Econometrics 2

Nikolas Kuschnig (nkuschnig@wu.ac.at)

Winter Semester

Vienna University for Economics and Business

Department of Economics

People

- I'm Nikolas and I'm a PhD student at WU.
- My interest in econometric methods is twofold — I
    1. apply them to *environmental* and *development* issues,
    2. develop them to help *learn more* from our data and environment.
- E.g., I study *deforestation* and model the *spillover effects* behind it.

If you have any questions you can send me a mail or ask after class.

People

- I'm Nikolas and I'm a PhD student at WU.
- My interest in econometric methods is twofold – I
    1. apply them to *environmental* and *development* issues,
    2. develop them to help *learn more* from our data and environment.
- E.g., I study *deforestation* and model the *spillover effects* behind it.

If you have any questions you can send me a mail or ask after class.

Tutor

We also have a tutor in Maximilian Heinze, who you can contact via mail if you have any questions or issues related to this class.

He will also hold tutorial sessions to help you with prerequisites.

Organisation

Plan

- Weekly class (attendance is compulsory)
  - schedule on vvz.wu.ac.at
- Assessment in three parts (each part must be positive)
  - 30% – assignments
  - 30% – midterm exam (2022-12-13)
  - 40% – final exam (2022-01-24)
- Grades are distributed as follows
  - $[90, 100] \rightarrow 1$
  - $[78, 89] \rightarrow 2$
  - $[65, 77] \rightarrow 3$
  - $[51, 64] \rightarrow 4$
  - $[0, 50] \rightarrow 5$

Outline

In the lectures, we will focus on **causal inference**. This means we have to cover a lot of (dry) theory — the assignments are designed for you *to apply your knowledge* to actual data, and incentivise you to think about **prediction** as well.

The midterm exam in December will cover theoretical underpinnings, the final exam in January will test your overall understanding.

Outline

In the lectures, we will focus on **causal inference**. This means we have to cover a lot of (dry) theory — the assignments are designed for you *to apply your knowledge* to actual data, and incentivise you to think about **prediction** as well.

The midterm exam in December will cover theoretical underpinnings, the final exam in January will test your overall understanding.

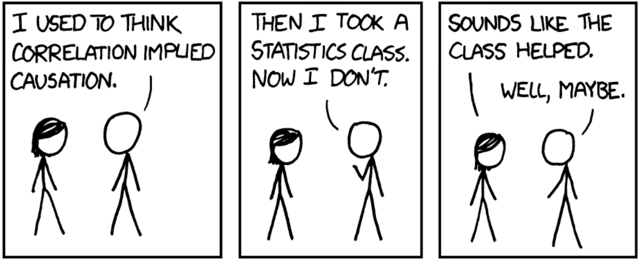


Figure 1: <xkcd.com> on causality versus correlation.

Forecasting competition

Econometrics is also useful for **prediction**.

- You can learn a lot about prediction via *trial-and-error*, so
- to facilitate that there will be a *voluntary* **forecast competition**.

You can find out more about the rules at kaggle.com/c/econometrics-2-w22; for now you should know that you will be able to earn **bonus points** on two deadlines:

First round 2022-12-08	**Second round** 2023-01-16
2 pts for places 1–3	4 pts for 1st
1 pt for places 4–10	3 pts for 2nd
	2 pts for 3rd
	1 pt for 4th

Content

Course requirements

You are expected to have **prior knowledge** of the following topics:

- multiple regression (application, interpretation),

These are covered in Econometrics I and you should have a solid understanding of them. It also helps to have working knowledge of **R**, e.g. from the Statistics with **R** course or the tutorial.

Course requirements

You are expected to have **prior knowledge** of the following topics:

- multiple regression (application, interpretation),
- estimators (least squares, classical assumptions, estimator properties),

These are covered in Econometrics I and you should have a solid understanding of them. It also helps to have working knowledge of **R**, e.g. from the Statistics with **R** course or the tutorial.

Course requirements

You are expected to have **prior knowledge** of the following topics:

- multiple regression (application, interpretation),
- estimators (least squares, classical assumptions, estimator properties),
- regression inference (hypothesis testing, confidence intervals),

These are covered in Econometrics I and you should have a solid understanding of them. It also helps to have working knowledge of **R**, e.g. from the Statistics with **R** course or the tutorial.

Course requirements

You are expected to have **prior knowledge** of the following topics:

- multiple regression (application, interpretation),
- estimators (least squares, classical assumptions, estimator properties),
- regression inference (hypothesis testing, confidence intervals),
- assumption failures (heteroskedasticity, correlation),

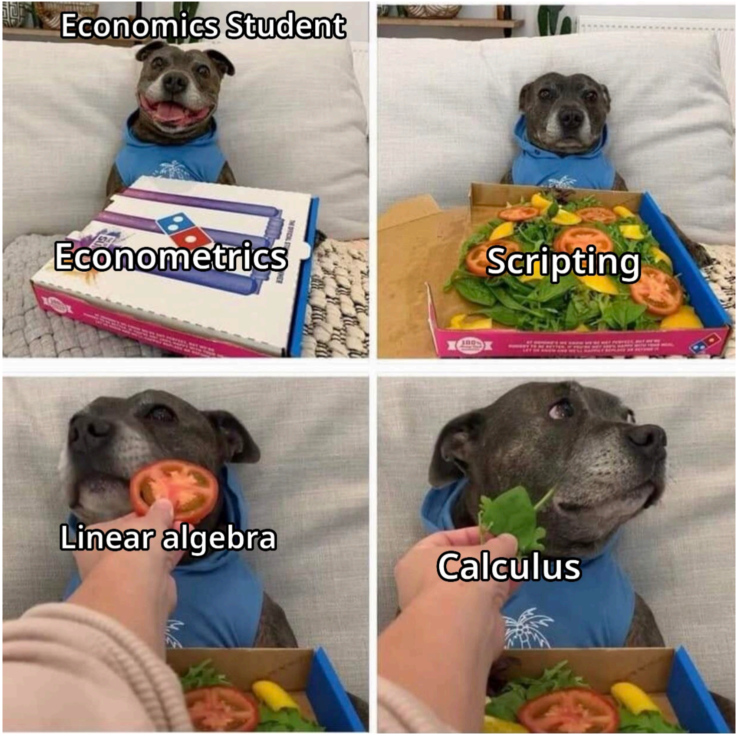
These are covered in Econometrics I and you should have a solid understanding of them. It also helps to have working knowledge of **R**, e.g. from the Statistics with **R** course or the tutorial.

Course requirements

You are expected to have **prior knowledge** of the following topics:

- multiple regression (application, interpretation),
- estimators (least squares, classical assumptions, estimator properties),
- regression inference (hypothesis testing, confidence intervals),
- assumption failures (heteroskedasticity, correlation),
- functional forms (dummy variables, interactions, log).

These are covered in Econometrics I and you should have a solid understanding of them. It also helps to have working knowledge of **R**, e.g. from the Statistics with **R** course or the tutorial.


Economics Student
Econometrics
Scripting
Linear algebra
Calculus

Study goals

After this course you should be

- equipped to *independently conduct econometric analyses*.

Study goals

After this course you should be

- equipped to *independently conduct econometric analyses*.
- aware of modelling *pitfalls* and how to address them.

Study goals

After this course you should be

- equipped to *independently conduct econometric analyses*.
- aware of modelling *pitfalls* and how to address them.
- have a solid understanding of *causal inference* — i.e. you will know
  - under which conditions we can interpret something causally,
  - how you could induce these conditions.

Study goals

After this course you should be

- equipped to *independently conduct econometric analyses*.
- aware of modelling *pitfalls* and how to address them.
- have a solid understanding of *causal inference* — i.e. you will know
  - under which conditions we can interpret something causally,
  - how you could induce these conditions.
- critically read and review applied research.

Materials

You only *need* the slides and material from class for this course. However, there's a lot of useful material that you can find online or in a library. For now, the following material might be interesting.

- Stock, J. H., and M. W. Watson (2015). *Introduction to Econometrics*. Book.
    - Hanck, C., Arnold, M., Gerber, A., and Schmelzer, M. (2021). *Introduction to Econometrics with **R***. Ebook.
- Wooldridge, J. (2015). *Introductory Econometrics: A Modern Approach*. Book.
- Gelman, A., Hill, J., and Vehtari, A. (2021). *Regression and Other Stories*. Book.
- Cunningham, S. (2021). *Causal Inference: The Mixtape*. Ebook.
- Venables, W. N., and D. M. Smith (2010). *An Introduction to **R***. Ebook.
- Lambert, B. (2014). *A Full Course in Undergraduate Econometrics*. YouTube Playlist (Part 1, Part 2).

An introduction to statistical learning

We observe the **samples** $\mathbf{y} \in \mathbb{R}^N$ (the *dependent*) and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_K) \in \mathbb{R}^{N \times K}$ (the *independent* variables), and assume that there is some relationship

$$\mathbf{y} = f(\mathbf{X}) + \mathbf{e}.$$

The unknown function f represents information that $\mathbf{X}$ provides about $\mathbf{y}$; all other relevant information is represented by the *error term* $\mathbf{e}$.

We observe the **samples** $\mathbf{y} \in \mathbb{R}^N$ (the *dependent*) and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_K) \in \mathbb{R}^{N \times K}$ (the *independent* variables), and assume that there is some relationship

$$\mathbf{y} = f(\mathbf{X}) + \mathbf{e}.$$

The unknown function f represents information that $\mathbf{X}$ provides about $\mathbf{y}$; all other relevant information is represented by the *error term* $\mathbf{e}$.

Naming conventions

The vector $\mathbf{y}$ is called *dependent*, *response*, or *output* variable and could, e.g., be **income**. The matrix $\mathbf{X}$ contains *independent*, *explanatory*, *control*, or *predictor* variables, or *features*. These could, e.g., be **occupation** and **ability**.

Why statistical learning?

We want an estimate $\hat{f}$ for two main reasons –

1. **prediction** – we want to learn about Y beyond our sample $\mathbf{y}$,
2. **inference** – we want to learn about the relation f between Y and X.

Why statistical learning?

We want an estimate $\hat{f}$ for two main reasons –

1. **prediction** – we want to learn about Y beyond our sample $\mathbf{y}$,
2. **inference** – we want to learn about the relation f between Y and X.

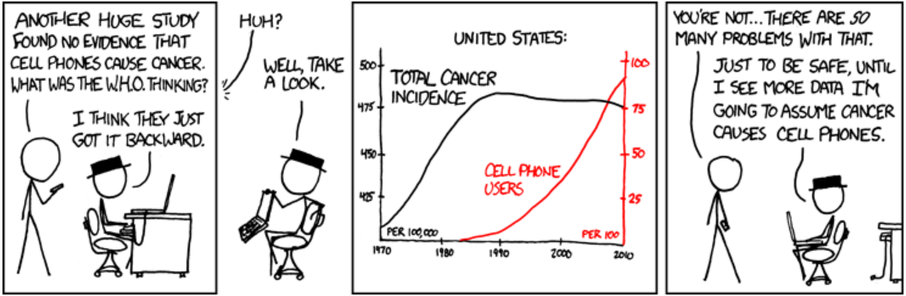


Figure 2: What can be learned from the data in this <xkcd.com> comic?

Prediction

Often, we can't obtain new observations of $\mathbf{y}$, but we can use other data $\mathbf{X}$ to **predict new values** of Y. We use our *estimate* $\hat{f}$ to obtain an estimate as

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{X}).$$

The h$\hat{\text{a}}$t indicates an **estimate**, so $\hat{\mathbf{y}}$ is our *in-sample* estimate of $\mathbf{y}$. With new data $\tilde{\mathbf{X}}$ we can get an *out-of-sample* estimate, i.e. a *prediction*.

For prediction, $\hat{f}$ can be a *black box* – as long as it works, we don't need to know how.

Example

Spotify may want to predict a (new) song that you would like to listen to. More concretely, they want to predict how to keep you engaged with Spotify.

Two important concepts in prediction

The accuracy of a prediction depends on the *reducible error* and *irreducible error*.

- **Reducible** error stems from **imperfect estimates** of f, i.e.

$$\hat{f} \approx f.$$

- **Irreducible** error are elements of $\mathbf{y}$ that **can't be explained** by $\mathbf{X}$. These elements are contained in the **error term e**, of

$$\mathbf{y} = f(\mathbf{X}) + \mathbf{e}.$$

Example

The Spotify algorithm can always be improved, but it cannot define you.

Decomposing predictive accuracy

We can decompose the **mean squared loss** into two parts.

$$\mathbb{E}\left[(\mathbf{y} - \hat{\mathbf{y}})^2\right] = \mathbb{E}\left[f(\mathbf{X}) + \mathbf{e} - \hat{f}(\mathbf{X})\right]^2$$

Decomposing predictive accuracy

We can decompose the **mean squared loss** into two parts.

$$
\begin{aligned}
\mathbb{E}\left[(\mathbf{y} - \hat{\mathbf{y}})^2\right] &= \mathbb{E}\left[f(\mathbf{X}) + \mathbf{e} - \hat{f}(\mathbf{X})\right]^2 \\
&= \mathbb{E}\left[\left[f(\mathbf{X}) - \hat{f}(\mathbf{X})\right]^2\right] + \mathbb{V}(\mathbf{e}).
\end{aligned}
$$

- We have the reducible error from our estimate $\hat{f}$ that can be improved,
- and the irreducible error from **e** that caps our accuracy. ▸ See the steps

Decomposing predictive accuracy

We can decompose the **mean squared loss** into two parts.

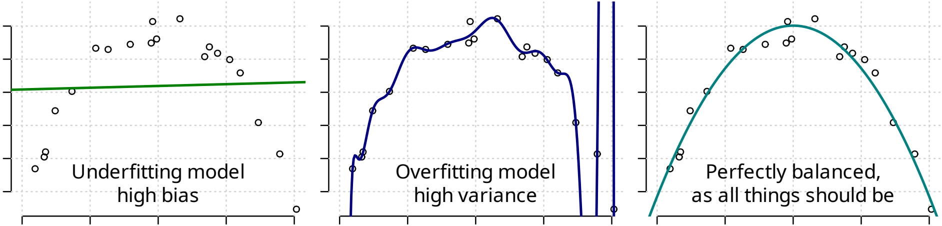
$$
\begin{aligned}
\mathbb{E}\left[(\mathbf{y}-\hat{\mathbf{y}})^2\right] &= \mathbb{E}\left[f(\mathbf{X})+\mathbf{e}-\hat{f}(\mathbf{X})\right]^2 \\
&= \mathbb{E}\left[\left[f(\mathbf{X})-\hat{f}(\mathbf{X})\right]^2\right]+\mathbb{V}(\mathbf{e}). \\
&= \mathrm{Bias}\left(\hat{f}(\mathbf{X})\right)^2+\mathbb{V}\left(\hat{f}(\mathbf{X})\right)+\mathbb{V}(\mathbf{e})\,.
\end{aligned}
$$

- We have the reducible error from our estimate $\hat{f}$ that can be improved,
  - and divided into the squared bias of our estimate $\hat{f}$,
  - and the variance of our model $\hat{f}$. ▸ See the steps
- and the irreducible error from **e** that caps our accuracy. ▸ See the steps

Overfitting and underfitting

For good prediction we want to minimise the reducible error — we do this by balancing the *bias* and the *variance* of our model $\hat{f}$. A useful distinction is between

- underfitting — $\hat{f}$ is not flexible enough to fit the data,
- overfitting — $\hat{f}$ follows the data (including irreducible errors) too closely.



Model bias and variance

We are talking about the bias and variance of a model $\hat{f}$, not a parameter (e.g. $\hat{\beta}$).

Inference

We want to **learn about the relationship** between random variables Y and X. We need to **understand** our model $\hat{f}$ to answer e.g. one of the following questions.

- Are X and Y **correlated**?
- What happens **if** we increase X by ten percent?
- Did the reduction in X **cause** higher Y?
- How does $\hat{f}$ map X to Y?

Examples

prosperity $\sim$ embargo

malaria $\sim$ insecticide use

cancer $\sim$ smoking

income $\sim$ discrimination

grade $\sim$ time studying

Causality

Many of these questions are *causal* — we want to learn about **causal effects**.

Consider the effect of a binary (i.e. yes or no) *treatment* $X \in \{0, 1\}$ on an *outcome* Y.
We can define the **potential outcomes** $Y(1)$ for treatment $X = 1$ and $Y(0)$ otherwise.
The difference gives us the causal effect of X on Y

$$\text{causal effect} = Y(1) - Y(0).$$

The fundamental problem of causal inference

In the real world, **we only ever observe $Y(1)$ or $Y(0)$**. The other one is an unobserved *counterfactual*.

Hurdles to causal inference

To **uncover causal effects**, we need to sidestep the fundamental problem of *causal inference* somehow. There's many challenges, but we'll definitely need

1. a definition of what constitutes a *causal effect*,
2. the *right* data, and the *right* model.

Example — discrimination

Income may be driven by discrimination (e.g. gender), but also experience or occupation. These factors could also be a pathway for discrimination.

Example — health

Non-smokers may be more conscious of their health than smokers – this may lead to lower cancer rates for reasons other than smoking.

Models of f

To learn about f, we need a **model** that suits the issue and the data — we care about, e.g., *flexibility* and *interpretability* — and a suitable way to estimate this model.

Some ways to characterise models is to distinguish between

- *parametric* ($\hat{f}$ has a finite number of parameters) and *non-parametric*,
- *supervised* and *unsupervised* (we don't have access to $\mathbf{y}$),
- *regression* ($\mathbf{y}$ is quantitative) and *classification* ($\mathbf{y}$ is qualitative).

"All models are wrong, but some are useful." — George Box

Models are an **approximation of reality** that **allows us to learn** about it.

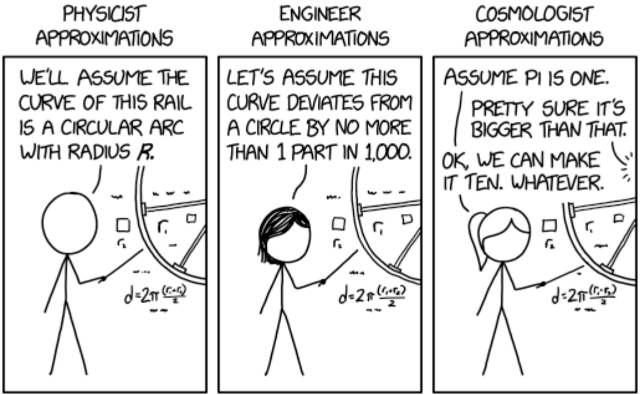


Figure 3: <xkcd.com>

Linear models

Linear models impose a certain **parametric** form on f. The dependent $\mathbf{y}$ should be a linear combination of $\mathbf{X}$, with parameters $\alpha \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^K$, as in

$$f(\mathbf{X}) = \alpha + \beta_1 \mathbf{x}_1 + \dots + \beta_K \mathbf{x}_K.$$

This way, we only need to estimate $K + 1$ parameters, which is usually

1. easy to do (e.g. using *least squares*),
2. easy to interpret (the partial effect of $\mathbf{x}_j$ is β_j),

and often yields good results that are not prone to *overfitting* (they usually don't follow the data too closely). In other cases, the linearity assumption may be too restrictive, and $\hat{f}$ may be far from the true f.

Non-parametric models

Non-parametric models do not impose a structure on f a-priori — instead, the structure is determined by *fitting as close as possible to the data* under certain other constraints. These methods can generally

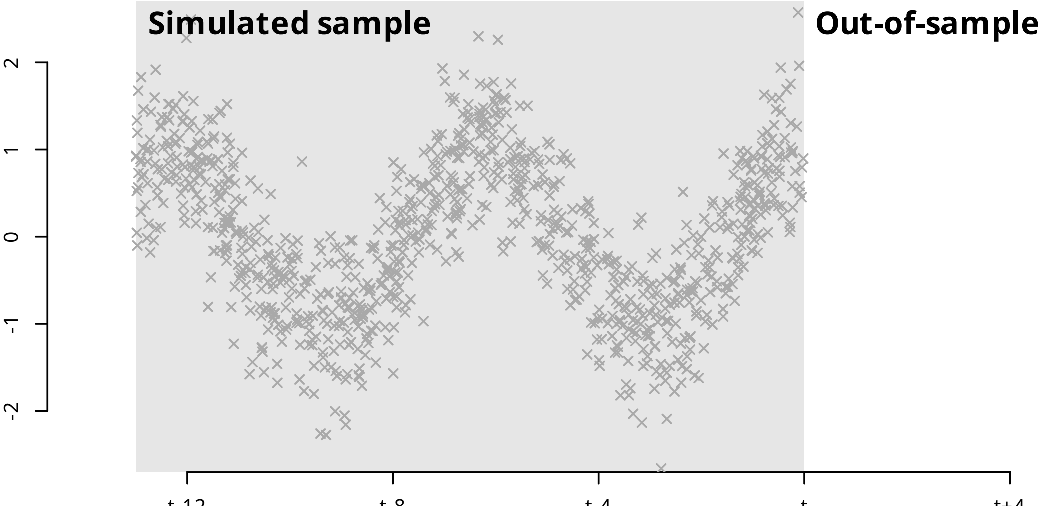
1. fit a wide range of possible forms of f,
2. tend to fit better to the data, and
3. are less reliant on model building.

However, non-parametric methods may require a lot more data, tend to be harder to interpret, and are susceptible to *overfitting*.

> *Counterintuitively, "non-parametric" does not imply that there are no parameters. Instead, the number and type of parameters are flexible (and potentially infinite).*

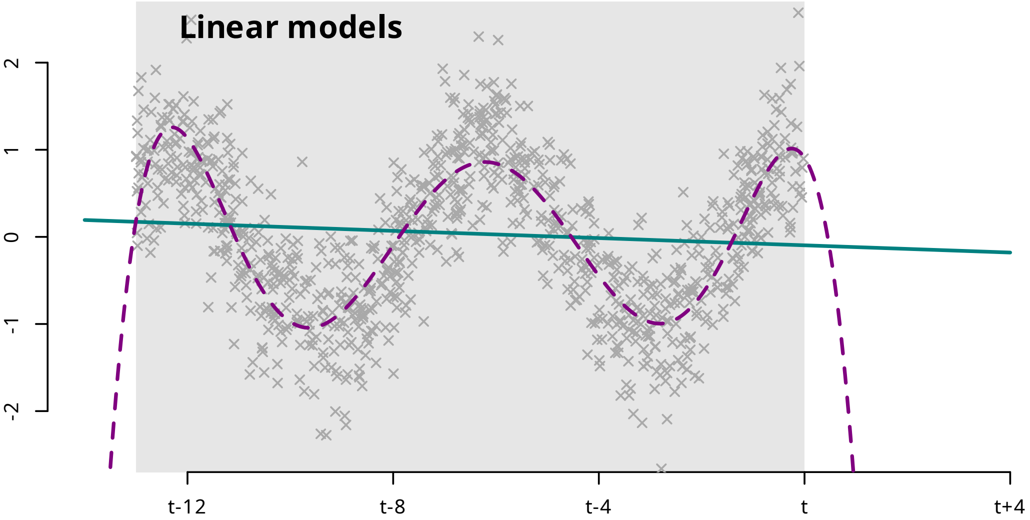
Parametric versus non-parametric model fit

We simulate some data from $Y \approx \sin X$ and compare model fit.



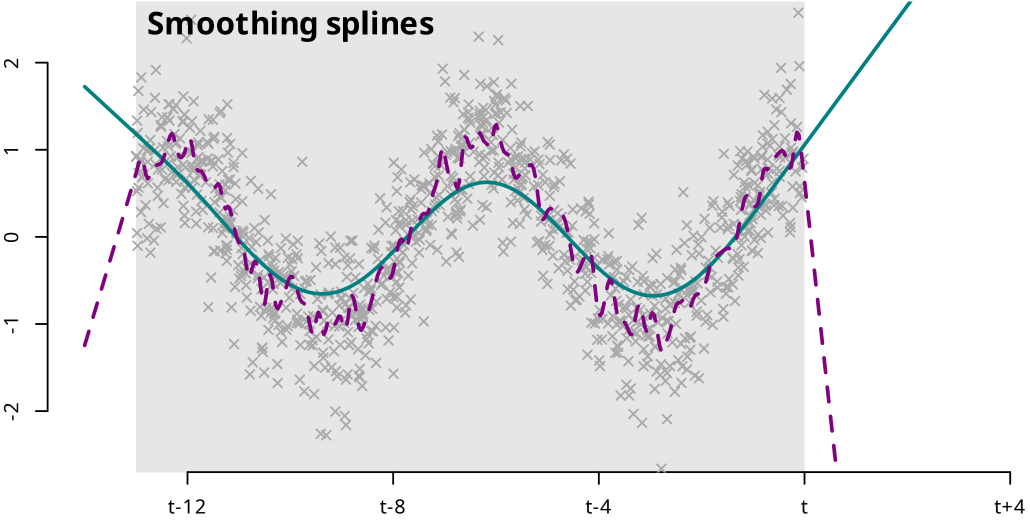
Parametric versus non-parametric model fit

Here, we fit two linear models — (1) $\mathbf{y} \approx \mathbf{x}\beta_1$, and (2) $\mathbf{y} \approx \mathbf{x}^1\beta_1 + \ldots + \mathbf{x}^6\beta_6$.



Parametric versus non-parametric model fit

Here, we fit two splines, with six and 100 degrees of freedom to the data.



Parametric versus non-parametric model fit

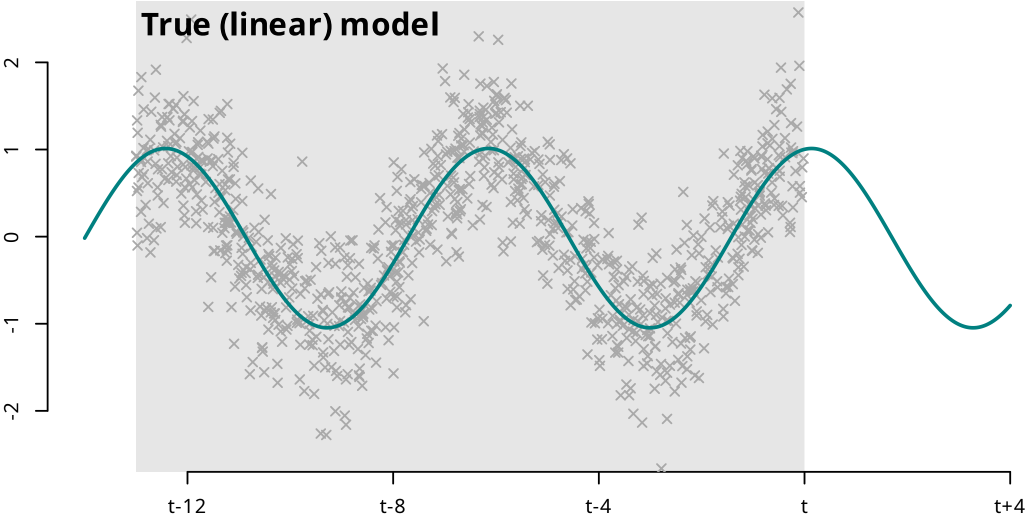


Figure 4: Parametric, non-parametric, and the true model $\mathbf{y} = \sin(\mathbf{x}) + \mathbf{e}$.

Supervised versus unsupervised

Unsupervised methods learn patterns from data *without a tagged outcome* Y via *mimicry*. Two examples are the *deep learning* models DALL·E and GPT3.

Image creation

```
> An astronaut, riding a horse in
a photorealistic style (Source).
```

Supervised versus unsupervised

Unsupervised methods learn patterns from data *without a tagged outcome* Y via *mimicry*. Two examples are the *deep learning* models DALL·E and GPT3.

Image creation

> An astronaut, riding a horse in a photorealistic style (Source).

Text creation

> Write a greentext story (Source).

Supervised versus unsupervised

Unsupervised methods learn patterns from data *without a tagged outcome* Y via *mimicry*. Two examples are the *deep learning* models DALL·E and GPT3.

Image creation

```
> An astronaut, riding a horse in a photorealistic style (Source).
```

Text creation

```
> Write a greentext story (Source).
```

```
> be me
> bottomless pit supervisor
> in charge of making sure the bottomless pit is, in fact, bottomless
> occasionally have to go down there and check if the bottomless pit is still bottomless
> one day i go down there and the bottomless put is no longer bottomless
> the bottom of the bottomless pit is now just a regular pit
> distress.jpg
> ask my boss what to do
> he says "just make it bottomless again"
> i say "how"
> he says "i don't know, you're the supervisor"
> rage.jpg
> quit my job
> become a regular pit supervisor
> first day on the job, go to the new hole
> it's bottomless
```

Interpretability versus flexibility

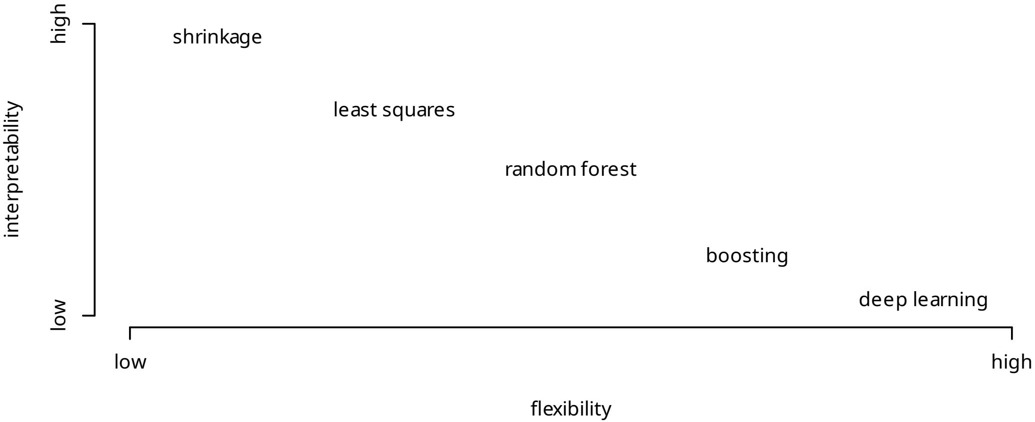


Figure 5: The interpretability–flexibility trade-off of methods, following James et al. (2021).

Choosing a suitable approach

We choose a model and estimation method depending on the issue of interest, and the available data. Central questions we may ask ourselves include the following.

- What is the *goal* of our analysis?
  - How easy to interpret should our estimate $\hat{f}$ be?
  - Do we need to generate accurate predictions?
- What does our *data* look like?
  - How much data do we have (observations N, and covariates K)?
  - Are we dealing with a regression or classification problem?

The role of econometrics

The role of econometrics

Econometrics seeks to *apply and develop statistical methods* to **learn about economic phenomena using empirical data**.

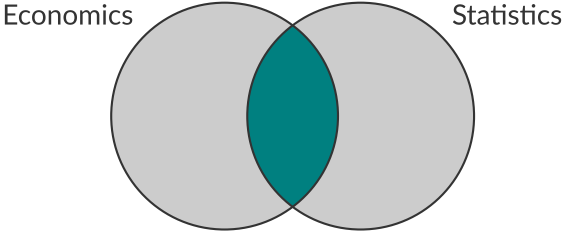


Figure 6: Econometrics lies at the intersection of economics and statistics.

An empirical shift

Econometrics plays an important role in an **empirical shift in economic research**, away from pure theory (Angrist et al. 2017; Hamermesh 2013). Today, economic theories are routinely confronted with real-world data.

> *"Experience has shown that each [...] of statistics, economic theory, and mathematics, is a necessary [...] condition for a real understanding of the quantitative relations in modern economic life." – Ragnar Frisch (1933)*

Empirical publications over time

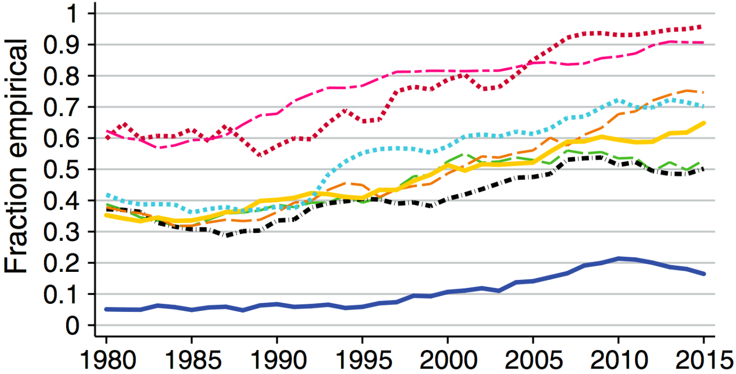


Figure 7: Weighted share of empirical publication in various economic fields (Angrist et al. (2017)).

A credibility revolution

Data and statistical methods *are not a panacea*. Econometrics has seen *considerable challenges and developments* since its inception. Important milestones concern

- uncertainty around model choice (e.g. Leamer 1983; Steel 2020),
- better research designs (e.g. Angrist and Pischke 2010),
- randomised experiments (see Athey and Imbens 2017),
- more flexible methods (Athey and Imbens 2019).

Many milestones build on some rather intuitive ideas; many open issues remain.

The econometric workhorse model

Consider how to transform the following *economic model* into an *econometric model*

$$\text{wage} \approx f(\text{education}, \text{experience}).$$

The econometric workhorse model

Consider how to transform the following *economic model* into an *econometric model*

$$\text{wage} \approx f(\text{education}, \text{experience}).$$

A sensible choice might be the following linear regression model

$$\mathbf{y}_{wage} = \mathbf{x}_1^{edu}\beta_1 + \mathbf{x}_2^{exp}\beta_2 + \mathbf{e}.$$

Linear models

Linear models are arguably the **workhorse models** of econometrics — they are valued for their *interpretability*, *parsimony*, and *extensibility*.

Goals of econometrics

The linear model's popularity is not surprising, given the classical tasks:

- testing a theory — Does class size affect grades?,
- evaluating a policy — What are impacts of an oil embargo?,
- forecasting the future — How quickly do stocks go up?

The central task is arguably **distilling causal effect** from *observational data*, since experimental data is rare (*why?*). When forecasting, economic theory can provide us with valuable **structural information** (*for example?*).

> *As you know by now — correlation does not need to imply causation. Consider the relation of sunburns and ice cream consumption (or one of many more examples).*

Linear algebra and the linear model

The linear model is an **essential building block**, and *linear algebra* gives us a very convenient way of expressing and dealing with these models. Let

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e},$$

where the $N \times 1$ vector $\mathbf{y}$ holds the dependent variable for all N observations, and the $N \times K$ matrix $\mathbf{X}$ contains all K explanatory variables. That is

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{pmatrix}.$$

The ordinary least squares estimator

The **ordinary least squares** (OLS) estimator minimises the *sum of squared residuals*, which is given by $\mathbf{e}'\mathbf{e}$ (i.e. $\sum_{n=1}^{N} e_n^2$). To find the estimate $\boldsymbol{\beta}_{OLS}$ we

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The ordinary least squares estimator

The **ordinary least squares** (OLS) estimator minimises the *sum of squared residuals*, which is given by $\mathbf{e}'\mathbf{e}$ (i.e. $\sum_{n=1}^{N} e_n^2$). To find the estimate $\boldsymbol{\beta}_{OLS}$ we

1. re-express the sum of squared residuals,

$$\begin{aligned}\mathbf{e}'\mathbf{e} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.\end{aligned}$$

The ordinary least squares estimator

The **ordinary least squares** (OLS) estimator minimises the *sum of squared residuals*, which is given by $\mathbf{e}'\mathbf{e}$ (i.e. $\sum_{n=1}^{N} e_n^2$). To find the estimate $\boldsymbol{\beta}_{OLS}$ we

1. re-express the sum of squared residuals,
2. find an *extreme value* via the partial derivative ($\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \boldsymbol{\beta}} = 0$),

$$\begin{aligned}\mathbf{e}'\mathbf{e} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.\end{aligned}$$

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta},$$

The estimator $\boldsymbol{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is directly available

The ordinary least squares estimator

The **ordinary least squares** (OLS) estimator minimises the *sum of squared residuals*, which is given by $\mathbf{e}'\mathbf{e}$ (i.e. $\sum_{n=1}^{N} e_n^2$). To find the estimate β_{OLS} we

1. re-express the sum of squared residuals,
2. find an *extreme value* via the partial derivative ($\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \beta} = 0$),
3. check whether we found a *minimum* via the second partial derivative.

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta. \end{aligned}$$

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta, \qquad \frac{\partial^2 \mathbf{e}'\mathbf{e}}{\partial^2 \beta} = 2\mathbf{X}'\mathbf{X}.$$

The estimator $\beta_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is directly available and a minimum. ▸ Show details

References i

Anderson, T. W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46–63. https://doi.org/10.1214/aoms/1177730090.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11 (1): 727–53. https://doi.org/10.1146/annurev-economics-080218-025643.

Angrist, Joshua D., Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. 2017. "Economic Research Evolves: Fields and Styles." *American Economic Review* 107 (5): 293–97. https://doi.org/10.1257/aer.p20171117.

Angrist, Joshua D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15 (4): 69–85. https://doi.org/10.1257/jep.15.4.69.

References ii

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30. https://doi.org/10.1257/jep.24.2.3.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32. https://doi.org/10.1257/jep.31.2.3.

———. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. “Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak.” *Journal of the American Statistical Association* 90 (430): 443–50. https://doi.org/10.1080/01621459.1995.10476536.

Buckles, Kasey S., and Daniel M. Hungerman. 2013. “Season of Birth and Later Outcomes: Old Questions, New Answers.” *Review of Economics and Statistics* 95 (3): 711–24. https://doi.org/10.1162/REST_a_00314.

Cunningham, Scott. 2021. *Causal Inference*. New Haven, CT, USA: Yale University Press. https://doi.org/10.12987/9780300255881.

Hamermesh, Daniel S. 2013. “Six Decades of Top Economics Publishing: Who and How?” *Journal of Economic Literature* 51 (1): 162–72. https://doi.org/10.1257/jel.51.1.162.

Imbens, Guido W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58 (4): 1129–79. https://doi.org/10.1257/jel.20191597.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1.

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27 (4): 435–54. https://doi.org/10.1017/pan.2019.11.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43. https://www.jstor.org/stable/1803924.

Pearl, Judea. 2009. *Causality*. *Cambridge Core*. Cambridge, England, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161.

References v

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

Steel, Mark F. J. 2020. “Model Averaging and Its Use in Economics.” *Journal of Economic Literature* 58 (3): 644–719. https://doi.org/10.1257/jel.20191385.

Reducible and irreducible error — decomposition steps

We have $\mathbf{y} = f(\mathbf{X}) + \mathbf{e}$, $\hat{\mathbf{y}} = \hat{f}(\mathbf{X})$, and $\mathbb{E}[\mathbf{e}] = 0$. Recall that $\mathbb{V}(\mathbf{e}) = \mathbb{E}\left[(\mathbf{e} - \mathbb{E}[\mathbf{e}])^2\right]$.

$$
\begin{aligned}
\mathbb{E}\left[(\mathbf{y} - \hat{\mathbf{y}})^2\right] &= \mathbb{E}\left[f(\mathbf{X}) + \mathbf{e} - \hat{f}(\mathbf{X})\right]^2 \\
&= \mathbb{E}\left[\left(f(\mathbf{X}) - \hat{f}(\mathbf{X})\right) + \mathbf{e}\right]^2 && \text{move terms and square} \\
&= \mathbb{E}\left[\left(f(\mathbf{X}) - \hat{f}(\mathbf{X})\right)^2 + 2\mathbf{e}\left(f(\mathbf{X}) - \hat{f}(\mathbf{X})\right) + \mathbf{e}^2\right] \\
&= \mathbb{E}\left[\left(f(\mathbf{X}) - \hat{f}(\mathbf{X})\right)^2\right] + \mathbb{E}\left[2\mathbf{e}\left(f(\mathbf{X}) - \hat{f}(\mathbf{X})\right)\right] + \mathbb{E}\left[\mathbf{e}^2\right] \\
&= \mathbb{E}\left[\left(f(\mathbf{X}) - \hat{f}(\mathbf{X})\right)^2\right] + 0 + \mathbb{E}\left[\mathbf{e}^2\right] && \text{simplify} \\
&= \mathbb{E}\left[\left(f(\mathbf{X}) - \hat{f}(\mathbf{X})\right)^2\right] + \mathbb{V}(\mathbf{e}).
\end{aligned}
$$

▸ Go back

Bias and variance — decomposition steps

We will use the shorthands $f = f(\mathbf{X})$, and $\hat{f} = \hat{f}(\mathbf{X})$. Recall that $\text{Bias}\left(\hat{f}\right) = \mathbb{E}\left[\hat{f}\right] - f$.

$$
\begin{aligned}
\mathbb{E}\left[(\mathbf{y} - \hat{\mathbf{y}})^2\right] &= \mathbb{E}\left[\left(f - \hat{f}\right)^2\right] + \mathbb{V}(\mathbf{e}) \\
&= \mathbb{E}\left[\left(f - \mathbb{E}\left[\hat{f}\right] + \mathbb{E}\left[\hat{f}\right] - \hat{f}\right)^2\right] + \mathbb{V}(\mathbf{e}) \qquad \text{add } 0 = (\mathbb{E}\left[\hat{f}\right] - \mathbb{E}\left[\hat{f}\right]) \\
&= \mathbb{E}\left[\left(\left(f - \mathbb{E}\left[\hat{f}\right]\right) + \left(\mathbb{E}\left[\hat{f}\right] - \hat{f}\right)\right)^2\right] + \mathbb{V}(\mathbf{e}) \qquad \text{square the terms} \\
&= \left(f - \mathbb{E}\left[\hat{f}\right]\right)^2 + \mathbb{E}\left[2\left(f - \mathbb{E}\left[\hat{f}\right]\right) \times \left(\mathbb{E}\left[\hat{f}\right] - \hat{f}\right)\right] + \mathbb{E}\left[\left(\mathbb{E}\left[\hat{f}\right] - \hat{f}\right)^2\right] + \mathbb{V}(\mathbf{e}) \\
&= \left(f - \mathbb{E}\left[\hat{f}\right]\right)^2 + 0 + \mathbb{E}\left[\left(\mathbb{E}\left[\hat{f}\right] - \hat{f}\right)^2\right] + \mathbb{V}(\mathbf{e}) \qquad \text{simplify} \\
&= \text{Bias}\left(\hat{f}\right)^2 + \mathbb{V}\left(\hat{f}\right) + \mathbb{V}(\mathbf{e}).
\end{aligned}
$$

▸ Go back

OLS estimator — derivation

We have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, which lets us re-express the sum of squared residuals as

$$\begin{aligned}\mathbf{e}'\mathbf{e} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y}' - \boldsymbol{\beta}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta},\end{aligned}$$

where we use the fact that for a scalar $\alpha = \alpha'$ to simplify $\mathbf{y}'\mathbf{X}\boldsymbol{\beta} = (\mathbf{y}'\mathbf{X}\boldsymbol{\beta})' = \boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$. Next, we set the first derivative $\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ to zero

$$\begin{aligned}-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= 0 \\ \mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}'\mathbf{y} \\ \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.\end{aligned}$$

The second partial derivative $2\mathbf{X}'\mathbf{X}$ is positive (definite) as long as it is invertible.

Go back

Causality

Causality

Causality is when one **cause** leads to some **effect**. The cause is partly responsible for the effect, and the effect partly depends on the cause. Questions of causality are of a *philosophical nature*, so a well-defined framework is important for discussions.

Consider a binary *treatment* (e.g. vaccination) X, and *outcome* (e.g. immunity) Y. We can think of the **causal effect** τ as the difference in **potential outcomes**

$$\tau = Y(X = 1) - Y(X = 0)$$

The fundamental problem of causal inference

In the real world **only one outcome is realised**; the other is a **counterfactual**. We have to *estimate this 'missing' outcome* to learn about the causal effect.

i	X_i	Y_i	$Y_i(1)$	$Y_i(0)$
1	0	1	**?**	1
2	0	1	**?**	1
3	1	1	1	**?**
4	1	0	0	**?**
$\vdots$				
N	1	1	1	**?**

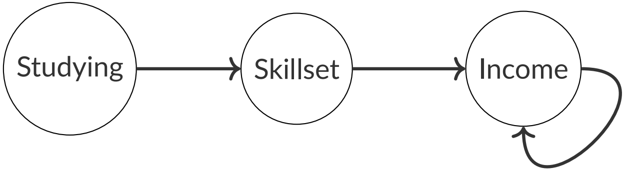
The potential outcomes framework is also called the Neyman–Rubin causal model.

We say an effect (estimate) is **causally identified** if we can *interpret it causally* in our chosen framework and scope — communicating the framework and scope is vital.

We may want to estimate a causal effect from $\mathbf{y}^{inc} = \mathbf{x}^{study}\beta + \mathbf{e}$.

Causal identification

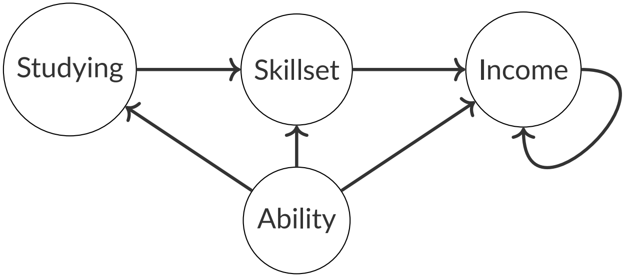
We say an effect (estimate) is **causally identified** if we can *interpret it causally* in our chosen framework and scope — communicating the framework and scope is vital.



However, you don't get paid directly for studying — skills are a *mediator*.

Causal identification

We say an effect (estimate) is **causally identified** if we can *interpret it causally* in our chosen framework and scope — communicating the framework and scope is vital.



Morever, ability may *confound* your effect estimates of $\mathbf{y}^{inc} = \mathbf{x}^{study}\beta + \mathbf{e}$.

Important causal quantities

Average treatment effect

The average causal effect is simply the mean of all treatment effects.

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[\tau_i] \\ &= \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]\,.\end{aligned}$$

Conditional average treatment effect

Often, we want to control for some third characteristic Z_i

$$\tau_{CATE} = \mathbb{E}[\tau_i | Z_i = z]\,.$$

A special case is the **average treatment effect on the treated** (ATT) where we condition on received treatment, $Z_i = X_i = 1$.

Estimating an average treatment effect

i	X_i	Y_i	$Y_i(1)$	$Y_i(0)$
1	0	1	–	1
2	0	0	–	0
3	0	0	–	0
4	0	0	–	0
5	1	1	1	–
6	1	1	1	–
7	1	1	1	–
8	1	0	0	–

- We could use $\mathbb{E}[Y_i(0)] = 0.25$ and $\mathbb{E}[Y_i(1)] = 0.75$, for

$$\tau_{ATE} = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = 0.5.$$

- We can also impose a linear model

$$\mathbf{y} = \mathbf{x}\tau + \mathbf{e},$$

and estimate τ_{ATE} using OLS.

Estimating an average treatment effect

i	X_i	Y_i	$Y_i(1)$	$Y_i(0)$
1	0	1	–	1
2	0	0	–	0
3	0	0	–	0
4	0	0	–	0
5	1	1	1	–
6	1	1	1	–
7	1	1	1	–
8	1	0	0	–

- We could use $\mathbb{E}[Y_i(0)] = 0.25$ and $\mathbb{E}[Y_i(1)] = 0.75$, for

$$\tau_{ATE} = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = 0.5.$$

- We can also impose a linear model

$$\mathbf{y} = \mathbf{x}\tau + \mathbf{e},$$

and estimate τ_{ATE} using OLS.

Ignorability

Ignorability

A treatment X is *ignorable* if

$$(Y(1), Y(0)) \perp\!\!\!\perp X.$$

This means that *both potential outcomes* are independent of X, the treatment.

When X is ignorable, the treatment is *randomly assigned* and only affects the outcome Y by either *realising* $Y(1)$ or $Y(0)$, i.e.

$$Y = Y(1)X + Y(0)(1 - X).$$

Example

The assignment of X should be ignorable — this is violated if, e.g., subjects are targeted (for vaccination), or select themselves (by responding to a survey).

Conditional ignorability

Conditional ignorability

A treatment X is *ignorable*, conditional on covariates Z, if

1. $(Y(1), Y(0)) \perp\!\!\!\perp X | Z$.
2. $\mathbb{P}(X = 1) \in (0, 1)$.

Potential outcomes are independent of X, conditional on Z, and there are both treated and untreated subjects.

If X is ignorable, we can use the sample averages $\mathbb{E}[Y_i(0)]$ and $\mathbb{E}[Y_i(1)]$ as estimates for $Y(0)$ and $Y(1)$ — the estimate of τ_{ATE} will be causally identified See the proof.

Randomised experiments

We learned that we can estimate a causal effect if

1. we have access to **parallel universes** (we can compare $Y(1)$ and $Y(0)$), or
2. the treatment is **ignorable** (we can compare the sample averages).

Until we figure out the first option, **experiments** (natural or designed), where the treatment is truly assigned randomly, are our best shot. However, even *with properly randomised* data, there are *threats to causal inference*.

> *Experiments are not always feasible, because (inter alia) they are expensive and often morally problematic. Thankfully, they're not our only option.*

Balance and Overlap

Assume that you have *perfectly randomised* data to investigate the effect of some treatment X – the treatment and control groups were **assigned randomly**.

For good inference, we want the treatment and control groups to be **comparable**, i.e.

- *balance* and
- *overlap* between the groups.

If the groups are imbalanced or there is a lack of overlap, we are forced to rely more on our model and assumptions, and less on the data.

Imbalance

An **imbalance** between the treated and control groups occurs when there are *differences between these groups*. This is problematic when there are differences in terms of third **variables that affect the outcome** Y.

If we have enough (e.g. ∞) data, these imbalances should disappear. Otherwise, we may want to account for them before comparing sample means of the groups.

Example — vaccination

You run an experiment to learn about the efficacy of vaccination and collect the randomised data to the right.

What do you have to watch out for?

$N_{treated}$	$N_{untreated}$	N_{total}
55	45	100

Imbalance

An **imbalance** between the treated and control groups occurs when there are *differences between these groups*. This is problematic when there are differences in terms of third **variables that affect the outcome** Y.

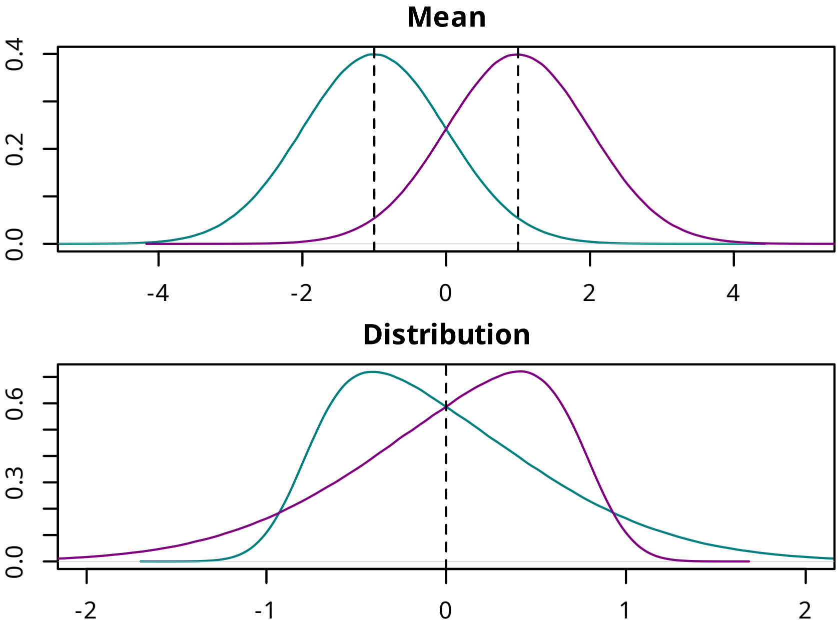
If we have enough (e.g. ∞) data, these imbalances should disappear. Otherwise, we may want to account for them before comparing sample means of the groups.

Example — vaccination

You run an experiment to learn about the efficacy of vaccination and collect the randomised data to the right.
What do you have to watch out for?

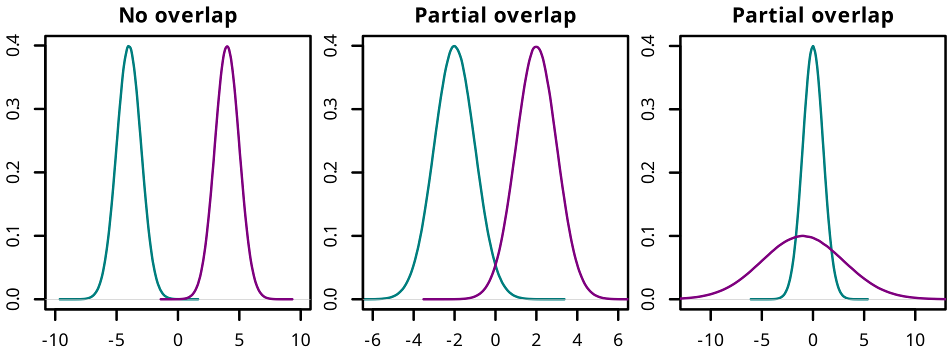
age	$N_{treated}$	$N_{untreated}$	N_{total}
0-69	10	15	25
70+	45	30	75
	55	45	100

Spotting imbalances



Overlap

The **overlap** describes how similar the *range of the data* is across groups. A lack of overlap means that there are no *equivalents in the two groups* (e.g. someone aged 90+) and we may have to **extrapolate beyond the support of the data**.



Experimental design — blocked experiments

When **designing an experiment**, we can use *prior information* to get more precise and accurate estimates – consider the vaccine efficacy experiment.

- We know that age probably *plays an important role*.

Experimental design — blocked experiments

When **designing an experiment**, we can use *prior information* to get more precise and accurate estimates — consider the vaccine efficacy experiment.

- We know that age probably *plays an important role*.
- We could divide the data into blocks.

Experimental design — blocked experiments

When **designing an experiment**, we can use *prior information* to get more precise and accurate estimates — consider the vaccine efficacy experiment.

- We know that age probably *plays an important role*.
- We could divide the data into blocks.
  - Subjects in a block should have *similar age*.
  - Random assignment of the treatment happens *within blocks*.

We minimise issues with balance and overlap by running many small experiments.

Estimates from a blocked experiment

If we conduct an experiment with B blocks, we can estimate the *average treatment effect* **within a block** $\mathcal{B}_b$ by comparing the sample averages

$$\hat{\tau}^b_{ATE} = \mathbb{E}\left[Y_j(1)\right] - \mathbb{E}\left[Y_j(0)\right] \text{ where } j \in \mathcal{B}_b.$$

Estimates from a blocked experiment

If we conduct an experiment with B blocks, we can estimate the *average treatment effect* **within a block** $\mathcal{B}_b$ by comparing the sample averages

$$\hat{\tau}^b_{ATE} = \mathbb{E}\left[Y_j(1)\right] - \mathbb{E}\left[Y_j(0)\right] \text{ where } j \in \mathcal{B}_b.$$

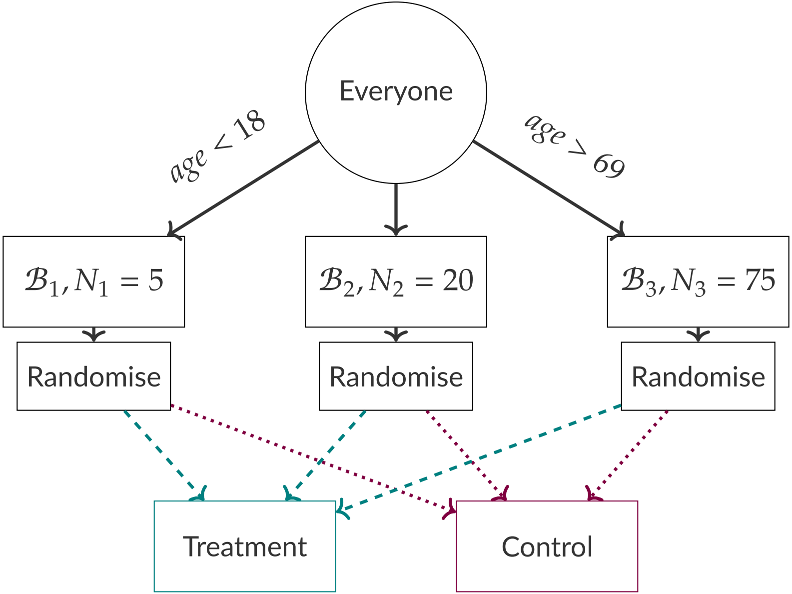
For an estimate of the *overall average treatment effect*, we take a weighted average

$$\hat{\tau}_{ATE} = \frac{\sum_i N_i \hat{\tau}^i}{\sum_i N_i},$$

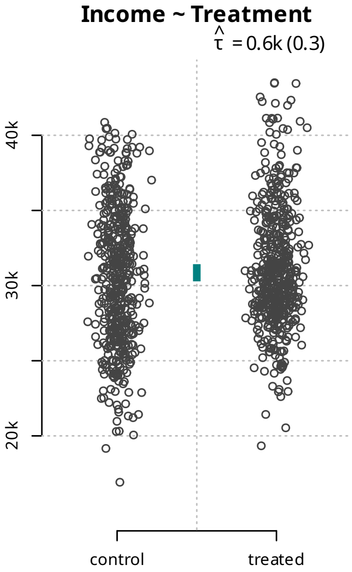
where N_i is the size of block $\mathcal{B}_i$, or estimate a *regression with block indicators*

$$y_i = \alpha + x_i \tau_{ATE} + \mathbb{1}(i \in \mathcal{B}_1)\gamma_1 + \ldots + \mathbb{1}(i \in \mathcal{B}_B)\gamma_B + e_i.$$

A blocked experiment visualised



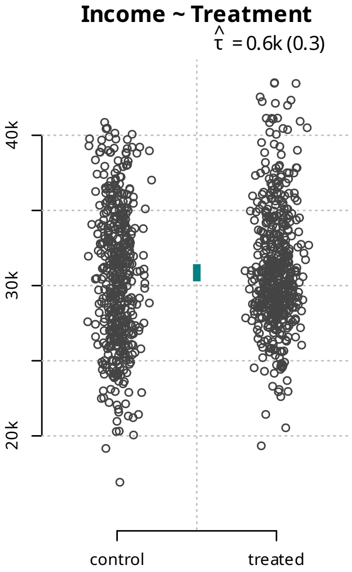
A note on randomisation and controls

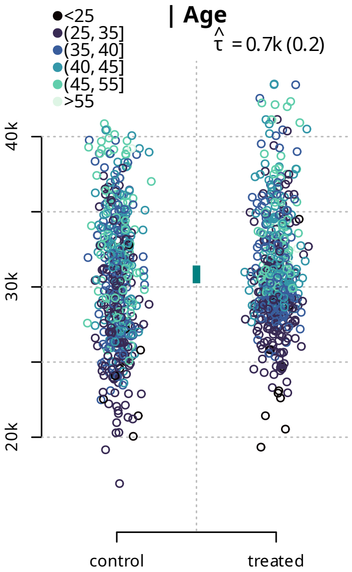


We evaluate a randomised experiment on the income effects of some treatment.

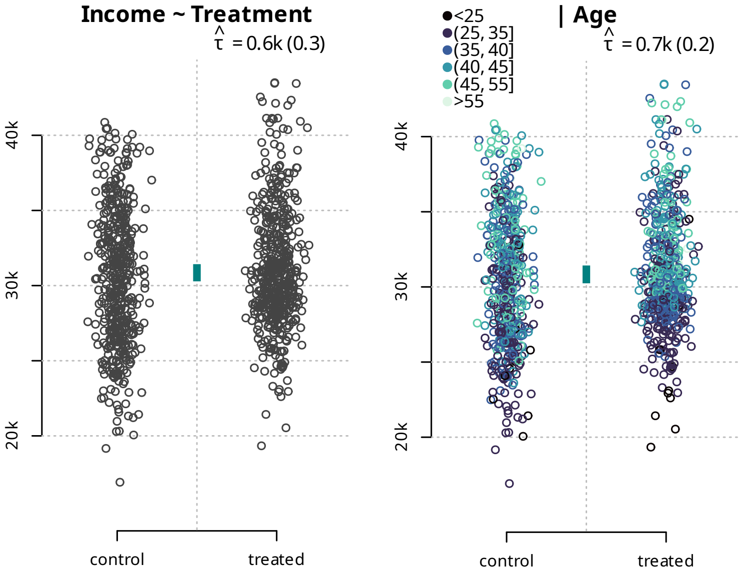
We also have information on age and education — how should we proceed?

A note on randomisation and controls



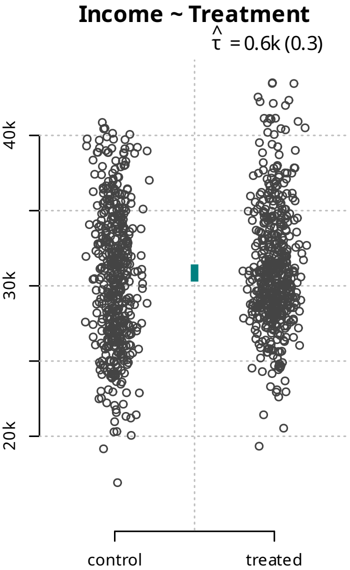


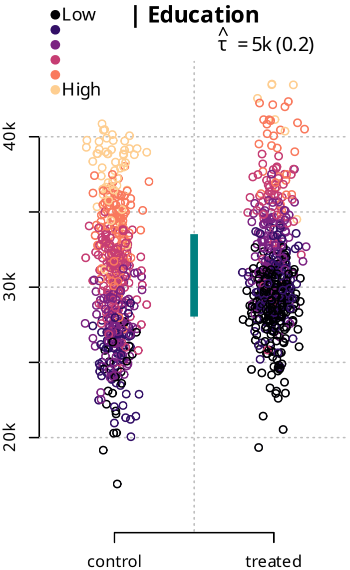
A note on randomisation and controls



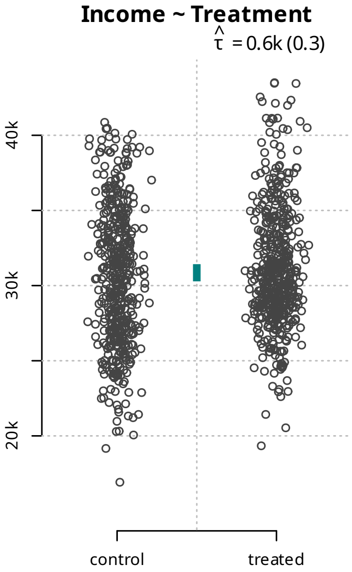
Controlling for covariates can help *improve the efficiency* of estimates.

A note on randomisation and controls

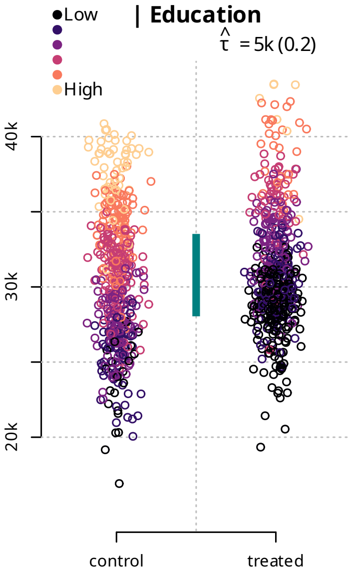




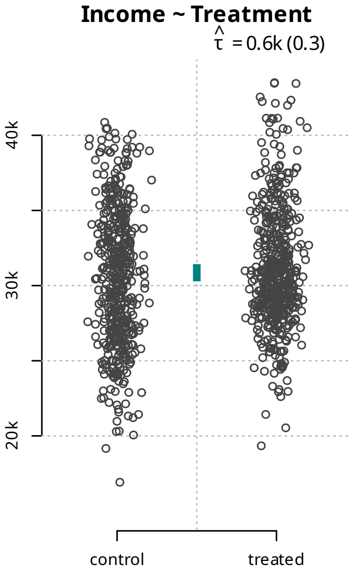
A note on randomisation and controls



Controlling for covariates can also **bias estimates**.



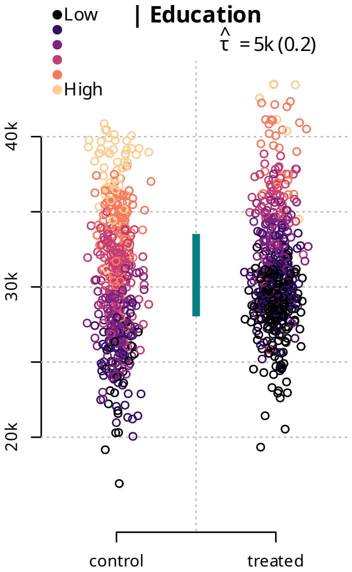
A note on randomisation and controls



Controlling for covariates can also **bias estimates**.

These covariates may still *carry useful information*.

causal inference ≠ prediction



Recap and outline

- To *identify a causal effect*, the treatment should be *ignorable*.
- This can be achieved in *experiments* with randomly assigned treatment.

Causal identification is **hard** — a lot can go wrong, e.g.:

- *imbalance*, which can mislead us,
- *lack of overlap*, which limits what we can learn,
- problematic *controls*, which can distort causal effects.

Before we get to *observational data*, we will proceed with

1. a *graphical framework to think about causality*, and then
2. look into some more *threats to causal inference*.

Ignorability and identification

Theorem

If X is ignorable conditional on Z, then

$$\mathbb{E}[\tau] = \sum_{z \in \operatorname{supp} Z} \left(\mathbb{E}[Y|X = 1, Z = z] - \mathbb{E}[Y|X = 0, Z = z] \right) \mathbb{P}(X = x) \,.$$

Proof: We know that $\mathbb{E}[Y(0)|Z] = \mathbb{E}[Y(0)|X = 0, Z] = \mathbb{E}[Y|X = 0, Z]$ by the ignorability of X, meaning we can treat counterfactuals and realised outcomes interchangeably, conditional on Z. The rest follows by the law of iterated expectations.

This implies that we can use averages to estimate counterfactuals.

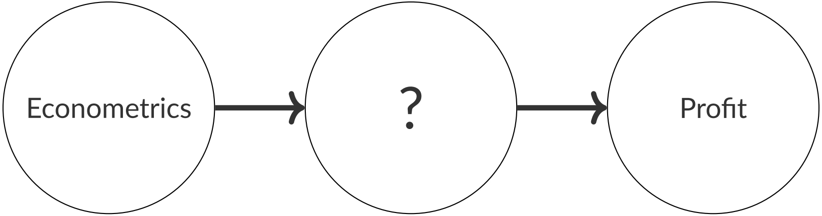
Go back

Causality and graphs

The directed acyclic graph

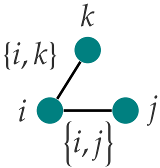
A directed acyclic graph (DAG) is a

- ~~fancy flowchart~~
- type of graph that we can use as a tool for causal modelling.



One slide of graph theory

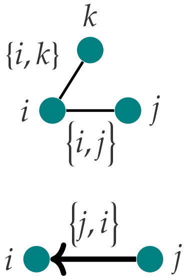
A **graph** $G(\mathcal{N}, \mathcal{E})$ consists of a set of **nodes** $\mathcal{N} = \{1, \ldots, N\}$, and a set of **edges** $\mathcal{E} = \big\{\{i, j\}, \{k, l\}, \ldots\big\}$, for $i, j, k, l \in \mathcal{N}$ between nodes.



One slide of graph theory

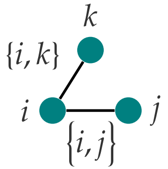
A **graph** $G(\mathcal{N}, \mathcal{E})$ consists of a set of **nodes** $\mathcal{N} = \{1, \ldots, N\}$, and a set of **edges** $\mathcal{E} = \left\{\{i,j\}, \{k,l\}, \ldots\right\}$, for $i,j,k,l \in \mathcal{N}$ between nodes.

In a **directed** graph, the set of edges is *ordered* — edges go from a tail to a head node. This means that $\{i,j\} \neq \{j,i\}$.

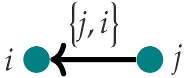


One slide of graph theory

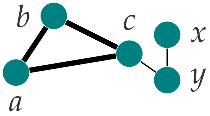
A **graph** $G(\mathcal{N}, \mathcal{E})$ consists of a set of **nodes** $\mathcal{N} = \{1, \dots, N\}$, and a set of **edges** $\mathcal{E} = \big\{\{i,j\}, \{k,l\}, \dots\big\}$, for $i, j, k, l \in \mathcal{N}$ between nodes.



In a **directed** graph, the set of edges is *ordered* — edges go from a tail to a head node. This means that $\{i,j\} \neq \{j,i\}$.



A *walk* is a sequence of edges which joins a sequence of nodes.
A **cycle** is a *walk* where all edges are distinct and the *first and the last node are equal*. A graph without cycles is an **acyclic** graph.

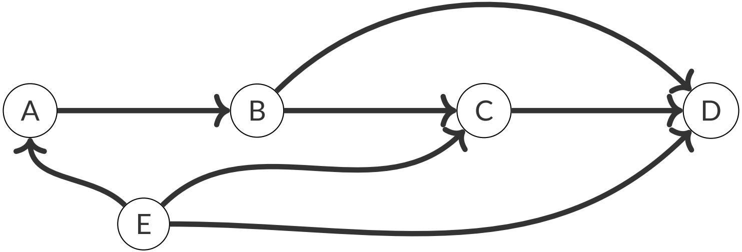


Because this isn't confusing enough, nodes are also referred to (inter alia) as 'vertices', 'agents', or 'points'. Edges are also called 'links', 'connections', or 'lines'.

Back to the DAG

DAGs have three layers of information that we can use:

1. nodes, to represent **random variables**,
2. directed edges that represent a **causal effect**,
3. missing edges, indicating the assumption of **no causal effect**.



Keep in mind that missing information can still be information.

DAGs and causal inference

DAGs are another **framework for causal inference** (similar to the *potential outcomes* framework that we already covered) that can be very helpful. They

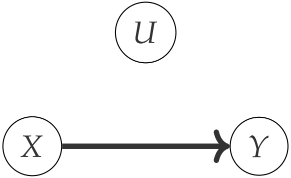
- **visualise** causal relationships between a number of variables,
  - allowing us to *transparently state our assumptions*,
- help us **identify** a causal effect,
  - showing which *variables to control for* to estimate the effect.

These types of graphs are commonly used to model many different kinds of information. Examples include family trees, version control systems, citations, project management, and object-oriented programs.

The basics of causal inference with DAGs

We want to learn about a **causal effect of education on income**. Let Y be *income*, X indicate *participation* in a course, and U be a measure of *aptitude*.

Let's construct a DAG to help isolate the causal effect of X on Y.

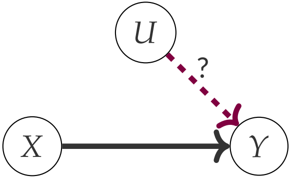


The basics of causal inference with DAGs

We want to learn about a **causal effect of education on income**. Let Y be *income*, X indicate *participation* in a course, and U be a measure of *aptitude*.

Let's construct a DAG to help isolate the causal effect of X on Y.

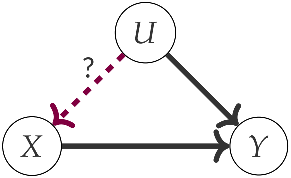
1. Is Y related to U?



We want to learn about a **causal effect of education on income**. Let Y be *income*, X indicate *participation* in a course, and U be a measure of *aptitude*.

Let's construct a DAG to help isolate the causal effect of X on Y.

1. Is Y related to U?
2. Is X related to U? Can we randomise treatment?

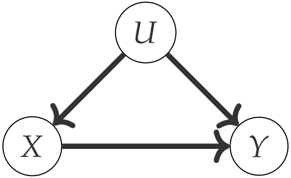


The basics of causal inference with DAGs

We want to learn about a **causal effect of education on income**. Let Y be *income*, X indicate *participation* in a course, and U be a measure of *aptitude*.

Let's construct a DAG to help isolate the causal effect of X on Y.

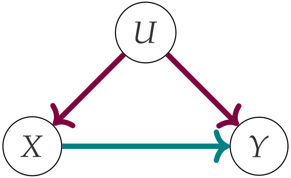
1. Is Y related to U?
2. Is X related to U? Can we randomise treatment?
3. Are there other important variables?



We want to learn about a **causal effect of education on income**. Let Y be *income*, X indicate *participation* in a course, and U be a measure of *aptitude*.

Let's construct a DAG to help isolate the causal effect of X on Y.

1. Is Y related to U?
2. Is X related to U? Can we randomise treatment?
3. Are there other important variables?

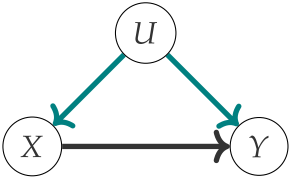


It turns out that there are *two paths* from X to Y,

1. one **direct path** $X \rightarrow Y$, and
2. one **backdoor path** $X \leftarrow U \rightarrow Y$.

Confounders and open backdoors

In our DAG, where we want to isolate $X \rightarrow Y$, we have an **open backdoor path** via U, which *confounds* the causal effect of interest.

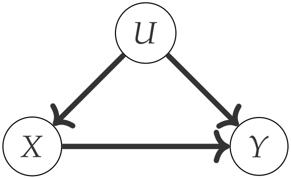


Confounder

A **confounder** is a variable that *influences both* the *dependent* and *explanatory* variables — effects of the confounder and the explanatory are mixed together.

Confounders and open backdoors

In our DAG, where we want to isolate $X \rightarrow Y$, we have an **open backdoor path** via U, which *confounds* the causal effect of interest.



We can *close the backdoor*, by controlling for U, e.g.

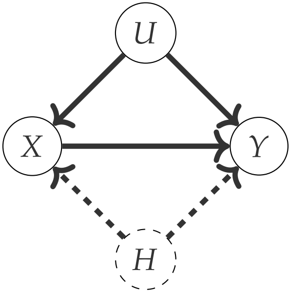
$$\mathbf{y} = \mathbf{x}\beta + \mathbf{u}\theta + \varepsilon.$$

Confounder

A **confounder** is a variable that *influences both* the *dependent* and *explanatory* variables — effects of the confounder and the explanatory are mixed together.

Confounders and open backdoors

In our DAG, where we want to isolate $X \to Y$, we have an **open backdoor path** via U, which *confounds* the causal effect of interest.



We can *close the backdoor*, by controlling for U, e.g.

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{u}\theta + \varepsilon.$$

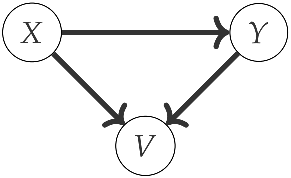
We have a problem if we *cannot* control for a confounder.

Confounder

A **confounder** is a variable that *influences both* the *dependent* and *explanatory* variables — effects of the confounder and the explanatory are mixed together.

Colliders and closed backdoors

Assume we can condition on the confounder from before, but we want to consider *social circles*, V. We assume they are caused by X and Y, giving us the DAG to the right.

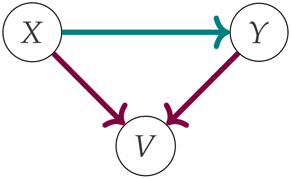


Collider

A **collider** is a variable that *is influenced by both* the *dependent* and *explanatory* variables – they act as a sink, and close backdoor paths.

Colliders and closed backdoors

Assume we can condition on the confounder from before, but we want to consider *social circles*, V. We assume they are caused by X and Y, giving us the DAG to the right.



There are two paths from X to Y,

1. one **direct path** $X \to Y$, and
2. one **backdoor path** $X \to V \leftarrow Y$.

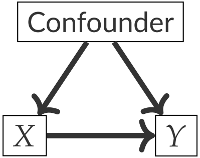
Because the backdoor *collides* at V, it **is already closed**.

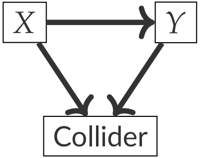
Collider

A **collider** is a variable that *is influenced by both* the *dependent* and *explanatory* variables – they act as a sink, and close backdoor paths.

An *open backdoor* between two variables creates systemic, *non-causal correlation* between them. To estimate a causal effect, we need to close backdoor paths, by

- **controlling for confounders** along the path,
- **leaving colliders** along the path **alone**.

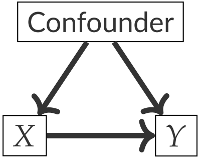


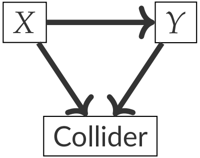


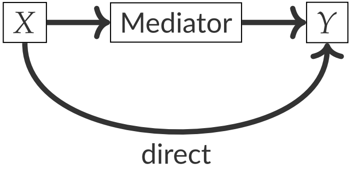
An *open backdoor* between two variables creates systemic, *non-causal correlation* between them. To estimate a causal effect, we need to close backdoor paths, by

- **controlling for confounders** along the path,
- **leaving colliders** along the path **alone**.

Another relevant type of variable is the **mediator**, which mediates (part of) the causal effect of X on Y. Controlling for a mediator removes the mediated effect.



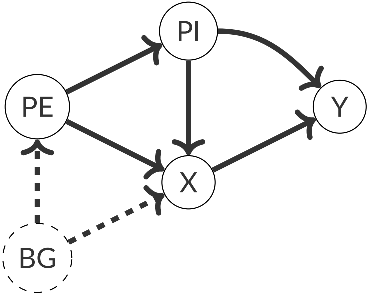




Example 1 — Education and income

We want to learn about the effects of education (X) on income (Y).

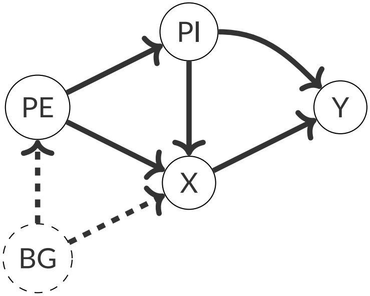
- Education is not chosen at random, but determined by other factors,
  - e.g. the education and income of parents (PE and PI),
  - and other unobserved background factors (BG).



Example 1 — Telling a story

Our DAG tells a story and encodes our assumptions — does this story make sense?

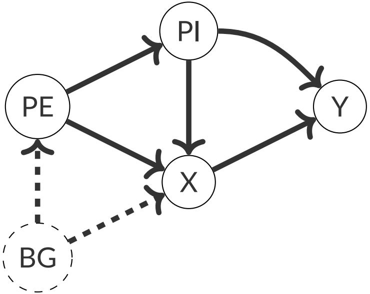
- We assumed that background factors, BG, only affect income via education.
- This means that e.g. ability, intelligence, motivation, and social environment have **no direct effect on income**.



Example 1 — Enumerating our DAG

If we still settle on this DAG, we proceed by *listing all paths between variables of interest* (in our case, these are X and Y).

1. $X \rightarrow Y$ (direct)
2. $X \leftarrow PI \rightarrow Y$ (backdoor 1)
3. $X \leftarrow PE \rightarrow PI \rightarrow Y$ (backdoor 2)
4. $X \leftarrow BG \rightarrow PE \rightarrow PI \rightarrow Y$ (backdoor 3)



Example 2 — Discrimination

Assume we want to investigate the **gender pay-gap** — i.e. whether, and, if so, to which extent it is caused by **discrimination**.

- But how **does** discrimination manifest?
  - Does discrimination directly lower income?
  - Does it affect the occupation chosen, hours worked, or promotions?

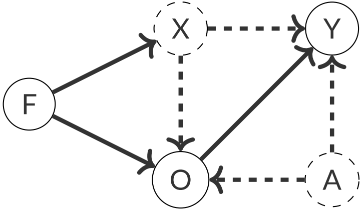
If we control for these factors, we will underestimate the effects of discrimination.

> *Some people may perceive that there is no gender pay-gap in their profession, especially after accounting for part-time work. This perspective is already conditional on occupation, level, hours, location, etc.*

Example 2 — To the drawing board

Let's consider a simple example — we are interested in the effect of

- gender-based (F) discrimination (X) on
- earnings (Y), accounting for
- occupation (O) and
- aptitude (A).

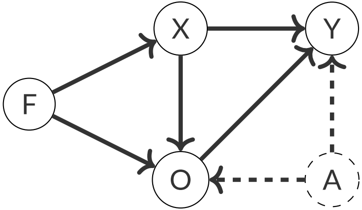


Example 2 — To the drawing board

Let's consider a simple example — we are interested in the effect of

- gender-based (F) discrimination (X) on
- earnings (Y), accounting for
- occupation (O) and
- aptitude (A).

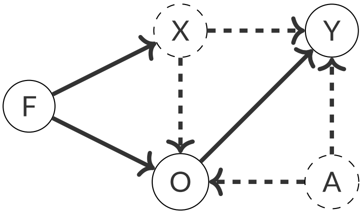
We will assume that we can observe and measure discrimination, but not aptitude.



Example 2 — Enumerating paths

The paths between X and Y are

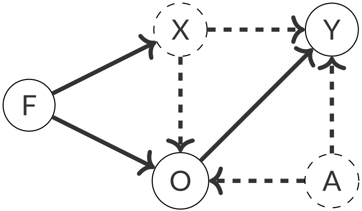
1. $X \rightarrow Y$,
2. $X \rightarrow O \rightarrow Y$,
3. $X \rightarrow O \leftarrow A \rightarrow Y$,
4. $X \leftarrow F \rightarrow O \rightarrow Y$,
5. $X \leftarrow F \rightarrow O \leftarrow A$.



Example 2 — Enumerating paths

The paths between X and Y are

1. $X \to Y$,
2. $X \to O \to Y$,
3. $X \to O \leftarrow A \to Y$,
4. $X \leftarrow F \to O \to Y$,
5. $X \leftarrow F \to O \leftarrow A$.



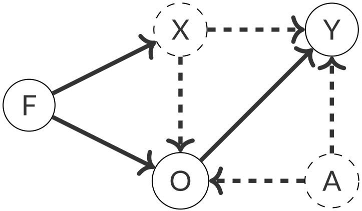
Consider these models to isolate paths 1 and 2.

- $Y \sim F$ – we get a compound effect of X and O (1, 2 and 4).

Example 2 — Enumerating paths

The paths between X and Y are

1. $X \to Y$,
2. $X \to O \to Y$,
3. $X \to O \leftarrow A \to Y$,
4. $X \leftarrow F \to O \to Y$,
5. $X \leftarrow F \to O \leftarrow A$.



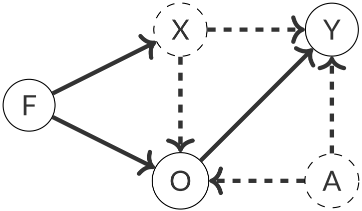
Consider these models to isolate paths 1 and 2.

- $Y \sim F$ – we get a compound effect of X and O (1, 2 and 4).
- $Y \sim X$ – we get the effects of X (1 and 2), but they are confounded by F (4).

Example 2 — Enumerating paths

The paths between X and Y are

1. $X \to Y$,
2. $X \to O \to Y$,
3. $X \to O \leftarrow A \to Y$,
4. $X \leftarrow F \to O \to Y$,
5. $X \leftarrow F \to O \leftarrow A$.

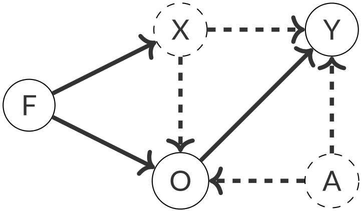


Consider these models to isolate paths 1 and 2.

- $Y \sim F$ – we get a compound effect of X and O (1, 2 and 4).
- $Y \sim X$ – we get the effects of X (1 and 2), but they are confounded by F (4).
- $Y \sim X, O$ – we get rid of the confounder F (4), and separate the effects of X (1 and 2), but are now confounded by A (3 and 5).

The paths between X and Y are

1. $X \to Y$,
2. $X \to O \to Y$,
3. $X \to O \leftarrow A \to Y$,
4. $X \leftarrow F \to O \to Y$,
5. $X \leftarrow F \to O \leftarrow A$.



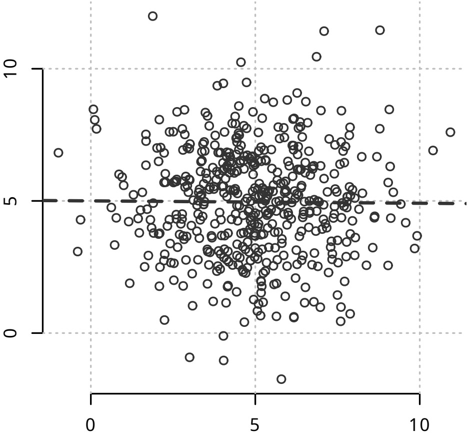
Consider these models to isolate paths 1 and 2.

- $Y \sim F$ – we get a compound effect of X and O (1, 2 and 4).
- $Y \sim X$ – we get the effects of X (1 and 2), but they are confounded by F (4).
- $Y \sim X, O$ – we get rid of the confounder F (4), and separate the effects of X (1 and 2), but are now confounded by A (3 and 5).

Without A, we cannot isolate the causal effect of X on Y in this model. DAGs can **highlight what cannot be done**.

Example 3 — Berkson's paradox

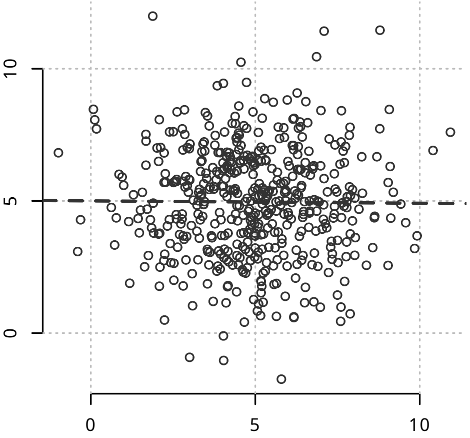
Your friend Alex postulates that, based on dating experience, *nice men are less handsome than rude ones*. You collect the data below, and find no correlation.



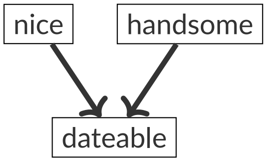
- Why could Alex still be right?

Example 3 — Berkson's paradox

Your friend Alex postulates that, based on dating experience, *nice men are less handsome than rude ones*. You collect the data below, and find no correlation.

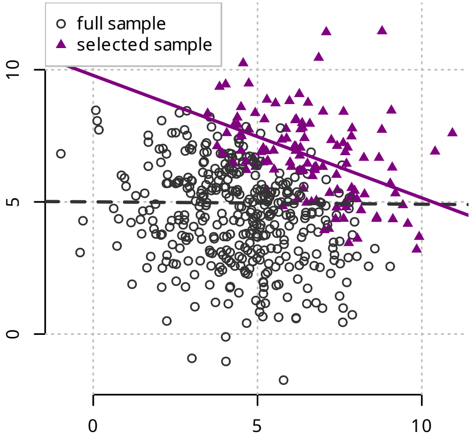


- Why could Alex still be right?
- Alex only dates someone if they are particularly nice and/or handsome.

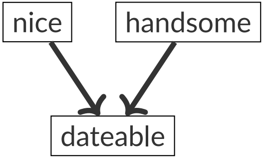


Example 3 — Berkson's paradox

Your friend Alex postulates that, based on dating experience, *nice men are less handsome than rude ones*. You collect the data below, and find no correlation.



- Why could Alex still be right?
- Alex only dates someone if they are particularly nice and/or handsome.
- *Dating experience is a collider* — conditioning on it causes bias.



Resources

- These slides are inspired by Cunningham (2021), who has a chapter on DAGs.
- Causal inference with DAGs is **covered comprehensively** by Pearl (2009).
- 'The Book of Why' (Pearl and Mackenzie 2018) takes a **more accessible** approach, covering the subject for a general audience.
- Imbens (2020) *reviews* DAG and potential outcome approaches to causality, with a focus on empiricial applications in economics.

You can create DAGs with pen and paper or specialised software, such as DAGitty or ggdag, or more general diagrams with PGF/TikZ in LaTeX, and diagrams.net / draw.io.

References i

Anderson, T. W., and Herman Rubin. 1949. “Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations.” *Annals of Mathematical Statistics* 20 (1): 46–63. https://doi.org/10.1214/aoms/1177730090.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. “Weak Instruments in Instrumental Variables Regression: Theory and Practice.” *Annual Review of Economics* 11 (1): 727–53. https://doi.org/10.1146/annurev-economics-080218-025643.

Angrist, Joshua D., Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. 2017. “Economic Research Evolves: Fields and Styles.” *American Economic Review* 107 (5): 293–97. https://doi.org/10.1257/aer.p20171117.

Angrist, Joshua D., and Alan B. Krueger. 2001. “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments.” *Journal of Economic Perspectives* 15 (4): 69–85. https://doi.org/10.1257/jep.15.4.69.

References ii

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30. https://doi.org/10.1257/jep.24.2.3.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32. https://doi.org/10.1257/jep.31.2.3.

———. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. “Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak.” *Journal of the American Statistical Association* 90 (430): 443–50. https://doi.org/10.1080/01621459.1995.10476536.

Buckles, Kasey S., and Daniel M. Hungerman. 2013. “Season of Birth and Later Outcomes: Old Questions, New Answers.” *Review of Economics and Statistics* 95 (3): 711–24. https://doi.org/10.1162/REST_a_00314.

Cunningham, Scott. 2021. *Causal Inference*. New Haven, CT, USA: Yale University Press. https://doi.org/10.12987/9780300255881.

Hamermesh, Daniel S. 2013. “Six Decades of Top Economics Publishing: Who and How?” *Journal of Economic Literature* 51 (1): 162–72. https://doi.org/10.1257/jel.51.1.162.

Imbens, Guido W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58 (4): 1129–79. https://doi.org/10.1257/jel.20191597.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1.

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27 (4): 435–54. https://doi.org/10.1017/pan.2019.11.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43. https://www.jstor.org/stable/1803924.

Pearl, Judea. 2009. *Causality*. *Cambridge Core*. Cambridge, England, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

Steel, Mark F. J. 2020. "Model Averaging and Its Use in Economics." *Journal of Economic Literature* 58 (3): 644–719. https://doi.org/10.1257/jel.20191385.

Threats to causal identification

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.

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* relates the analysis to the investigated theoretical construct.
- *Content validity* relates the analysed aspects to the relevant real-world aspects.
- *Predictive validity* concerns the utility for prediction.

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* relates the analysis to the investigated theoretical construct.
- *Content validity* relates the analysed aspects to the relevant real-world aspects.
- *Predictive validity* concerns the utility for prediction.

- **External validity** determines whether an insight can be *generalised*.
- **Internal validity** qualifies the *causal interpretation* of an inference.

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.*

External validity

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

External validity

External validity is the validity of an analysis *outside its own context*, telling us whether **findings can be generalised** 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 experiment may control important factors tightly.
  - A poor analysis limits what we learn at all.

External validity and testing code

```
what_day_is_it <- function() return("Monday")
```

I tested this function several times when I wrote it, and it worked every time.

Threats to external validity

Sample size and population

- The individuals in your sample may not represent the population —
  - e.g. a US study on unemployment may not generalise to Austria.
- Your sample size may be too small for the issue —
  - small or rare effects could be too small to measure or could not appear.

Situations

- Your analysis may be specific to a point in time —
  - e.g. due to politics, weather, or other circumstances.
- Insights may be bound to a specific location —
  - geography may affect how the analysis turns out.

Dealing with external validity

We can often solve issues with external validity by **reprocessing** the collected data.

Generalisability and imbalance

Age plays an important role in vaccine effectiveness. If individuals in the sample are younger than the overall population, insights from a study may be biased. To fix this, we could re-weigh the age-specific effect using age distribution of the population.

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

> *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, ...*

Learning 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, perhaps, with different methods.

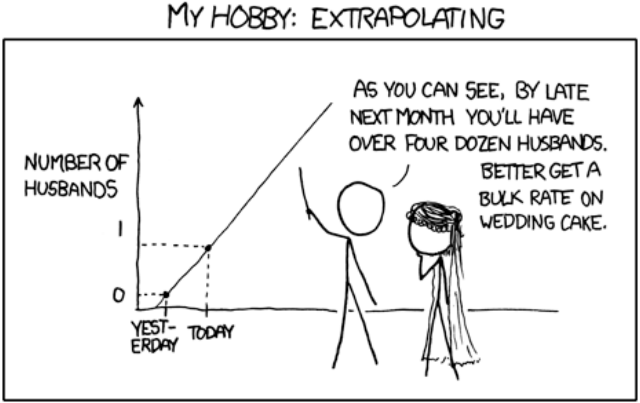


Figure 8: <xkcd.com>.

Internal validity

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

Internal validity

Internal validity is the validity of an analysis *within its own context*, i.e. 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*.
  - What could have gone wrong during an experiment?
  - What other explanations do we have for a correlation?

Occam's razor, or the principle of parsimony

There may be incomprehensibly many alternatives for each explanation. The idea of **Occam's razor** is to give preference to the *simplest explanation* (that cannot be refuted), i.e. the one with the fewest parameters and/or assumptions.

Revisiting the Gauss-Markov theorem

Ordingary 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. $\mathbb{E}[\mathbf{e}|\mathbf{X}] = \mathbb{E}[\mathbf{e}] = 0$.
- The model is *linear in parameters*, e.g. $f(\mathbf{X}) = \beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_K \mathbf{x}_K$.
- No *perfect collinearity*, i.e. $\mathbf{X}$ has full rank and we can compute $(\mathbf{X}'\mathbf{X})^{-1}$.
- *Homoskedasticity* and no *serial correlation*, i.e. $\mathbb{V}(\mathbf{e}|\mathbf{X}) = \mathbf{I}\sigma^2$.

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* (that is focused on the expectation). The exogeneity assumption $\mathbb{E}[\mathbf{e}|\mathbf{X}] = 0$ is sometimes substituted with **weak exogeneity** – $\mathrm{Cov}\left(\mathbf{X}, \mathbf{e}\right) = 0$. This guarantees consistency, but not unbiasedness of the estimator.

A failure of exogeneity is called **endogeneity** and causes bias and inconsistency by confounding the effects of our regressors $\mathbf{X}$ and the true errors $\mathbf{e}$ on $\mathbf{y}$.

Parameter bias and consistency

An estimate $\hat{\theta}$ is **unbiased** if $\mathbb{E}\left[\hat{\theta}\right] = \theta$ ▸ See proof for OLS. It is **consistent** if it converges in probability to the true parameter with increasing data, i.e.

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

The effect of endogeneity

Consider the effect of adjusting $\mathbf{x}_1$ to $\mathbf{x}_1^*$ – we have

$$\mathbb{E}\left[\mathbf{y}|\mathbf{X}^*\right] - \mathbb{E}\left[\mathbf{y}|\mathbf{X}\right] = \beta_1(\mathbf{x}_1^* - \mathbf{x}_1) + (\mathbb{E}[\mathbf{e}|\mathbf{X}^*] - \mathbb{E}[\mathbf{e}|\mathbf{X}]).$$

Under exogeneity, we get the correct effect since the second term is zero.

The effect of endogeneity

Consider the effect of adjusting $\mathbf{x}_1$ to $\mathbf{x}_1^*$ – we have

$$\mathbb{E}\left[\mathbf{y}|\mathbf{X}^*\right] - \mathbb{E}\left[\mathbf{y}|\mathbf{X}\right] = \beta_1(\mathbf{x}_1^* - \mathbf{x}_1) + (\mathbb{E}[\mathbf{e}|\mathbf{X}^*] - \mathbb{E}[\mathbf{e}|\mathbf{X}]).$$

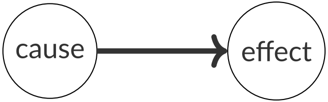
Under exogeneity, we get the correct effect since the second term is zero.

However, if $\mathbf{x}_1$ and $\mathbf{e}$ are correlated, we have $\mathbb{E}[\mathbf{e}|\mathbf{X}] = \theta_1\mathbf{x}_1 + \theta_0$ with $\theta_1 \neq 0$. We cannot separate the effects of observed factors (β_1) and unobserved ones (θ_1) and estimate

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

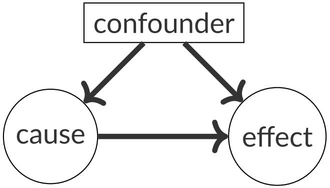
Threats to internal validity

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



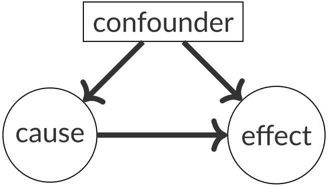
Confounders and omitted variables

We already learned that a **confounder**, a third variable that drives both the cause and effect, can cloud causal effects — if it is not accounted for.



Confounders and omitted variables

We already learned that a **confounder**, a third variable that drives both the cause and effect, can cloud causal effects — if it is not accounted for.



Consider the following *true* model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \mathbf{e}.$$

What are the implications of estimating $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \mathbf{e}$ instead?

Omitted variable bias

Bias from a confounder is also called **omitted variable bias**. It occurs if

1. The omitted variable is correlated with the regressors, and
2. it is also a determinant of $\mathbf{y}$.

In our example, the bias is given by

$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 + \frac{\mathrm{Cov}\left(\mathbf{x}_1, \mathbf{x}_2\right)}{\mathbb{V}\left(\mathbf{x}_1\right)}\beta_2.$$

Omitted variable bias

Bias from a confounder is also called **omitted variable bias**. It occurs if

1. The omitted variable is correlated with the regressors, and
2. it is also a determinant of $\mathbf{y}$.

In our example, the bias is given by

$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 + \frac{\operatorname{Cov}(\mathbf{x}_1, \mathbf{x}_2)}{\mathbb{V}(\mathbf{x}_1)}\beta_2.$$

Income, education, and ability

Assume you're interested in the effects of education ($\mathbf{x}_1$) on income ($\mathbf{y}$). Ability ($\mathbf{x}_2$) affects income ($\beta_2 \neq 0$) and is correlated with education ($\operatorname{Cov}(\mathbf{x}_1, \mathbf{x}_2) \neq 0$) — to *causally identify* β_1, we need to control for ability.

Omitted variables and proxies

Many (potentially) omitted variables *cannot be observed*. Instead, we may be able to use a **proxy variable**. Recall the true model:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \mathbf{e},$$

where we cannot observe $\mathbf{x}_2$. Instead, we could control for a proxy, $\mathbf{z}$, that fulfils

$$\mathbf{z} = \theta_0 + \theta_1 \mathbf{x}_2 + \mathbf{u}.$$

Ability and IQ

To *causally identify* the effect of education on income, we could use the results of an IQ test as proxy variable for ability.

Using proxy variables

In order to use a proxy variable to identify a causal effect, it must

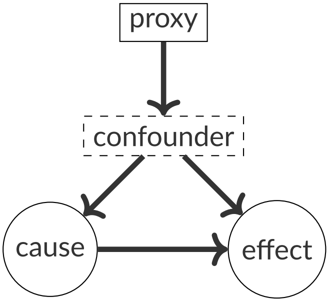
1. correlate with the omitted variable ($\theta_1 \neq 0$),
2. not correlate with other explanatory variables ($\text{Cov}\left(\mathbf{X}, \mathbf{u}\right) = 0$),
3. have no direct impact on the dependent variable ($\text{Cov}\left(\mathbf{z}, \mathbf{e}\right) = 0$).

Using proxy variables

In order 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 ($\text{Cov}(\mathbf{X}, \mathbf{u}) = 0$),
3. have no direct impact on the dependent variable ($\text{Cov}(\mathbf{z}, \mathbf{e}) = 0$).

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 useful type of proxy variables ('instrumental variables') at a later stage.



Selection bias

If our sample is not random, we may speak of **selection bias** — some subjects are more/less prone to be *selected for our sample*, thus distorting statistical insights.

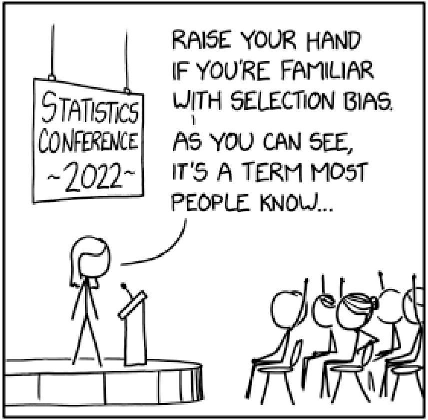


Figure 9: <xkcd.com>.

Types of selection biases

Selection bias is related to *sample issues* that may plague external validity, but also threatens (supposedly) in-sample inference. There are many *types of selection bias*; some notable examples are listed below.

- Doctors *prescribe treatment if* they think patients will *benefit*.
- 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 $p < .001$).
- We like to focus on evidence that makes sense to us and *confirms our priors*.
- Successful individuals give advice that is *conditional on their experience*.

Why could the introduction of steel helmets lead to higher rates of head injury?

Selection bias and spillover effects?

Figure 10: Rainbow crosswalk in Vienna, <wien.gv.at>.

Data issues

Data may be *subject to various issues*, e.g. due to errors during collection. This may affect our ability to analyse the data.

- Can we use **survey data** of savings or income?
- Are there potential issues when tracking *development over time*?
- Can we ignore *satellite images with clouds* when classifying forests?
- **How do we quantify** ability? How to measure gross domestic product?
- What could go wrong during **data collection**?
  - There will definitely be *typos*, there could be malice, and
  - our computers have *finite precision*, and cosmic rays can cause *bit flips*.

If you're on the fence – now is the time to argue about what can truly be known.

Missing data

Consider a true f describing a population of size N, but we only observe $M(< N)$ subjects. What can we learn from our subset?

- We are fine, if our M samples are a **random subset of the population** — the selection process is *ignorable*.
- Otherwise, there may be *selection bias* — we differentiate between
  1. **endogenous** sample selection, related to the dependent variable, and
  2. **exogenous** sample selection, based on explanatory or third variables.

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

Non-random missingness, censoring, and truncation

If there is a pattern to missingness we,

- may have to account for it to avoid bias (e.g. self-reported income), or
- can benefit from accounting for it (more information yields better estimates).

Censoring and truncation

When only parts of a sample are known, we speak of **censoring**. E.g. 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.

If samples where a value exceeds some threshold are missing, it is **truncated**.

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 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 β_{OLS} an **influential observation**, i, has a combination of *high residual* (e_i) and *high leverage* ($h_i = \left[\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]_{ii}$); its influence is given by

$$\beta - \beta_{(i)} = \frac{\left(\mathbf{X}'\mathbf{X}\right)^{-1} x_i' e_i}{1 - h_i}.$$

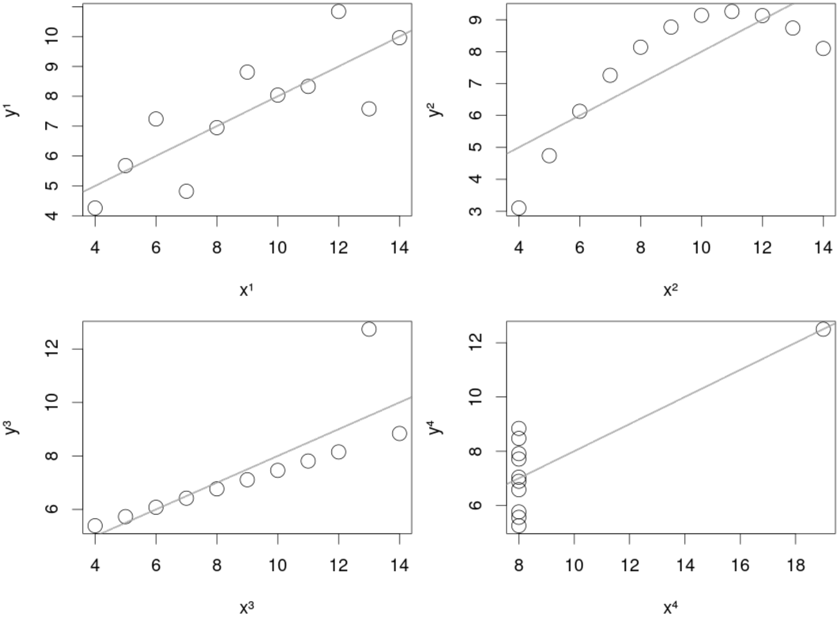


Figure 11: **Anscombe's quartet** — four different datasets with equal means, variance, and regression lines — emphasises the importance of in-depth analysis (see Anscombe, 1973).

Dealing with outliers and influential observations

We may discover outliers early on, when *exploring the data* (e.g. via summary statistics or plots) or later when evaluating the model (e.g. the residual values).

- It can be *tempting to remove outliers* from the analysis, as supposed errors.
- However, they may **convey the most interesting aspects** of the problem.
- A good model allows us to learn, and accommodates exceptional cases.

Dealing with outliers and influential observations

We may discover outliers early on, when *exploring the data* (e.g. via summary statistics or plots) or later when evaluating the model (e.g. the residual values).

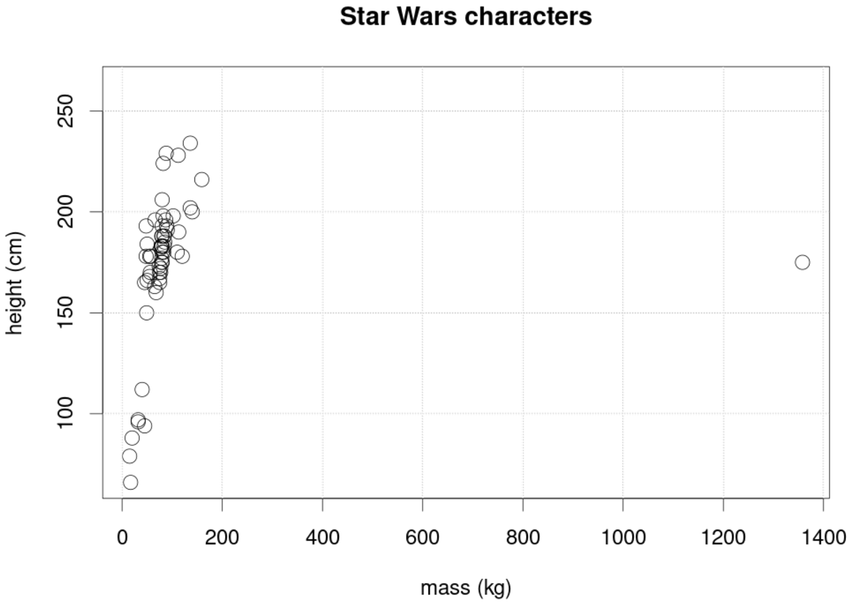
- It can be *tempting to remove outliers* from the analysis, as supposed errors.
- However, they may **convey the most interesting aspects** of the problem.
- A good model allows us to learn, and accommodates exceptional cases.

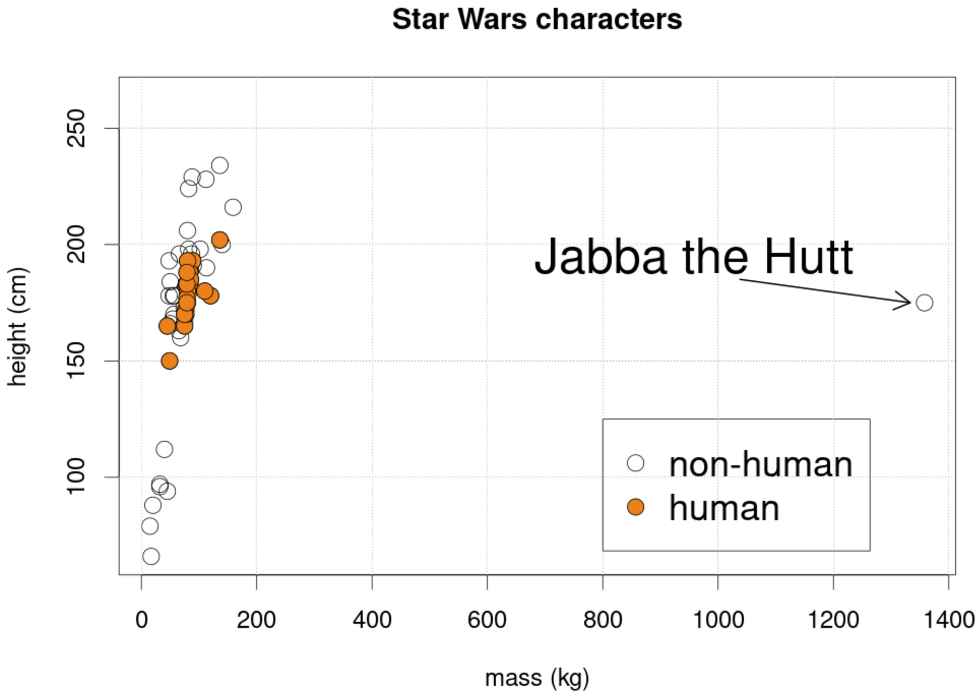
Robust methods

There are many estimation methods that are **more robust to few observations**. Examples include M-, S-, or **least absolute deviation** (LAD) estimation. There, we minimise *absolute residuals* as

$$\beta_{LAD} = \arg\min_{\beta} \left\{ |\mathbf{y} - \mathbf{X}\beta| \right\}.$$

Spotting an outlier — error or information?




Star Wars characters
Jabba the Hutt
non-human
human
height (cm)
mass (kg)
250
200
150
100
0
200
400
600
800
1000
1200
1400

Measurement errors in the dependent variable

Consider a true f with one explanatory variable, where the dependent variable ($\mathbf{y}$) is **observed with additional errors** ($\mathbf{u}$). We only observe $\mathbf{z} = \mathbf{y} + \mathbf{u}$, and estimate

$$\mathbf{z} = \beta\mathbf{x} + \mathbf{e} + \mathbf{u}.$$

What happens?

- If the error ($\mathbf{u}$) is random, we let $\tilde{\mathbf{e}} = \mathbf{e} + \mathbf{u}$ and can proceed as usual.
  - Measurement error is just more error – estimates are valid, but less precise.
- However, if the error is **not independent** of $\mathbf{x}$, we will suffer from bias.

Errors in the explanatory variable

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

$$\mathbf{y} = \beta\left(\mathbf{z} - \mathbf{u}\right) + \mathbf{e}.$$

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

$$\mathbf{y} = \beta\mathbf{z} + \tilde{\mathbf{e}}.$$

What happens?

- Our estimates will suffer from **attenuation bias**.

Attenuation bias

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

With measurement error in $\mathbf{x}$, we estimate $\mathbf{y} = \beta\mathbf{z} + \tilde{\mathbf{e}}$, and find that

$$\begin{aligned} \mathrm{Cov}\left(\mathbf{z}, \tilde{\mathbf{e}}\right) &= \mathrm{Cov}\left(\mathbf{z}, \mathbf{e} - \beta\mathbf{u}\right) \\ &= \mathrm{Cov}\left(\mathbf{x} + \mathbf{u}, \mathbf{e} - \beta\mathbf{u}\right) \neq 0. \end{aligned}$$

Attenuation bias

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

With measurement error in $\mathbf{x}$, we estimate $\mathbf{y} = \beta\mathbf{z} + \tilde{\mathbf{e}}$, and find that

$$\begin{aligned}\operatorname{Cov}(\mathbf{z}, \tilde{\mathbf{e}}) &= \operatorname{Cov}\left(\mathbf{z}, \mathbf{e} - \beta\mathbf{u}\right)\\ &= \operatorname{Cov}\left(\mathbf{x} + \mathbf{u}, \mathbf{e} - \beta\mathbf{u}\right) \neq 0.\end{aligned}$$

We may assume (1) $\operatorname{Cov}(\mathbf{x}, \mathbf{e}) = 0$, (2) $\operatorname{Cov}(\mathbf{x}, \mathbf{u}) = 0$, (3) $\operatorname{Cov}(\mathbf{u}, \mathbf{e}) = 0$, but

$$\operatorname{Cov}\left(\mathbf{u}, -\beta\mathbf{u}\right) = -\beta\,\mathbb{E}\left[\mathbf{u}^2\right].$$

Attenuation bias

Consider a *weaker version of ignorability* of the treatment — we want $\operatorname{Cov}(\mathbf{x}, \mathbf{e}) = 0$. With measurement error in $\mathbf{x}$, we estimate $\mathbf{y} = \beta\mathbf{z} + \tilde{\mathbf{e}}$, and find that

$$\begin{aligned}\operatorname{Cov}(\mathbf{z}, \tilde{\mathbf{e}}) &= \operatorname{Cov}\left(\mathbf{z}, \mathbf{e} - \beta\mathbf{u}\right) \\ &= \operatorname{Cov}\left(\mathbf{x} + \mathbf{u}, \mathbf{e} - \beta\mathbf{u}\right) \neq 0.\end{aligned}$$

We may assume (1) $\operatorname{Cov}(\mathbf{x}, \mathbf{e}) = 0$, (2) $\operatorname{Cov}(\mathbf{x}, \mathbf{u}) = 0$, (3) $\operatorname{Cov}(\mathbf{u}, \mathbf{e}) = 0$, but

$$\operatorname{Cov}\left(\mathbf{u}, -\beta\mathbf{u}\right) = -\beta\,\mathbb{E}\left[\mathbf{u}^2\right].$$

Here, the bias is given by ▸ See details

$$\mathbb{E}\left[\hat{\beta}\right] = \beta\frac{\sigma_{\mathbf{x}}^2}{\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{u}}^2}.$$

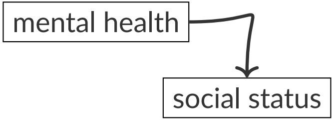
The bias goes towards zero ($a/(a+b) \leqslant 1$), and reduces the size of estimates.

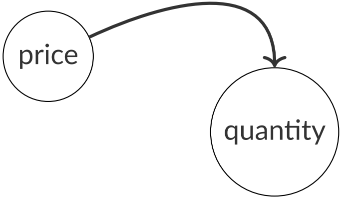
Simultaneity and reverse causality

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 $X \leftarrow Y$, and
- **simultaneity**, where $X \leftrightarrow Y$

With *pure reverse causality*, the issue is determining the direction of causation. With simultaneity, we want to disentangle the effects. Consider the following DAGs.



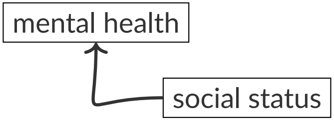


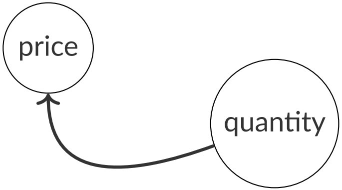
Simultaneity and reverse causality

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 $X \leftarrow Y$, and
- **simultaneity**, where $X \leftrightarrow Y$

With *pure reverse causality*, the issue is determining the direction of causation. With simultaneity, we want to disentangle the effects. Consider the following DAGs.



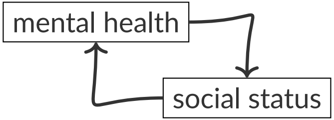


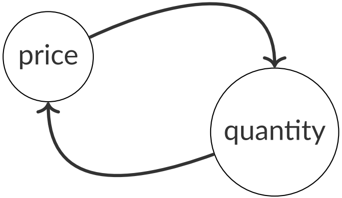
Simultaneity and reverse causality

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

- **reverse causality**, where $X \leftarrow Y$, and
- **simultaneity**, where $X \leftrightarrow Y$

With *pure reverse causality*, the issue is determining the direction of causation. With simultaneity, we want to disentangle the effects. Consider the following DGs.





Simultaneity in demand and supply

Consider the following supply and demand functions, driven by the price $\mathbf{p}$.

$$\mathbf{d} = \beta^d \mathbf{p} + \mathbf{e}^d,$$
$$\mathbf{s} = \beta^s \mathbf{p} + \mathbf{e}^s.$$

Usually, we **cannot measure supply and demand**. Instead, we observe the **quantity sold** $\mathbf{q}$ (from the equilibrium $\mathbf{q} = \mathbf{d} = \mathbf{s}$). We have

$$\mathbf{q} = \beta^d \mathbf{p} + \mathbf{e}^d = \beta^s \mathbf{p} + \mathbf{e}^s.$$

In this model, we **cannot 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 $\mathbf{p}$.

$$\begin{aligned}
\beta^d \mathbf{p} + \mathbf{e}^d &= \beta^s \mathbf{p} + \mathbf{e}^s \\
\beta^d \mathbf{p} &= \beta^s \mathbf{p} + \mathbf{e}^s - \mathbf{e}^d \\
\beta^d \mathbf{p} - \beta^s \mathbf{p} &= \mathbf{e}^s - \mathbf{e}^d \\
\mathbf{p} \left(\beta^d - \beta^s \right) &= \mathbf{e}^s - \mathbf{e}^d \\
\mathbf{p} &= \frac{\mathbf{e}^s - \mathbf{e}^d}{\beta^d - \beta^s}.
\end{aligned}$$

The effect of interest, $\mathbf{p}$, is a **function of the errors** — we can't disentangle its effects. If we regress $\mathbf{q}$ on $\mathbf{p}$, we can't tell whether the effect stems from the demand or supply function.

Structural equations and simultaneity bias

Consider the following **structural equations**

$$\mathbf{y} = \beta_1 \mathbf{z} + \beta_2 \mathbf{x}_1 + \mathbf{u},$$
$$\mathbf{z} = \theta_1 \mathbf{y} + \theta_2 \mathbf{x}_2 + \mathbf{v}.$$

We can derive a **reduced form** equation by solving for $\mathbf{z}$

$$\mathbf{z} = \gamma_1 \mathbf{x}_1 + \gamma_2 \mathbf{x}_2 + \boldsymbol{\varepsilon},$$

where

$$\gamma_1 = \frac{\theta_1 \beta_2}{1 - \theta_1 \beta_1} \quad \gamma_2 = \frac{\theta_2}{1 - \theta_1 \beta_1}$$
$$\boldsymbol{\varepsilon} = \frac{\theta_1 \mathbf{u} + \mathbf{v}}{1 - \theta_1 \beta_1}.$$

Simultaneity bias

The *reduced form* of our *structural equations* make two issues clear

- The reduced form parameters γ_1 and γ_2 are non-linear functions of the structural parameters, $\boldsymbol{\beta}, \boldsymbol{\theta}$.
- The structural parameters are not ignorable – $\mathbf{z}$ and $\mathbf{u}$ are correlated via $\mathbf{y}$.

In the reduced form, the error term is

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

where the correlation between $\theta_1 \mathbf{u}$ and the structural regressor $\mathbf{y}$ causes bias in

$$\mathbf{z} = \theta_1 \mathbf{y} + \theta_2 \mathbf{x}_2 + \mathbf{v}.$$

Outlook

Going forward, we will cover methods for dealing with these issues, including

- instrumental variable models, simultaneous equation models,
- matching procedures, flexible estimation methods, and quasi-experiments.

Other threats to internal validity

There are **countless other threats** to internal validity. These can generally be seen as variants of the concepts we already considered. Examples include

- historical bias, due to events outside our control,
- experimenter bias, where the conductor affects the experiment (inadvertently),
- diffusion, where spillover effects between subjects complicate inference,
- reversion to the mean, where larger samples tend to be less extreme.

Unbiased OLS estimates

The OLS estimate of β is unbiased under the Gauss-Markov assumptions.

$$
\begin{aligned}
\beta_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} && \text{fill in for } \mathbf{y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{e}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}\beta + \mathbf{X}'\mathbf{e}) && \text{split the sum} \\
&= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} \\
&= \mathbf{I}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} && \text{take expectation} \\
\mathbb{E}\left[\beta_{OLS}\right] &= \beta + \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}|\mathbf{X}\right] && \text{condition on } \mathbf{X} \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbb{E}[\mathbf{e}|\mathbf{X}] && \text{note that } \mathbb{E}[\mathbf{e}|\mathbf{X}] = 0 \\
&= \beta
\end{aligned}
$$

Go back

Attenuation bias

We show the attenuation bias from estimating $\mathbf{y} = \beta\mathbf{x} + \mathbf{e}$ with $\mathbf{z} = \mathbf{x} + \mathbf{u}$, i.e.

$$\mathbf{y} = \beta(\mathbf{z} - \mathbf{u}) + \mathbf{e} = \beta\mathbf{z} + \mathbf{e} - \beta\mathbf{u}, \mathbf{y} = \beta\mathbf{z} + \tilde{\mathbf{e}},$$

$$\hat{\beta} = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y} = \beta + (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\tilde{\mathbf{e}},$$

$$\hat{\beta} = \beta + (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{e} - (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\beta\mathbf{u},$$

$$\hat{\beta} = \beta + 0 - \beta\,(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{u},$$

$$\hat{\beta} = \beta - \beta\left((\mathbf{x} + \mathbf{u})'(\mathbf{x} + \mathbf{u})\right)^{-1}(\mathbf{x} + \mathbf{u})'\mathbf{u},$$

$$\mathbb{E}\left[\hat{\beta}\right] = \beta\left(1 - \frac{\mathrm{Cov}\,(\mathbf{x}, \mathbf{u}) + \mathbb{V}(\mathbf{u})}{\mathbb{V}(\mathbf{x}) + \mathrm{Cov}\,(\mathbf{x}, \mathbf{u}) + \mathbb{V}(\mathbf{u})}\right),$$

where we assume $\mathrm{Cov}\,(\mathbf{x}, \mathbf{u}) = 0$ to reformulate as $\mathbb{E}\left[\hat{\beta}\right] = \beta\,\sigma^2_{\mathbf{x}}\left(\sigma^2_{\mathbf{x}} + \sigma^2_{\mathbf{u}}\right)^{-1}$.

Go back

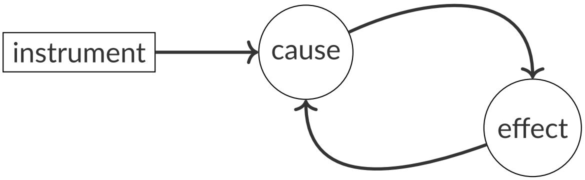
Instrumental variable regression

Why instrumental variables?

Instrumental variables (IV) allow us to *isolate a causal effect* from observational data. This is particularly important when

- there is *simultaneous causality*, or
- *omitted* variables are *unobtainable*.

We can use instruments with the **two-stage least squares** (2SLS) estimator, which allows us to obtain *consistent* estimates in such settings.



How does instrumental variable regression work?

Consider a model with *endogenous* regressors, $\mathbf{X}$, that are correlated with the error term, $\mathbf{e}$. With IV regression, we use *instrumental variables*, $\mathbf{Z}$, to **consistently** estimate the effects of the endogenous regressors.

For this to work, an *instrument* must satisfy two conditions.

1. **Exogeneity condition** — the instrument must be **uncorrelated with the error term**, $\mathbf{e}$; otherwise, the instrument is **invalid**.
2. **Relevance condition** — $\mathbf{X}$ and $\mathbf{Z}$ **must be correlated**; if the correlation is low or non-existent, the instrument is **weak**.

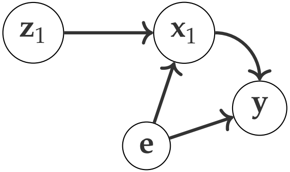
   We can test the relevance of an instrument, but not its exogeneity (we don't observe $\mathbf{e}$*).*

Illustration — first stage

Consider a model with one endogenous variable

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \mathbf{e},$$

where $\mathrm{Cov}(\mathbf{x}_1, \mathbf{e}) \neq 0$, e.g. due to an omitted variable.



In the **first stage** we use the *instrument* $\mathbf{z}_1$ to estimate

$$\begin{aligned} \mathbf{x}_1 &= \theta_0 + \theta_1 \mathbf{z}_1 + \mathbf{u} \\ &= \hat{\mathbf{x}}_1 + \mathbf{u}, \end{aligned}$$

i.e. we use the instrument to **predict the endogenous variable** $\hat{\mathbf{x}}_1$.

Illustration — second stage

In the **second stage**, we use our prediction $\hat{\mathbf{x}}_1$ instead of $\mathbf{x}_1$. If the instrument is valid, it is **exogenous by design** – it only depends on the instrument that is uncorrelated with $\mathbf{e}$. We estimate

$$\mathbf{y} = \beta_0 + \beta_1 \hat{\mathbf{x}}_1 + \mathbf{e},$$

and obtain a **biased, but consistent estimate** of β_1.

Recap — consistency

An estimate $\hat{\theta}$ is **consistent** if it converges in probability to the true parameter with increasing N, i.e. $\text{plim}_{N \to \infty} |\hat{\theta} - \theta| > \varepsilon = 0$ – also denoted by $\hat{\theta} \xrightarrow{\text{p}} \theta$.

The instrumental variable regression model

Consider a more general model

$$\mathbf{y} = \mathbf{U}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{U} = [\mathbf{W}\,\mathbf{X}]$, with $\text{Cov}\,(\mathbf{W}, \mathbf{e}) = 0$ and $\text{Cov}\,(\mathbf{X}, \mathbf{e}) \neq 0$ – we have

- $\mathbf{W}$ containing L **exogenous** regressors, and
- $\mathbf{X}$ with K **endogenous** regressors.

Assume we have M instrumental variables, in $\mathbf{Z}$.

- If $M \geqslant K$ we can *identify* the effect of the endogenous regressors.
- There is (at least) one instrument per endogenous variable to isolate its effect.

2SLS — the concept

The concept behind the 2SLS estimator is similar to before. First, we **regress** the *endogenous regressors*, $\mathbf{X}$, **on** the *exogenous variables* $\mathbf{W}$ and the *instruments* $\mathbf{Z}$. We will assume that there are no exogenous regressors for simplicity.

$$\mathbf{X} = \mathbf{Z}\delta + \mathbf{v},$$
$$\hat{\delta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.$$

We can now obtain a prediction $\hat{\mathbf{X}} = \mathbf{Z}'\hat{\delta}$ for the next stage. We can also express this prediction as $\hat{\mathbf{X}} = \mathbf{P_Z}\mathbf{X}$, where we use the **projection matrix**

$$\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'.$$

2SLS — the estimator

Next, we replace the endogenous variables with their prediction $\hat{\mathbf{X}} = \mathbf{P_Z X}$. We obtain the 2SLS estimator of the model as follows.

$$
\begin{aligned}
\mathbf{y} &= \hat{\mathbf{X}}\beta + \mathbf{e}, \\
\hat{\beta} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{P}_\mathbf{Z}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y} \\
\beta_{2SLS} &= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y}.
\end{aligned}
$$

This works since $\mathbf{P}_\mathbf{Z}$ is *symmetric* ($\mathbf{P}_\mathbf{Z}' = \mathbf{P}_\mathbf{Z}$) and *idempotent* ($\mathbf{P}_\mathbf{Z}\mathbf{P}_\mathbf{Z} = \mathbf{P}_\mathbf{Z}$).

The covariance matrix of the 2SLS estimator is $Cov(\beta_{2SLS}) = \sigma^2(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}$.

2SLS — the estimator

Next, we replace the endogenous variables with their prediction $\hat{\mathbf{X}} = \mathbf{P_Z X}$. We obtain the 2SLS estimator of the model as follows.

$$
\begin{aligned}
\mathbf{y} &= \hat{\mathbf{X}}\beta + \mathbf{e}, \\
\hat{\beta} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{P}_\mathbf{Z}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y} \\
\beta_{2SLS} &= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y}.
\end{aligned}
$$

This works since $\mathbf{P}_\mathbf{Z}$ is *symmetric* ($\mathbf{P}_\mathbf{Z}' = \mathbf{P}_\mathbf{Z}$) and *idempotent* ($\mathbf{P}_\mathbf{Z}\mathbf{P}_\mathbf{Z} = \mathbf{P}_\mathbf{Z}$).

The covariance matrix of the 2SLS estimator is $Cov(\beta_{2SLS}) = \sigma^2(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}$.

2SLS — the estimator

Next, we replace the endogenous variables with their prediction $\hat{\mathbf{X}} = \mathbf{P_Z X}$. We obtain the 2SLS estimator of the model as follows.

$$
\begin{aligned}
\mathbf{y} &= \hat{\mathbf{X}}\beta + \mathbf{e}, \\
\hat{\beta} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{P}_\mathbf{Z}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y} \\
\beta_{2SLS} &= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y}.
\end{aligned}
$$

This works since $\mathbf{P}_\mathbf{Z}$ is *symmetric* ($\mathbf{P}_\mathbf{Z}' = \mathbf{P}_\mathbf{Z}$) and *idempotent* ($\mathbf{P}_\mathbf{Z}\mathbf{P}_\mathbf{Z} = \mathbf{P}_\mathbf{Z}$).

The covariance matrix of the 2SLS estimator is $Cov(\beta_{2SLS}) = \sigma^2(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}$.

A special case — the IV estimator

When the coefficients are *just identified* ($M = K$), we can use the **IV estimator**

$$\beta_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

We can derive it by pre-multiplying $\mathbf{Z}'$ in the standard model.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{e} \\ \mathbf{Z}'\mathbf{y} &= \mathbf{Z}'\mathbf{X}\beta + \mathbf{Z}'\mathbf{e} \\ \mathbf{Z}'\mathbf{X}\beta_{IV} &= \mathbf{Z}'\mathbf{y} \\ \beta_{IV} &= \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y}. \end{aligned}$$

$M = K$ *means that the dimensions of* $\left(\mathbf{Z}'\mathbf{X}\right)^{-1} \in \mathbb{R}^{M\times K}$ *and* $\mathbf{Z}'\mathbf{y} \in \mathbb{R}^{M\times 1}$ *match.*

A special case — the IV estimator

When the coefficients are *just identified* ($M = K$), we can use the **IV estimator**

$$\beta_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.$$

We can derive it by pre-multiplying $\mathbf{Z}'$ in the standard model.

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\beta + \mathbf{e} \\
\mathbf{Z}'\mathbf{y} &= \mathbf{Z}'\mathbf{X}\beta + \mathbf{Z}'\mathbf{e} \\
\mathbf{Z}'\mathbf{X}\beta_{IV} &= \mathbf{Z}'\mathbf{y} \\
\beta_{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.
\end{aligned}$$

$M = K$ means that the dimensions of $(\mathbf{Z}'\mathbf{X})^{-1} \in \mathbb{R}^{M\times K}$ and $\mathbf{Z}'\mathbf{y} \in \mathbb{R}^{M\times 1}$ match.

Proving consistency of the IV estimator

$$
\begin{aligned}
\beta_{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \\
&= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{X}\beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{e} \\
&= \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{e}.
\end{aligned}
$$

Proving consistency of the IV estimator

$$
\begin{aligned}
\beta_{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \\
&= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{X}\beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{e} \\
&= \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{e} = \beta + \left(\mathbf{Z}'\mathbf{X}N^{-1}\right)^{-1}\mathbf{Z}'\mathbf{e}N^{-1}.
\end{aligned}
$$

We can factor in $\frac{N}{N}$, and from the *exogeneity* and *relevance* conditions we get

- $\text{Cov}\,(\mathbf{Z}, \mathbf{e}) = 0$ implying that $\mathbf{Z}'\mathbf{e}N^{-1} \xrightarrow{p} 0$,
- $\text{Cov}\,(\mathbf{Z}, \mathbf{X}) \neq 0$ implying that $\mathbf{Z}'\mathbf{X}N^{-1} \xrightarrow{p} c = \mathbb{E}[\mathbf{Z}'\mathbf{X}]$.

Proving consistency of the IV estimator

$$\begin{aligned}\beta_{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \\ &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{X}\beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{e} \\ &= \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{e}.\end{aligned}$$

We can factor in $\frac{N}{N}$, and from the *exogeneity* and *relevance* conditions we get

- $\text{Cov}\left(\mathbf{Z}, \mathbf{e}\right) = 0$ implying that $\mathbf{Z}'\mathbf{e}N^{-1} \xrightarrow{\text{p}} 0$,
- $\text{Cov}\left(\mathbf{Z}, \mathbf{X}\right) \neq 0$ implying that $\mathbf{Z}'\mathbf{X}N^{-1} \xrightarrow{\text{p}} c = \mathbb{E}[\mathbf{Z}'\mathbf{X}]$.

We see that $\beta_{IV} \xrightarrow{\text{p}} \beta + \frac{0}{c} = \beta$ as $N \to \infty$.

This proof relies on the fact that $\text{plim}\,\frac{a}{b} = \frac{\text{plim}\,a}{\text{plim}\,b}$, *which is not the case for expectations.*

Small-sample bias of the IV estimator

The IV estimator is consistent, but *almost certainly biased*.

$$\begin{aligned} \beta_{IV} &= \beta + \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{e} \\ \mathbb{E}\left[\beta_{IV}\right] &= \beta + \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{e}\right]. \end{aligned}$$

Small-sample bias of the IV estimator

The IV estimator is consistent, but *almost certainly biased*.

$$\beta_{IV} = \beta + \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{e}$$
$$\mathbb{E}\left[\beta_{IV}\right] = \beta + \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{e}\right].$$

We rely on $N \to \infty$, since we cannot separate the second term —

1. if we conditioned on $\mathbf{Z}$, we'd be stuck with $\left(\mathbf{Z}'\mathbf{X}\right)^{-1}$,
2. if we conditioned on $\mathbf{X}$ and $\mathbf{Z}$, we'd open up $\mathbb{E}[\mathbf{e}|\mathbf{Z},\mathbf{X}]$, as in

$$\beta + \mathbb{E}\left[\mathbb{E}\left[\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{e}|\mathbf{Z},\mathbf{X}\right]\right] = \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbb{E}[\mathbf{e}|\mathbf{Z},\mathbf{X}]\right].$$

Small-sample bias of the IV estimator

The IV estimator is consistent, but *almost certainly biased*.

$$\beta_{IV} = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{e}$$
$$\mathbb{E}\left[\beta_{IV}\right] = \beta + \mathbb{E}\left[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{e}\right].$$

We rely on $N \to \infty$, since we cannot separate the second term –

1. if we conditioned on $\mathbf{Z}$, we'd be stuck with $(\mathbf{Z}'\mathbf{X})^{-1}$,
2. if we conditioned on $\mathbf{X}$ and $\mathbf{Z}$, we'd open up $\mathbb{E}[\mathbf{e}|\mathbf{Z},\mathbf{X}]$, as in

$$\beta + \mathbb{E}\left[\mathbb{E}\left[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{e}|\mathbf{Z},\mathbf{X}\right]\right] = \mathbb{E}\left[(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbb{E}[\mathbf{e}|\mathbf{Z},\mathbf{X}]\right].$$

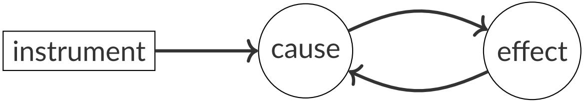
Summary — instrumental variables

We use **instrumental variables** to isolate the causal effect of an endogenous variable. The instruments must be

1. exogenous or **valid** (uncorrelated with the error term),
2. relevant or **strong** (correlated with the endogenous variable).

We need *at least* one instrument per endogenous variable, and use the *2SLS* or *IV* estimators to get **consistent**, but **biased** estimates. The size of the bias depends on

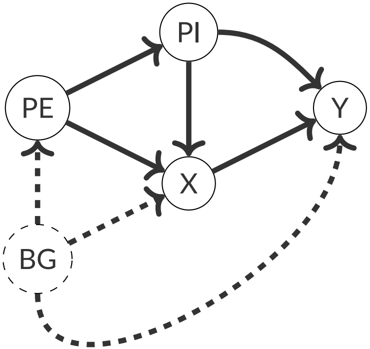
- the *exogeneity* (for $\mathbf{Z}'\mathbf{e}$) and *relevance* (for $\mathbf{Z}'\mathbf{X}$) of the instrument, and
- the size, N, of the sample.



Examples for instrumental variables

Consider the effect of **education** (X) on **income** (Y). Last time, we assumed

- the education (PI) and income (PI) of parents play a role,
- there is no causal effects of background factors (BF) such as ability.

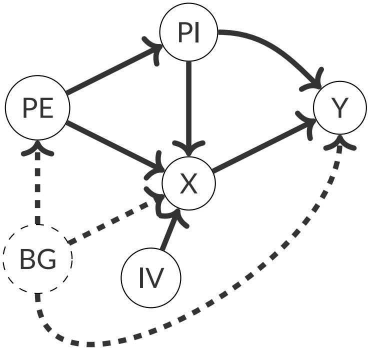


Examples for instrumental variables

Consider the effect of **education** (X) on **income** (Y). Last time, we assumed

- the education (PI) and income (PI) of parents play a role,
- there is no causal effects of background factors (BF) such as ability.

With an IV for education, we could bypass this restriction.



An IV for omitted variables

The background factors are an *omitted variable* that we *cannot obtain*. To distill a causal effect, Angrist and Krueger (2001) use the **quarter of birth** as an *instrument for education*. Why and how?

An IV for omitted variables

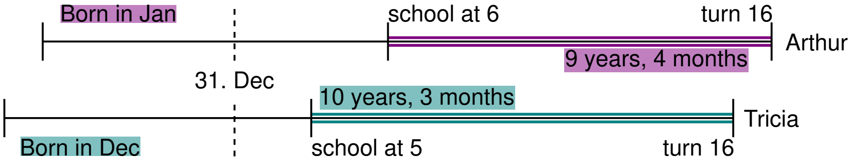
The background factors are an *omitted variable* that we *cannot obtain*. To distill a causal effect, Angrist and Krueger (2001) use the **quarter of birth** as an *instrument for education*. Why and how?

- In the United States, students must attend school from *the calendar year in which they turn six* until *their 16th* birthday.
- School entry is once per year, so *the length of schooling* at age 16 differs, and students who drop out at 16 create variation in education.

An IV for omitted variables

The background factors are an *omitted variable* that we *cannot obtain*. To distill a causal effect, Angrist and Krueger (2001) use the **quarter of birth** as an *instrument for education*. Why and how?

- In the United States, students must attend school from *the calendar year in which they turn six* until *their 16th* birthday.
- School entry is once per year, so *the length of schooling* at age 16 differs, and students who drop out at 16 create variation in education.



The date of birth as instrument

As an instrument, the quarter of birth, should

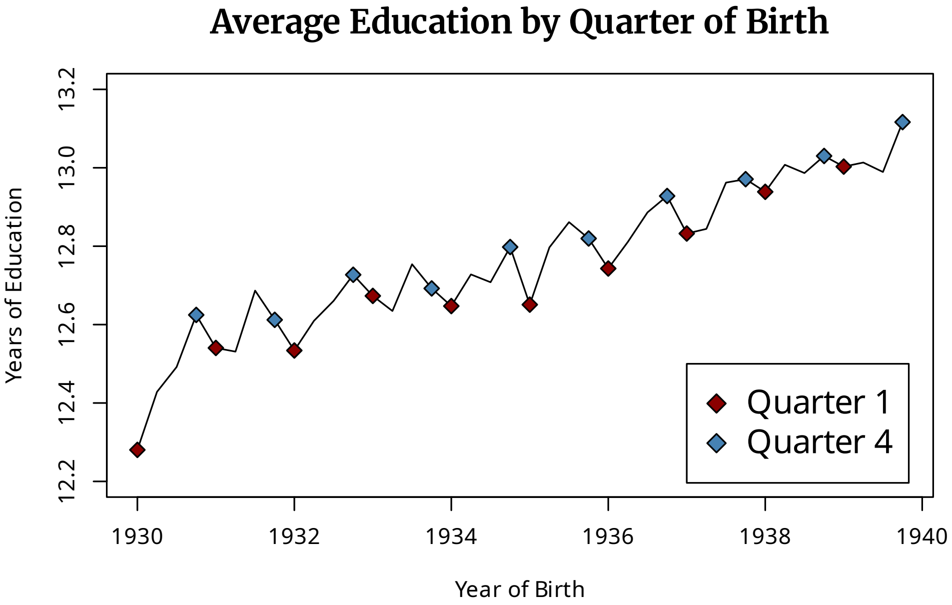
1. not affect income directly (be *valid*), and
2. affect education (be *relevant*).

Validity is always up for discussion, but regarding **relevance**, Angrist and Krueger (2001) *show that men born earlier* in the year tend to *have lower education* on average.

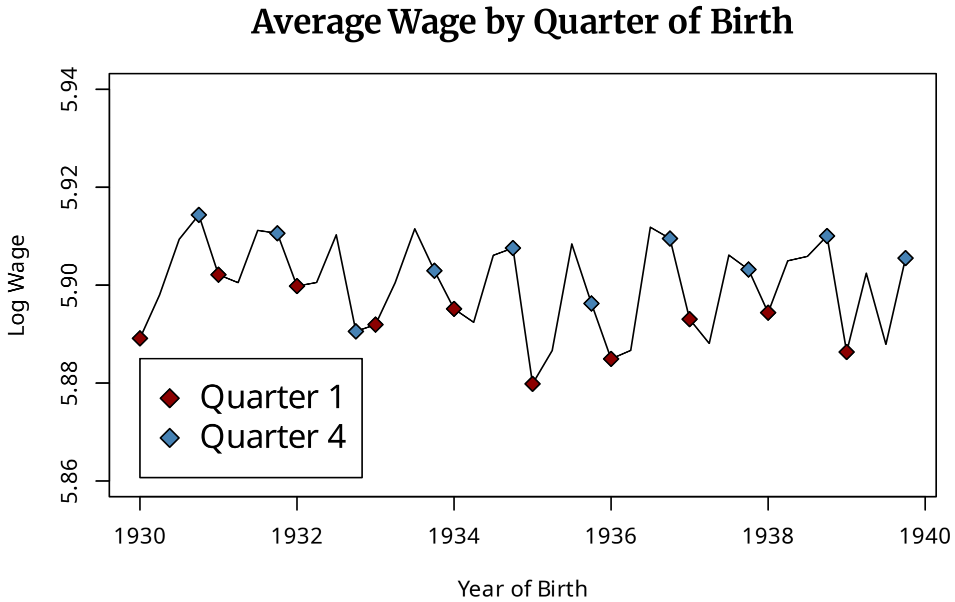
Let's replicate their work using census data from 1980 of 300k men in their 40s, we

- want to know whether the quarter of birth affects education, and
- whether this variation affects income.

Education and quarter of birth

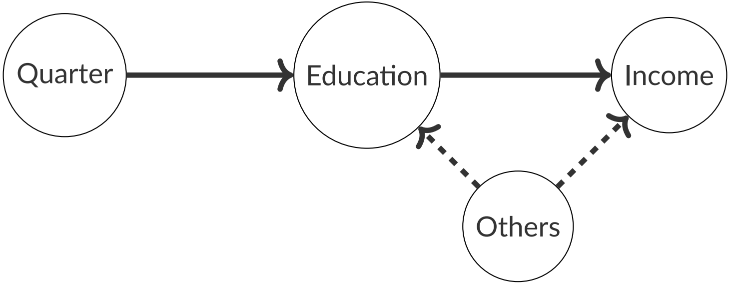


Wages and quarter of birth



- Men born earlier in the year tend to have less education.
- This seems to translate to a relation between wages and dates of birth.

Angrist and Krueger (2001) use these figures to *motivate that wage differences by quarter of birth* are **due to educational differences**.



Assessing the relevance of the instrument

Specifically, Angrist and Krueger (2001) use an interaction of quarter and year born as instruments. We can *assess the relevance* of instrument by

- computing the F statistic of the first stage,
- or the t value of a single instrument.

In our example, we find $F = 4.91$. The results of a simplified dummy-only version are

Education $\sim$	Estimate	Standard error
2nd quarter	0.057	0.0163
3rd quarter	0.113	0.0160
4th quarter	0.149	0.0162

Estimation results

We replicate a basic specification of Angrist and Krueger (2001) using OLS and 2SLS.

Wage $\sim$	LS	(SE)	IV	(SE)
Education	0.071	(0.0003)	0.089	(0.0161)

They **find similar estimates** when using OLS and IV models. If their instrument works as intended, we learn that

- *omitted variable bias* is relatively *limited*,
- *omitted variables* reduce the impact of education on wages.

In their paper, they extend this simple setup with covariates for ethnic group, region of residence, marital status, and age.

Assessing IV approaches

Do our our IV regression works as intended? Standard diagnostics include

- the *Durbin-Wu-Hausman test*, which compares the consistency of *OLS* estimator to the *less efficient, but consistent* IV estimator,
- F *and* t *statistics*, which indicate the *strength of instruments*.

Testing *exogeneity*, i.e. the validity, of instruments is not as straightforward.

- If we have multiple instruments, we can use *Sargan's* J *test* for overidentification.
- In general, we have to rely on intellectual work.

Choosing between IV and OLS

- 2SLS is **consistent**, *assuming valid and relevant instruments*.
- OLS is **more efficient**, and *may not suffer from endogeneity*.
- We prefer OLS, if there is no issue with endogeneity.

The *Durbin-Wu-Hausman test* compares an (assumed) consistent estimator to a more efficient one that may be inconsistent. The idea is to

1. Use the *first stage residuals* as explanatory in the original model.
2. Test the relevance of this variable, i.e. $\beta_j = 0$ for variable j.
3. If we reject the null hypothesis that $\beta_j = 0$, we **reject exogeneity of the explanatory** and thus, we *reject the consistency* of OLS.

Relevance of instruments

Weak instruments can be fatal for IV regression. Recall that

$$\beta_{IV} = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\,\mathbf{Z}'\mathbf{e},$$

where the second term should disappear as $N \to \infty$. If the instrument is

- irrelevant then $\mathbf{Z}'\mathbf{X}$ is small, which amplifies $\mathbf{Z}'\mathbf{e}$,
- completely irrelevant then the limit is not defined (we don't like 0^{-1}).

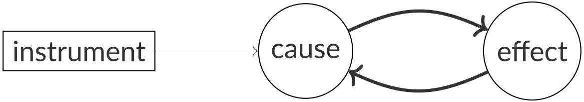
IV with weak instruments can be a lot worse than OLS, since

- **inconsistency** from *small violations* of the exogeneity condition is magnified,
- the **small-sample bias** of the 2SLS estimator is large,
- *confidence intervals* will be too tight.

Checking for weak instruments

Weak instruments can be a large problem. To assess them, we can check their explanatory power using F and t tests.

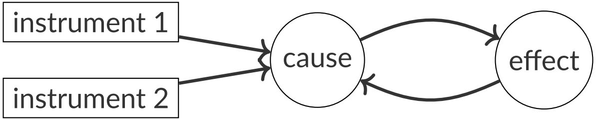
- $F > 10$ (and even $F > 100$) has been suggested as a rule of thumb.
- As an alternative, it makes sense to report *Anderson-Rubin confidence sets* (Anderson and Rubin 1949), which are robust to identification.



There is still a lot to learn about weak instruments, especially about multiple weak instruments for identification. For a recent review see Andrews, Stock, and Sun (2019).

Overidentification

If we have *more instruments than endogenous regressors*, we have **overidentification**.



With overidentification we can use *Sargan's J test*. The idea is to

- *compare estimates* using different instruments –
- if they are exogenous, estimates should be the same.
- The test's null hypothesis is that all instruments are valid.

The issue is that *we don't learn **which** instrument is not valid*, and estimates could always be similar or different by chance.

Assessing the results of Angrist and Krueger (2001)

Test	Statistic	p value
Weak instrument F test	4.907	0.000
Sargan's J test	25.442	0.655

Relevance

Bound, Jaeger, and Baker (1995) argue that **instruments are weak** (supported by $F = 4.9$). They show that an *irrelevant* instrument leads to similar results.

Exogeneity

Buckles and Hungerman (2013) see **exogeneity** as **violated** (not indicated by $J = 25.4$) – there is *seasonality in mother's characteristics*, which may affect the income of their children. Women that give birth in winter are *younger*, *less educated*, and are *less likely to be married*.

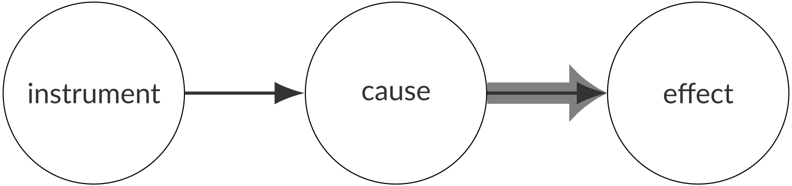
Understanding instruments

Instruments help us **isolate the causal effect** in a confounded relationship; we want

- *strong instruments*, so we have sufficient statistical power, and
- *exogenous instruments*.

Evaluating their **exogeneity** is arguably the complicated part, requiring

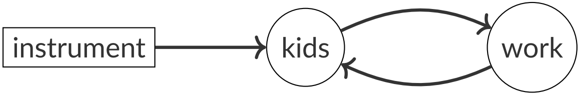
- **in-depth knowledge** about the phenomenon under study, and
- **creativity** for coming up with an instrument that is *confusing enough* for it to be *exogenous, yet relevant*.



Examples — family size and female labour

We want to learn about the way **family size** affects the **labour supply** of women — e.g. to better understand discrimination or design policies for more equality.

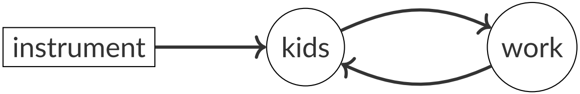
- Women with *more children* tend to *work less*.
- This is *unlikely to be exogenous* — kid's are not randomly assigned.



Examples — family size and female labour

We want to learn about the way **family size** affects the **labour supply** of women — e.g. to better understand discrimination or design policies for more equality.

- Women with *more children* tend to *work less*.
- This is *unlikely to be exogenous* — kid's are not randomly assigned.

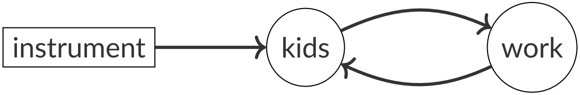


Now consider the fact that mothers whose **first two children are of the same gender** work less (our of the home) than others. **How is this related to labour supply**?

Examples — family size and female labour

We want to learn about the way **family size** affects the **labour supply** of women — e.g. to better understand discrimination or design policies for more equality.

- Women with *more children* tend to *work less*.
- This is *unlikely to be exogenous* — kid's are not randomly assigned.



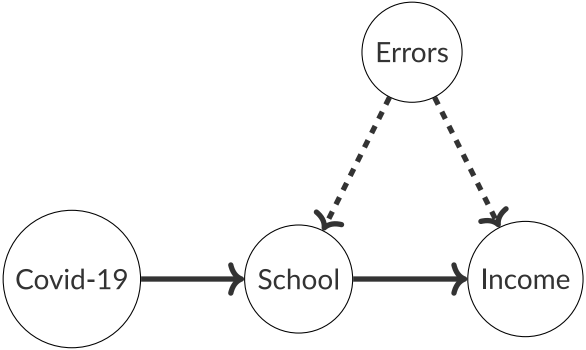
Now consider the fact that mothers whose **first two children are of the same gender** work less (our of the home) than others. **How is this related to labour supply**?

It **probably isn't** — however, it may be *related to family size.* Parents may have a preference for mixed genders and choose to have a third kid.

Examples — the elusiveness of instruments

Instruments are elusive, and ought to be specific to a situation — if they *are exogenous*, they *should not be relevant for most other* applications.

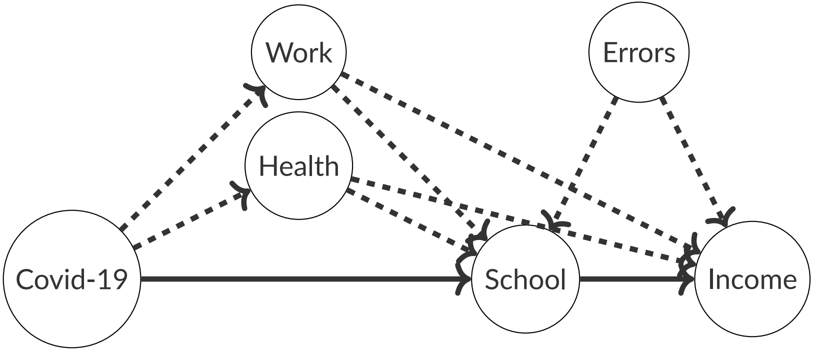
- The weather (e.g. rainfall) is a well-known, and commonly-used instrument.
- Covid-19 may seem like an instrument, e.g. for income effects of schooling.



Examples — the elusiveness of instruments

Instruments are elusive, and ought to be specific to a situation — if they *are exogenous*, they *should not be relevant for most other* applications.

- The weather (e.g. rainfall) is a well-known, and commonly-used instrument.
- Covid-19 may seem like an instrument, e.g. for income effects of schooling.



Examples — further ones

There are countless studies using interesting instruments. Some instruments can work in many settings — two notable examples are explained below.

Shift-share instruments

The *shift-share* (or Bartik) instrument combines aggregate changes (shifts) with initial values of individuals (shares), one of which has to be exogenous.

Judge fixed effects

If all subjects have to pass a *randomly assigned* judge (e.g.), who assigns a treatment, the *different characteristics* of judges will create random variation that we can use.

The name stems from the random assignment of judges in the United States.

Anderson, T. W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46–63. https://doi.org/10.1214/aoms/1177730090.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11 (1): 727–53. https://doi.org/10.1146/annurev-economics-080218-025643.

Angrist, Joshua D., Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. 2017. "Economic Research Evolves: Fields and Styles." *American Economic Review* 107 (5): 293–97. https://doi.org/10.1257/aer.p20171117.

Angrist, Joshua D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15 (4): 69–85. https://doi.org/10.1257/jep.15.4.69.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30. https://doi.org/10.1257/jep.24.2.3.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32. https://doi.org/10.1257/jep.31.2.3.

———. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

References iii

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90 (430): 443–50. https://doi.org/10.1080/01621459.1995.10476536.

Buckles, Kasey S., and Daniel M. Hungerman. 2013. "Season of Birth and Later Outcomes: Old Questions, New Answers." *Review of Economics and Statistics* 95 (3): 711–24. https://doi.org/10.1162/REST_a_00314.

Cunningham, Scott. 2021. *Causal Inference.* New Haven, CT, USA: Yale University Press. https://doi.org/10.12987/9780300255881.

Hamermesh, Daniel S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature* 51 (1): 162–72. https://doi.org/10.1257/jel.51.1.162.

Imbens, Guido W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58 (4): 1129–79. https://doi.org/10.1257/jel.20191597.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1.

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27 (4): 435–54. https://doi.org/10.1017/pan.2019.11.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43. https://www.jstor.org/stable/1803924.

Pearl, Judea. 2009. *Causality*. *Cambridge Core*. Cambridge, England, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

Steel, Mark F. J. 2020. "Model Averaging and Its Use in Economics." *Journal of Economic Literature* 58 (3): 644–719. https://doi.org/10.1257/jel.20191385.

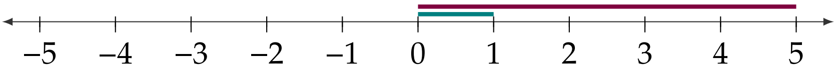
Non-linear models

Limited dependent variables

So far, we have only dealt with **continuous and unconstrained** *dependent variables*, i.e. $Y \in \mathbb{R}$, but many interesting variables are **limited** in some form, e.g.

- probabilities range from zero to one,
- GDP is a positive variable.

We can treat these **limited variables** as approximately continuous, but this may cause severe issues. Instead, we can turn to specialised *limited dependent variable* (LDV) models.



Examples for LDVs

We may distinguish between **regression** and **classification** tasks.

Examples for LDVs

We may distinguish between **regression** and **classification** tasks.

Classification

We speak of a classification model, if the outcome is

- *binary* (e.g. passed or failed, good boy or not),
- *categorical* (e.g. nationality, breed of dog),
  - *ordinal* (e.g. good – okay – bad).

Examples for LDVs

We may distinguish between **regression** and **classification** tasks.

Classification

We speak of a classification model, if the outcome is

- *binary* (e.g. passed or failed, good boy or not),
- *categorical* (e.g. nationality, breed of dog),
    - *ordinal* (e.g. good – okay – bad).

Regression

We speak of a regression model, if the outcome is

- *censored*, *truncated*, or *positive* (e.g. wages, wealth, time, forest loss),
- *count data* (e.g. the number of votes, days since an accident),

Regression is generally used in a much broader sense, and may encompass classification.

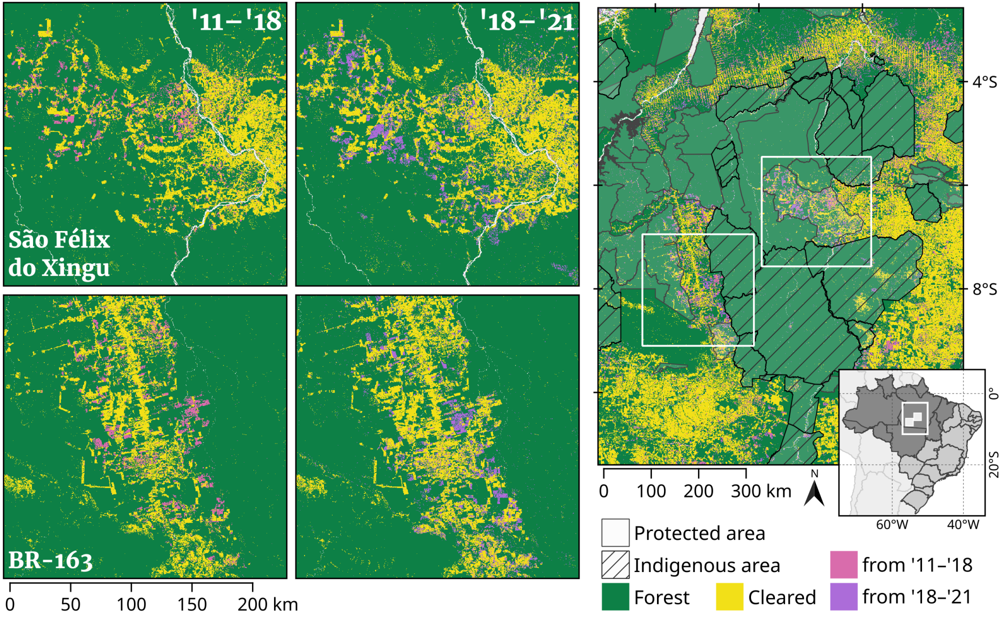


Figure 12: Land use change in the Brazilian Amazon.

The linear probability model

First, consider the implications of using the **linear probability model** (LPM)

$$\mathbf{y} = f(\mathbf{X}) + \mathbf{e} = \mathbf{X}\beta + \mathbf{e},$$

where the dependent variable is a **probability**, i.e. $\mathbf{y} \in [0, 1]$.

The linear probability model

First, consider the implications of using the **linear probability model** (LPM)

$$\mathbf{y} = f(\mathbf{X}) + \mathbf{e} = \mathbf{X}\beta + \mathbf{e},$$

where the dependent variable is a **probability**, i.e. $\mathbf{y} \in [0, 1]$.

For a *binary dependent* the expected value is equal the probability that $y_i = 1$.

$$\mathbb{E}\left[\mathbf{y}\right] = 0 \cdot \mathbb{P}\left(\mathbf{y} = 0\right) + 1 \cdot \mathbb{P}\left(\mathbf{y} = 1\right).$$

Conditional on the regressor, $\mathbf{X}$, we have

$$\mathbb{E}\left[\mathbf{y} \,|\, \mathbf{X}\right] = \mathbb{P}\left(\mathbf{y} = 1 \,|\, \mathbf{X}\right) = \mathbf{X}\beta.$$

Understanding the LPM

$$\mathbb{P}\left(\mathbf{y} \mid \mathbf{X}\right) = \beta_0 + \mathbf{x}_1\beta_1 + \ldots + \mathbf{x}_K\beta_K.$$

The LPM implies that the coefficient β_j gives us the *expected absolute change of probability* if $\mathbf{x}_j$ is changed by 1. This *linearity assumption* can be a **major limitation**.

Consider, e.g., a model of the probability of a cell of **land being deforested**.

Understanding the LPM

$$\mathbb{P}\left(\mathbf{y} \mid \mathbf{X}\right) = \beta_0 + \mathbf{x}_1\beta_1 + \ldots + \mathbf{x}_K\beta_K.$$

The LPM implies that the coefficient β_j gives us the *expected absolute change of probability* if $\mathbf{x}_j$ is changed by 1. This *linearity assumption* can be a **major limitation**.

Consider, e.g., a model of the probability of a cell of **land being deforested**.
We have $Y \in \{0, 1\}$ and covariates on

- population density in the area,
- distance to the nearest city,
- distance to agricultural land,
- precipitation, and temperature.

Understanding the LPM

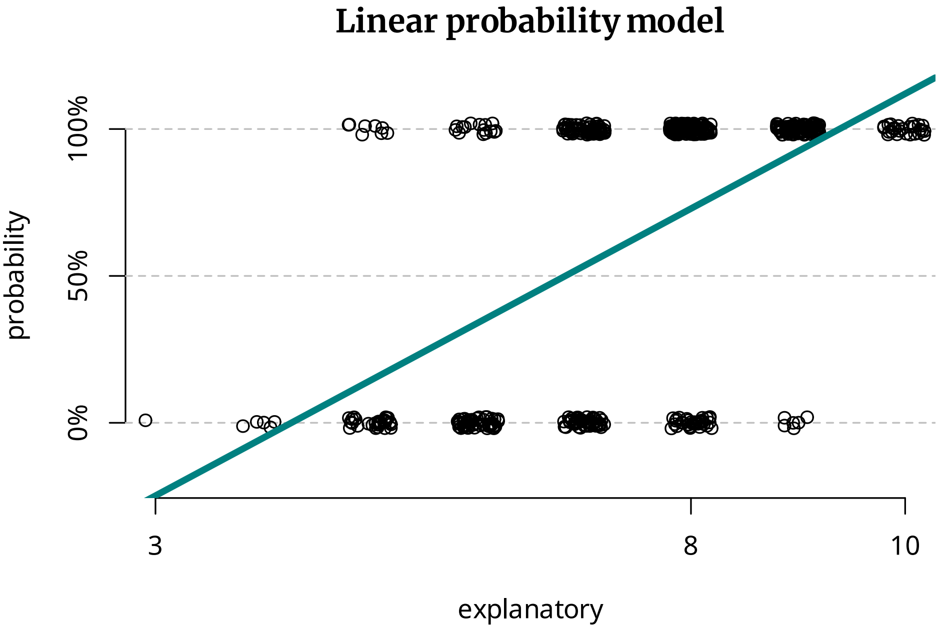
$$\mathbb{P}\left(\mathbf{y} \,|\, \mathbf{X}\right) = \beta_0 + \mathbf{x}_1\beta_1 + \ldots + \mathbf{x}_K\beta_K.$$

The LPM implies that the coefficient β_j gives us the *expected absolute change of probability* if $\mathbf{x}_j$ is changed by 1. This *linearity assumption* can be a **major limitation**.

Consider, e.g., a model of the probability of a cell of **land being deforested**. Or a turtle identifier. We have $Y \in \{0, 1\}$ and covariates on

- population density in the area,
- distance to the nearest city,
- distance to agricultural land,
- precipitation, and temperature.

Figure 13: What kind of turtle is this?


Linear probability model
probability
100%
50%
0%
3
8
10
explanatory

Modelling probabilities

When dealing with **probabilities**, the *linearity* assumption for f may be too strong – we need another approach.

- Consider a function G that satisfies $0 < G(z) < 1$.
- We could use G to adapt our model to

$$\mathbb{P}\left(\mathbf{y} \,|\, \mathbf{X}\right) = G(\mathbf{X}\beta).$$

This way, we can model a **latent variable**, $\mathbf{z} = \mathbf{X}\beta$, using a linear model, and link it to the dependent $\mathbf{y}$ via the *non-linear function* G, giving us $\mathbf{y} = G(\mathbf{z})$.

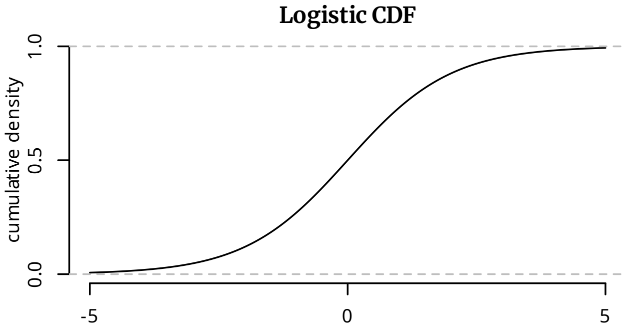
Link function

The inverse function $G^{-1}(z)$ is called the *link function*.

The logit model

For the *logit model*, we use the the cumulative distribution function (CDF) of a logistic variable — the *logistic function* — for G. The link function are *log-odds*, $\log \frac{p}{1-p}$.

$$G(z) = \frac{e^z}{e^z + 1}.$$

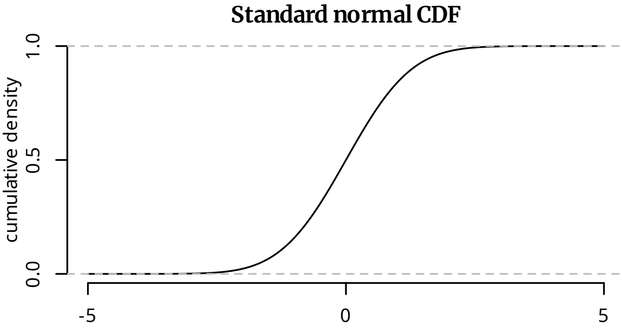


Probit model

For the *probit model*, we use the CDF of a *standard normal distribution*,

$$G(z) = \Phi(z) = \mathbb{P}(Z \leqslant z)\,, \text{ where } Z \sim \mathcal{N}(0,1),$$

which gives us the probability that the standard normal variable Z is smaller than z.



Interpretation

The interpretation of logit and probit models is not as straightforward as in linear models due to their non-linearity.

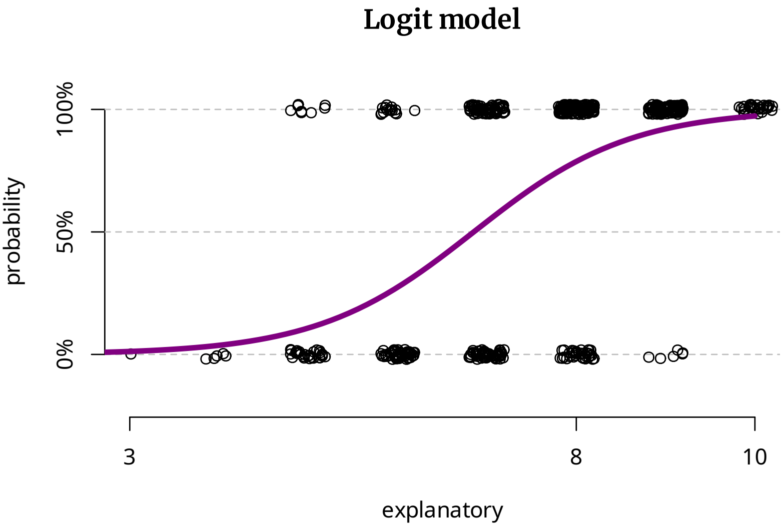
- We *can* interpret the
  - sign of coefficients, i.e. the direction of the expected change, and
  - significance of coefficients.

So if $\beta_j > 0$ we expect the probability to increase with $\mathbf{x}_j$ and vice versa.

However, we **cannot interpret the magnitude** of coefficients as magnitude of the effect of $\mathbf{X}$ on $\mathbf{y}$. Instead, it captures the effects of $\mathbf{X}$ on the latent $\mathbf{z}$, which we rarely care about.

Interpreting predictions

We can interpret *predicted probabilities* or differences in certain scenarios.



Partial effects

The problem with interpreting coefficients is that **partial effects** of $\mathbf{x}_j$ are *affected by all other variables*. Assume $\mathbf{x}_1$ is a dummy, then

$$\mathbb{P}\Big(\mathbf{y} \,|\, \mathbf{x}_1 = 1, \mathbf{x}_2, \cdot\Big) = G(\beta_0 + \beta_1 + \mathbf{x}_2\beta_2 + ...)$$
$$\mathbb{P}\Big(\mathbf{y} \,|\, \mathbf{x}_1 = 0, \mathbf{x}_2, \cdot\Big) = G(\beta_0 + \mathbf{x}_2\beta_2 + ...)$$

The *change depends on the level of* $\mathbf{x}_2$ and other variables. The same holds for continuous variables, with the partial effect given by

$$\frac{\partial\, \mathbb{P}\Big(\mathbf{y} \,|\, \mathbf{x}_j = x_j, \cdot\Big)}{\partial \mathbf{x}_j} = g(\mathbf{X}\beta)\, \beta_j,$$

where $g(z) = G'(z)$, i.e. the first derivative.

Reporting partial effects

We can use summary measures to help *interpret partial effects* in non-linear models.

Partial effect at the average

The *partial effect at the average* (PEA) is given by

$$g(\bar{\mathbf{X}}\hat{\beta})\,\hat{\beta}_j,$$

and gives partial effects where **explanatory variables are at their mean**.

Average partial effect

The *average partial effect* (APE) is given by

$$\frac{\sum_{j=1}^{N} g(\mathbf{X}\hat{\beta})}{N}\,\hat{\beta}_j.$$

We calculate the **partial effect for each observation** and take the average.

Inference — testing

To *test the significance* of single coefficients, we can use t values. For *multiple coefficients* we can use the **likelihood ratio** test

$$\mathrm{LR} = 2(\log \mathcal{L}_u - \log \mathcal{L}_r).$$

We compare the **likelihood** of the unrestricted ($\mathcal{L}_u$) and restricted ($\mathcal{L}_r$) models, where the models are required to be *nested* (the complex model nests the simpler one).

Likelihood

The likelihood function is the *joint probability of the observed data*, viewed as a function of the parameters.

The statistic converges asymptotically to a χ^2 distribution – if the null hypothesis happens to be true. Finite sample behaviour is generally unknown.

Inference — comparing models

We can compare model specifications using

- R^2, the proportion of explained variance,
  - for non-linear models there are various pseudo R^2 measures,
- the likelihood, $\mathcal{L}$, of a given model, or
- **information criteria** (IC).

Inference — comparing models

We can compare model specifications using

- R^2, the proportion of explained variance,
  - for non-linear models there are various pseudo R^2 measures,
- the likelihood, $\mathcal{L}$, of a given model, or
- **information criteria** (IC).

Many measures of model fit *always increase with complexity* — IC prefer parsimony.

Akaike information criterion

$$\text{AIC} = 2K - 2\log\hat{\mathcal{L}}$$

Bayesian (or Schwarz) information criterion

$$\text{BIC} = K\log N - 2\log\hat{\mathcal{L}}$$

Other probability models

Probabilities are not the only **limited dependent variables**, and there is a range of other specialised models. This includes the

- **poisson model** for *count* variables,
  - e.g. $Y \in \{0, 1, 2, ...\}$ with votes,
- **tobit model** for *censored* variables,
  - e.g. $Y > 0$ with forest loss,
- **heckit model** for *non-random samples*,
  - which uses the *Heckman correction*, modeling the sampling probability,
- **multinomial probit/logit model** for *categorical* variables,
  - e.g. $Y \in \{\text{agree}, \text{disagree}, \text{unsure}\}$.

Count data

Count data takes on non-negative integer values ($0, 1, 2, ...$) and often has a substantial number of *zero outcomes* ('zero-inflated').

To build a model for this kind of data, we could

- think of a latent Normal variable behind Y,
- or use a *discrete* probability distribution.

Count data

Count data takes on non-negative integer values ($0, 1, 2, ...$) and often has a substantial number of *zero outcomes* (‘zero-inflated’).

To build a model for this kind of data, we could

- think of a latent Normal variable behind Y,
- or use a *discrete* probability distribution.

The **Poisson distribution** is an example; we can use it to express the probability that a *given number of events occurrs in a fixed interval.*

> *In 1898, Ladislaus Bortkiewicz used the Poisson distribution when investigating the number of soldiers in the Prussian army that were killed by horse kicks.*

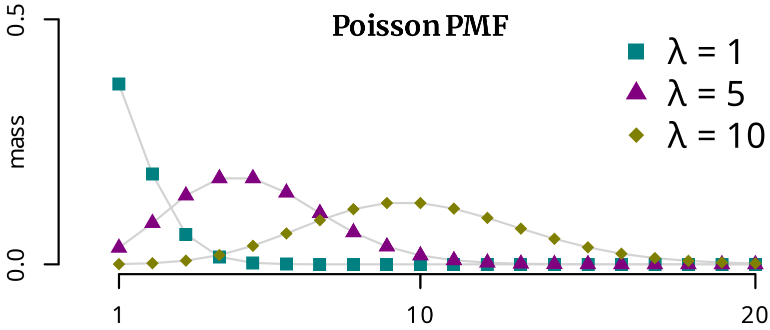
Figure 14: The Count von Count.

Poisson distribution

The *probability mass function* (PMF) of the **Poisson distribution** is

$$\mathbb{P}\big(Y = y_i \,|\, \lambda\big) = \frac{\lambda^{y_i} \exp^{-\lambda}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots,$$

where the parameter λ is also the expectation $\mathbb{E}[Y]$ and variance $\mathbb{V}(Y)$.



Poisson model

We generally expect that the expectation, i.e. the mean $\lambda = \mathbb{E}\left[\mathbf{y}\right]$, depends on other variables. Consider a *Poisson model* with dependent mean; let

$$\lambda = \mathbb{E}\left[\mathbf{y} \,|\, \mathbf{X}; \beta\right] = \exp\left\{\mathbf{X}\beta\right\},$$

where we use the exponential function to ensure that $\mathbb{E}\left[\mathbf{y} \,|\, \mathbf{X}\right] > 0$. We get

$$\mathbb{P}\left(Y = y_i \,|\, x_i; \beta\right) = \frac{\exp\left\{x_i\beta\right\}^{y_i} \exp^{-\exp\left\{x_i\beta\right\}}}{y_i!}, \quad y_i = 0, 1, 2, \dots,$$

describing the probability of each observation.

Censored and truncated data

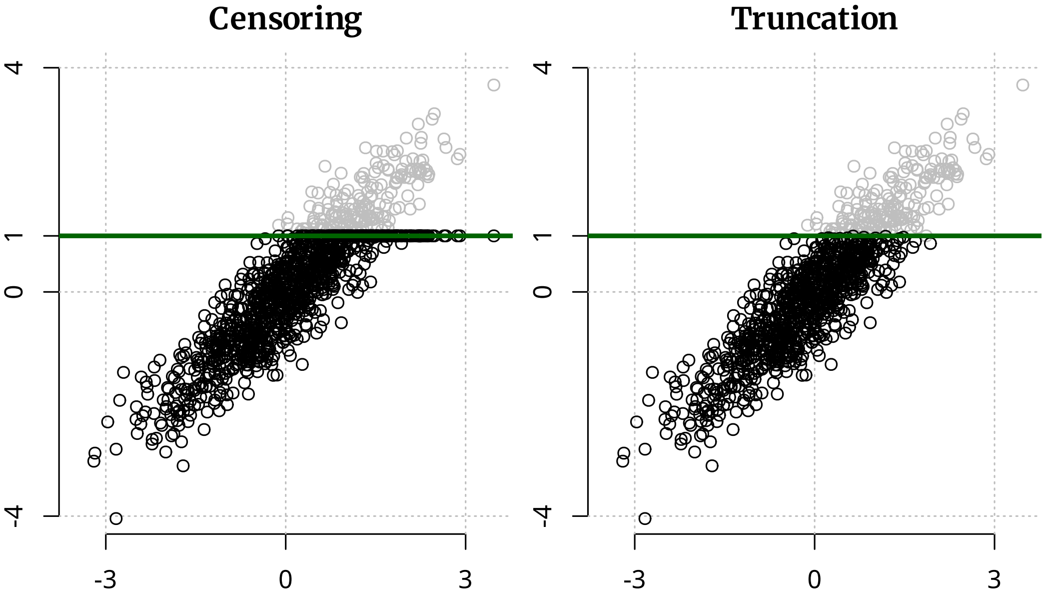
We speak of **censored** (truncated) data if the *data is censored* (truncated) at some *threshold* for some reason. This could be

- square meters in a 30m² cell ($Y \in [0, 30]$),
- wages ($Y \in [0, \infty]$),
- temperature in degree Celsius ($Y \in [-273.15, \infty]$), et cetera.

Censoring can be

- *absolute* (no values beyond the threshold), or
- *relative* (selection problem beyond the threshold).

It can happen by design or due to missing data. Censored (and especially truncated) data often has a substantial number of *observations at the threshold*.


Censoring
Truncation
4
1
0
-4
-3
0
3

Outlook

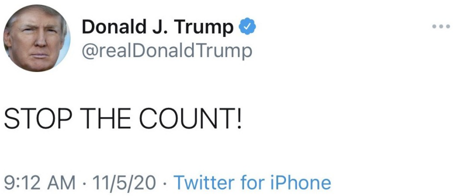
We covered a number of *limited dependent variables*, why they are important, and how we can **learn more** using more general, non-linear models.

Next up, we will learn how to conduct **maximum likelihood estimation**, which

- allows us to efficiently *estimate general models*,
- provides a connection to more advanced topics (such as shrinkage).

Afterwards, we'll proceed with more options and methods for **causal inference**, including *matching*, *quasi-experiments*, and *regression discontinuities*.

Figure 15: Poisson models are contentious.



Maximum likelihood estimation

Estimation of general models

We need a good way of **estimating** more *general models*, such as

$$\mathbf{y} = G(\mathbf{X}, \boldsymbol{\beta}) + \mathbf{e}.$$

These models (e.g. the logit model) are *not linear in parameters* — OLS isn't even BLUE.

- When minimising $\mathbf{e}'\mathbf{e}$, we have to consider K ($\boldsymbol{\beta} \in \mathbb{R}^K$) partial derivatives,
- $\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \beta_j}$ generally involves all $\boldsymbol{\beta}$, and there is no closed form solution.

Non-linear least squares

Non-linear least squares estimation is a conceptually straightforward approach. First, we *approximate with a linear model*, and the refine the estimates iteratively. However, estimates are generally *not unique* and *inefficient*.

Maximum likelihood estimation

Maximum likelihood (ML) estimation is a method for estimating parameters. It works by maximising a **likelihood function**, the *joint probability distribution of the data* as a function of the parameters, given by

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \mathbb{P}\Big(\mathbf{y} \,|\, X; \boldsymbol{\beta}\Big) .$$

- We set $\boldsymbol{\beta}_{ML}$ so the observed data is *most probable* within our model.
- The resulting ML estimator is *consistent*, *asymptotically normal*, and *asymptotically efficient* in most cases.

The likelihood $\mathcal{L}(\theta|X)$ itself is not a probability – we allow θ to vary, not X.

The ML estimator

For computational convenience, we usually work with the **log-likelihood**

$$\ell(\boldsymbol{\beta}) = \log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log \mathbb{P}\Big(\mathbf{y} \,|\, \mathbf{X}; \boldsymbol{\beta}\Big).$$

- $\boldsymbol{\beta}_{ML}$ is then the estimate that maximises the log-likelihood function.
- The equation $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$ generally has *no closed form solution*, and *iterative optimization algorithms* are used instead.

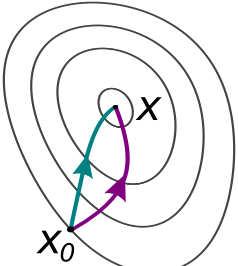
The ML estimator

For computational convenience, we usually work with the **log-likelihood**

$$\ell(\boldsymbol{\beta}) = \log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log \mathbb{P}\left(\mathbf{y} \,|\, \mathbf{X}; \boldsymbol{\beta}\right).$$

- $\boldsymbol{\beta}_{ML}$ is then the estimate that maximises the log-likelihood function.
- The equation $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$ generally has *no closed form solution*, and *iterative optimization algorithms* are used instead.

Examples for iterative optimization are Gradient Descent (based on the first derivative) and Newton's method (which also uses the second derivative).



ML estimation for binary outcomes

A **distributional assumption** lies at the center of ML estimation. For *binary outcomes*, where $Y \in \{0, 1\}$), we can use the *Bernoulli distribution* with probability mass function

$$f(y_i \,|\, p) = p^{y_i}(1 - p)^{1 - y_i}.$$

We have $\mathbb{P}(Y = 1) = p = 1 - \mathbb{P}(Y = 0)$ for the parameter p, and — due to independence of observations — the joint probability is

$$f(y_1, y_2, \dots, y_N \,|\, p) = \prod_{i=1}^{N} p^{y_i}(1 - p)^{1 - y_i}.$$

The Bernoulli distribution is a special case of the Binomial distribution with a single trial.

The log-likelihood for Bernoulli variables

With a Bernoulli outcome, we can use the likelihood

$$\mathcal{L}(p) = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i}.$$

- To find p_{ML}, we need to maximise the likelihood by solving for $\frac{\partial L}{\partial p} = 0$.
- The product is difficult to differentiate — we'd prefer a sum.
- We can use properties of the logarithm, and maximise the log-likelihood instead.

  We need to solve

$$\frac{\partial \ell}{\partial p} = \frac{\partial \sum_i \log[p^{y_i}(1-p)^{1-y_i}]}{\partial p} = 0.$$

Deriving the ML estimator i

To obtain the ML estimate, we first reformulate the log-likelihood as

$$\begin{aligned}
\ell(p) &= \sum_{i=1}^{N} \log[p^{y_i}(1-p)^{1-y_i}] \\
&= \sum_{i=1}^{N} y_i \log p + (1-y_i)\log(1-p) \\
&= N\bar{\mathbf{y}} \log p + N(1-\bar{\mathbf{y}})\log(1-p).
\end{aligned}$$

Where the last step relates the summation to the mean — $\sum_i y_i = N\bar{\mathbf{y}}$).

Next, we need to differentiate with respect to p.

Deriving the ML estimator ii

We know that

$$\ell(p) = N\bar{\mathbf{y}}\log p + N(1-\bar{\mathbf{y}})\log(1-p),$$

which we need to differentiate with respect to p, and solve for p_{ML}.

$$\begin{aligned}
\frac{\partial \ell(p)}{\partial p} = \frac{N\bar{\mathbf{y}}}{p} - \frac{N(1-\bar{\mathbf{y}})}{1-p} = 0 \\
\frac{N\bar{\mathbf{y}}}{p} &= \frac{N(1-\bar{\mathbf{y}})}{1-p} \\
\bar{\mathbf{y}}(1-p) &= p(1-\bar{\mathbf{y}}) \\
p_{ML} &= \bar{\mathbf{y}}.
\end{aligned}$$

The *maximum likelihood estimate* is the **average number of occurences in the sample**.

ML estimation for logit models

With **logit models**, we have a *Bernoulli outcome* Y, and *model the probability* p using the logistic function. We have the following PMF

$$\begin{aligned}\mathbb{P}\Big(Y = y_i \,|\, x_i\Big) &= p^{y_i}\,(1-p)^{1-y_i} \\ &= \left(\frac{e^{x_i\boldsymbol{\beta}}}{1+e^{x_i\boldsymbol{\beta}}}\right)^{y_i}\left(1-\frac{e^{x_i\boldsymbol{\beta}}}{1+e^{x_i\boldsymbol{\beta}}}\right)^{1-y_i}\end{aligned}$$

and set $\boldsymbol{\beta}_{ML}$ by (numerically) maximising the log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N}\Big[-\log\Big(1+e^{x_i\boldsymbol{\beta}}\Big) + y_i x_i\boldsymbol{\beta}\Big].$$

ML estimation for Poisson models

With **Poisson models**, we have a *Poisson outcome*, and *model the mean* λ using an exponential function. We have the following PMF

$$\mathbb{P}\left(Y = y_i \,|\, x_i\right) = \frac{\exp\left\{x_i\boldsymbol{\beta}\right\}^{y_i} \exp^{-\exp\left\{x_i\boldsymbol{\beta}\right\}}}{y_i!}.$$

and set $\boldsymbol{\beta}_{ML}$ by (numerically) maximising the log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i x_i \boldsymbol{\beta} - \exp\left\{x_i\boldsymbol{\beta}\right\}.$$

The log-likelihood measures fit, relating the fitted ($x_i\boldsymbol{\beta}$) to the observed value (y_i).

The linear model and ML estimation

Consider the **standard linear model** with normally distributed errors, given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad \mathbf{e} \sim \mathcal{N}(0, \sigma^2).$$

This implies that $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$. So far, we've used *ordinary least squares* to estimate the parameters — now we can also use *maximum likelihood estimation*.

Normal distribution

The Normal distribution, denoted by $\mathcal{N}(\mu, \sigma^2)$, has the probability density function

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Normal probability density function

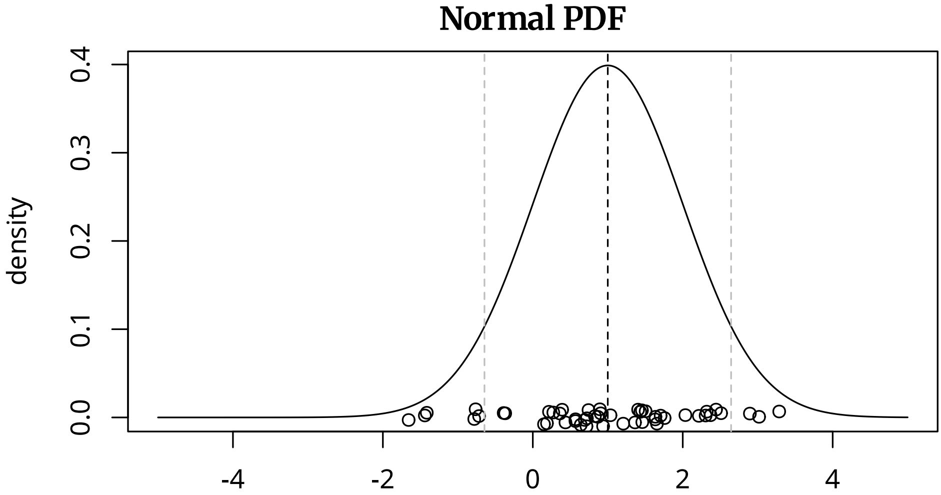


Figure 16: An $\mathcal{N}(1, 1)$ density, and 50 draws from it.

Deriving the ML estimator

We can get the likelihood function from the PDF

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left\{ \frac{-1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}.$$

To obtain estimates, we will need the log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2) = \frac{N}{2} \log(2\pi) - N \log \sigma - \frac{1}{2\sigma^2} \Big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\Big)' \Big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\Big).$$

Deriving the ML estimator

We can get the likelihood function from the PDF

$$\mathcal{L}(\boldsymbol{\beta},\sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left\{\frac{-1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}.$$

To obtain estimates, we will need the log-likelihood

$$\ell(\boldsymbol{\beta},\sigma^2) = \frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right).$$

We will focus on $\boldsymbol{\beta}_{ML}$ — notice how the last term measures the squared deviations.

Likelihood function with one coefficient

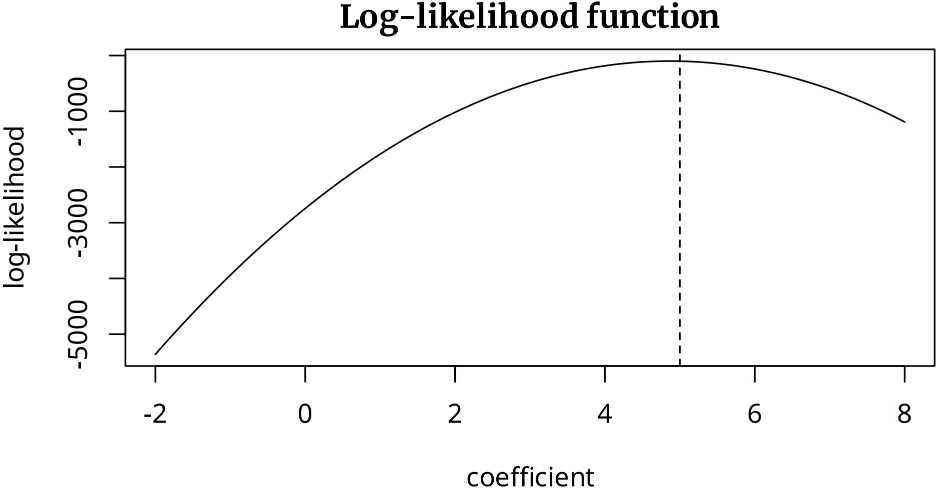


Figure 17: Visualisation of the log-likelihood for simulated data with one coefficient — $\ell(\beta)$.

Likelihood functions as a contour plot

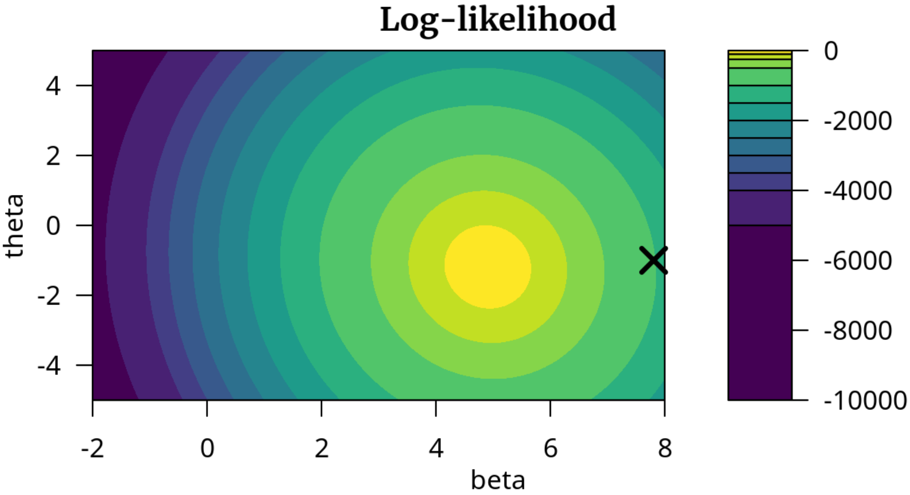


Figure 18: Visualisation of the log-likelihood for simulated data with two coefficients – $\ell(\beta, \theta)$.

The maximum likelihood

To find $\boldsymbol{\beta}_{ML}$, we need to maximise the log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2) = \frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right).$$

The maximum likelihood

To find $\boldsymbol{\beta}_{ML}$, we need to maximise the log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2) = \frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right).$$

When taking the derivative, the first two elements drop out, and we have

$$\frac{\partial\ell(\boldsymbol{\beta}, \sigma^2)}{\partial\boldsymbol{\beta}} = -2\sigma^{-2}\left(-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right).$$

The maximum likelihood

To find $\boldsymbol{\beta}_{ML}$, we need to maximise the log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2) = \frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right).$$

When taking the derivative, the first two elements drop out, and we have

$$\frac{\partial\ell(\boldsymbol{\beta}, \sigma^2)}{\partial\boldsymbol{\beta}} = -2\sigma^{-2}\left(-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right).$$

- We obtain $\boldsymbol{\beta}_{ML}$ from $\frac{\partial\ell(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}} = 0$, and
- check whether $\ell(\boldsymbol{\beta}, \sigma^2)$ is maximal by checking the second derivative.

OLS and ML estimation

For the linear model with Normal errors, the OLS and ML estimates of $\boldsymbol{\beta}$ coincide.

Shrinkage estimators — the LASSO

Let's discard the **constraint of unbiased estimators**.

- Theoretically, there's an *unlimited number of regressors*; most are irrelevant.
- We only want to *keep important regressors*, and pull coefficients towards the mean – recall the phenomenon of *regression to the mean*.

How can we achieve this in the linear model?

$$\hat{\beta} = \min_{\beta} \left\{ \left(\mathbf{y} - \mathbf{X}\beta\right)' \left(\mathbf{y} - \mathbf{X}\beta\right) \right\}$$

Shrinkage estimators — the LASSO

Let's discard the **constraint of unbiased estimators**.

- Theoretically, there's an *unlimited number of regressors*; most are irrelevant.
- We only want to *keep important regressors*, and pull coefficients towards the mean — recall the phenomenon of *regression to the mean*.

How can we achieve this in the linear model?

$$\begin{aligned}\hat{\beta} &= \min_{\beta} \left\{ \left(\mathbf{y} - \mathbf{X}\beta\right)' \left(\mathbf{y} - \mathbf{X}\beta\right) \right\} \\ &= \min_{\beta} \left\{ \left(\mathbf{y} - \mathbf{X}\beta\right)' \left(\mathbf{y} - \mathbf{X}\beta\right) + \lambda|\beta| \right\}.\end{aligned}$$

We can introduce various penalty terms to punish larger coefficent values.

Maximum likelihood estimation

- ML estimators are based on the *probability distribution* of Y.
- We learn about the
  - *parameters of this underlying distribution*,
  - **conditional on** the *data* we observe and the chosen *distribution*.

To find a ML estimator we

1. model the *probability of each observation*,
2. derive the *joint probability* of all observations,
3. consider the joint probability as a function of its parameters θ, conditional on the data $\mathcal{D}$ – this gives us the *likelihood function* $\mathcal{L}$,
4. *maximise the log-likelihood*, $\ell(\theta|\mathcal{D})$, with respect to θ.

Matching

Matching observations

Recall the fundamental problem of causal inference — *we can't observe the counterfactual* to our treatment. With **matching**, we try to find *close matches to the treated units* within the data. Specifically, we

- divide the dataset in *treated* and *control* units,
- find the ones with the **closest matching** characteristics between each of them,
- prune away unmatched observations without creating selection bias,
- perform our analysis with the matched dataset.

This procedure allows us to create a *sample with balanced confounders*, emulating the balance induced by *completely randomized* or *blocked* experiments. Matcing is an *intuitive* and *parsimonious* alternative to highly elaborate specifications and can aid with causal inference.

Methods for matching

There are many methods for *matching* that differ in their notion of **closeness**.

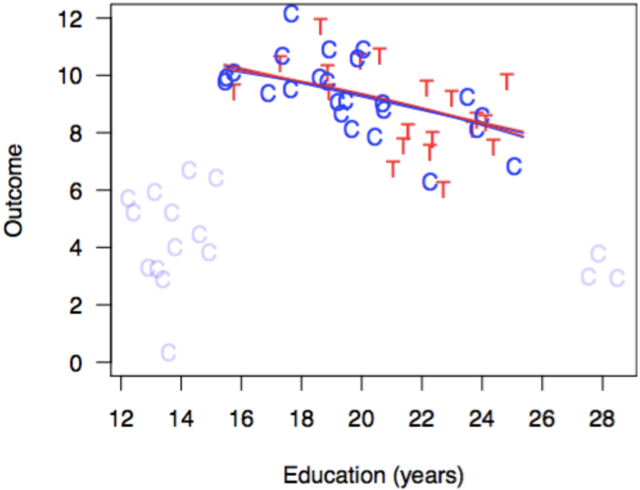


Figure 19: Illustration of a full and matched sample (by King, 2015).

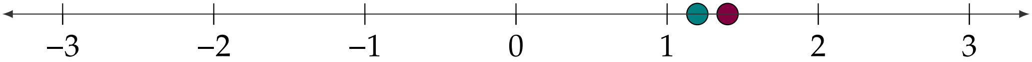
Propensity score matching

Assume you know the **propensity of being treated** for every unit. We could use this information to counteract any selection biases.

Propensity score matching

Assume you know the **propensity of being treated** for every unit. We could use this information to counteract any selection biases.

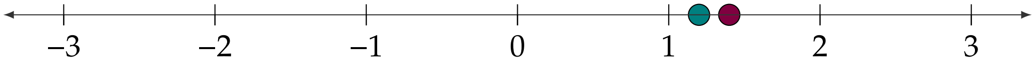
Propensity score matching (PSM) looks to estimate this propensity, and use it to match observations. We *estimate the treatment propensity*, and match control and treatment units with similar **propensity scores**, emulating a fully randomised experiment.



Propensity score matching

Assume you know the **propensity of being treated** for every unit. We could use this information to counteract any selection biases.

Propensity score matching (PSM) looks to estimate this propensity, and use it to match observations. We *estimate the treatment propensity*, and match control and treatment units with similar **propensity scores**, emulating a fully randomised experiment.



However, PSM is a problematic method for matching (King and Nielsen 2019), it

1. throws away information by using only a single dimension — the propensity score,
2. suffers from the *propensity score paradox* — random pruning causes imbalance.

Some alternatives, such as *Mahalanobis distance matching* (MDM), use some *distance between observations* to find matches.

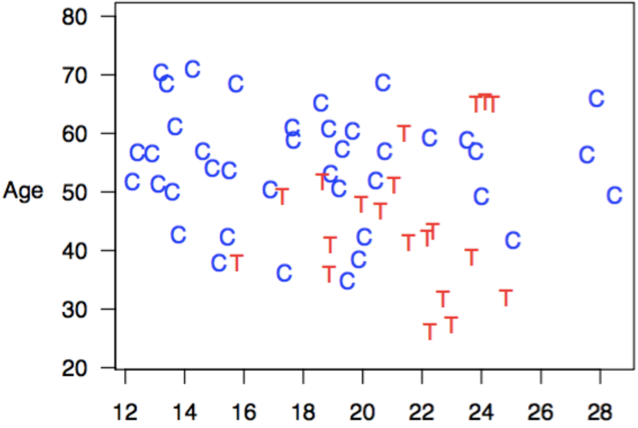


Figure 20: MDM matching (by King, 2015)

MDM matches

MDM uses the Mahalanobis distance; observations further than some boundary, or *caliper*, are pruned.

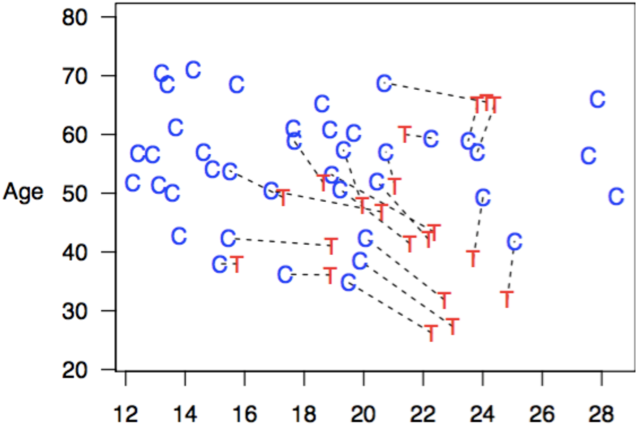


Figure 21: MDM matching (by King, 2015)

Coarsened exact matching

Coarsened exact matching (CEM) approximates a fully-blocked experiment. It works by coarsening explanatory variables to some degree, i.e. separating values into bins.

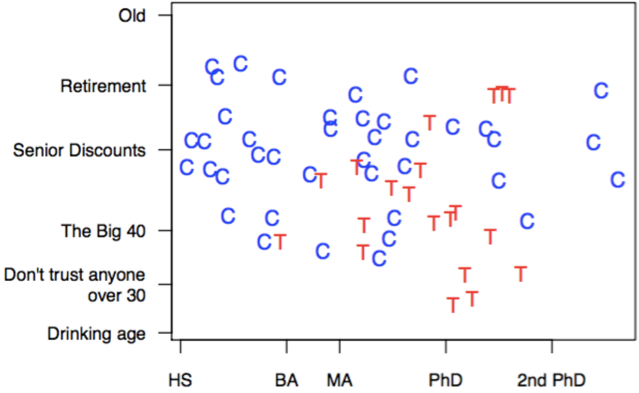


Figure 22: CEM matching (by King, 2015)

CEM matches

CEM then sorts observations into strata with unique values for all variables on the coarsened scale. Strata without treated or controlled observations are then pruned.

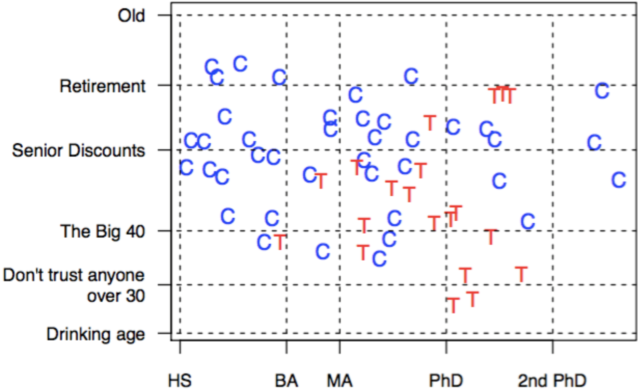


Figure 23: CEM matching (by King, 2015)

Quasi-experiments

Natural experiments

A **natural experiment** is a study where an *experimental setting is induced by nature* or other factors outside our control.

- It is an *observational study* with properties of randomised experiments.
- This provides a good basis for causal inference, and
- doesn't suffer from potential issues of a conducting an experiment, such as
  - cost,
  - ethics
  - ...

Economic research often relies on natural experiments.

> *The sickle cell trait can be seen as a long-run natural experiment for the health effects of* ***malaria*** *– it provides some protection against it, but leads to sickle cell disease.*

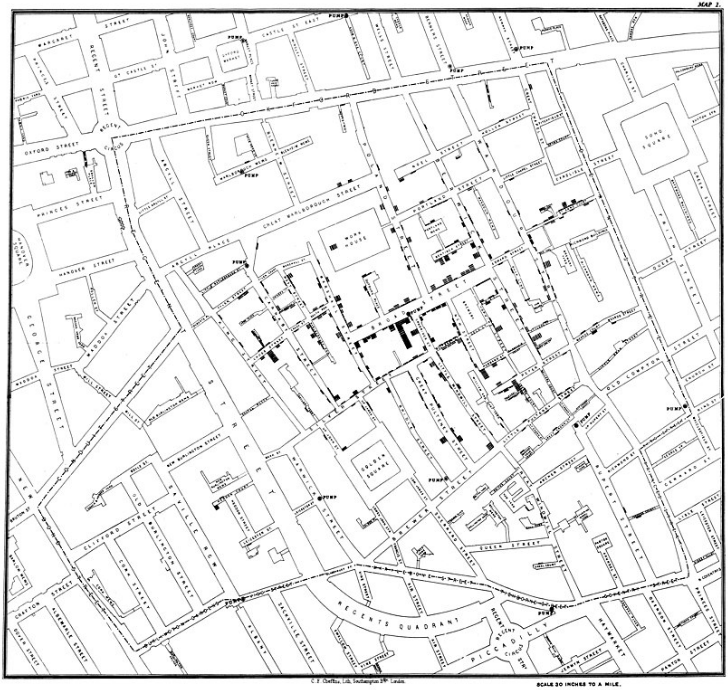
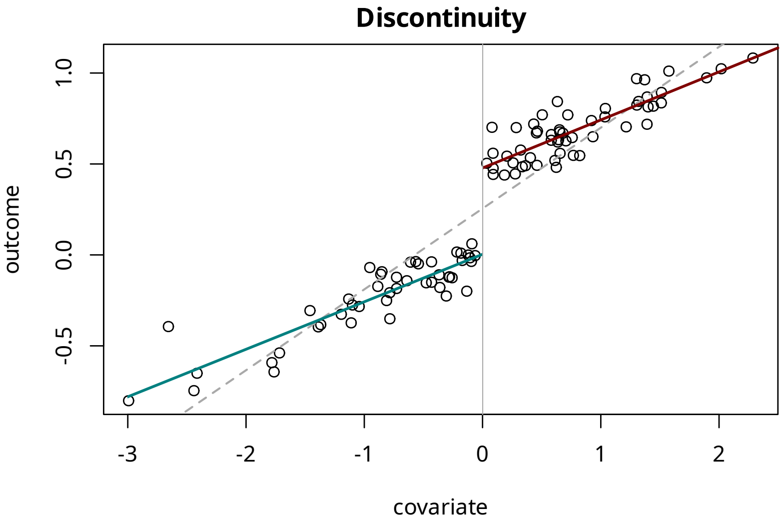


Figure 24: Map of cholera cases and the Broad Street water pump in London (Snow, 1954).

Figure 25: Alexander Pirnie drawing the first number of the Vietnam draft lottery in 1969.

Regression discontinuity

A *regression discontinuity design* (RDD) is another *quasi-experimental* design.



How does RDD work?

When there is a **sharp cutoff** in treatment assignment, we may be able to

- compare observations on either side of this discontinuity.
- We learn about the *local treatment effect*.

Example — scholarships

Consider a **merit-based scholarship** as an example.

- We cannot compare recipients and non-recipients, since *high-performers are more likely* to receive the scholarship.
- If the scholarship is awarded at a cutoff grade of 1.5 we might be able to use this cutoff to compare students near it.

Requirements for a RDD

- For an ideal RDD, all *other relevant variables* are continuous at the cutoff,
- and there is sufficient *randomness* in the assignment around the cutoff.

Moreover, we need to **correctly model** the *functional form*.

Issues

In practice, these requirements are hard to check, since

- effects are often *contaminated* by other factors, and
- we never truly know the functional form.

A common problem are RDD studies that "discover" a discontinuity by *overfitting* the data. Many potential *discontinuities act on multiple factors* (e.g. age thresholds) and treatment can often be influenced (e.g. in exams).

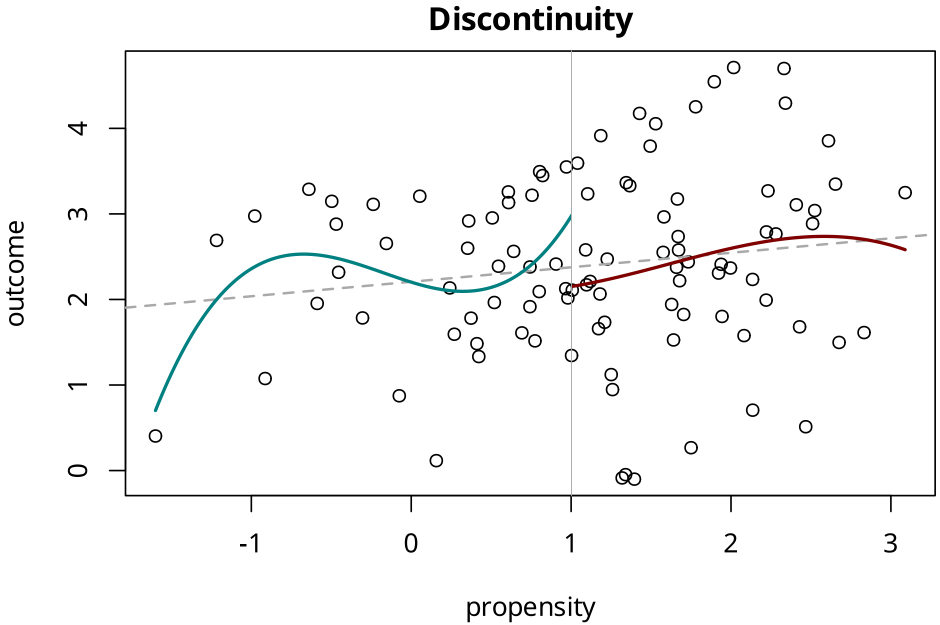


Figure 26: An artificial discontinuity by overfitting with a polynomial regression.

Panel data

Panel data

We talk of **panel** (or longitudinal) data when we have *repeated measurements* of our individual units over time. This means, we have three dimensions of data — variables, individual units, and time.

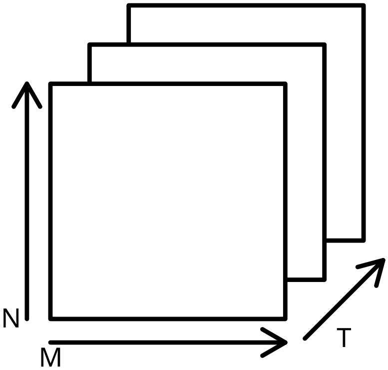


Figure 27: Panel structure.

Examples

Individual	Date	Income	Age	Education
A	2020	1200	20	medium
A	2021	1300	21	medium
B	2020	1800	24	medium
B	2021	2600	25	high

Some panel datasets are the EU-SILC (Statistics on Income and Living Conditions), HFCS (Household Finance and Consumption Survey), and Google's data on you.

Why panel data?

Panel data and models have some useful *advantages*:

- *more* data (more is more),
- potential *efficiency* gains,
- follows *relationships over time*,
- considers *unobserved* individual- or time-specific effects.

Some potential drawbacks include panel mortality (individuals drop out), panel effects (impacts of repeated data collection), cross-section dependency, decreasing marginal returns of observations.

Pooled cross sections

We can imagine a panel model by stacking cross sectional models as:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_T \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{pmatrix} + \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_T \end{pmatrix}, \tag{1}$$

We repeat the model for every date and usually assume constant coefficients, i.e.

$$y_{it} = x_{it}\beta + e_{it}.$$

The result is referred to as *pooled cross-sections*.

Applications of panel data

Pooled cross-sections can be very useful for causal inference — we can

- isolate *individual-specific*, and
- *time-specific* effects.

Moreover, panel data opens up an additional research design.

Example — deforestation

Consider the effects of opening up a mine in the Amazon on deforestation.

- We have a treatment group nearby, and a control group of unaffected forest.
- In an experiment, we'd randomly assign the mine for comparability.

Difference-in-differences

If we have panel data, we can use a **difference-in-differences** (diff-in-diff) approach.

For this, we divide our data in four and estimate

$$y_{it} = \alpha + x_{\text{after}}\phi + x_{\text{treated}}\theta + x_{\text{interacted}}\delta + \dots$$

–	Before	After	**Difference**
Control	α	$\alpha + \phi$	ϕ
Treatment	$\alpha + \theta$	$\alpha + \theta + \phi + \delta$	$\phi + \delta$
Difference	θ	$\theta + \delta$	δ

We obtain the treatment effect $\hat{\delta}$ from the *difference of the differences*.

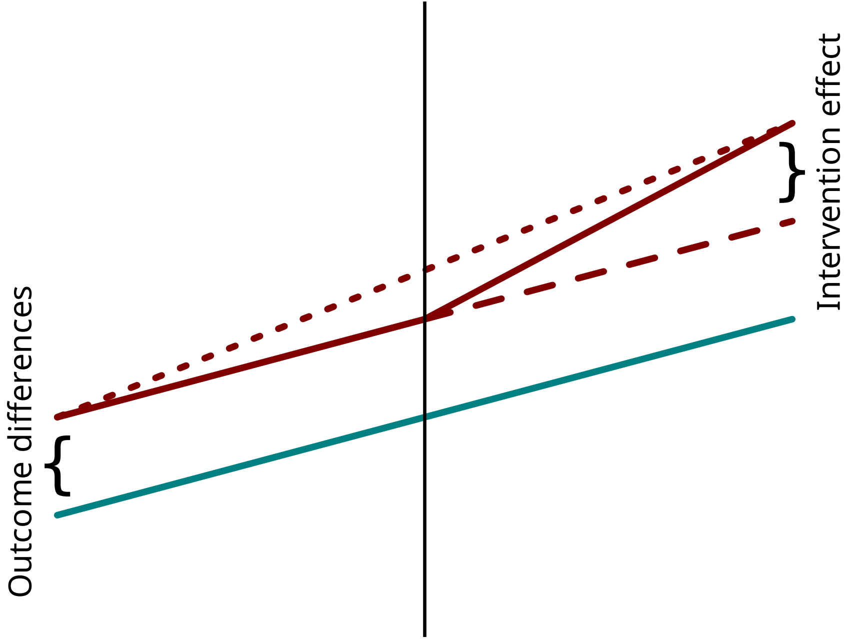


Figure 28: Effect estimation using diff-in-diff. Outcome of the control group below in teal, of the treatment group in red.

Controlling for unobservables

With panel data we can account for *unobserved or unobservable variables*. Consider

$$y_{it} = \alpha + \psi_t^{\text{period}} + \mu_i^{\text{individual}} + \dots + \varepsilon_{it}.$$

We can *include intercepts for each period and individual* — the **fixed effects**. The baseline for individual i at time t is

$$\alpha + \psi_t + \mu_i.$$

The error ε_{it} only contains unobserved factors that **vary over time and individual**. The parameter μ_2, e.g., captures *all effects on individual 2* that do not vary over time, even if they are unobservable.

Fixed effect model

Consider a *fixed effect* model of crime rates in the US for 1982 and 1987

$$y_{it}^{\text{crim}} = \alpha + \psi_t^{1987} + \mu_i^{\text{state}} + x_{it}^{\text{unemp}}\beta + \varepsilon_{it}.$$

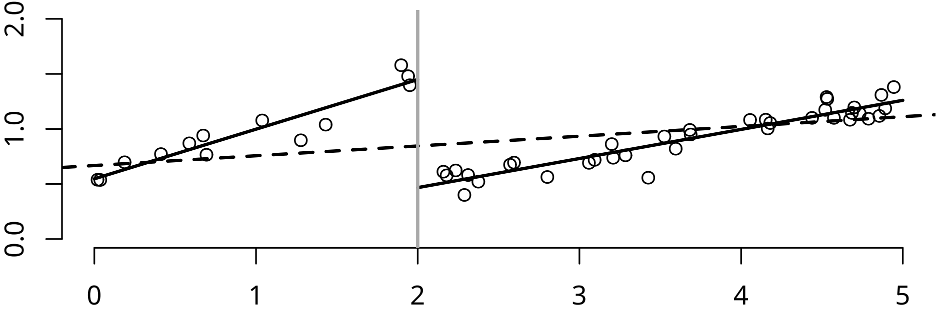
The *time-parameter* ψ captures the effect of the year 1987 against 1982. The *individual-parameter* μ_i captures the effect of states.

Unobservable effects may correlate with explanatory variables without violating the exogeneity assumption — this means we may only need to *control for variables that vary over time and individuals*.

Changing relationships

Panel data also allows us to investigate whether *coefficients differ over certain groups*, e.g. time. The *Chow test* allows this by dividing the data into two groups a and b and checking whether $\beta_a = \beta_b$.

One example are *structural breaks* over time, where relations change after some event.



Anderson, T. W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46–63. https://doi.org/10.1214/aoms/1177730090.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11 (1): 727–53. https://doi.org/10.1146/annurev-economics-080218-025643.

Angrist, Joshua D., Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. 2017. "Economic Research Evolves: Fields and Styles." *American Economic Review* 107 (5): 293–97. https://doi.org/10.1257/aer.p20171117.

Angrist, Joshua D., and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15 (4): 69–85. https://doi.org/10.1257/jep.15.4.69.

References ii

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30. https://doi.org/10.1257/jep.24.2.3.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32. https://doi.org/10.1257/jep.31.2.3.

———. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

References iii

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90 (430): 443–50. https://doi.org/10.1080/01621459.1995.10476536.

Buckles, Kasey S., and Daniel M. Hungerman. 2013. "Season of Birth and Later Outcomes: Old Questions, New Answers." *Review of Economics and Statistics* 95 (3): 711–24. https://doi.org/10.1162/REST_a_00314.

Cunningham, Scott. 2021. *Causal Inference*. New Haven, CT, USA: Yale University Press. https://doi.org/10.12987/9780300255881.

Hamermesh, Daniel S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature* 51 (1): 162–72. https://doi.org/10.1257/jel.51.1.162.

Imbens, Guido W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58 (4): 1129–79. https://doi.org/10.1257/jel.20191597.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1.

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27 (4): 435–54. https://doi.org/10.1017/pan.2019.11.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43. https://www.jstor.org/stable/1803924.

Pearl, Judea. 2009. *Causality*. *Cambridge Core*. Cambridge, England, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

Steel, Mark F. J. 2020. "Model Averaging and Its Use in Economics." *Journal of Economic Literature* 58 (3): 644–719. https://doi.org/10.1257/jel.20191385.