# 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*.
- 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) casual model: $X \to Y$.
  - *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:

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

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

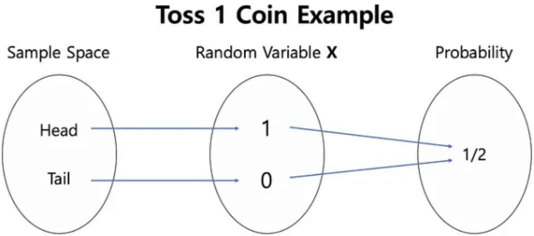
$$\text{average treatment effect} = \bar{Y}(1) - \bar{Y}(0) \tag{2}$$

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

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \tag{3}$$

# Causation in science

- Think about random assignment as flipping a coin.
  - In **expectation** (as $n \to \infty$), 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.



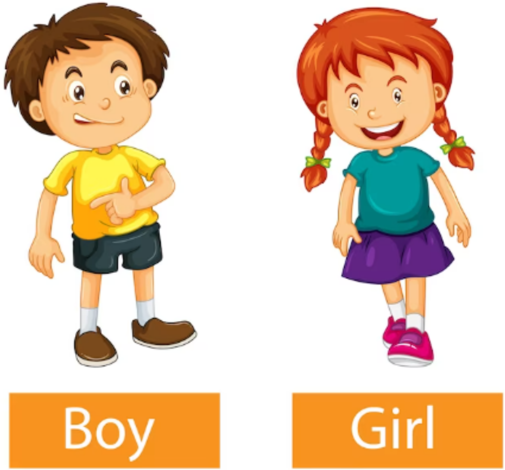

- **What if**, in expectation, a coin has a probability of 0.7 ?

# Causation in science

- Is labor union membership a random occurrence?

## Causation in science

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

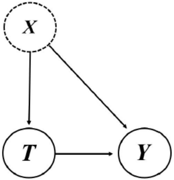


# Causation in science

- **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**.

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i. \tag{4}$$

## 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 devitation 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 ($S$).

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

- $S^2$ 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.

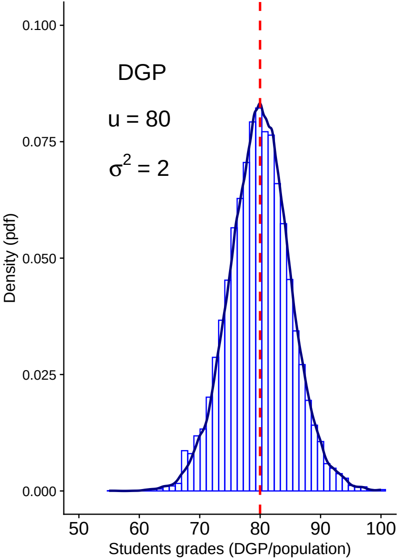
# DGP vs. sample

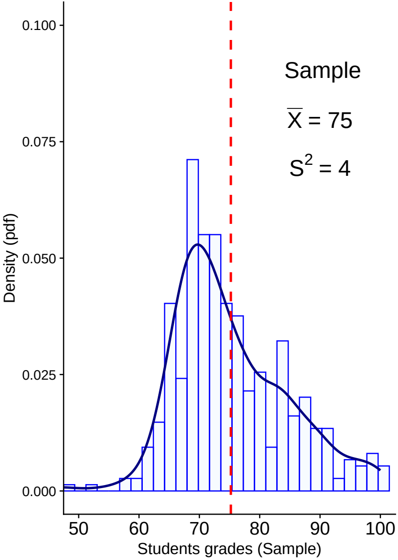
We distinguish between the **Data Generating Process** (DGP) and the data **sample**.

- 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 ($\mu$) and variance ($\sigma^2$), 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

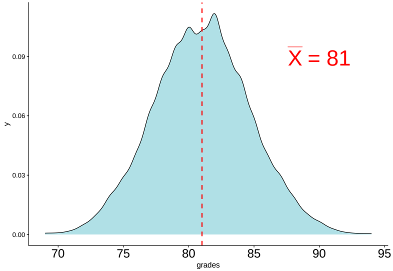
- The **expectation** $E[.]$ of a random variable X, denoted as $E[X]$, is a useful measure of central tendency of the DGP.
  - The expectation is also called the **expected value** or **mean**.
  - In the case of the normal distribution, the expectation is the first **central moment** and is denoted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

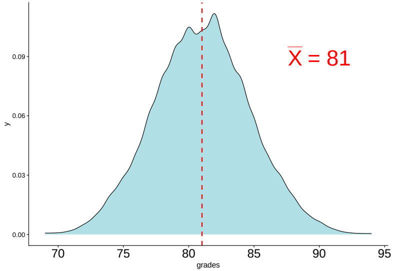
$$E[grades] = ?$$



# Unconditional distributions

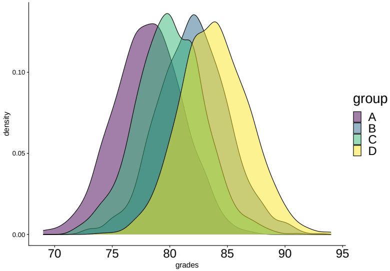
- We have a sample of UW students' grades.
- What may be a good candidate to estimate the mean of this population?

$$E[grades] = 81$$



# Conditional distributions

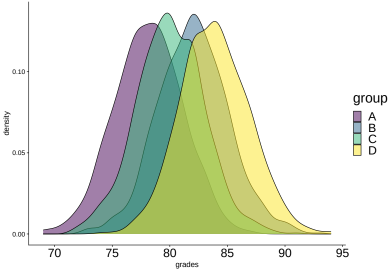
- We can compare the grade distribution for these different **sub-populations**.
  - Group A
  - Group B
  - Group C
  - Group D



# Conditional distributions

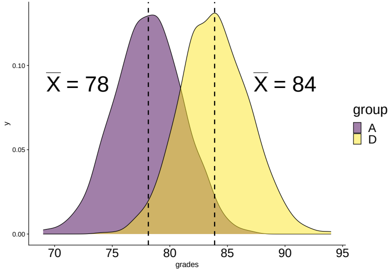
- We can **condition** grades on on a fixed value ($x$) of the group random variable.
- We call this the **conditional mean** (or **conditional expectation**).

$$E[grades \mid group = x]$$



# Conditional distributions

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




$$E[grades \mid group = A] = 78$$

$$E[grades \mid group = D] = 84$$

# 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*).

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

- 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

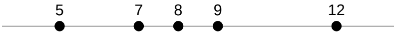
- The **best predictor**, denoted as $\theta$, minimizes **prediction error** ($e$), which is the distance of each data point from our best guess: $e = Y_i - \theta$.
- **Mean Squared Error** (MSE) quantifies the magnitude of prediction error.

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

*Note*: The notation $\theta$ is arbitrary and denotes the optimal or best predictor.

# Prediction error: first guess

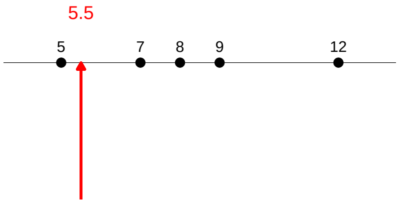
What is your **best guess** ($\theta$) that **minimizes** the prediction error ($MSE$)?




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

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

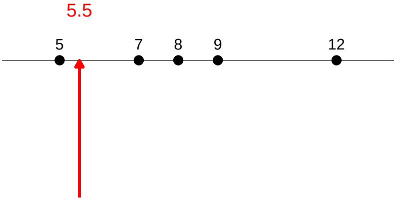
# Prediction error: first guess




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

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

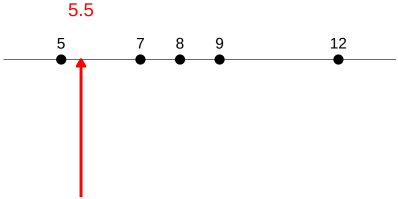
# Prediction error: first guess




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

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

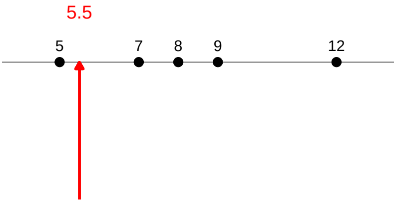
# Prediction error: first guess




| $N_i$ | $Y_i$ | $\theta$ | $Y_i - \theta$ | *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 |

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

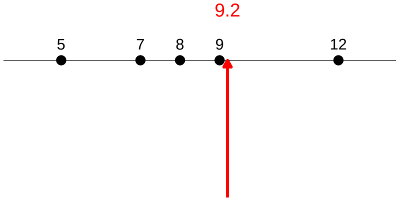
# Prediction error: first guess




| $N_i$ | $Y_i$ | $\theta$ | $Y_i - \theta$ | *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 |

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

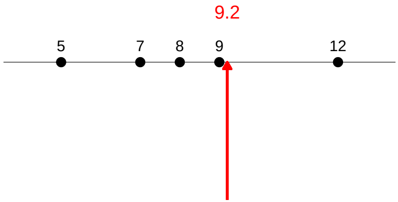
# Prediction error: second guess




| $N_i$ | $Y_i$ | $\theta$ | $Y_i - \theta$ | *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 | |

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

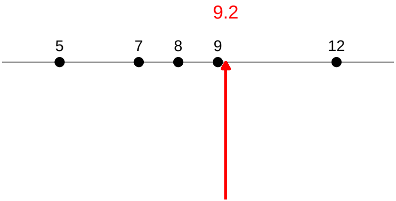
# Prediction error: second guess




| $N_i$ | $Y_i$ | $\theta$ | $Y_i - \theta$ | 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




| $N_i$ | $Y_i$ | $\theta$ | $Y_i - \theta$ | 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 |

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

# 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**, $\theta_1$ or $\theta_2$?
  - 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:

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \tag{5}$$

- 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:

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

- 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

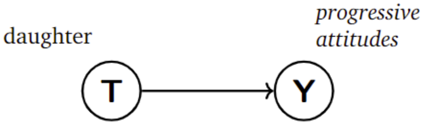
- In observational research designs, we cannot randomize an intervention, but we can identify causal effects by **conditioning** on confounders and making some (*heroic*) assumptions.
  - **Unconfoundedness** or selection on observable assumption.

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1) | X_i) \tag{6}$$

- 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.

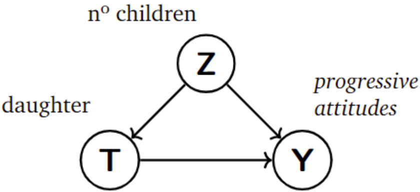
# Causality review: PS2, Q5

- 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.

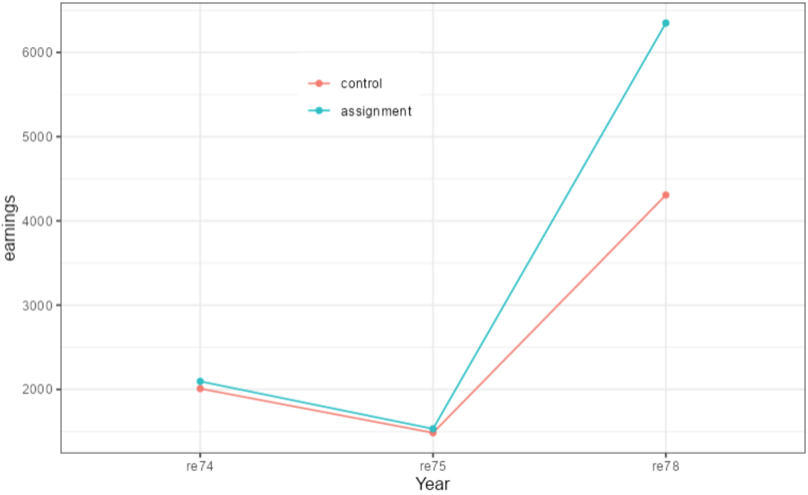
# Research design: Differences-in-Differences

$$\text{DiD} = [\bar{Y}(1)_{after} - \bar{Y}(0)_{after}] - [\bar{Y}(1)_{before} - \bar{Y}(0)_{before}] \quad (7)$$

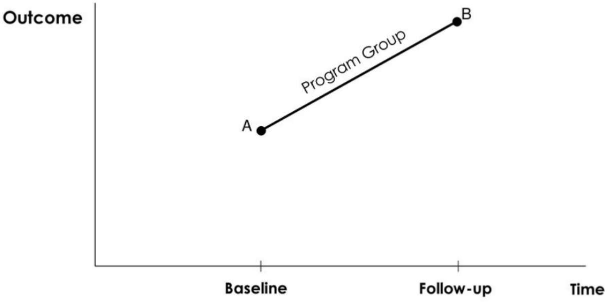
- Where
  - DiD is the differences-in-differences estimator,
  - $\bar{Y}(1)_{after}$ is the average unemployment for New Jersey **after** increasing the MW,
  - $\bar{Y}(0)_{after}$ is the average unemployment for Pennsylvania **after** *not* increasing the MW,
  - $\bar{Y}(1)_{before}$ is the average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

# Research design: Differences-in-Differences

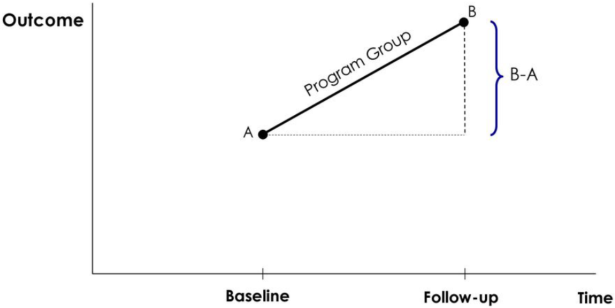
# Research design: Parallel trends



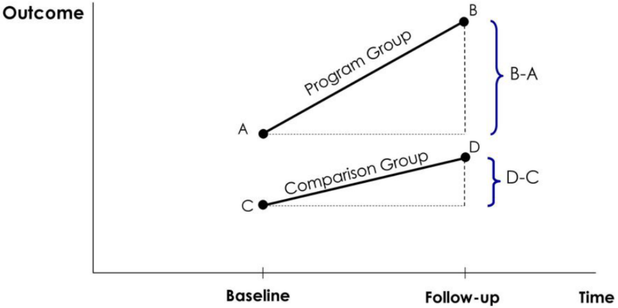
# Research design: Differences-in-Differences



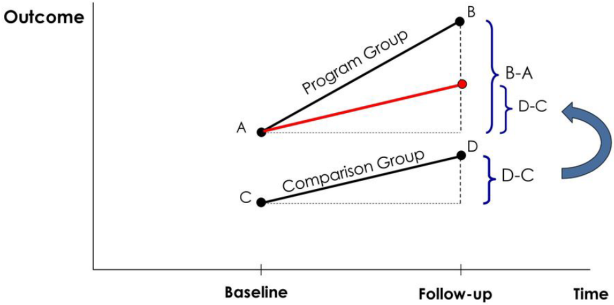
# Research design: Differences-in-Differences



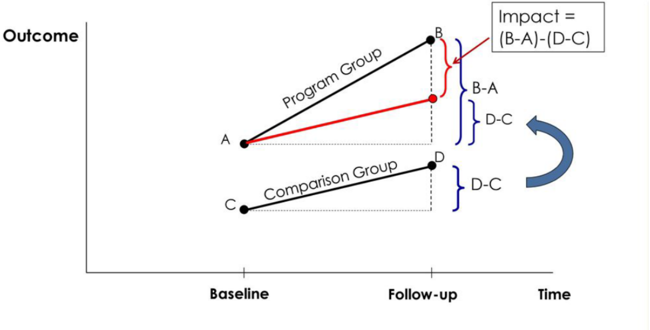
# Research design: Differences-in-Differences



# 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).

## Statistics: recap

So far, we have seen:

- The **population** mean and variance:

$$\mu = E[X] \quad ; \quad \sigma^2 = V[X] = E[(X-\mu)^2]$$

- The **sample** mean and variance:

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

*Note*:

- the **expectation** $E[.]$ is an operator that calculates the **average value** of a function of a random variable.
- *disclaimer*: it is actually more than an average, but for now it is *"fine"*.

## Statistics: recap

- the **population** mean ($\mu$) and variance ($\sigma^2$) are fixed quantities (*TRUEs*) of a **data generating process**.
- the **sample** mean ($\bar{X}$) and variance ($S^2$) are random variables, and **estimators** of the population parameters ($\mu$ and $\sigma^2$).

In addition, we have seen:

- The conditional expectation (or mean): $E[Y|X]$.
- The mean squared error, a measurement of **prediction error**:
  - MSE : $E[\,(Y_i - \theta)^2\,]$

## Statistics: covariance

Note:

$$V[X] = E[(X - \mu)^2] = E[(X - \mu)(X - \mu)]$$

We can ask how much **two variables** vary together with the covariance:

$$Cov[Y, X] = E[(Y - \mu_Y)(X - \mu_X)]$$

- **Covariance** measures the degree to which two random variables change (*vary*) together.
- It quantifies the extent of **linear association** between two variables.

## Statistics: correlation

- A drawback of **covariance** is its sensitivity to the original numeric **scale** of each variable (Y and X).
- To normalize its scale, we can compute the ratio of each variable's **standard deviation**, resulting in Pearson's correlation:

$$\rho = \frac{Cov[Y, X]}{S(Y)S(X)}$$

- It offers a standardized measure of the **strength** and **direction** of the linear relationship between two variables.

# 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 $\mu$!

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

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

# Bivariate regression

$$Y_i = \alpha + \beta X_i + e_i \quad (8)$$

Notation:

- $Y$ is the **outcome** or dependent variable.
- $X$ (or $T$) is the **predictor**, covariate, or independent variable.
- $\alpha$ (or sometimes $\beta_0$) is the **intercept**.
- $\beta$ are **coefficients** or slopes of linear relationships.
- $e$ is the **error** term or disturbance.
- Subscript $i$ refers to each observation (row).

*Research question*: what is the relationship between fertility and education?

- $Y$: Fertility rates.
- $X$: Education Beyond Primary School.

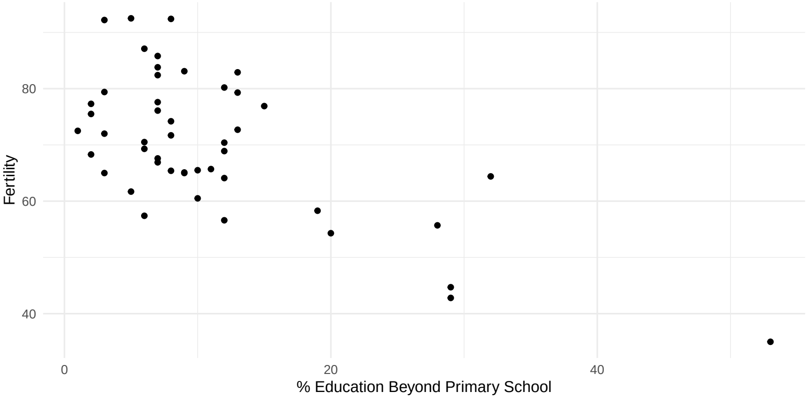
# Bivariate regression

- **Note**: in this case, each *i*th refers to a municipality from Switzerland.

```
##              Fertility Education Agriculture Examination Catholic
## Courtelary        80.2        12        17.0          15     9.96
## Delemont          83.1         9        45.1           6    84.84
## Franches-Mnt      92.5         5        39.7           5    93.40
## Moutier           85.8         7        36.5          12    33.77
## Neuveville        76.9        15        43.5          17     5.16
## Porrentruy        76.1         7        35.3           9    90.57
```

# Bivariate regression

Is there a negative or positive relationship between education and fertility? How strong is this relationship? What would “no relationship” look like visually?



# Bivariate regression: correlation

We can quantify this direction and strength by **correlation**:

```
cor(swiss$Education, swiss$Fertility)
```

```
## [1] -0.6637889
```

```
swiss %>%
  select(Education,Fertility) %>%
  cor() %>% round(digits=2)
```

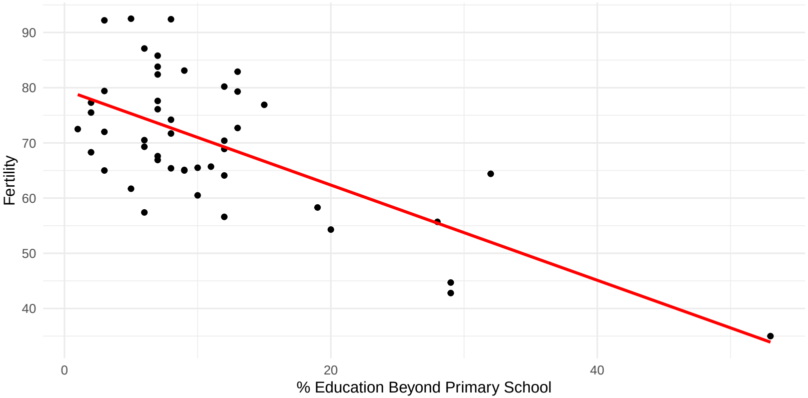
```
##           Education Fertility
## Education      1.00     -0.66
## Fertility     -0.66      1.00
```

# Bivariate regression: correlation

- Assumes **linear** relationship: it measures the **strength** and **direction** of a linear association between variables.
  - Not optimal for **non-linear** relationships.
- Values range from -1 to 1.
- Interpreting the magnitude of the coefficient: in general, **larger** absolute values of the correlation coefficient indicate **stronger** relationships.

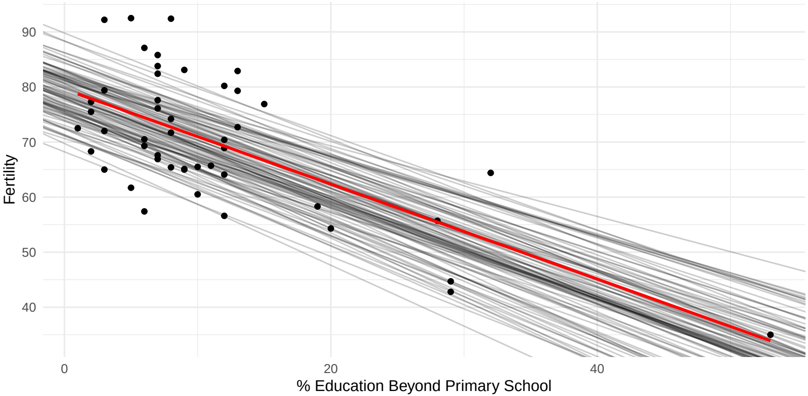
# Bivariate regression: ggplot

- We can ask ggplot to plot a regression line fit on top of our scatter.
  - `geom_smooth(method="lm")`



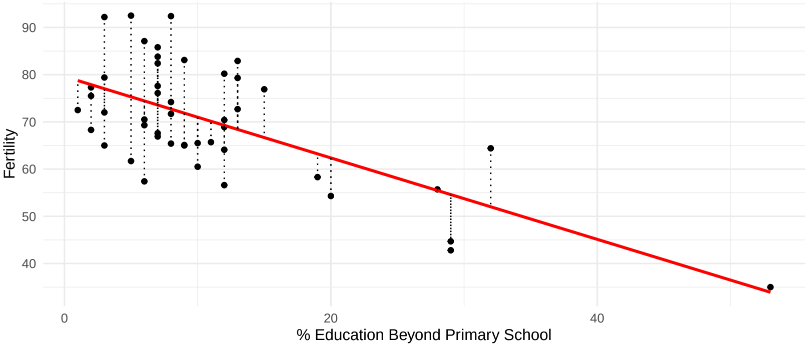
# Bivariate regression: OLS

- How does R draw this regression line?
  - In fact, you can draw many lines that "pass through" those points:



# Bivariate regression: OLS

- If I ask you to draw only one line that *"best predicts the relationship."* How do we pick the *"best fitting"* line?
  - The answer is in the **OLS** (ordinary least squares) estimator
  - OLS is the line that **minimizes** the **sum of squared distance** (*error*) of all points.



# Bivariate regression: `lm()` function

- How do we run regression to produce the best fitting line?

```r
res <- lm(Fertility ~ Education, data = swiss)
coef(res)
```

```
## (Intercept)   Education
##  79.6100585  -0.8623503
```

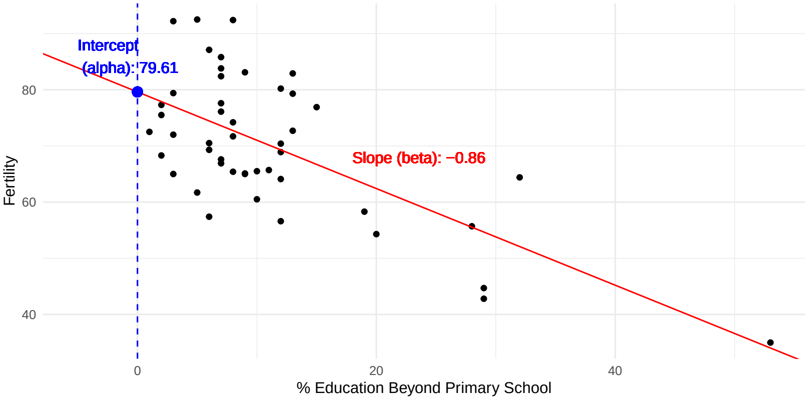
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_i \tag{9}$$

$$\widehat{Fertility}_i = (79.61) + (-0.86) Education_i \tag{10}$$

- **Prediction**: If education increases in 1 unit, *all else equal*, fertility ($\hat{Y}$) deceases in -0.86 units.

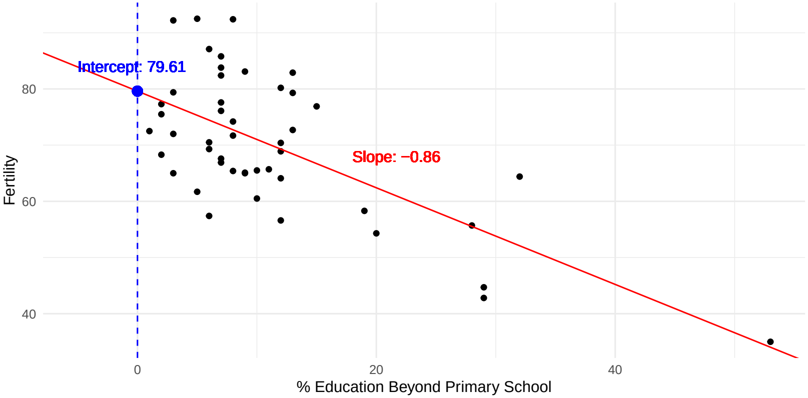
# Bivariate regression: estimates

- Visualizing $\hat{\alpha}$ and $\hat{\beta}$:



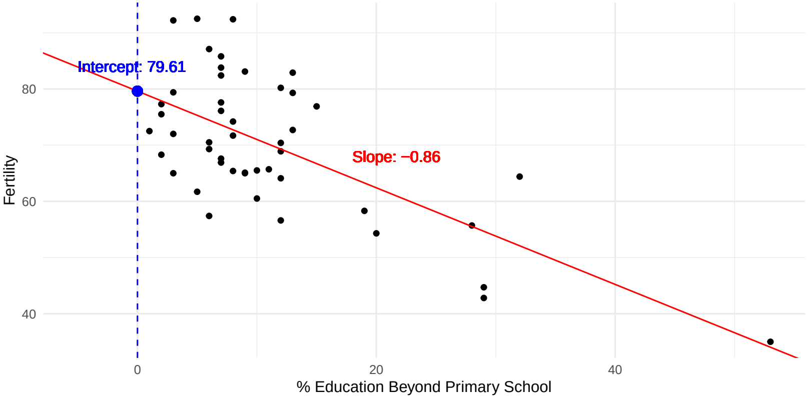
# Bivariate regression: prediction

Estimated model: $\hat{Fertility}_i = \hat{\alpha} + \hat{\beta}_1 Education_i$



# Bivariate regression: prediction

Empirical model: $\hat{Fertility}_i = 79.61 - 0.86 * Education_i$

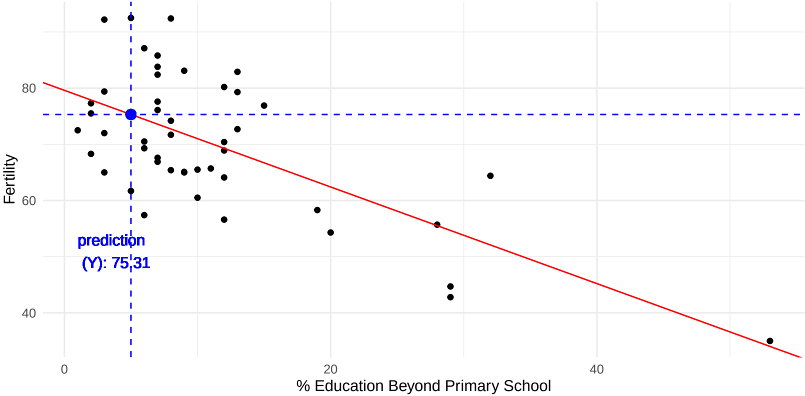


# Bivariate regression: prediction

What is the predicted fertility rate when education is at 5?

$\hat{Fertility}_i = 79.61 - 0.86 * Education_i$

$75.31 = 79.61 - 0.86 * 5$

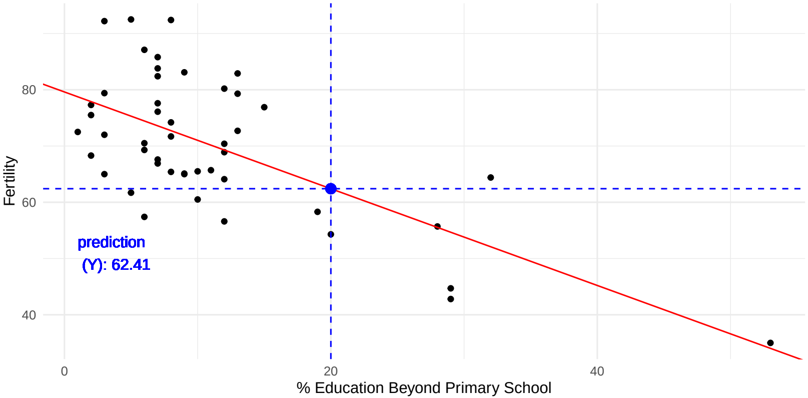


# Bivariate regression: prediction

What is the predicted fertility rate when education is at 20?

$\hat{Fertility}_i = 79.61 - 0.86 * Education_i$

$62.41 = 79.61 - 0.86 * 20$

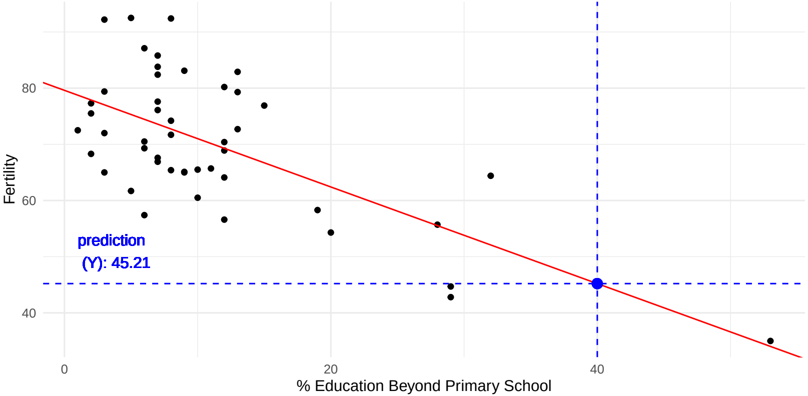


# Bivariate regression: prediction

What is the predicted fertility rate when education is at 40?

$\hat{Fertility}_i = 79.61 - 0.86 * Education_i$

$45.21 = 79.61 - 0.86 * 40$

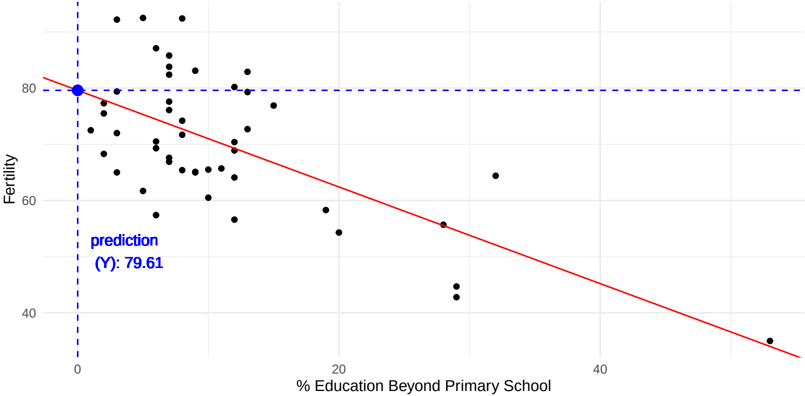


# Bivariate regression: prediction

What is the predicted fertility rate when education is at 0?

$\hat{Fertility}_i = 79.61 - 0.86 * Education_i$

$79.61 = 79.61 - 0$



# Bivariate regression: DGP/population and sample

- A **population model** that represents a data generating process:

$$Y_i = \alpha + \beta X_i + e_i \tag{11}$$

- The **sample model** that we estimate:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_i \tag{12}$$

- We can use the predicted outcomes ($\hat{Y}$) from our empirical model to estimate the **prediction error** or **residuals**:

$$\hat{e}_i = Y_i - \hat{Y}_i \tag{13}$$

## Best fitting model: MSE

Recall that $\theta$ was our best predictor that *minimizes* the **sum of squared errors** (SSE) and we take the mean (*expectation*) to compute the **mean squared error** (MSE).

$$SSE : (Y - \theta)^2 \qquad MSE : E[(Y - \theta)^2]$$

We can calculate the MSE from the regression analysis, where $\theta$ concerns now each model parameter.

$$\text{Intercept-only model} : E[(Y - \mu)^2]$$

$$\text{Bivariate model} : E[(Y - (\alpha + \beta_1))^2]$$

**Model comparison**: the model with the lowest **MSE** is the one that provides the **best fit**.

## Bivariate regression: intercept.

- The intercept, denoted by $\alpha$, represents the **predicted value** of the outcome variable $\hat{Y}$ when all covariates on the left-hand side of the equation are set to 0.
- The intercept is estimated as function of the **estimated slopes** and **sample means**:

Bivariate model:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} * \bar{X}_1 \tag{14}$$

- The intercept is **not equivalent** to the sample mean value of the outcome, $\bar{Y}$, when all covariates are 0.
  - **Exception**: If the covariates are **centered**, which means they are transformed to have a mean of 0. E.g., $X_i - \bar{X}$.

# Bivariate regression: intercept.

Bivariate model: $\hat{\alpha} = \bar{Y} - (\hat{\beta} * \bar{X}_1)$

```
(Y_mean <- mean(swiss$Fertility)) # sample mean of Y
```

```
## [1] 70.14255
```

```
(X_mean <- mean(swiss$Education)) # sample mean of X
```

```
## [1] 10.97872
```

```
Y_mean - (beta * X_mean) # estimating the intercept
```

```
## Education
##  79.61006
```

```
intercept
```

```
## (Intercept)
##    79.61006
```

# Bivariate regression: slope.

- In a **bivariate regression**, the estimated slope coefficient represents the change in the dependent variable (Y) associated with a unit increase in the independent variable (X).
  - Empirical model: $\hat{Fertility}_i = 79.61 - 0.86 * Education_i$.
  - Interpretation: $\beta$ has a slope of $-0.86$, and represents the **average** change in *Fertility* for every unit of increase in *Education*.
- In the **multivariate regression**, the estimated slope gives us the expected change in Y for each unit increase in X, holding all other variables constant (at their means).

# Multivariate regression: slope.

The slope in a **multivariate analysis** is influenced by the inclusion of variables and their relationships with the outcome variable.

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + e_i \tag{15}$$

- Including the variable Z affects the estimated coefficient of X.
- In the presence of Z, the interpretation of the coefficient of X changes from the bivariate case.
  - It now reflects the effect of X on Y while controlling for the impact of Z on Y and keeping Z values constant.

# Multivariate regression: slope.

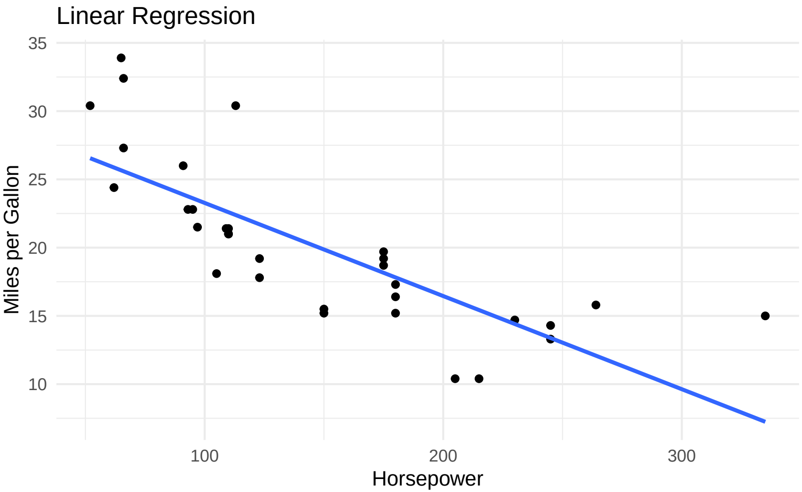
The formula for the estimated coefficient of X in the multivariate regression is:

$$\hat{\beta} = \frac{Cov(X, Y|Z)}{Var(X|Z)} \tag{16}$$

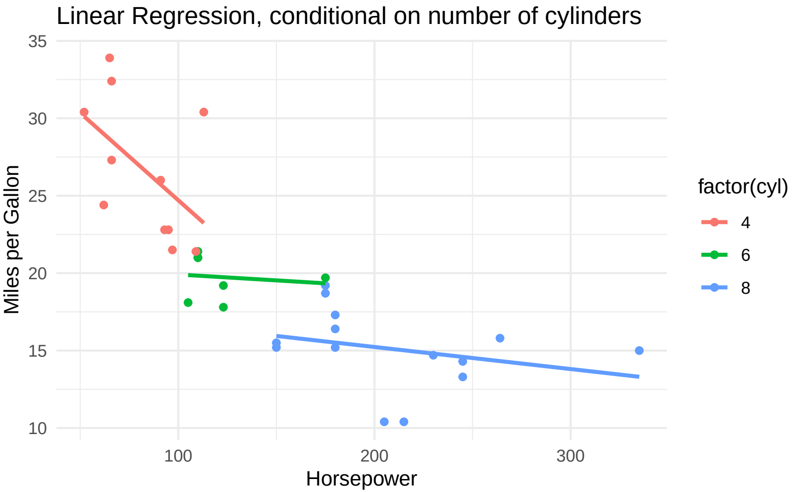
where $cov(X, Y|Z)$ is the **conditional covariance** between X and Y given Z, and $var(X|Z)$ is the **conditional variance** of X given Z.

**Bottom line**: in a multivariate regression, the inclusion of each variable affects the estimation of other parameters, including coefficients and intercept, due to interdependence among variable variations.

# Confounding



# Confounding



# Function: `stargazer()`

- To present results from several models in a output table, use the function `stargazer()`.
  - In the `RMarkdown`, you will need to activate the code chunk option `results='asis'`

```
library(stargazer)
m1 <- lm(mpg ~ hp, data=mtcars)
m2 <- lm(mpg ~ hp + cyl, data=mtcars)
```

# Function: `stargazer()`

```
stargazer(m1,m2,header = FALSE,typ="latex") # type="text" for R console
```

**Table 1:**

| | *Dependent variable:* | |
|---|---|---|
| | mpg | |
| | (1) | (2) |
| hp | $-0.068^{***}$ | $-0.019$ |
| | (0.010) | (0.015) |
| cyl | | $-2.265^{***}$ |
| | | (0.576) |
| Constant | $30.099^{***}$ | $36.908^{***}$ |
| | (1.634) | (2.191) |
| Observations | 32 | 32 |
| $R^2$ | 0.602 | 0.741 |
| Adjusted $R^2$ | 0.589 | 0.723 |

## Time to code a little bit!

- Complete the activity `Regression.rmd`