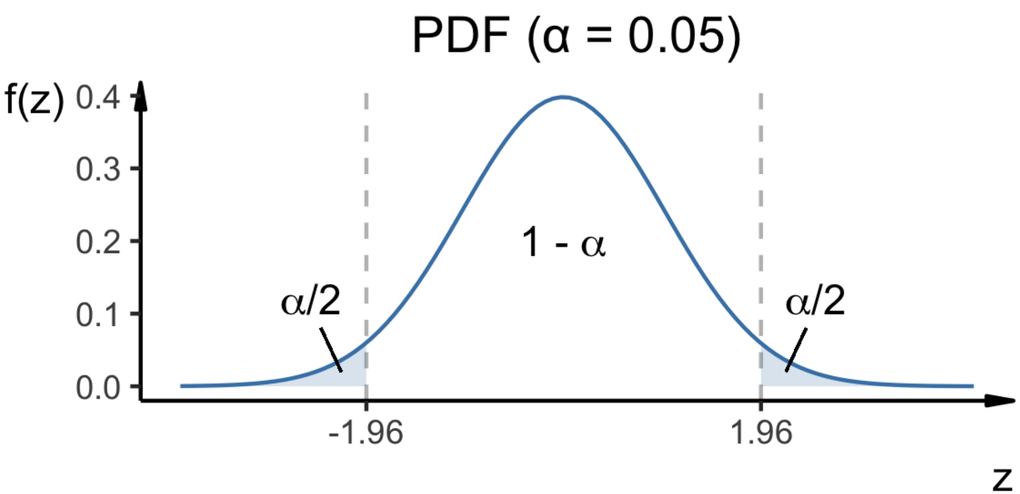
If $|t| > 2.58$ (or $|t| > 1.96$ or $|t| > 1.64$), we can reject the null hypothesis H_0 at the 1% (or 5% or 10%) level.

The term **significance level** refers to a pre-chosen probability (e.g., 5% significance level or $\alpha = 0.05$).

Many authors refer to *statistically significant* if $|t| > 1.96$ and *statistically highly significant* if $|t| > 2.58$.

Hypothesis Testing





Hypothesis Testing

A ***p*-value**, or probability value, is the probability of finding the observed (or a more extreme) t-statistic when the null hypothesis would indeed be true.

You can interpret the p-value as an **empirical significance level**: it denotes the significance level at which we can just reject the null hypothesis.

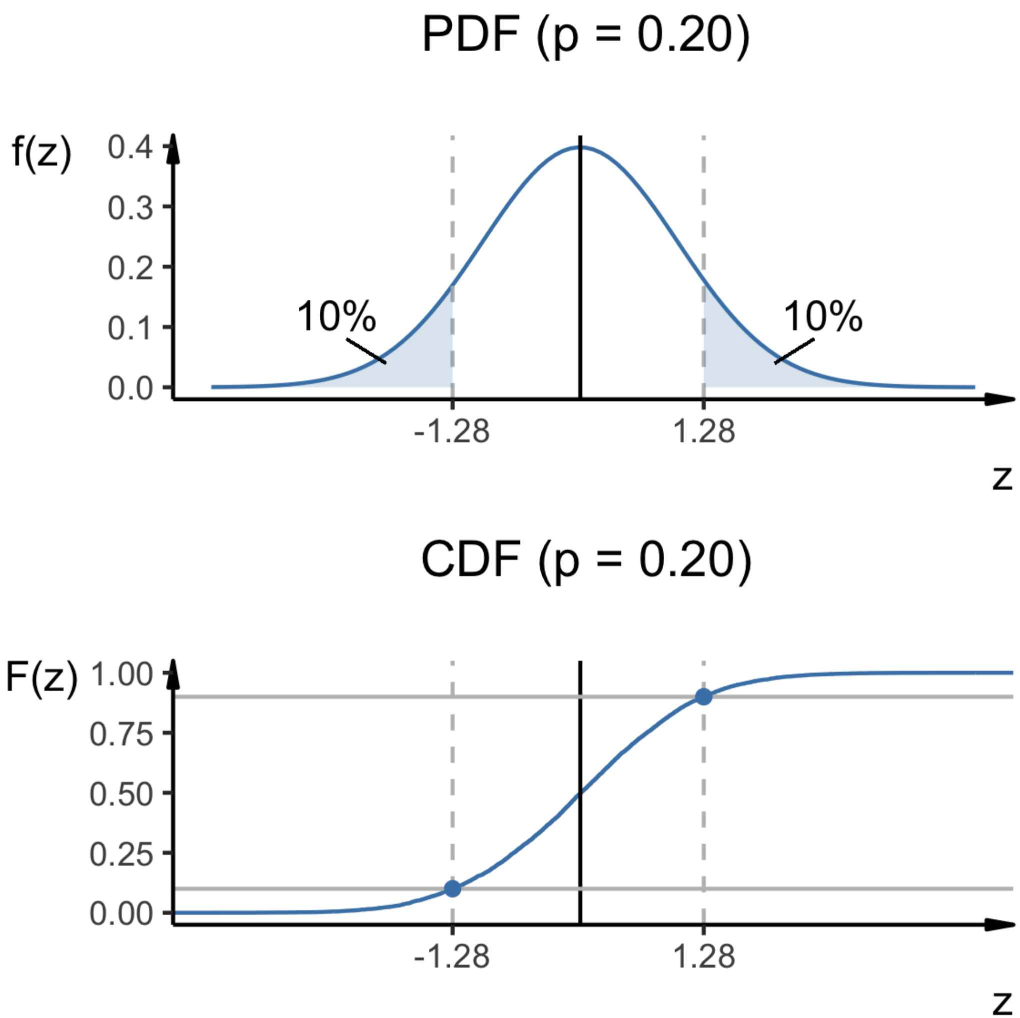
To calculate a p-value on your own, you need:

- the absolute value of the t-statistic: $|t|$
- the standard normal cdf: $\Phi(\cdot)$

$$p = 2\left(1 - \Phi\left(|t|\right)\right)$$

A small p-value is evidence against the null.

Hypothesis Testing



Hypothesis Testing

A **confidence interval** (*CI*) is the point estimate plus and minus a measure for the (likely) variation in that estimate.

To calculate a *CI* on your own, you need:

- the point estimate: $\widehat{\beta}$
- the standard error of this point estimate: SE
- the critical value of the intended significance level: $c = \Phi^{-1}(1-\alpha/2)$

$$CI = \widehat{\beta} \pm c \cdot \mathrm{SE}$$

Remember the critical values: 2.58 (1% level); 1.96 (5% level); 1.64 (10% level)

If the hypothesized value β_{H_0} is outside the confidence interval, we can interpret this as evidence against the null.

Hypothesis Testing

Different ways to convey similar information

Different ways to convey similar information

1. Asterisks to indicate the level of significance, e.g., *** significant at 1% level; ** significant at 5% level; * significant at 10% level
2. Standard errors in parentheses
3. Confidence intervals
4. p-values in parentheses
5. t-statistics in parentheses

Different ways to convey similar information

Which way should I pick? Depends on your audience (and changes over time).

More and more journals rule out fixed thresholds of significance (e.g., see SMJ Author Guidelines in the subsection “Reporting Results of Statistical Analyses”)

Guidelines Regarding Empirical Research in *SMJ*

Reporting Results of Statistical Analyses

SMJ no longer accepts papers for publication that report or refer to cutoff levels of statistical significance (p-values). In statistical studies, authors should report either standard errors or exact p-values (without asterisks) or both, and should interpret these values appropriately in the text. Rather than referring to specific cutoff points, the discussion could report confidence intervals, explain the standard errors and/or the probability of observing the results in the particular sample, and assess the implications for the research questions or hypotheses tested.

Different ways to convey similar information

More and more journals rule out fixed thresholds of significance (e.g., see AER Guidelines in the subsection “Style Guide”)

aeaweb.org

Tables

-- Columns must be in vertical (or portrait) orientation.
-- Tables must be no more than 9 columns wide including row headings.
-- Number your tables consecutively with Arabic numerals.
-- Use only horizontal lines and additional blank space to show space distinction.
-- Do not use shading.
-- Do not abbreviate in column headings.
-- To denote sections of a table, use panel A, panel B, etc.
-- Place a zero in front of the decimal point in all decimal fractions (e.g., 0.357, not .357).
-- For footnotes pertaining to specific table entries, footnote keys should be lowercase letters (a, b, c, etc.). Do not use asterisks to denote significance of estimation results. Report the standard errors in parentheses.
-- Place source lines after the footnotes. If the citation is a complete sentence, place a period at the end of the source line.
-- Include full citations of sources in the references.

Different ways to convey similar information

Potential pitfalls: *p*-hacking and publication bias.

Addressing these problems may be more important than the way how we present our results.

Lengthy discussions:
Statistical Significance, p-Values, and the Reporting of Uncertainty (Imbens, JEP 2021)
or here:
In Praise of Confidence Intervals (Romer, AER P&P 2020)

Additional discussions (not that important from an applied point of view):
in Statistics
in Operations Research

Sidenote 1: Significance vs Relevance

Statistical significance

- Measured by t-tests, p-values, confidence intervals, standard errors
- Addresses concerns about precision of estimates

Economic relevance

- Measured by coefficient magnitudes (i.e. the absolute value of $\widehat{\beta}$)
- Addresses concerns about policy relevance

Potential pitfalls

Overemphasize **economically irrelevant** variables that are statistically significant (if your sample size is large, you can easily find tiny effects).

Discard **economically relevant** variables that are statistically insignificant (if your sample size is small, you can easily miss important variables).

Sidenote 2: Multiple testing

Testing a **single** parameter of a *multivariate* regression works in the same way as described on the previous slides.

If we want to test an hypothesis that involves **multiple** parameters of a multivariate regression, we need more sophisticated methods.

If you want to know more, join my lecture “Microeconometric Methods for Big Data”.

Recommended reading

For next week please read chapter:

8.2 Panel Data: Estimation

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

Illustration asterisks

Go back

Dickstein et al. (AER P&P, 2015): Market Size and Health Insurance Premiums

AEA PAPERS AND PROCEEDINGS

TABLE 2—REGION-LEVEL ANALYSES

	Number of insurers		Premium	
	(1)	(2)	(3)	(4)
log population	0.652***	0.645***	−108.9***	−137.2**
	(0.187)	(0.221)	(24.09)	(61.65)
log land area (100s of sq. miles)		−0.212*		203.0***
		(0.129)		(68.83)
Fraction population urban		−3.015***		1095.8**
		(1.030)		(510.1)
Fraction pop urban squared		3.094***		−1,047.1*
		(1.095)		(597.4)
Observations	398	398	398	398
R^2	0.619	0.659	0.621	0.656

Notes: Specifications 2 and 4 include as controls: median income, share of households with income 25K–100K, Medicare Geographic Adjustment Factor, share of adult population in 40–64 age bin, percent of employed population working in establishments with fewer than 10 employees, and number of short-term general hospitals. Price regressions include as an additional control the deductible of the second lowest priced silver plan. All regressions include state fixed effects. Standard errors (clustered at the state level) in parentheses.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Illustration standard errors

Go back

Cubas et al. (AER P&P, 2021): Work-Care Balance and the Gender Wage Gap

TABLE 1—GENDER GAP IN WORK AND HOUSEHOLD CARE

	Day dummies (1)	+ Dem. controls (2)	+ Usual hours (3)	+<50 hours (4)
Panel A. Working hours				
Female gap in work hours	−0.746	−0.742	−0.317	−0.279
	(0.0424)	(0.0429)	(0.0415)	(0.0427)
Observations	16,073	16,073	16,073	16,073
Average hours, men	8.691			
Average hours, women	7.943			
Average hours, total	8.441			
Panel B. Household care				
Female gap in household care hours	0.388	0.387	0.310	0.272
	(0.0196)	(0.0196)	(0.0190)	(0.0194)
Observations	16073	16073	16073	15483
Average hours, men	0.726			
Average hours, women	1.114			
Average hours, total	0.856			
Panel C. Household care in prime time				
Incidence of household care 8 to 5	0.150	0.150	0.135	0.117
	(0.00755)	(0.00755)	(0.00757)	(0.00787)
Observations	14,386	14,386	14,386	13896
Average, men	0.197			
Average, women	0.347			
Average, total	0.246			

Notes: The table is based on ATUS respondents who are 18–65 years old, who report usual weekly hours >=35 in the CPS, who are married with at least one child in the household, and whose diary day is a weekday. "Work" corresponds to hours spent on "work and work-related activities," which does not include travel or commuting time. "Household care" corresponds to hours spent on "caring for and helping household members," which does not include housework. "Incidence of household care 8–5" is equal to one if the respondent reported nonzero household care between 8 AM and 5 PM. For work and household care hours, we restrict the sample to those who report nonzero time spent on work-related activities. For the incidence measure, we restrict the sample to those who report nonzero time spent on work-related activities at the work site. Each column reports the coefficient on the "female" dummy with various controls. Column 1 includes day and year fixed effects. Column 2 includes age, education category, and race fixed effects. Column 3 adds usual weekly hours reported in the CPS. Column 4 only includes workers who reported usual weekly hours of less than 50.

Source: Data are from the 2003–2018 ATUS

Illustration confidence intervals

Go back

TABLE 2—PATENTS AND FOLLOW-ON INNOVATION ON HUMAN GENES CLAIMED IN ACCEPTED/REJECTED PATENT APPLICATIONS: REGRESSION ESTIMATES

	log of follow-on innovation in 2011–2012	Any follow-on innovation in 2011–2012
	(1)	(2)
Panel A. Scientific publications		
Patent granted	0.0019	−0.0014
	(0.0060)	(0.0054)
Mean of dependent variable	0.1104	0.1094
Observations	15,524	15,524
Panel B. Clinical trials		
Patent granted	0.0006	−0.0015
	(0.0080)	(0.0043)
Mean of dependent variable	0.1038	0.0659
Observations	15,524	15,524
Panel C. Diagnostic test		
Patent granted	–	−0.0092
	–	(0.0056)
Mean of dependent variable	–	0.1199
Observations	–	15,524

Notes: This table estimates differences in follow-on innovation on genes claimed in at least one granted patent relative to genes claimed in at least one patent application but never in a granted patent. The sample for these regressions is constructed from gene-level data, and includes genes claimed in at least one patent application in our USPTO human gene patent application sample ($N = 15{,}524$). Each coefficient is from a separate regression. Estimates are from ordinary-least-squares models. Heteroskedasticity robust standard errors.

more econometrically appropriate than modeling the outcome in levels. We focus on the log of follow-on innovation and (separately) an indicator for any follow-on innovation.

Given the absence of strong visual evidence for a difference in follow-on innovation across patented and non-patented genes, our focus here is on what magnitudes of effects can be ruled out by our confidence intervals. Across these specifications, our 95 percent confidence intervals tend to reject declines or increases in follow-on innovation on the order of more than 5–15 percent. For brevity, we focus on interpreting the log coefficients. For our measures of follow-on scientific research (publications; panel A of Table 2) and commercialization (clinical trials; panel B of Table 2), the 95 percent confidence intervals can reject declines or increases of more than 2 percent. For our measure of diagnostic test availability (only measured as a

Illustration p-value

Go back

Müller and Schwieren (J. of Business Economics, 2020)

Table 5 Tobit regression on y, the amount returned by Player 2 for those players who received a positive amount from Player 1

Variable	Coeff.	SE	*p*-value
x	1.645***	0.269	0.000
Neuroticism	– 0.004	0.043	0.933
Extraversion	0.005	0.046	0.915
Opennness	0.078	0.065	0.236
Agreeableness	– 0.027	0.049	0.580
Conscientiousness	0.080	0.049	0.220
Female	– 0.061	2.133	0.904
Age	0.039	0.369	0.915
n	47		
pseudo R^2	0.103		

This table shows the coefficients (2nd column), standard errors (3rd column) and *p*-values (4th column) from a tobit regression. The dependent variable is the amount returned by Player 2

*, **, *** Indicate significance at the 10%, 5% and 1% level respectively

Applied Econometrics

Illustration t-statistics

Go back

Cawley (JHR, 2004): The Impact of Obesity on Wages

Table 2
Coefficients and t-Statistics from Log Wage Regressions for Males

	White Males			Black Males			Hispanic Males		
	OLS	OLS with Lag Weight	Fixed Effects	OLS	OLS with Lag Weight	Fixed Effects	OLS	OLS with Lag Weight	Fixed Effects
Column Number	1	2	3	4	5	6	7	8	9
BMI	−0.001	−0.003	−0.0001	0.004	0.005	0.003	−0.007	−0.009	−0.002
	(−0.83)	(−1.54)	(−0.21)	(2.19)	(2.04)	(1.96)	(−3.12)	(−3.15)	(−1.03)
Weight in pounds	−0.0002	0.0005	0.0001	0.0006	0.0007	0.0005	−0.0011	−0.001	−0.0003
	(−1.01)	(−1.68)	(0.07)	(2.21)	(2.06)	(1.96)	(−3.47)	(−3.18)	(−0.90)
Underweight	−0.14	0.005	−0.035	−0.099	−0.046	0.013	0.029	0.107	−0.005
	(−3.05)	(0.07)	(−1.12)	(−2.75)	(−0.70)	(0.28)	(0.44)	(1.13)	(−0.12)
Overweight	0.039	0.016	0.022	0.031	0.019	0.014	−0.025	−0.021	0.018
	(3.04)	(1.05)	(2.63)	(1.87)	(0.94)	(1.14)	(−1.13)	(−0.82)	(1.15)
Obese	−0.033	−0.075	0.013	0.043	0.042	0.031	−0.066	−0.100	0.023
	(−1.73)	(−3.05)	(0.89)	(1.80)	(1.29)	(1.64)	(−2.21)	(−2.41)	(0.91)
Number of observations	29,410	12,410	29,410	13,414	6,128	13,414	9,070	4,079	9,070

Notes:
1) Data: NLSY males.
2) One of three measures of weight is used: BMI, weight in pounds (controlling for height in inches) or the three indicator variables for clinical weight classification: underweight, overweight, and obese (where healthy weight is the excluded category).
3) For BMI and weight in pounds, coefficients and *t* statistics are listed. For indicators of clinical weight calssification, the percent change in log wages associated with a change in the indicator variable from 0 to 1 and *t* statistics are listed.
4) Other regressors include: number of children ever born, age of youngest child, general intelligence, highest grade completed, mother's highest grade completed, father's highest grade completed, years of actual work experience, job tenure, age, year, and indicator variables for marital status, county unemployment rate, current school enrollment, part-time job, white collar job, and region of residence.

Cawley

Applied Econometrics

Contact

Helmut Farbmacher
office.econometrics@mgt.tum.de