

CS&SS 321 - Data Science and Statistics for Social Sciences

Module III - Introduction to causal inference and linear models

Ramses Llobet

Module III

- ▶ This module introduces and reviews the topic of causation in science.
 - ▶ *randomization*.
 - ▶ *applied causal inference*.
 - ▶ *causal modeling* (module IV).
- ▶ It also introduces the **linear regression model** and the method of **least squares** (LS).

The statistics war of the late XXth century



The statistics war of the XXIth century

- Causal inferences requires a model outside of the statistical model.



Causes in, causes out

- ▶ Why do experiments work? When do they work?
- ▶ What if treatment is imperfect assigned?
- ▶ Should you *control* for anything? Everything?

Answers depend upon **causal assumptions** (\rightarrow).

- ▶ An **assumption** is a premise or supposition that is accepted *without direct evidence*, often forming the basis for reasoning or an argument.

Causes in, causes out

- ▶ Causal assumptions requires **causal knowledge** of social systems.
- ▶ For example, where X represents **rain** and Y represents **puddles**.
 - ▶ What **causal assumption** (\rightarrow) you find more reasonable?
 - (i) $X \leftarrow Y$
 - (ii) $X \rightarrow Y$



Causal design

- ▶ **Step 1:** sketch a (scientific) causal model: $X \rightarrow Y$.
 - ▶ *Causes in:* assumptions reflect **background knowledge** (*theory and literature review*).
- ▶ **Step 2:** use the model to design **data collection** and **statistical procedures**.
- ▶ **Step 3:** use statistical analyses to **hypothesis test** and report results.
 - ▶ *Causes out:* test assumptions' implications about the **causal mechanism**.

Causal design: intervention

- ▶ In causal inference, an **intervention** is a deliberate and controlled manipulation of one or more variables in a system to assess their **causal impact** on the outcome of interest.
 - ▶ *Example:* Pouring a bucket of water on the floor creates a puddle; does rain follow?
- ▶ We formalize this via the **potential outcomes** framework.



Causation in science

Treatment indicator: $T_i \in \{0, 1\}$, where i refers respondents.

► (1) example:

- $T_i = 0$ indicates no membership in a union.
- $T_i = 1$ indicates membership in a union.

► (2) example:

- $T_i = 0$ indicates no daughters.
- $T_i = 1$ indicates having daughters.

Outcome: Y_i

► (1) example: redistribution attitudes (*gincdif*).

► (2) example: pro-feminist attitudes (*progressive.vote*).

Causation in science

- ▶ Consider the treatments' (T) **causal mechanisms** (\rightarrow) that drives the **outcome** (Y).
 - ▶ **Why** does labor **union membership** increase support for redistribution?
 - ▶ **Why** does having a **daughter** increase pro-feminist attitudes?

Potential outcomes $Y_i(0)$, $Y_i(1)$, where:

- ▶ **(1) example:**
 - ▶ $Y_i(0)$ represents redistribution attitudes *without* membership.
 - ▶ $Y_i(1)$ represents redistribution attitudes *with* membership.
- ▶ **(2) example:**
 - ▶ $Y_i(0)$ represents pro-feminist attitudes *without* daughters.
 - ▶ $Y_i(1)$ represents pro-feminist attitudes *with* daughters.

Causation in science

The **fundamental problem of causality** posits that we cannot observe two outcomes at the same time:

$$\text{individual treatment effect} = Y_{\text{Ramses}}(1) - Y_{\text{Ramses}}(0) \quad (1)$$

Instead, we **estimate** group-level effects by taking the differences in means between **treatment**, $\bar{Y}(1)$, and **control**, $\bar{Y}(0)$, groups.

$$\text{average treatment effect} = \bar{Y}(1) - \bar{Y}(0) \quad (2)$$

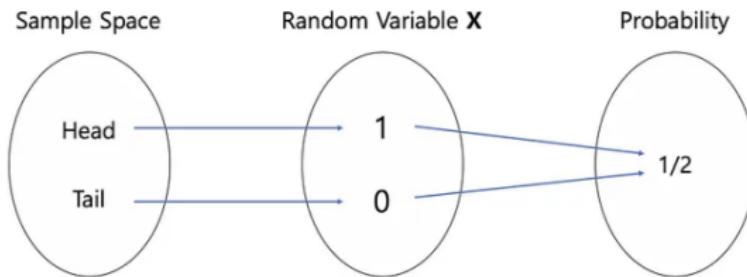
However, we can identify **ATE** if, and only if, the treatment D has been **randomly assigned** to each respondent i. Formally,

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \quad (3)$$

Causation in science

- Think about random assignment as flipping a coin.
 - In **expectation** (as $n \rightarrow \infty$), a fair coin has a probability of 0.5 to show tails (0) or heads (1).
 - By definition, a random event has a probability of 0.5.

Toss 1 Coin Example



- What if, in expectation, a coin has a probability of 0.7 ?

Causation in science

- ▶ Is labor union membership a random occurrence?



Causation in science

- Is having a girl (instead of a boy) a random occurrence?



Boy

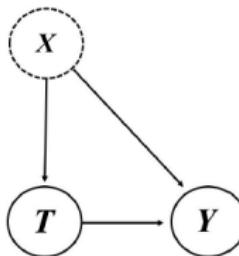
Girl

Causation in science

- ▶ **Selection bias:** Self-selection and unbalanced factors introduce bias in our statistical estimations.
 - ▶ *Self-selection:* Left-wing individuals are more likely to become labor union activists.
 - ▶ *Unbalanced factors:* Labor union members may systematically differ from non-union members in terms of other variables such as occupation and income.

Causation in science

- ▶ In observational studies, unconditional treatment effects are unlikely due to the influence of **confounding** factors, both **observed** and **unobserved**.



- ▶ However, sometimes we can assume **conditional random effects**.

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i. \quad (4)$$

Causation in science

- ▶ Let's work a short coding example.
- ▶ Open the file `unions_sweden.Rmd`, we will do only the **first** section.
- ▶ We will finish the remaining section next week.

From previous model: Data Generating Process

- ▶ Two very useful pieces of information from a DGP are its **mean** and **standard deviation**.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^N X_i \quad ; \quad S = \sqrt{\frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})^2}$$

where

- ▶ \bar{X} represents the **sample mean**.
- ▶ N is the number of **observations** in the sample.
- ▶ X_i represents **values** from a variable in the sample.
- ▶ S represents the **sample standard deviation**.

Standard deviation and variance

- The **standard deviation** and **variance** are both measures of the spread of a distribution.
 - To estimate the variance (S^2), we simply take the **square** of the standard deviation (S).

$$S^2 = \left(\sqrt{\frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})^2} \right)^2$$

$$S^2 = \frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})^2$$

- S^2 is the **sample** variance.
- Q: Why choose the standard deviation over the variance to report **summary statistics**?

Mean and variance

$$\bar{X} = \frac{1}{N} \sum_{i=1}^n X_i \quad ; \quad S^2 = \frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})^2$$

- ▶ The **sample mean** (\bar{X}) describes the location (*the center*) of the data (*distribution*).
- ▶ The **sample variance** (S^2) measures the variability in the data (*distribution*).
 - ▶ The variance describes the **average deviation** in a distribution.

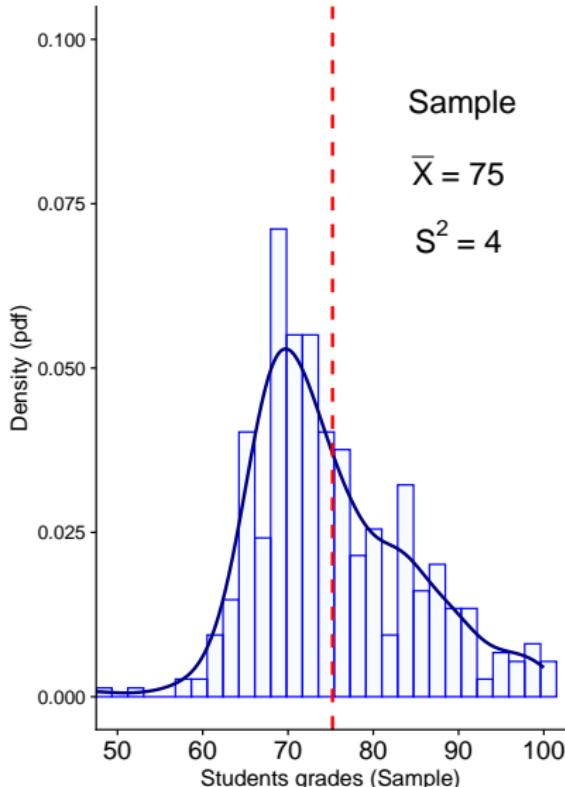
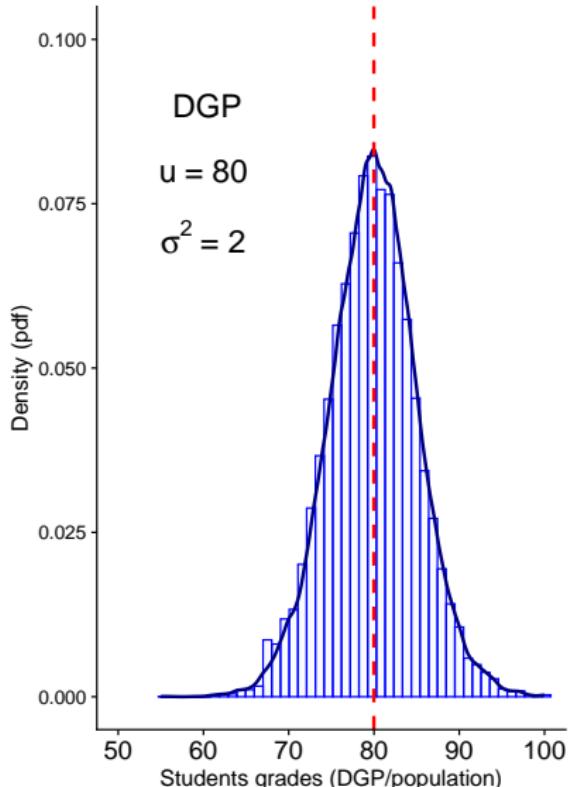
DGP vs. sample

We distinguish between the **Data Generating Process** (DGP) and the data **sample**.

- ▶ DGP or *population* is a **theoretical** concept describing how observed/sampled data is generated.
 - ▶ It follows a **distribution**, typically depicted as the *TRUE* (!?).
 - ▶ Its parameters, mean (μ) and variance (σ^2), are **fixed**.
- ▶ The sample is an **empirical** construct, representing realizations/occurrences of a data process.
 - ▶ Sample data maps into **distributions** of *random variables*.
 - ▶ Its parameters, mean (\bar{X}) and variance (S^2), are **random**.

Note: we use the sample to infer (**approach**) the underlying *TRUE* of a DGP.

DGP vs. sample



Unconditional distributions

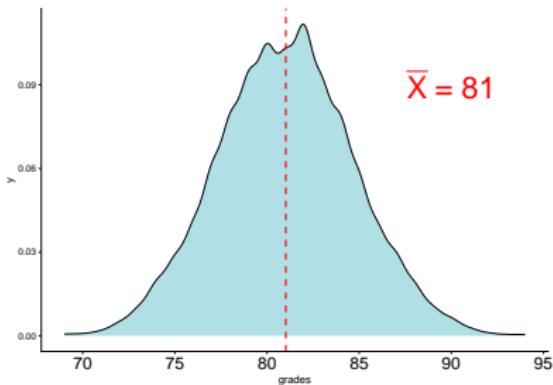
- ▶ The **expectation** $E[.]$ of a random variable X , denoted as $E[X]$, is a useful measure of central tendency of the DGP.
 - ▶ The expectation is also called the **expected value** or **mean**.
 - ▶ In the case of the normal distribution, the expectation is the first **central moment** and is denoted as μ .
- ▶ In general, a natural estimator of the expectation is the **sample mean**.

$$\mu = E[X] = \bar{X} = \frac{1}{n} \sum_{i=1}^N X_i$$

Unconditional distributions

- ▶ We have a sample of UW students' grades.
- ▶ What may be a good candidate to estimate the mean of this population?

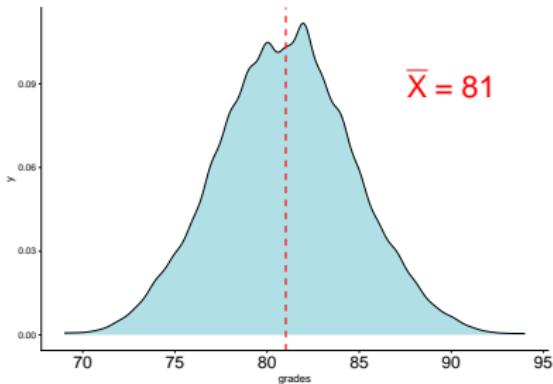
$$E[\text{grades}] = ?$$



Unconditional distributions

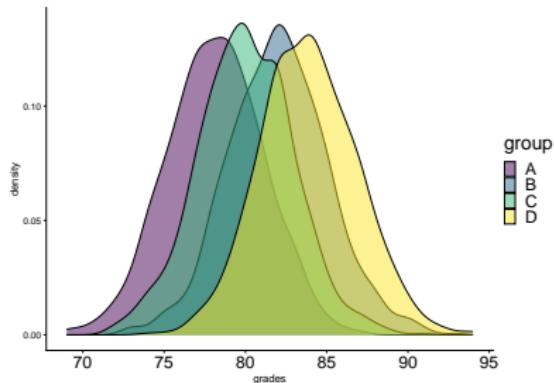
- We have a sample of UW students' grades.
- What may be a good candidate to estimate the mean of this population?

$$E[\text{grades}] = 81$$



Conditional distributions

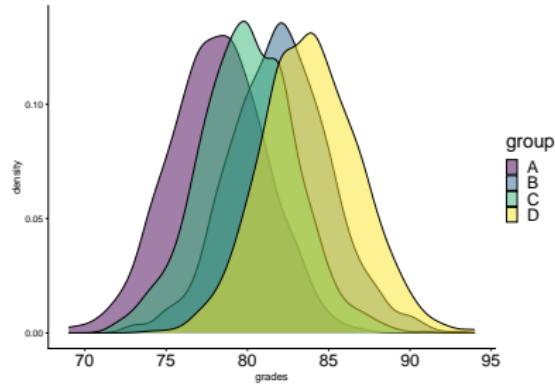
- We can compare the grade distribution for these different **sub-populations**.
 - Group A
 - Group B
 - Group C
 - Group D



Conditional distributions

- We can **condition** grades on a fixed value (x) of the group random variable.
- We call this the **conditional mean** (or **conditional expectation**).

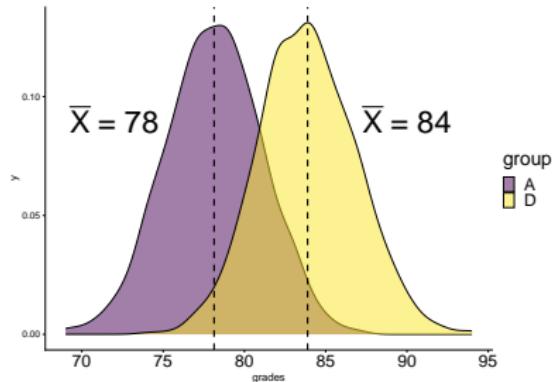
$$E[\text{grades} \mid \text{group} = x]$$



Conditional distributions

- ▶ For example, take the conditional mean of groups A and D.

$$E[\text{grades} \mid \text{group} = A] = 78$$



$$E[\text{grades} \mid \text{group} = D] = 84$$

Conditional distributions

- ▶ When **conditioning** a distribution (*grades*), we **adjust** it to a second variable (*group*).
- ▶ This offers more insight into the **variance** of the outcome (*grades*).

$$E[\text{grades} \mid \text{group} = D] - E[\text{grades} \mid \text{group} = A] = 84 - 78 = 6$$

- ▶ However, it is crucial to note that we **cannot** attribute *causality* or interpretation to these differences.
- ▶ Conditioning helps in **describing variation** but does not constitute a **model** or explanation by itself.

Best predictor

- ▶ In statistics, we model data to **predict quantities** of interest.
 - ▶ *What is the causal effect of a cancer treatment?*
 - ▶ *What will be the stock market price next month?*
- ▶ Prediction is the closest **best guess** (*estimate*) among all data realizations in a distribution.
 - ▶ *What is the best estimate in predicting the midterm grades of all students in CS&SS321?*

Best predictor

- ▶ The **best predictor**, denoted as θ , minimizes **prediction error** (e), which is the distance of each data point from our best guess: $e = Y_i - \theta$.
- ▶ **Mean Squared Error** (MSE) quantifies the magnitude of prediction error.

$$\text{MSE} : E[(Y_i - \theta)^2]$$

Note: The notation θ is arbitrary and denotes the optimal or best predictor.

Prediction error: first guess

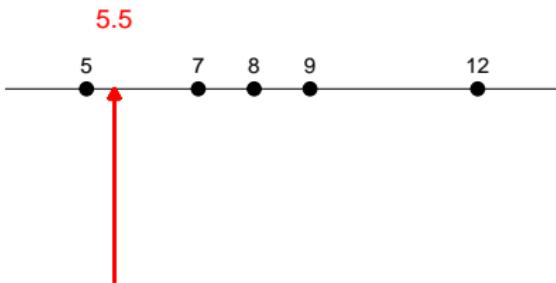
What is your **best guess** (θ) that **minimizes** the prediction error (MSE)?



N_i	Y_i	θ	$Y_i - \theta$	error
1	5			
2	7			
3	8			
4	9			
5	12			

$$MSE = E[(Y_i - \theta)^2]$$

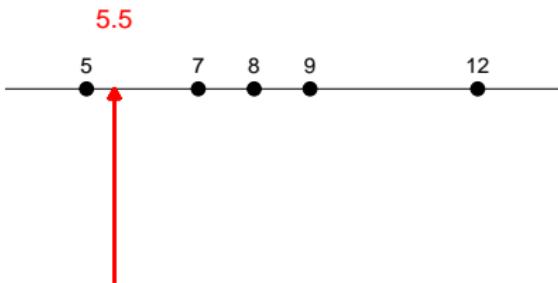
Prediction error: first guess



N_i	Y_i	θ	$Y_i - \theta$	error
1	5	5.5	-0.5	-0.5
2	7	5.5	1.5	1.5
3	8	5.5	2.5	2.5
4	9	5.5	3.5	3.5
5	12	5.5	6.5	6.5

$$MSE = E[(Y_i - 5.5)^2]$$

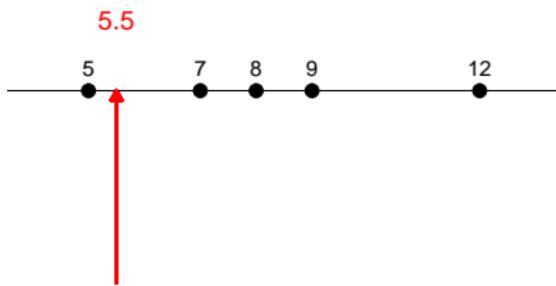
Prediction error: first guess



N_i	Y_i	θ	$Y_i - \theta$	error
1	5	5.5	5-5.5	5.5
2	7	5.5	7-5.5	5.5
3	8	5.5	8-5.5	5.5
4	9	5.5	9-5.5	5.5
5	12	5.5	12-5.5	5.5

$$MSE = E[(Y_i - 5.5)^2]$$

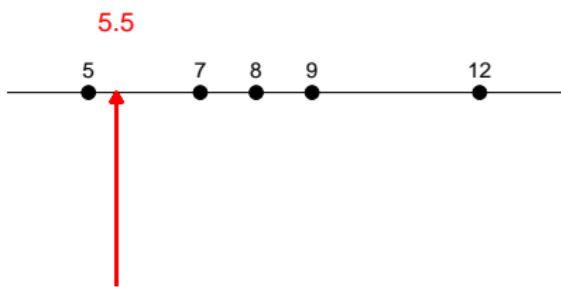
Prediction error: first guess



N_i	Y_i	θ	$Y_i - \theta$	error
1	5	5.5	5-5.5	-0.5
2	7	5.5	7-5.5	1.5
3	8	5.5	8-5.5	2.5
4	9	5.5	9-5.5	3.5
5	12	5.5	12-5.5	6.5

$$MSE_1 = \frac{1}{5}(-0.5 + 1.5 + 2.5 + 3.5 + 6.5)^2$$

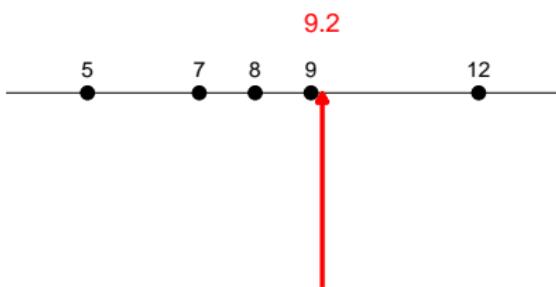
Prediction error: first guess



N_i	Y_i	θ	$Y_i - \theta$	error
1	5	5.5	5-5.5	-0.5
2	7	5.5	7-5.5	1.5
3	8	5.5	8-5.5	2.5
4	9	5.5	9-5.5	3.5
5	12	5.5	12-5.5	6.5

$$\begin{aligned}MSE_1 &= \frac{1}{5}(-0.5 + 1.5 + 2.5 + 3.5 + 6.5)^2 \\&= \frac{(13.5)^2}{5} = \frac{182.25}{5} = 36.45\end{aligned}$$

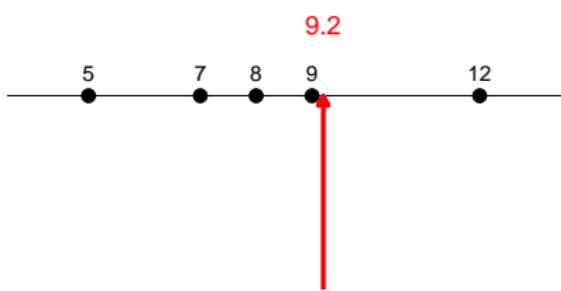
Prediction error: second guess



N_i	Y_i	θ	$Y_i - \theta$	error
1	5	9.2	5-9.2	
2	7	9.2	7-9.2	
3	8	9.2	8-9.2	
4	9	9.2	9-9.2	
5	12	9.2	12-9.2	

$$MSE_2 = E[(Y_i - 9.2)^2]$$

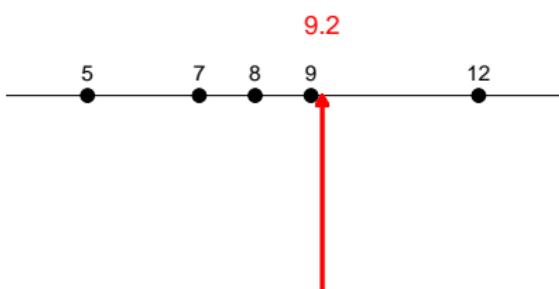
Prediction error: second guess



N_i	Y_i	θ	$Y_i - \theta$	result
1	5	9.2	5-9.2	-4.2
2	7	9.2	7-9.2	-2.2
3	8	9.2	8-9.2	-1.2
4	9	9.2	9-9.2	-0.2
5	12	9.2	12-9.2	2.8

$$MSE_2 = \frac{1}{5}(-4.2 + -2.2 + -1.2 + -0.2 + 2.8)^2$$

Prediction error: second guess



N_i	Y_i	θ	$Y_i - \theta$	result
1	5	9.2	5-9.2	-4.2
2	7	9.2	7-9.2	-2.2
3	8	9.2	8-9.2	-1.2
4	9	9.2	9-9.2	-0.2
5	12	9.2	12-9.2	2.8

$$\begin{aligned}MSE_2 &= \frac{1}{5}(-4.2 - 2.2 - 1.2 - 0.2 + 2.8)^2 \\&= \frac{(-5)^2}{5} = \frac{25}{5} = 5\end{aligned}$$

Best predictor and prediction error

- ▶ Two best guesses are provided: $\theta_1 = 5.5$ and $\theta_2 = 9.2$.
- ▶ From these best guesses, two measures of prediction error are retrieved: $MSE_1 = 36.45$ and $MSE_2 = 5$.
- ▶ The best predictor minimizes prediction error given the data.
 - ▶ Which was the **best predictor**, θ_1 or θ_2 ?
 - ▶ It's evident that $MSE_1 > MSE_2$.
 - ▶ Therefore, 9.2 better predicts this DGP than 5.5.

Best predictor and conditional means

- ▶ Let's work a short coding example.
- ▶ Open the file BestGuess.Rmd, and complete all the exercises.

Causality review

- ▶ Effective research designs can aid in identifying **causal effects** from **associations**, but they also come with their own set of **assumptions**.
- ▶ Experimental designs:
 - ▶ Randomization (e.g., RCT).
- ▶ Observational studies:
 - ▶ Confounding adjustment (via causal modeling).
 - ▶ “Natural” experiments (as if random).
- ▶ Even if **assumptions** are met, and often can **never** be completely confirmed, there is a trade-off in **conclusions validity**.

Coding exercise

- ▶ Open the file CausRev.Rmd and complete as many sections as possible.
 - ▶ The four sections are **not cumulative**; you can proceed to the next one if you feel stuck or encounter unfamiliar functions.
 - ▶ Refer to my **Module 2 slides** for explanations and detailed examples of any new functions.

Causality review: randomization

- ▶ In randomized experiments, we can identify average treatment effects (**ATE**) only if the **intervention** and treatment T are randomly assigned to each respondent i .
 - ▶ This relies on the **exchangeability** or exogeneity assumption:

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \quad (5)$$

- ▶ This assumption implies that all other variables/factors, both observables (like income) and non-observables (like ideology), are **balanced**.
- ▶ However, in practice, randomization is never perfectly implemented, and some imbalance may occur.

Causality review: randomization

- If, and only if, **randomization** has been *perfectly* implemented **and** there is **covariate balance**, we can **estimate** the causal effect of the treatment by computing the following:

$$\begin{aligned}\text{DiD} &= E[Y | T = 1] - E[Y | T = 0] \\ &= \bar{Y}_{1T} - \bar{Y}_{0T}\end{aligned}$$

- Under **ideal** randomization, no statistical modeling is necessary.
- A simple **differences-in-means** (*conditional means*) estimator provides the causal effect of interest.

Causality review: observational studies

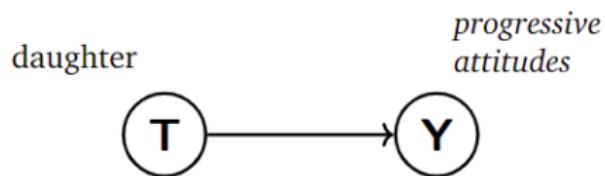
- ▶ In observational research designs, we cannot randomize an intervention, but we can identify causal effects by **conditioning** on confounders and making some (*heroic*) assumptions.
 - ▶ **Unconfoundedness** or selection on observable assumption.

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1) | X_i) \quad (6)$$

- ▶ Unconfoundedness implies that causal effects can be identified if we **adjust** for a set of variables that bias the causal effect.
- ▶ **Causal modeling** (Module 4) can help identify unconfoundedness, but it is practically impossible to meet in most applications.

Causality review: PS2, Q5

- ▶ Think about the **causal assumptions/mechanism.**
- ▶ Can someone be **biased** to have girls (instead of boys)?
- ▶ Having a girl is an **event** (*coin flip*), however, what is a **pre-condition** to having a daughter?



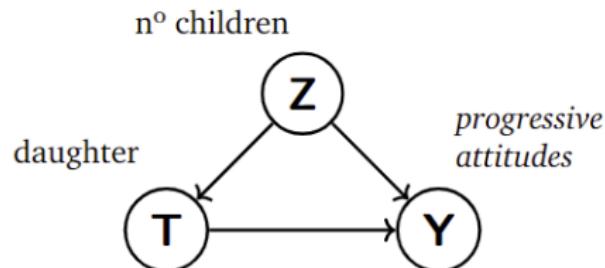
Causality review: PS2, Q5

- ▶ Conditional on children, having a daughter *may* be a random occurrence.

$$girl_i \perp\!\!\!\perp (PA_i(0), PA_i(1) | child_i)$$

PA: *Progressive Attitudes*.

- ▶ However, we need to provide **evidence** that supports this assumption.



Research design: natural experiments

- ▶ Over the past two decades, there has been an explosion in **applied causal inference**.
 - ▶ It relies on finding observational research designs with features that make it easier to assume *as-if randomness*.
 - ▶ Instrumental regression.
 - ▶ Discontinuous regression.
 - ▶ Differences-in-Differences, etc.
 - ▶ These are known as **natural experiments** because *nature* randomly assigns the **intervention**.
 - ▶ **Strong (heroic!) assumptions** must be met to infer causality.
 - ▶ For example, in time-series/panel studies, causal estimation requires the assumption of **parallel trends**.

Research design: Differences-in-Differences

- ▶ A study conducted by [Card and Krueger \(1994\)](#) analyzed the impact of minimum wage laws (**T**) on unemployment (**Y**) in two neighboring American states.
 - ▶ *Natural experiment*: New Jersey increased its minimum wage (MW) while Pennsylvania did not.
- ▶ The underlying **assumption** is that New Jersey and Pennsylvania have **similar** economic systems (**as-if random**).
- ▶ If the assumption holds, and the **only** difference between the states is the intervention (minimum wage law), we can estimate the causal effect with a **Diff-in-Diff** estimator.

Research design: Differences-in-Differences

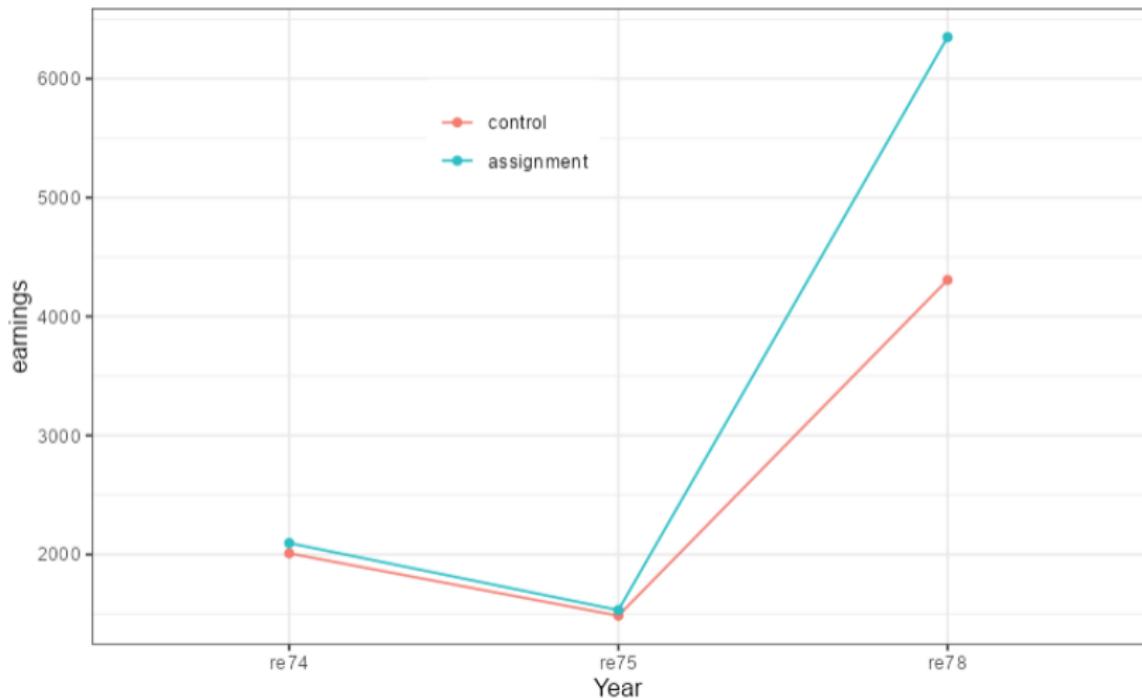
$$\text{DiD} = (\bar{Y}_{1T} - \bar{Y}_{1C}) - (\bar{Y}_{0T} - \bar{Y}_{0C}) \quad (7)$$

- Where
 - DiD is the differences-in-differences estimator,
 - \bar{Y}_{1T} is the average unemployment for New Jersey **after** increasing the MW,
 - \bar{Y}_{1C} is the average unemployment for Pennsylvania **before not** increasing the MW,
 - \bar{Y}_{0T} is the average unemployment for New Jersey **after not** increasing the MW,
 - \bar{Y}_{0C} is the average unemployment for Pennsylvania **before not** increasing the MW.

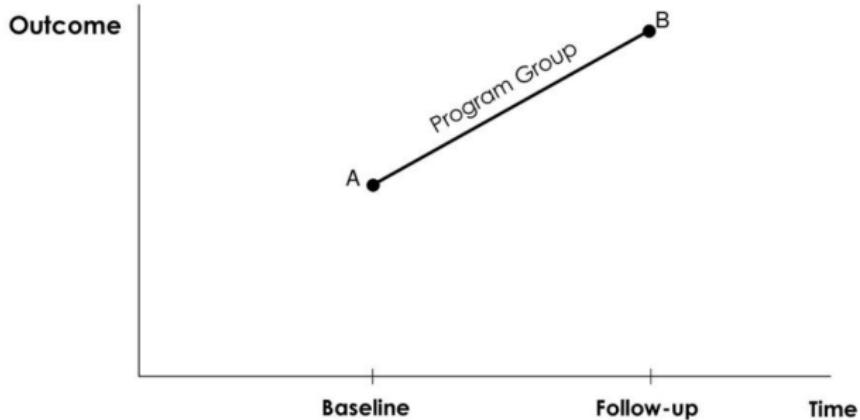
Research design: Differences-in-Differences



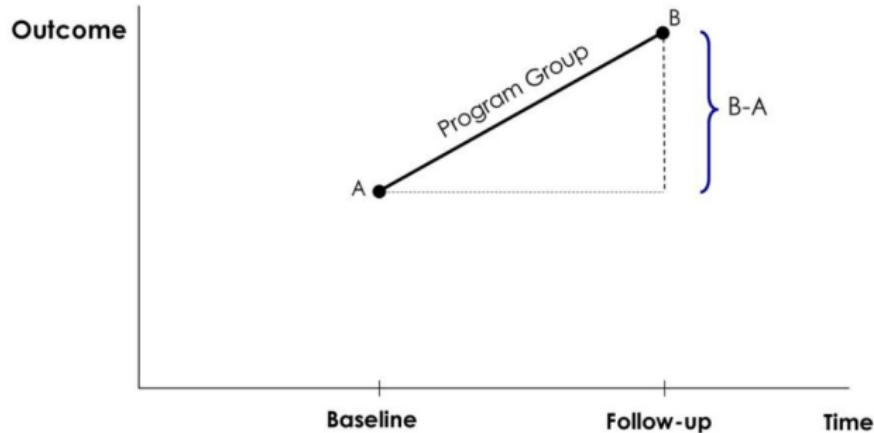
Research design: Parallel trends



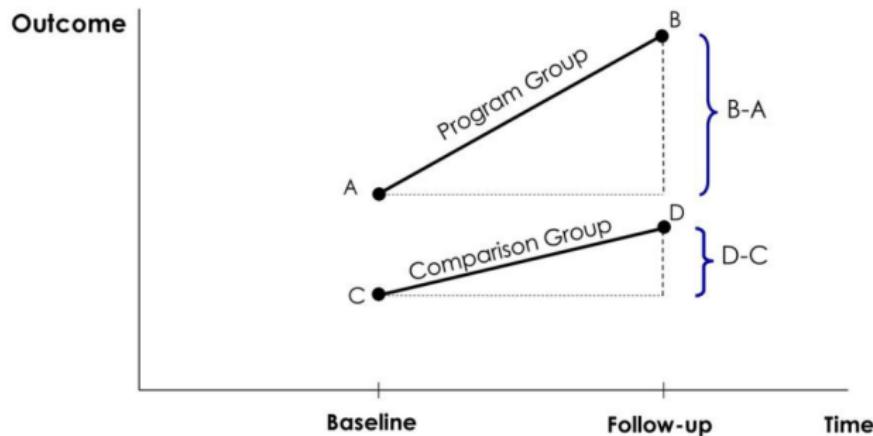
Research design: Differences-in-Differences



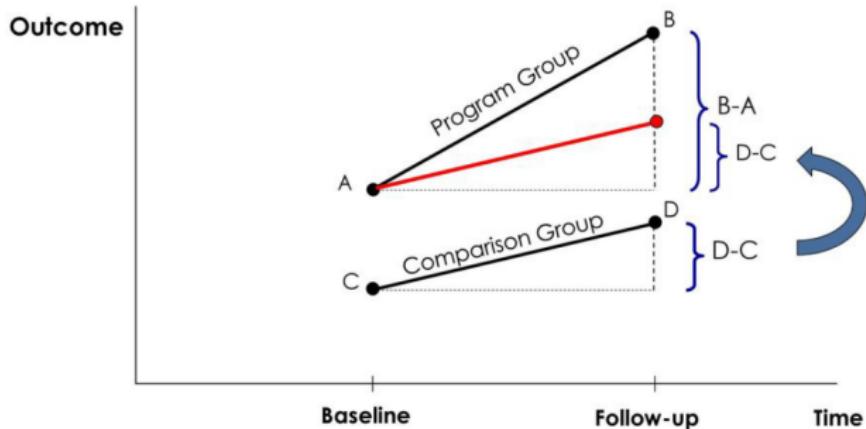
Research design: Differences-in-Differences



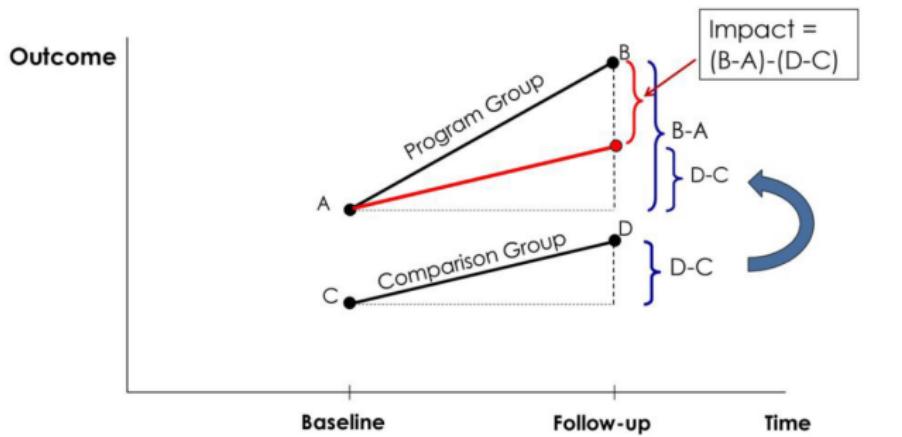
Research design: Differences-in-Differences



Research design: Differences-in-Differences



Research design: Differences-in-Differences



Causality review: key points

- ▶ To identify a causation in experimental settings, *perfect randomization* provides **covariate balance** between treatment and control groups (*exchangeability*), considering both **observed** and **unobserved** variables.
 - ▶ However, in practice, even in experimental designs, practitioners often **adjust** by conditioning on **unbalanced confounders** (*unconfoundedness*).
- ▶ In some **observational studies**, it is possible to estimate causal effects in **research designs** that mimic *natural experiments*.

Causality review: key points

- ▶ Understand the trade-offs between **internal** and **external** validity when interpreting research design and statistical results (see Professor Ainsley's Week 3 slides).
- ▶ Adjusting for **confounding** in observational studies through linear regression does not guarantee identification of a causal effect.
 - ▶ Identification of a causal effect requires **balanced unobservable** characteristics or assumptions as **as-if random**, like in [Card and Krueger \(1994\)](#).

Linear model: intercept only

A special case of the regression model is when there are no regressors

$$Y = \mu + e$$

In the **intercept only model**, we find out that the best predictor is μ !

Hence, the best predictor of an unconditional distribution is its **mean**. We can show this by computing the MSE:

$$\text{MSE} : E[(Y - \theta)^2] = E[(Y - \mu)^2]$$