Module 3: Threats to Causal Identification

Econometrics II

Sannah Tijani (stijani@wu.ac.at)
Department of Economics, WU Vienna

Max Heinze (mheinze@wu.ac.at)
Department of Economics, WU Vienna

November 6, 2025

Validity

External Validity

Internal Validity

Confounders and OVB

Validity

In order to assess the quality of causal inferences, it helps to think of the validity of a statistical analysis. Different concepts of validity include the following:

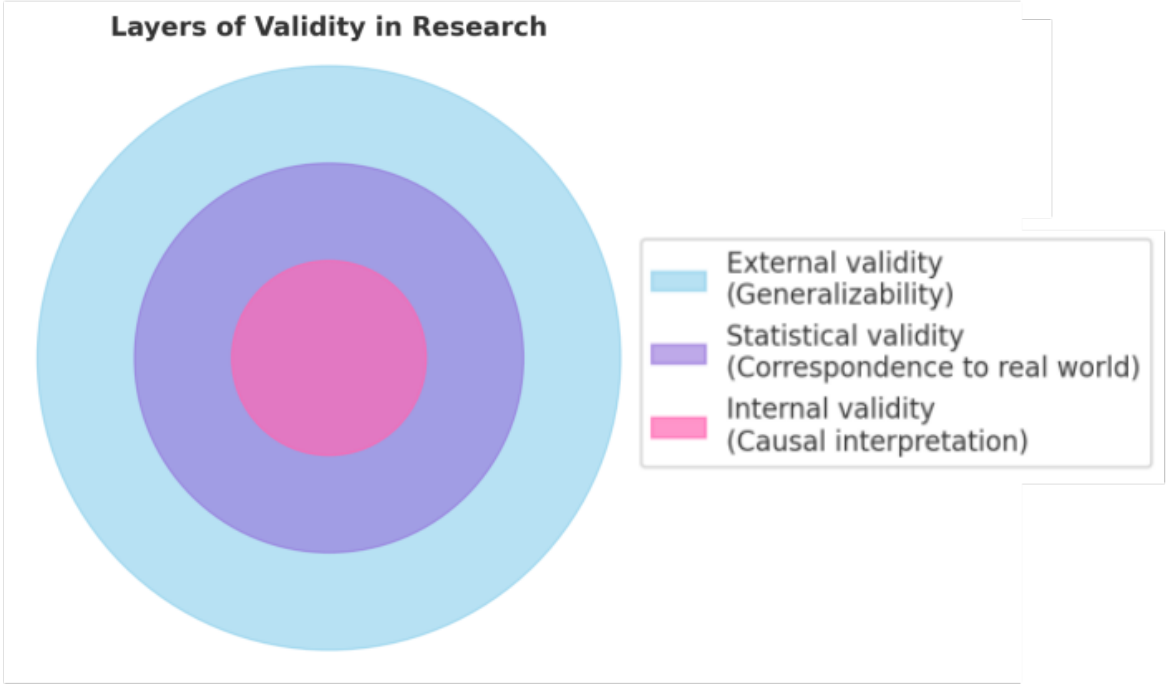
- **Construct validity**: Refers to whether the analysis, test, or measurement actually captures the theoretical concept (or "construct") it claims to measure. Without construct validity, your results don't really connect back to the **theory** you want to test.
- **Content validity**: Refers to whether the measure covers all the important aspects of the real-world concept you are trying to capture. Ensures your analysis is not just theoretically sound but also practically **comprehensive—covering the domain of the phenomenon in the real world**.
- **Predictive validity**: Refers to how well your measure or model can predict future outcomes related to the concept. Strong predictive validity means your analysis isn't just descriptive—it can be used for **decision-making** and **policy design**.

Internal vs External Validity

External validity: determines whether an insight can be generalized.

Statistical validity: the validity of an analysis can be thought of as the extent to which the analysis corresponds to the relevant aspects of the real world.

Internal validity: qualify the causal interpretation of an inference.



Validity

External Validity

Internal Validity

Confounders and OVB

Selection Bias

External Validity

Statistical validity is the validity of an analysis **outside its own context**, telling us whether **findings can be generalized** across situations, people, time, regions etc...

- Analyses may yield insights that are highly specific to their circumstances
- There can be **trade-offs** between external and other types of validity: A perfect analysis may control important factors tightly and a poor analysis limits what we learn at all

Example

Imagine we study cooperation using a lab experiment with students playing a public goods game. We tightly control the environment: same stakes, same instructions, no distractions. Result: we can confidently say “in this precise setting, people contribute 50% on average.”

What we can we say about external validity?

Threats to External Validity

- **Population**: The individual selected in your sample should be representative of the population.

E.g. Lab experiment on dictator game with WEIRD students

- **Sample Size**: The sample you are using may just be too small to observe an effect

E.g. Meta-Analysis by McKenzie on the effect of training on management practices

- **Situations**: Your analysis may be specific to a point in time and/or a specific location

E.g. Card & Krueger (1994) Minimum Wage Study: study comparing fast-food restaurants in New Jersey vs. Pennsylvania after New Jersey raised its minimum wage. Contrary to textbook predictions, employment did not fall in New Jersey relative to Pennsylvania.

Dealing with External Validity

- **Population and Sample Size**: We can reprocess/weight the data we are using

Example

Sample: For decades, drug trials were conducted only on men (often young, white).

Problem: Findings about dosage, side effects, and efficacy were applied to women and older adults, despite metabolic and hormonal differences.

Solutions?

Dealing with External Validity (2)

- **Situations**: Issues with external validity ultimately stem from the interactions between (uncountably many) factors that may (or may not) be relevant.

Example

The effects of studying on academic performance may also be (slightly) affected by:whether you eat breakfast, the type of breakfast, your diet, your social life, the incidence of an armed conflict abroad, a game being published,...

Which one would be relevant?

Generalisable Facts

- There are many generalisable insights that we can learn, and that are worth learning.
- A good test of external validity is the replication of an analysis in **different settings** and/or with different methods.

Case	Original Insight	What Replication Found	What It Shows About External Validity
Sampson & Cohen (1988)	Proactive policing reduces robbery rates	Similar negative correlation across multiple U.S. cities, with further nuance	Suggests original insight holds across many U.S. cities (some generalizability)
Minneapolis Domestic Violence (1981)	Arrest reduces repeat domestic violence	null, opposite or smaller effect sizes, vary by location, method, measurement	Demonstrates strong internal validity in one context does not guarantee generalizability across locations, times, or institutions

Validity

External Validity

Internal Validity

Confounders and OVB

Selection Bias

Measurement Error

Internal Validity

Internal validity is the validity of an analysis within its own context. It is the extent to which the analysis allows for **causal inference**.

- Empirical evidence may support various different interpretations.
- We want to be able to credibly eliminate non-causal interpretations.
- **The Principle of Parsimony** (Occam's razor): There may be incomprehensibly many alternatives for each explanation. The idea is to give preference to the simplest explanation (that cannot be refuted), i.e the one with the fewest parameters and/or assumptions.

The Gauss-Markov Theorem

Ordinary least-squares (OLS) estimation yields the best, linear, unbiased estimator (BLUE) under the following conditions.

- The data stems from a **random sample** of the population.
- **Exogeneity** (zero conditional mean of errors), i.e. $\mathbb{E}[\boldsymbol{u} \mid X] = \mathbb{E}[\boldsymbol{u}] = 0$.
- The model is **linear in parameters**, e.g. $f(X) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$.
- No **perfect collinearity**, i.e. $\boldsymbol{X}$ has full rank and we can compute $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
- **Homoskedasticity** and no **serial correlation**, i.e. $\mathrm{Var}(\boldsymbol{u} \mid X) = \sigma^2 \boldsymbol{I}$.

The first four assumptions imply that $\hat{\boldsymbol{\beta}}$ is **unbiased**, the last one implies that $\hat{\sigma}^2$ is **unbiased** and, hence, that the estimate is **efficient**.

Exogeneity

- Exogeneity is a weaker form of **ignorability**
- The exogeneity assumption $\mathbb{E}[\boldsymbol{u} \mid X] = 0$ is sometimes substituted with **weak exogeneity** $\mathrm{Cov}(X, \boldsymbol{u}) = 0$
- This guarantees consistency, but not unbiasedness of the estimator.
- A failure of exogeneity is called **endogeneity**
- It causes bias and inconsistency by confounding the effects of our regressors **X** and the true errors **u** on **y**
- **Parameter bias and consistency**: An estimate $\hat{\theta}$ is **unbiased** if $\mathbb{E}[\hat{\theta}] = \theta$. It is **consistent** if it converges in probability to the true parameter with increasing data:

$$\mathrm{plim}_{N \to \infty} |\hat{\theta} - \theta| > \varepsilon = 0$$

The Effect of Endogeneity

- Consider the effect of adjusting $\boldsymbol{x}_1$ to $\boldsymbol{x}^*$:

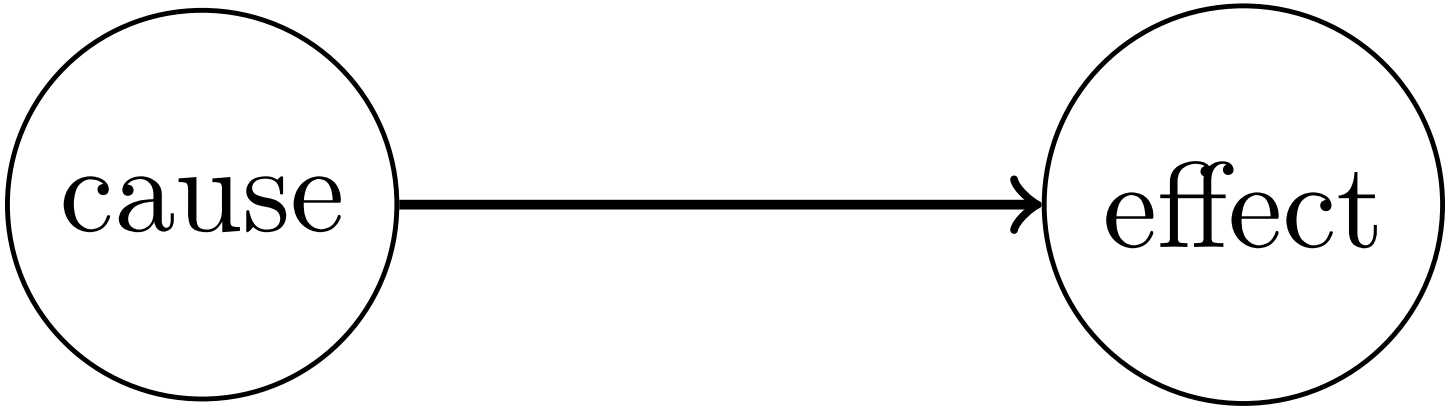
$$\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}^*] - \mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}] = \beta_1(\boldsymbol{x}_1^* - \boldsymbol{x}_1) + \big(\mathbb{E}[\boldsymbol{u} \mid \boldsymbol{X}^*] - \mathbb{E}[\boldsymbol{u} \mid \boldsymbol{X}]\big).$$

- Under exogeneity, we get the correct effect since the second term is zero.
- However, if x_1 and e are correlated, we have $\mathbb{E}[\boldsymbol{u} \mid X] = \theta_1 x_1 + \theta_0 \quad \text{with } \theta_1 \neq 0$
  We cannot separate the effects of observed factors (β_1) and unobserved ones (θ_1) and estimate

$$\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}^*] - \mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}] = \beta_1 \left(\boldsymbol{x}_1^* - \boldsymbol{x}_1\right) + \theta_1 \left(\boldsymbol{x}_1^* - \boldsymbol{x}_1\right).$$

Threats to Internal Validity

- There are many **threats** to internal validity.
- It can help to think in terms of **frameworks** for causal inference.
- You can use either **directed acyclic graphs**
- And/or **potential outcomes** and **ignorability** of a treatment.
- There are many **common issues** that we'll cover in more details



Validity

External Validity

Internal Validity

Confounders and OVB

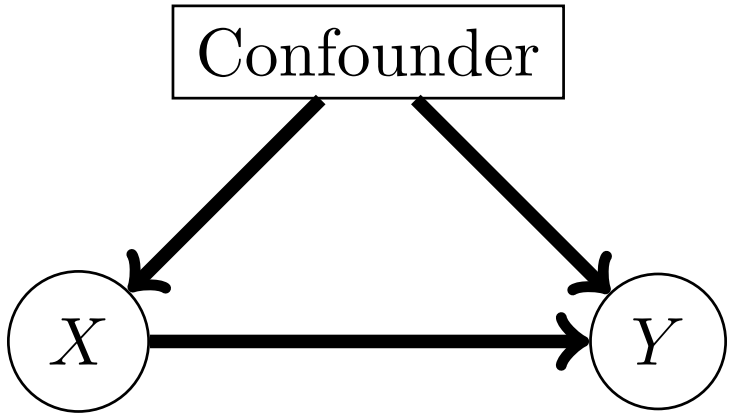
Selection Bias

Measurement Error

Simultaneity

Confounders

- A **confounder** is an additional variable that drives both the cause and effect
- If we don't account for that variable we can't provide a causal effect explanation for what we observe.



- Consider the following true model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

What are the implications of estimating $y = \beta_0 + \beta_1 x_1 + e$ instead ? (Econometrics 1)

Omitted Variable Bias

Bias from a confounder is also called **omitted variable bias** it occurs when:

- The omitted variable is correlated with the regressors ($\text{Cov}(x_1, x_2) \neq 0$)
- It is also a determinant of y ($\beta_2 \neq 0$)

From the previous slide, the bias is given by:$\mathbb{E}[\hat{\beta}_1] = \beta_1 + \frac{\text{Cov}(x_1,x_2)}{\text{Var}(x_1)}\beta_2$

Practice task

Imagine a simple OLS with 1 explanatory variable X on Y, what could be the issue in the following case:

(1) The effect of plasma donation centers on crime rates?

(2) The effect of social media on mental health?

Proxy Variable (1)

- Many (potentially) **omitted variables** cannot be observed.
- To solve this we might be able to use **proxy variables**
- Recall the true model: $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \boldsymbol{u}$
- We cannot observe $\boldsymbol{x}_2$, but we could control for a proxy, , that fulfills:

$$\boldsymbol{z} = \theta_0 + \theta_1 \boldsymbol{x}_2 + \boldsymbol{e}$$

- E.g if you are interested in understanding the effect of alcohol consumption on wage, you won't be able to observe self-control
- Self-control can have an effect on both wage and alcohol consumption
- You can proxy self-control using credit score, or tardiness at work

Proxy Variables (2)

To use a **proxy variable** to identify a causal effect, it must:

(1) Correlate with the **omitted variable**: $\theta_1 \neq 0$

(2) Not correlate with other **explanatory variables** $Cov(\boldsymbol{X}, \boldsymbol{e}) = 0$

(3) have no direct impact on the **dependent variable** $Cov(\boldsymbol{z}, \boldsymbol{u}) = 0$

Condition 1 calls for an edge from the proxy to the confounder, while conditions 2 and 3 imply a lack of other (relevant) edges.

We will revisit another type of proxy variables (Instrumental variables) later.

External Validity

Internal Validity

Confounders and OVB

Selection Bias

Measurement Error

Simultaneity

Outlook

Type of Selection Bias

- If the sample is not **random**, we may speak of **selection bias**
- It is the idea that some subjects are more or less likely to be selected for our sample than others, distorting statistical insight
- Selection bias is related to sample issues that may plague **external validity**, but also threatens in-sample inference.
- There are many **types of selection bias**

Examples:

- Subjects may drop out of the sample (or even the population) for many reasons.
- Subjects may self-select (i.e. volunteer) for certain treatments.
- Journals like to publish groundbreaking results (shocking and significant).
- We like to focus on evidence that makes sense to us and confirms our priors.

Internal Validity

Confounders and OVB

Selection Bias

Measurement Error

Simultaneity

Outlook

Practice

Data Issues

Data may be subject to **various issues**, due to errors in collection, which may affect our ability to analyse it.

- Can we use **survey data** of savings or income?
- Can we ignore **satellite images** with clouds when classifying forest?
- How do we **quantify** ability?
- How to **measure** gross domestic product?
- What can go wrong when **collecting** data? Typos, malice ...

Missing Data

Consider a true f describing a population of size N, but we only observe $M(< N)$. Can we learn something using our subset?

- We can if our M represents a **random subset of the population**, then the selection process is **ignorable**
- Otherwise there may be **selection bias**

We can differentiate between **selection bias**:

- **endogenous** sample selection, related to the dependent variable
- **exogenous** sample selection, related to the explanatory or third variables

We need to account for **endogenous** sample selection to guarantee **internal validity**; **exogenous** selection limits **external validity**.

Non-random missingness

If there seems to be a pattern to missingness we may have to account for it to **avoid bias** or to benefit from accounting for it.

- **Truncation**: If samples where a value exceeds some threshold are missing, it is truncated
- **Censoring**: when only parts of the sample are known, we speak of censoring if

(1) Values are too low/high for our instruments to measure

(2) We stop measuring at a predetermined time (or after a number of events)

(3) There are incentives for reporting certain values

Outliers and Influential Observations

Outliers are observations that are very different from the rest, and may stem from:

- an inappropriate model
- data errors
- heterogeneity in the sample
- random chance

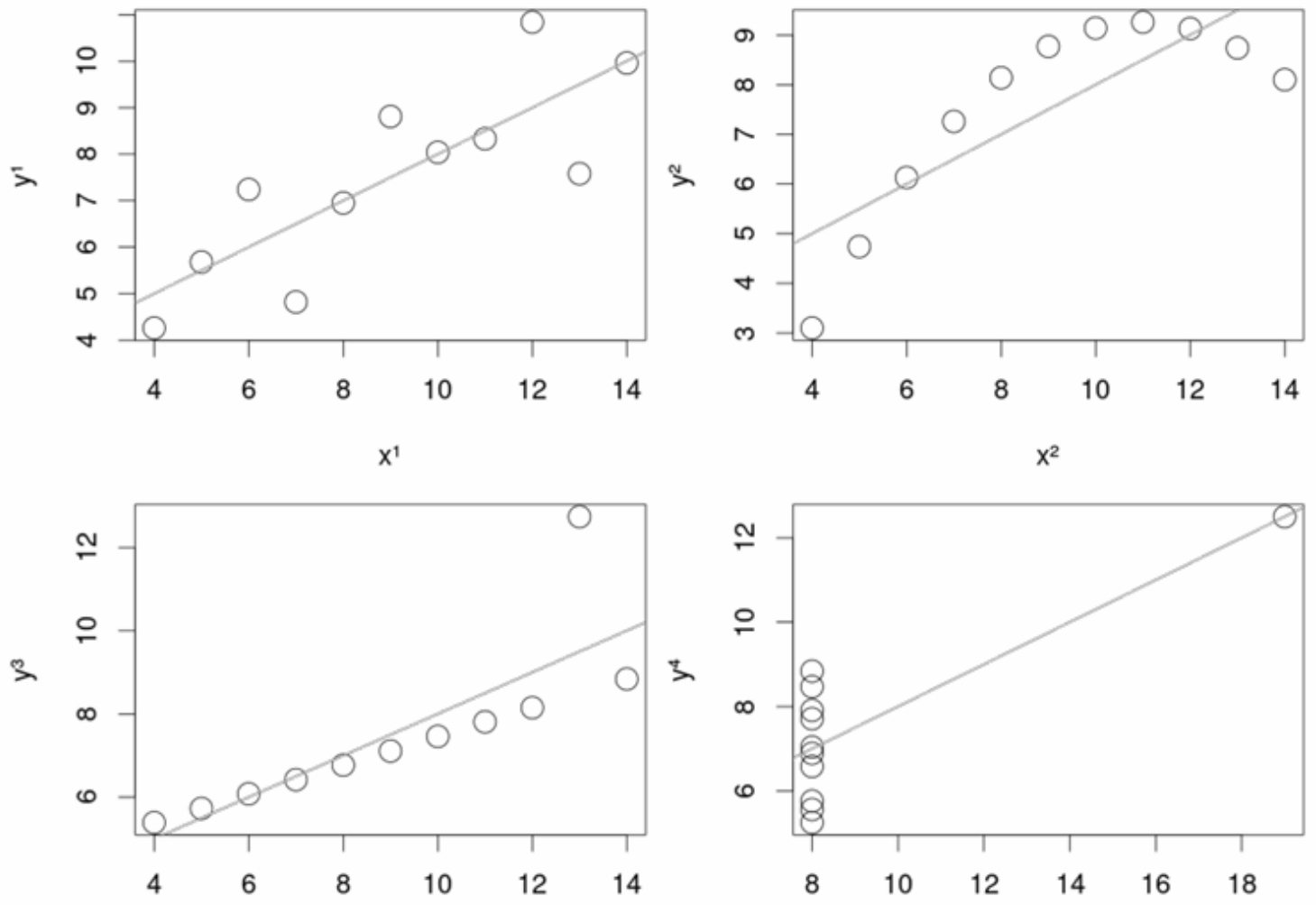
Outliers may have a large impact on estimates, i.e **high influence**.

For $\boldsymbol{\beta}_{OLS}$ an **influential observation**, i, has a combination of **high residual** $\boldsymbol{e}_i$ and a **high leverage** $\boldsymbol{h}_i = [\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']_{ii}$.

Its influence is given by: $\boldsymbol{\beta} - \boldsymbol{\beta}_{(i)} = \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'\boldsymbol{e}_i}{1-\boldsymbol{h}_i}$

Example

Anscombe's quartet — four different datasets with equal means, variance, and regression lines (Anscombe, 1973).



Dealing with Outliers

- When **exploring the data** (which you should always do, e.g via summary statistics or plots) or later when **evaluating the model** (e.g the residual values) you may discover an outlier early.
- It can be tempting to remove outliers from the analysis as supposed errors
- But, they may convey **the most interesting aspects** of the problem
- A good model allows us to learn, and accommodates exceptional cases
- There are many estimation methods that are **more robust to few observations**
- Example: M-, S-, or Least Absolute Deviation estimation, where we minimize absolute residuals $\boldsymbol{\beta}_{LAD} = \arg\min_{\beta} |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|$

Measurement Errors in the Dependent Variable

- Consider a true f with one explanatory variable $\boldsymbol{x}$, where the dependent variable $\boldsymbol{y}$ is **observed with additional errors ()**.
- We only observe $\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{u}$
- We estimate: $\boldsymbol{z} = \beta\boldsymbol{x} + \boldsymbol{e} + \boldsymbol{u}$
- What happens? It depends on whether $\boldsymbol{u}$ is random.

Errors in th Explanatory Variable

- Now, consider a true f with one explanatory variable $\boldsymbol{x}$ that is itself **observed with errors**
- We want $\boldsymbol{y} = \beta(\boldsymbol{z} - \boldsymbol{u}) + \boldsymbol{e}$, but only observe $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{u}$ and estimate

$$\boldsymbol{y} = \beta(\boldsymbol{z} - \boldsymbol{u}) + \boldsymbol{e}$$

- We can collect the errors in $\boldsymbol{a} = \boldsymbol{e} - \beta\boldsymbol{u}$ and rewrite as

$$\boldsymbol{y} = \beta\boldsymbol{z} + \boldsymbol{a}$$

- What happens? Our estimates will suffer from the **attenuation bias**

Attenuation Bias

Consider a weaker version of **ignorability** of the treatment — we want $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{e}) = 0$

With the measurement error in $\boldsymbol{x}$, we estimate $\boldsymbol{y} = \beta \boldsymbol{z} + \boldsymbol{a}$ and find that

$$\mathrm{Cov}(\boldsymbol{z}, \boldsymbol{a}) = \mathrm{Cov}(\boldsymbol{z}, \boldsymbol{e} - \beta \boldsymbol{u}) = \mathrm{Cov}(\boldsymbol{x} + \boldsymbol{u}, \boldsymbol{e} - \beta \boldsymbol{u}) \neq 0$$

We may assume:

(1) $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{e}) = 0$

(2) $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{u}) = 0$

(3) $\mathrm{Cov}(\boldsymbol{u}, \boldsymbol{e}) = 0$

But $\mathrm{Cov}(\boldsymbol{u}, -\beta \boldsymbol{u}) = -\beta \mathbb{E}[\boldsymbol{u}^2]$.

Attenuation Bias (2)

Here the bias is given by:

$$\mathbb{E}[\hat{\beta}] = \beta \frac{\sigma_{\boldsymbol{x}}^2}{\sigma_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{u}}^2}$$

The bias goes toward 0, and reduce the size of the estimates.

Attenuation bias proof

We can show the attenuation bias from estimating $\boldsymbol{y} = \beta\boldsymbol{x} + \boldsymbol{e}$ with $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{u}$

$$
\begin{aligned}
\boldsymbol{y} = \beta(\boldsymbol{z} - \boldsymbol{u}) + \boldsymbol{e} = \beta\boldsymbol{z} + \boldsymbol{e} - \beta\boldsymbol{u} = \beta\boldsymbol{z} + \tilde{\boldsymbol{e}},& \\
\hat{\beta} = (\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{y} = \beta + (\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\tilde{\boldsymbol{e}},& \\
\hat{\beta} = \beta + (\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{e} - (\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\beta\boldsymbol{u},& \\
\hat{\beta} = \beta + 0 - \beta(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{u},& \\
\hat{\beta} = \beta - \beta\big[(\boldsymbol{x} + \boldsymbol{u})'(\boldsymbol{x} + \boldsymbol{u})\big]^{-1}(\boldsymbol{x} + \boldsymbol{u})'\boldsymbol{u},& \\
\mathbb{E}[\hat{\beta}] = \beta\left(1 - \frac{\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{u}) + \mathbb{V}(\boldsymbol{u})}{\mathbb{V}(\boldsymbol{x}) + \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{u}) + \mathbb{V}(\boldsymbol{u})}\right),&
\end{aligned}
$$

Confounders and OVB

Selection Bias

Measurement Error

Simultaneity

Outlook

Practice

Simultaneity and Reverse Causality

The causal effect of interest, i.e, $X \to Y$, is not always as straightforward as we would like. Instead, we may encounter:

- **Reverse Causality**, where $Y \to X$
  - (1) Happiness and income: being happy increases productivity → income
  - (2) Health and exercise: poor health reduces ability to exercise → low exercise correlates with poor health
  - (3) Education and growth: richer countries invest more in education
- **Simultaneity**, where $Y \leftrightarrow X$
  - (1) Police and crime: crime increases → more police deployed → less crime → fewer police later.
  - (2) Price and quantity: market clears instantly — price and quantity determined together.

Reverse Causality

With **pure reverse causality**, the issue is determining the direction of causation.

Example: Can the civil tribunals' ineffectiveness in enforcing contracts explain the presence of the Italian Mafia today in Italy? (Braccioli, 2025)

It is not clear whether:

- The mafia is able to expand because the State Capacity is low

OR

- The State Capacity is low because of the mafia

Simultaneity

With **simulatneity** we need to disentangle the effects. Consider the following supply and demand functions, driven by the price p:

$$d = \beta^d p + u^d$$

$$s = \beta^s p + u^s$$

We can't observe supply and demand, but, we observed the **quantity sold q**, at equilibrium ($q = d = s$):

$$q = \beta^d p + e^d = \beta^s p + u^s$$

In this setting it is impossible to differentiate between the effect of price on supply or demand.

Parameter Identification

To see why the parameters β^d and β^s are **unidentified**, we can solve for p.

$$\beta^d p + u^d = \beta^s p + u^s,$$

$$\beta^d p = \beta^s p + u^s - u^d,$$

$$\beta^d p - \beta^s p = u^s - u^d,$$

$$p(\beta^d - \beta^s) = u^s - u^d,$$

$$p = \frac{u^s - u^d}{\beta^d - \beta^s}.$$

The effect of interest, p, is a **function of the errors**. Thus, we can't distinguish the effects. If we regress q on p, we can't tell whether the effect stems from the demand or supply.

Structural equations and simultaneity bias

Consider the following **structural equations**:

$$y = \beta_1 z + \beta_2 x_1 + u$$

$$z = \theta_1 y + \theta_2 x_2 + v$$

We can derive a **reduced form** equation by solving for z:

$$z = \gamma_1 x_1 + \gamma_2 x_2 + \varepsilon$$

Where:

$$\gamma_1 = \frac{\theta_1 \beta_2}{1 - \theta_1 \beta_2}; \gamma_2 = \frac{\theta_2}{1 - \theta_1 \beta_1}; \varepsilon = \frac{\theta_1 u + v}{1 - \theta_1 \beta_1}$$

Simultaneity Bias

The **reduced form** of our **structural parameters**, makes two issues clear:

- the reduced form parameters γ_1, γ_2 are non-linear functions of the structural parameters β, θ
- The structural parameters are not ignorable - z and u are correlated via y

In this reduced form, te error term is

$$\varepsilon = \frac{\theta_1 u + v}{1 - \theta_1 \beta_1}$$

where, the correlation between $\theta_1 u$ and the structural regressor y causes bias in

$$z = \theta_1 y + \theta_2 x_2 + v$$

Selection Bias

Measurement Error

Simultaneity

Outlook

Practice

Outlook

Now that we have seen what can go **wrong**, we will start seeing how we can make it **right**.

This includes: Instrumental variable models, simultaneous equations models, matching procedures, flexible estimation methods, and quasi-experiments. However, there are many threats to internal validity we did not mention but that are very relevant:

- Historical bias, due to events outside our control
- Experimenter bias, where the conductor affects the experiment
- Diffusion, spillover effects of treatment, where spillover effects between subjects complicate inference
- Reversion to the mean, where larger samples tend to be less extreme

Measurement Error

Simultaneity

Outlook

Practice

Exercise 1

You want to estimate whether time spent walking your dog improves mental health, and you have access to GPS tracker attached to the dog's collar.

- What regression would you run?
- What potential problem could arise?
- How would you solve it?

Exercise 2

What are the potential issue with the following questions that could arise:

- Does having health insurance improve health?
- Does living near a high-performing school increase housing prices?
- Does foreign aid promote economic growth?

References

Braccioli, F. (2025). *The institutional role of the italian mafia: Enforcing contracts when the state does not*. https://federicabraccioli.github.io/files/Braccioli_MafiaInstitution_JMP.pdf

Cunningham, S. (2021). *Causal inference*. Yale University Press. https://doi.org/10.12987/9780300255881