SciencesPo
DEPARTMENT OF ECONOMICS

ScPoEconometrics

Multiple Regression Model

Florian Oswald, Gustave Kenedi, Mylène Feuillade and Pierre Villedieu
SciencesPo Paris
2022-03-08

Quick "Quiz" on Last Week's Material

1. From your ***computer*** 👉 connect to ***www.wooclap.com/SCPOCAUSALITY***

   OR

2. From your ***phone*** 👉 flash QR code below

Today - Multiple Regression Model

- Multiple independent variables in our model
- Interpretation for continuous and dummy regressors
- Dummy variable trap
- Omitted variable bias
- Adjusted R_2
- Empirical applications:
    - *Class size* and *student performance*

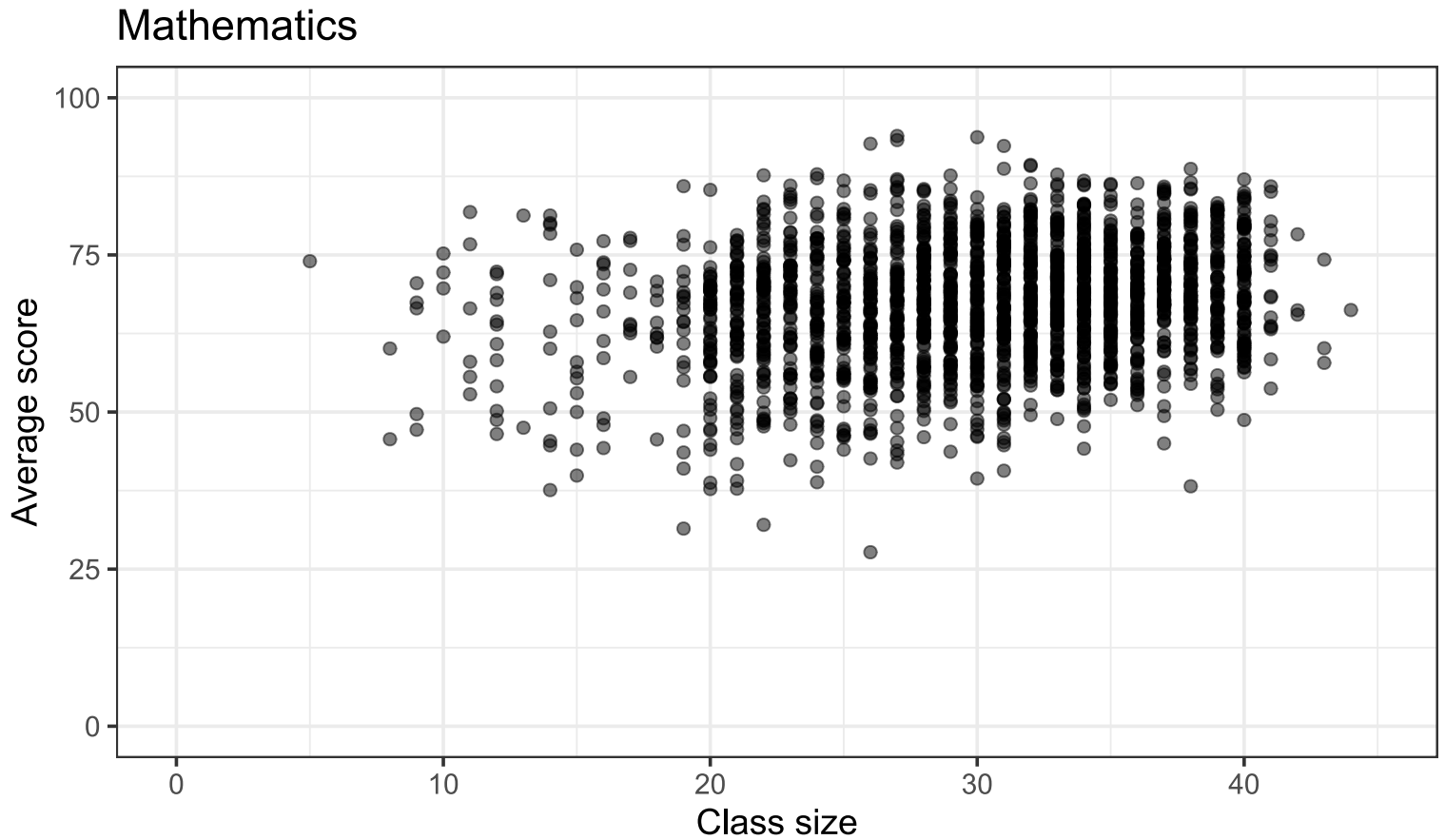
Class size and student performance

- Let's go back to Angrist and Lavy's (1999)'s analysis of the effect of class size on student performance in Israel.

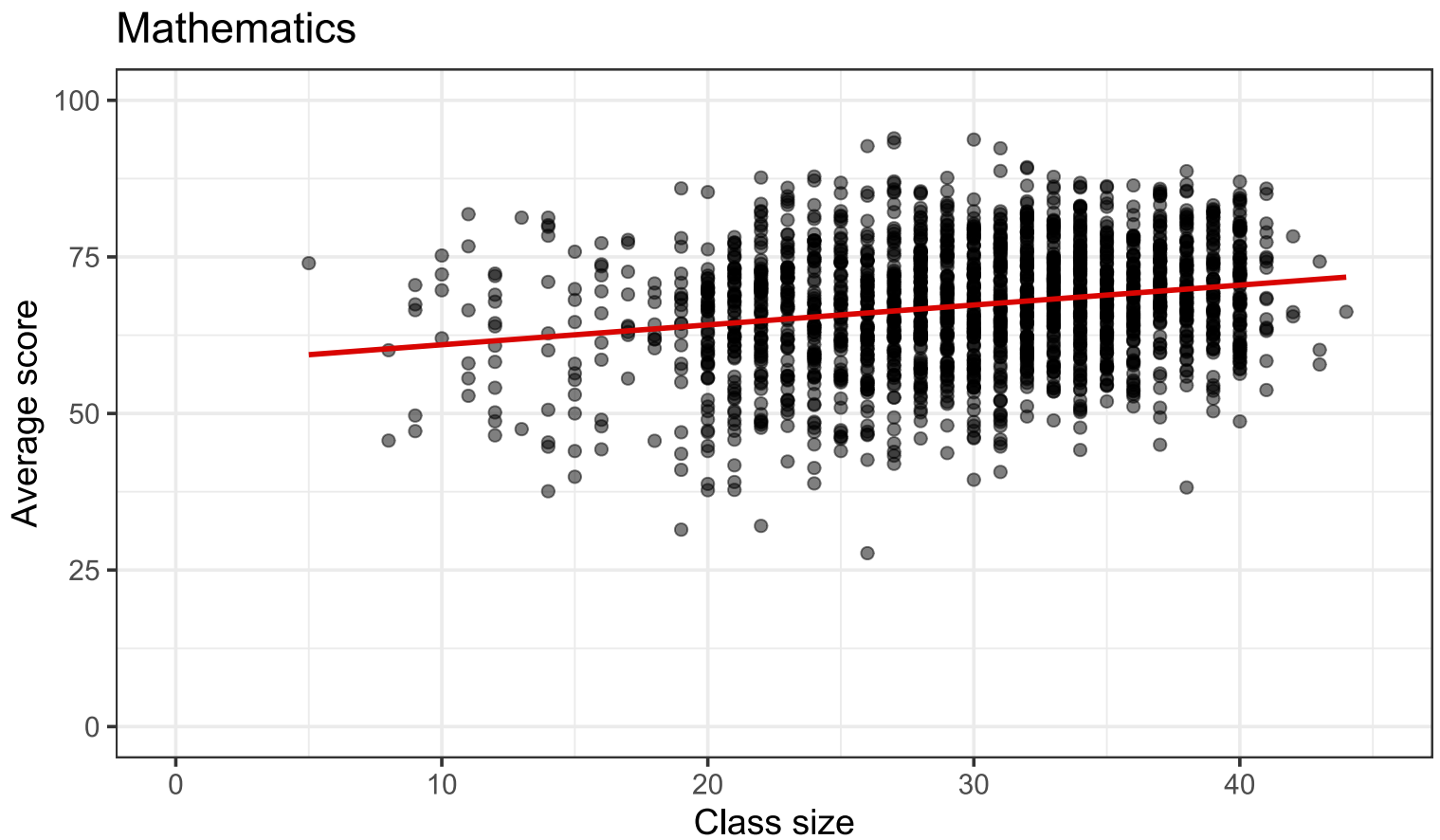
Class size and student performance

- Let's go back to Angrist and Lavy's (1999)'s analysis of the effect of class size on student performance in Israel.
- With a **simple linear regression**, we found that class size was positively ***associated*** with students' scores in maths and reading.

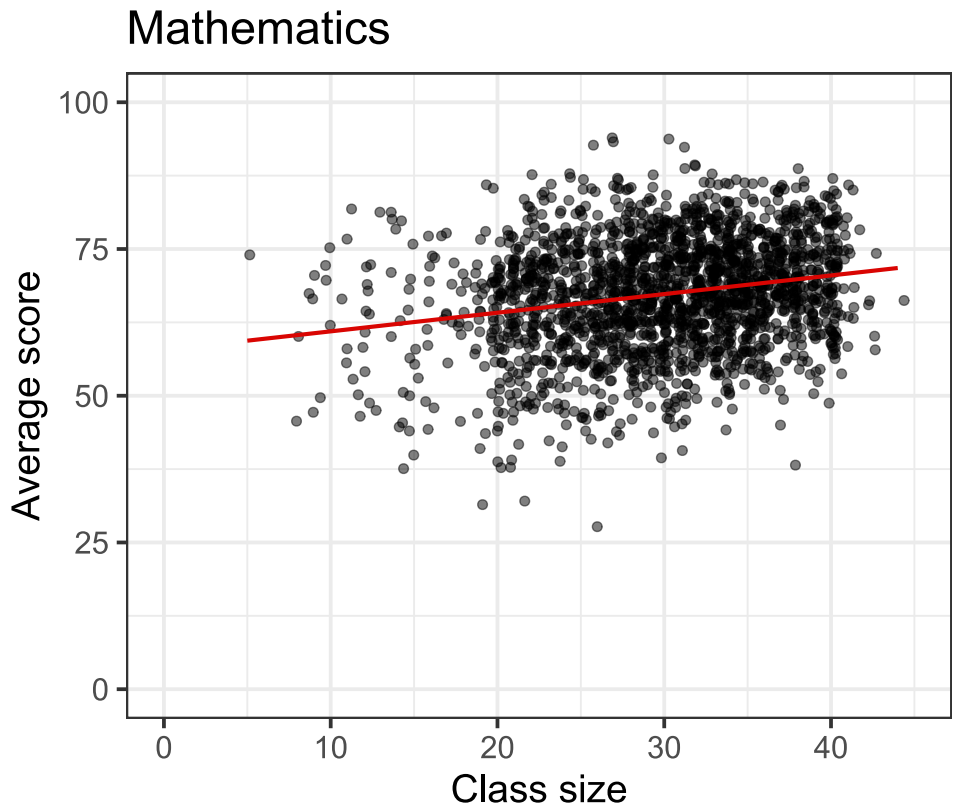
Class size and student performance: Raw relationship



Class size and student performance: Raw relationship



Class size and student performance: Raw relationship




```
lm(avgmath ~ classize, grades)
```

```
##
## Call:
## lm(formula = avgmath ~ classize, data = grades)
##
## Coefficients:
## (Intercept)     classize
##     57.7939       0.3175
```

Class size and student performance

- Let's go back to Angrist and Lavy's (1999)'s analysis of the effect of class size on student performance in Israel.
- With a **simple linear regression**, we found that class size was positively ***associated*** with students' scores in maths and reading.
- This is intuitively unexpected, and contrasts with the simple results from the *STAR* randomized experiment.

Class size and student performance

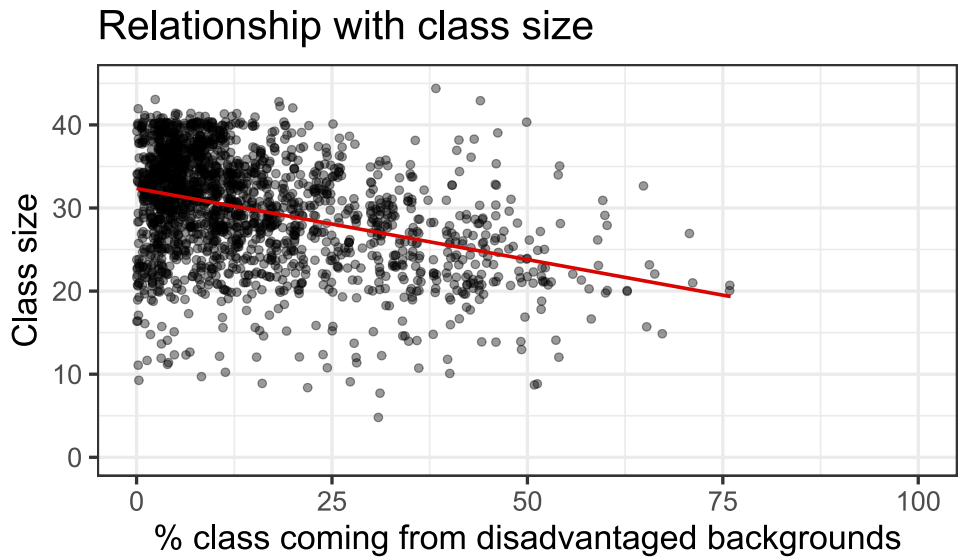
- Let's go back to Angrist and Lavy's (1999)'s analysis of the effect of class size on student performance in Israel.
- With a **simple linear regression**, we found that class size was positively ***associated*** with students' scores in maths and reading.
- This is intuitively unexpected, and contrasts with the simple results from the *STAR* randomized experiment.
- Could it be that some other variable may be related to class size ***as well as*** students' performance?
- In particular, we mentioned the **location effect**: large classes may be more common in wealthier and bigger cities, while small classes may be more likely in poorer rural areas.

Class size and student performance

- Let's go back to Angrist and Lavy's (1999)'s analysis of the effect of class size on student performance in Israel.
- With a **simple linear regression**, we found that class size was positively ***associated*** with students' scores in maths and reading.
- This is intuitively unexpected, and contrasts with the simple results from the *STAR* randomized experiment.
- Could it be that some other variable may be related to class size ***as well as*** students' performance?
- In particular, we mentioned the **location effect**: large classes may be more common in wealthier and bigger cities, while small classes may be more likely in poorer rural areas.
- Let's investigate this hypothesis.

Class size and student performance: Confounders

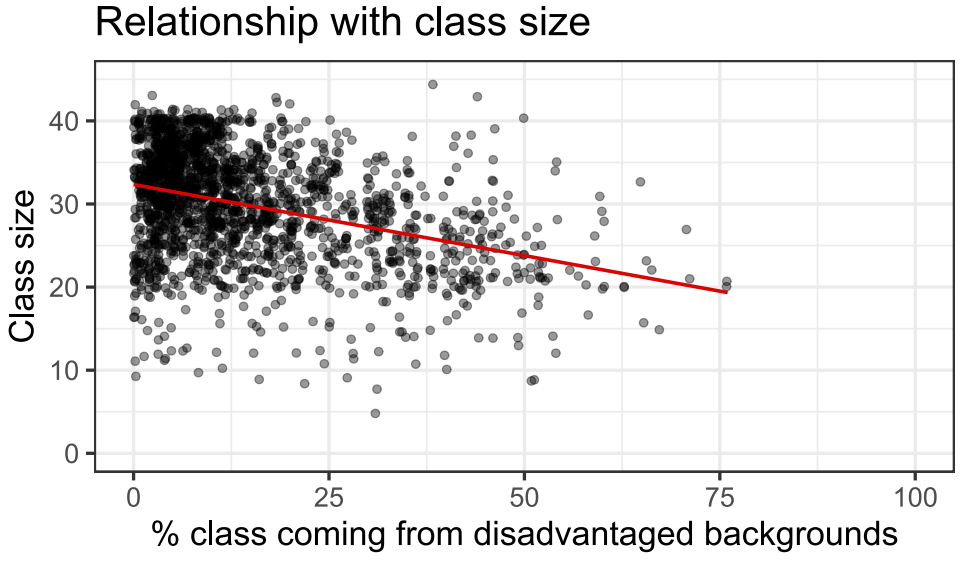
Link between **class size** and **the share of students who come from** `disadvantaged` **backgrounds** in the class.



👉 On average, there is a greater % of disadvantaged students in smaller classes.

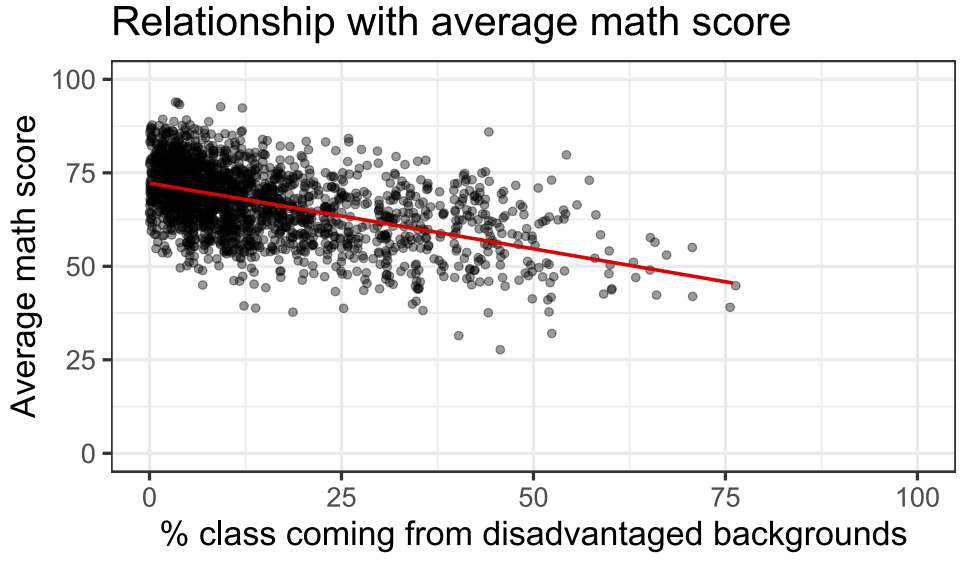
Class size and student performance: Confounders

Link between **class size** and **the share of students who come from** `disadvantaged` **backgrounds** in the class.



👉 On average, there is a greater % of disadvantaged students in smaller classes.

Link between **average math score** and **the share of students who come from** `disadvantaged` **backgrounds** in the class.



👉 On average, the greater the % of students coming from a disadvantaged background, the lower the average math score.

Class size and student performance: Multiple regression

- Suppose we want to know the effect of class size on average math scores, ***controlling for*** the fact that there is a negative relationship between the % of disadvantaged students and class size **AND** average math score.

Class size and student performance: Multiple regression

- Suppose we want to know the effect of class size on average math scores, ***controlling for*** the fact that there is a negative relationship between the % of disadvantaged students and class size **AND** average math score.
- To do so, we have to include both `classize` and `disadvantaged` variables as *regressors* in the regression.

Class size and student performance: Multiple regression

- Suppose we want to know the effect of class size on average math scores, ***controlling for*** the fact that there is a negative relationship between the % of disadvantaged students and class size **AND** average math score.
- To do so, we have to include both `classize` and `disadvantaged` variables as *regressors* in the regression.
- As such we can obtain an estimate of the effect of class size on average math score, ***purged* of the effect of the disadvantaged variable**.

Class size and student performance: Multiple regression

- Suppose we want to know the effect of class size on average math scores, ***controlling for*** the fact that there is a negative relationship between the % of disadvantaged students and class size **AND** average math score.
- To do so, we have to include both `classize` and `disadvantaged` variables as *regressors* in the regression.
- As such we can obtain an estimate of the effect of class size on average math score, ***purged* of the effect of the disadvantaged variable**.
- The model we want to estimate becomes:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \% \text{ disadvantaged}_i + e_i$$

Class size and student performance: Multiple regression

- Suppose we want to know the effect of class size on average math scores, ***controlling for*** the fact that there is a negative relationship between the % of disadvantaged students and class size **AND** average math score.
- To do so, we have to include both `classize` and `disadvantaged` variables as *regressors* in the regression.
- As such we can obtain an estimate of the effect of class size on average math score, ***purged* of the effect of the disadvantaged variable**.
- The model we want to estimate becomes:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \% \text{ disadvantaged}_i + e_i$$

- This is ***multiple regression***! We will estimate this model in a few slides. Let's formalize what we have seen so far.

Multiple Regression's Purpose

- Recall from two weeks ago, the **Simple Linear Model** can be written as

$$y_i = b_0 + b_1 x_i + e_i,$$

where y_i is the ***dependent variable*** and x_i is the ***independent variable***.

Multiple Regression's Purpose

- Recall from two weeks ago, the **Simple Linear Model** can be written as

$$y_i = b_0 + b_1 x_i + e_i,$$

where y_i is the ***dependent variable*** and x_i is the ***independent variable***.

- Remember: We say that `X` *causes* `Y` when if we were to intervene and change the value of `X` ***without changing anything else*** then `Y` would also change as a result.

Multiple Regression's Purpose

- Recall from two weeks ago, the **Simple Linear Model** can be written as

$$y_i = b_0 + b_1 x_i + e_i,$$

where y_i is the ***dependent variable*** and x_i is the ***independent variable***.

- Remember: We say that `X` *causes* `Y` when if we were to intervene and change the value of `X` ***without changing anything else*** then `Y` would also change as a result.

⚠ Unless all other factors affecting y_i are uncorrelated with x_i, b_1 **cannot be interpreted as a causal effect**.

Multiple Regression's Purpose

- Recall from two weeks ago, the **Simple Linear Model** can be written as

$$y_i = b_0 + b_1 x_i + e_i,$$

where y_i is the ***dependent variable*** and x_i is the ***independent variable***.

- Remember: We say that `X` *causes* `Y` when if we were to intervene and change the value of `X` ***without changing anything else*** then `Y` would also change as a result.

⚠ Unless all other factors affecting y_i are uncorrelated with x_i, b_1 **cannot be interpreted as a causal effect**.

We need to **enrich the model** and take into account factors that are simultaneously related to y_i **and** x_i.

Multiple Regression Model

The expanded model can be written as:

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + b_3 x_{3,i} + \cdots + b_k x_{k,i} + e_i,$$

where x_1, x_2, ..., x_k are k regressors, and b_1, b_2, ..., b_k are the associated k coefficients.

Multiple Regression Model

The expanded model can be written as:

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + b_3 x_{3,i} + \cdots + b_k x_{k,i} + e_i,$$

where x_1, x_2, ..., x_k are k regressors, and b_1, b_2, ..., b_k are the associated k coefficients.

Estimation: We obtain the values for $(b_0, b_1, b_2, \ldots, b_k)$ in the same way as before, using **OLS**.

- $(b_0^{OLS}, b_1^{OLS}, b_2^{OLS}, \ldots, b_k^{OLS})$ are the values that minimize the **Sum of Squared Residuals**.
- That is they minimize

$$\begin{aligned}\sum_i e_i^2 &= \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum_i \left[y_i - (b_0 + b_1 x_{1,i} + b_2 x_{2,i} + b_3 x_{3,i} + \cdots + b_k x_{k,i})\right]^2\end{aligned}$$

Multiple Regression Model: Interpretation

For now assume both the dependent variable (y_i) and the independent variables (x_k) are numeric.

> Intercept (b_0): **The predicted value of y $(\hat{y})$ if all the regressors $(x_1, x_2, x_3, \dots)$ are equal to 0.**

Multiple Regression Model: Interpretation

For now assume both the dependent variable (y_i) and the independent variables (x_k) are numeric.

> Intercept (b_0): **The predicted value of y $(\hat{y})$ if all the regressors $(x_1, x_2, x_3, \dots)$ are equal to 0.**
>
> Slope (b_k): **The predicted change, on average, in the value of y *associated* to a one-unit increase in x_k...**
>
> **... keeping all the other regressors constant!**

Multiple Regression Model: Interpretation

For now assume both the dependent variable (y_i) and the independent variables (x_k) are numeric.

> Intercept (b_0): **The predicted value of y $(\hat{y})$ if all the regressors $(x_1, x_2, x_3, \dots)$ are equal to 0.**
>
> Slope (b_k): **The predicted change, on average, in the value of y *associated* to a one-unit increase in x_k...**
>
> **... keeping all the other regressors constant!**

- Notice that the *keeping all the other regressors constant* is the only part that changes compared to SLM.
- In other words, you are considering the individual effect of the variable x_k on y **in isolation** of the effect that the other regressors might have on y.

Multiple Regression Model: Interpretation

For now assume both the dependent variable (y_i) and the independent variables (x_k) are numeric.

> Intercept (b_0): **The predicted value of y $(\hat{y})$ if all the regressors $(x_1, x_2, x_3, \dots)$ are equal to 0.**
>
> Slope (b_k): **The predicted change, on average, in the value of y *associated* to a one-unit increase in x_k...**
>
> **... keeping all the other regressors constant!**

- Notice that the *keeping all the other regressors constant* is the only part that changes compared to SLM.
- In other words, you are considering the individual effect of the variable x_k on y **in isolation** of the effect that the other regressors might have on y.
- **Link with causal inference**: Only the regressors included in the model are held constant, those that are not in the model can still vary and "bias" your estimates.

Multiple Regression with R

- Very similar to simple linear regression:

```
lm(formula = dependent variable ~  independent variable 1 + independent variable 2 + ...,
   data = data.frame containing the data)
```

Multiple Regression with R

- Very similar to simple linear regression:

```
lm(formula = dependent variable ~  independent variable 1 + independent variable 2 + ...,
   data = data.frame containing the data)
```

Class size and student performance: Multiple regression

Let's estimate the model from earlier on by OLS:
$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \% \text{ disadvantaged}_i + e_i$

```
lm(avgmath ~ classize + disadvantaged, grades)
```

```
##
## Call:
## lm(formula = avgmath ~ classize + disadvantaged, data = grades)
##
## Coefficients:
##   (Intercept)       classize  disadvantaged
##      69.94438        0.07168       -0.33958
```

Class size and student performance: Multiple regression

```
##
## Call:
## lm(formula = avgmath ~ classize + disadvantaged, data = grades)
##
## Coefficients:
##   (Intercept)       classize  disadvantaged
##      69.94438        0.07168       -0.33958
```

Questions

1. How do you interpret each of these coefficients?
2. How do you explain the change in the `classize` coefficient compared to the SLM case?

Class size and student performance: Multiple regression

```
##
## Call:
## lm(formula = avgmath ~ classize + disadvantaged, data = grades)
##
## Coefficients:
##   (Intercept)       classize  disadvantaged
##      69.94438        0.07168       -0.33958
```

Answers

1. How do you interpret each of these coefficients?

- b_0 = 69.94: When `class size` and `disadvantaged` are set to 0, the *predicted* value of the average math score is 69.94.

Class size and student performance: Multiple regression

```
##
## Call:
## lm(formula = avgmath ~ classize + disadvantaged, data = grades)
##
## Coefficients:
##   (Intercept)       classize  disadvantaged
##      69.94438        0.07168       -0.33958
```

Answers

1. How do you interpret each of these coefficients?

- b_0 = 69.94: When `class size` and `disadvantaged` are set to 0, the *predicted* value of the average math score is 69.94.
- b_1 = 0.07: Keeping the percentage of *disadvantaged students* constant in the class, a 1-student increase in class size is ***associated, on average,*** with a 0.07 point increase in average math score.

Class size and student performance: Multiple regression

```
##
## Call:
## lm(formula = avgmath ~ classize + disadvantaged, data = grades)
##
## Coefficients:
##   (Intercept)       classize  disadvantaged
##      69.94438        0.07168       -0.33958
```

Answers

1. How do you interpret each of these coefficients?

- b_0 = 69.94: When `class size` and `disadvantaged` are set to 0, the *predicted* value of the average math score is 69.94.
- b_1 = 0.07: Keeping the percentage of *disadvantaged students* constant in the class, a 1-student increase in class size is ***associated, on average,*** with a 0.07 point increase in average math score.
- b_2 = - 0.34: Keeping the *class size* constant, a 1-***percentage point*** increase in the share of *disadvantaged students* is ***associated, on average,*** with a 0.34 point decrease in average math score.

Class size and student performance: Multiple regression

```
##
## Call:
## lm(formula = avgmath ~ classize + disadvantaged, data = grades)
##
## Coefficients:
##   (Intercept)       classize  disadvantaged
##      69.94438        0.07168       -0.33958
```

Answers

2. How do you explain the change in the `classize` coefficient compared to the SLM case?

Class size and student performance: Multiple regression

```
##
## Call:
## lm(formula = avgmath ~ classize + disadvantaged, data = grades)
##
## Coefficients:
##   (Intercept)       classize  disadvantaged
##      69.94438        0.07168       -0.33958
```

Answers

2. How do you explain the change in the `classize` coefficient compared to the SLM case?

- b_1 decreases when the `disadvantaged` variable is taken into account. This was expected since part of the positive effect of class size was partly due to the smaller share of disadvantaged students in bigger classes.

Percentage vs. Percentage Point: A Primer

Example: % of disadvantaged students in class increases from 10 to 25 %?

Percentage vs. Percentage Point: A Primer

Example: % of disadvantaged students in class increases from 10 to 25 %?

Questions:

1. What's the *percentage point* change?
2. What's the *percentage* change?

Percentage vs. Percentage Point: A Primer

Example: % of disadvantaged students in class increases from 10 to 25 %?

Answers:

1. This is a $25 - 10 = 15$ ***percentage points*** increase.

Percentage vs. Percentage Point: A Primer

Example: % of disadvantaged students in class increases from 10 to 25 %?

Answers:

1. This is a $25 - 10 = 15$ ***percentage points*** increase.
2. This is a $\frac{25-10}{10}\% = 150$ ***percent*** increase.

Percentage vs. Percentage Point: A Primer

Example: % of disadvantaged students in class increases from 10 to 25 %?

Answers:

1. This is a $25 - 10 = 15$ ***percentage points*** increase.
2. This is a $\frac{25-10}{10}\% = 150$ ***percent*** increase.

You ***need*** to pay attention to whether you are talking about ***percentage points*** or ***percentage*** changes! They imply drastically different magnitudes!

Task 1

10:00

Let's analyse the regression results using **reading** score as the dependent variable.

1. Load the data from here using the `read_dta()` function from the `haven` package. Assign it to an object `grades`.

2. Regress `avgverb` on `classize` and `disadvantaged` and assign the output to a new object `reg`. Interpret the coefficients. How do they compare to the simple linear regression? How do they compare with the math score regression coefficients?

3. (Optional) What are the other available variables that we may add in the regression?

    - Run the regression with all these variables and assign it to `reg_full`.
    - Look at the coefficients.
    - Discuss all coefficients: sign and magnitude.

A Numeric and a Dummy Regressor: Interpretation

You know how to interpret coefficients when the variable is numeric (i.e. continuous).

A Numeric and a Dummy Regressor: Interpretation

You know how to interpret coefficients when the variable is numeric (i.e. continuous).

What if one of the regressor is a ***dummy variable***, that is it takes a value 1 if some condition is `TRUE` and 0 otherwise?

A Numeric and a Dummy Regressor: Interpretation

You know how to interpret coefficients when the variable is numeric (i.e. continuous).

What if one of the regressor is a ***dummy variable***, that is it takes a value 1 if some condition is `TRUE` and 0 otherwise?

Example: How do I interpret the coefficients in the following model

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

`religious` is a dummy variable equal to 1 if the school is a religious school, 0 if it isn't

A Numeric and a Dummy Regressor: Interpretation

You know how to interpret coefficients when the variable is numeric (i.e. continuous).

What if one of the regressor is a ***dummy variable***, that is it takes a value 1 if some condition is `TRUE` and 0 otherwise?

Example: How do I interpret the coefficients in the following model

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

`religious` is a dummy variable equal to 1 if the school is a religious school, 0 if it isn't

```
lm(avgmath ~ classize + religious, grades)
```

```
##
## Call:
## lm(formula = avgmath ~ classize + religious, data = grades)
##
## Coefficients:
## (Intercept)     classize    religious
##     61.3092       0.2311      -3.7800
```

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} = 0) &= b_0 + b_1 \times 0 + b_2 \times 0 \\ &= b_0\end{aligned}$$

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} = 0) &= b_0 + b_1 \times 0 + b_2 \times 0 \\ &= b_0\end{aligned}$$

$\rightarrow b_0$ corresponds to the expected average math score when class size is 0 and the school is not religious

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1) = b_0 + b_1 \times n_1 + b_2 \times \text{religious}$$

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1) = b_0 + b_1 \times n_1 + b_2 \times \text{religious}$$

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1 + 1) = b_0 + b_1 \times (n_1 + 1) + b_2 \times \text{religious}$$

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1) = b_0 + b_1 \times n_1 + b_2 \times \text{religious}$$

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1 + 1) = b_0 + b_1 \times (n_1 + 1) + b_2 \times \text{religious}$$

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1 + 1) - \mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1)$$

$$= b_0 + b_1 \times (n_1 + 1) + b_2 \times \text{religious} - (b_0 + b_1 \times n_1 + b_2 \times \text{religious}) = b_1$$

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0,1\} \ \& \ \text{class size} = n_1) = b_0 + b_1 \times n_1 + b_2 \times \text{religious}$$

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0,1\} \ \& \ \text{class size} = n_1 + 1) = b_0 + b_1 \times (n_1 + 1) + b_2 \times \text{religious}$$

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0,1\} \ \& \ \text{class size} = n_1 + 1) - \mathbb{E}(\text{average math score} | \text{religious} \in \{0,1\} \ \& \ \text{class size} = n_1)$$

$$= b_0 + b_1 \times (n_1 + 1) + b_2 \times \text{religious} - (b_0 + b_1 \times n_1 + b_2 \times \text{religious}) = b_1$$

$\rightarrow b_1$ corresponds to the expected change in average math score associated, on average, with a 1 student increase in class size, controlling for the religious status of the school (= keeping the religious status constant)

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) &= b_0 + b_1 \times \text{class size} + b_2 \times 1 \\ &= b_0 + b_1 \times \text{class size} + b_2\end{aligned}$$

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\begin{aligned}
\mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) &= b_0 + b_1 \times \text{class size} + b_2 \times 1 \\
&= b_0 + b_1 \times \text{class size} + b_2 \\
\mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N}) &= b_0 + b_1 \times \text{class size} + b_2 \times 0 \\
&= b_0 + b_1 \times \text{class size}
\end{aligned}$$

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) &= b_0 + b_1 \times \text{class size} + b_2 \times 1 \\ &= b_0 + b_1 \times \text{class size} + b_2\end{aligned}$$

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N}) &= b_0 + b_1 \times \text{class size} + b_2 \times 0 \\ &= b_0 + b_1 \times \text{class size}\end{aligned}$$

$$\begin{aligned}&\mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) - \\ &\quad \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N}) \\ &\quad = b_0 + b_1 \times \text{class size} + b_2 - (b_0 + b_1 \times \text{class size}) = b_2\end{aligned}$$

A Numeric and a Dummy Regressor: Formally

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) &= b_0 + b_1 \times \text{class size} + b_2 \times 1 \\ &= b_0 + b_1 \times \text{class size} + b_2\end{aligned}$$

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N}) &= b_0 + b_1 \times \text{class size} + b_2 \times 0 \\ &= b_0 + b_1 \times \text{class size}\end{aligned}$$

$$\begin{aligned}&\mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) - \\ &\quad \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N}) \\ &\quad = b_0 + b_1 \times \text{class size} + b_2 - (b_0 + b_1 \times \text{class size}) = b_2\end{aligned}$$

$\rightarrow b_2$ corresponds to the expected difference in average math score between religious and non-religious schools, keeping class size constant.

A Numeric and a Dummy Regressor: Summary

Our model is:

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

We have the following equalities:

$$b_0 = \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} = 0)$$

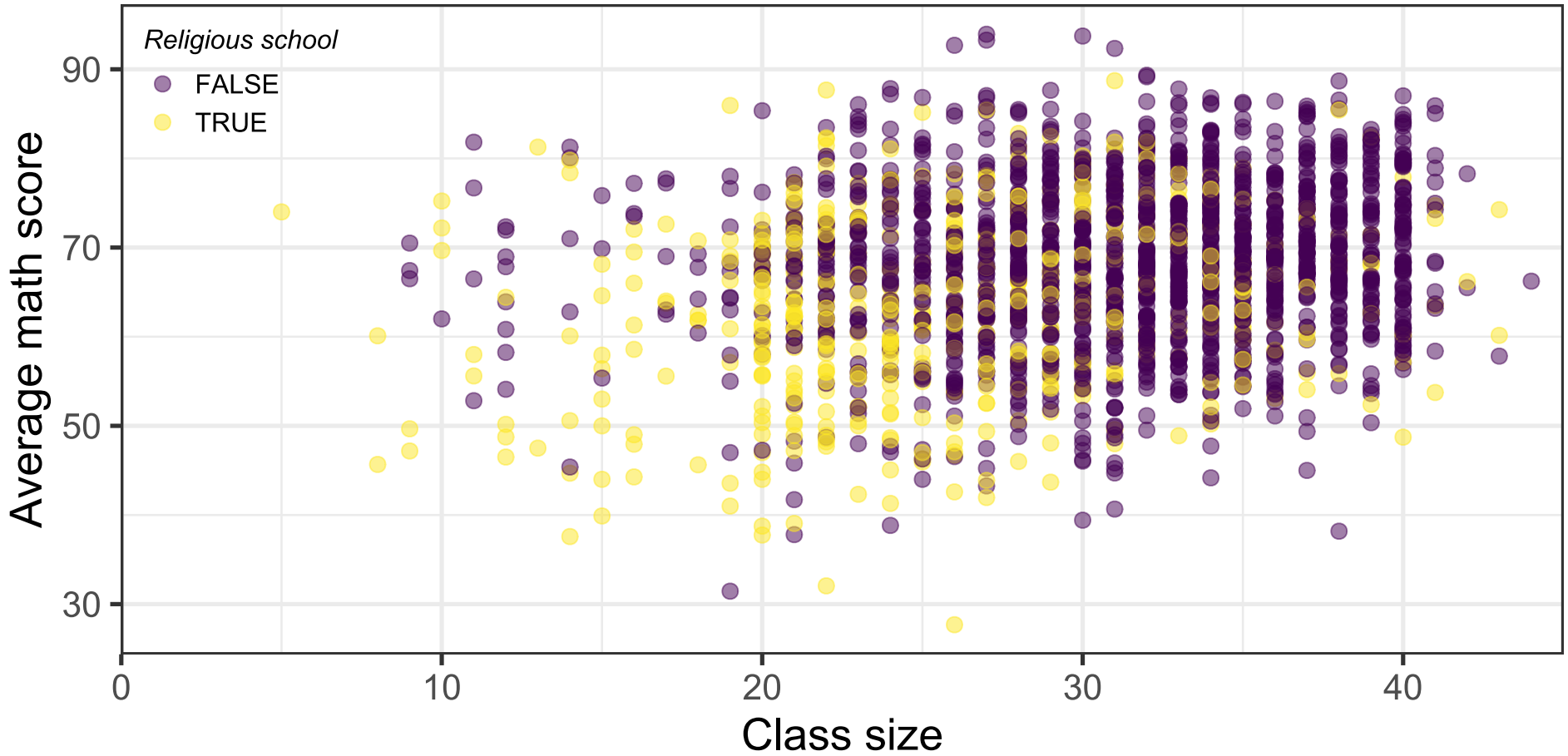
$$b_1 = \mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1 + 1) - \mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1)$$

$$b_2 = \mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) - \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N})$$

$$b_0 + b_2 = \mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} = 0)$$

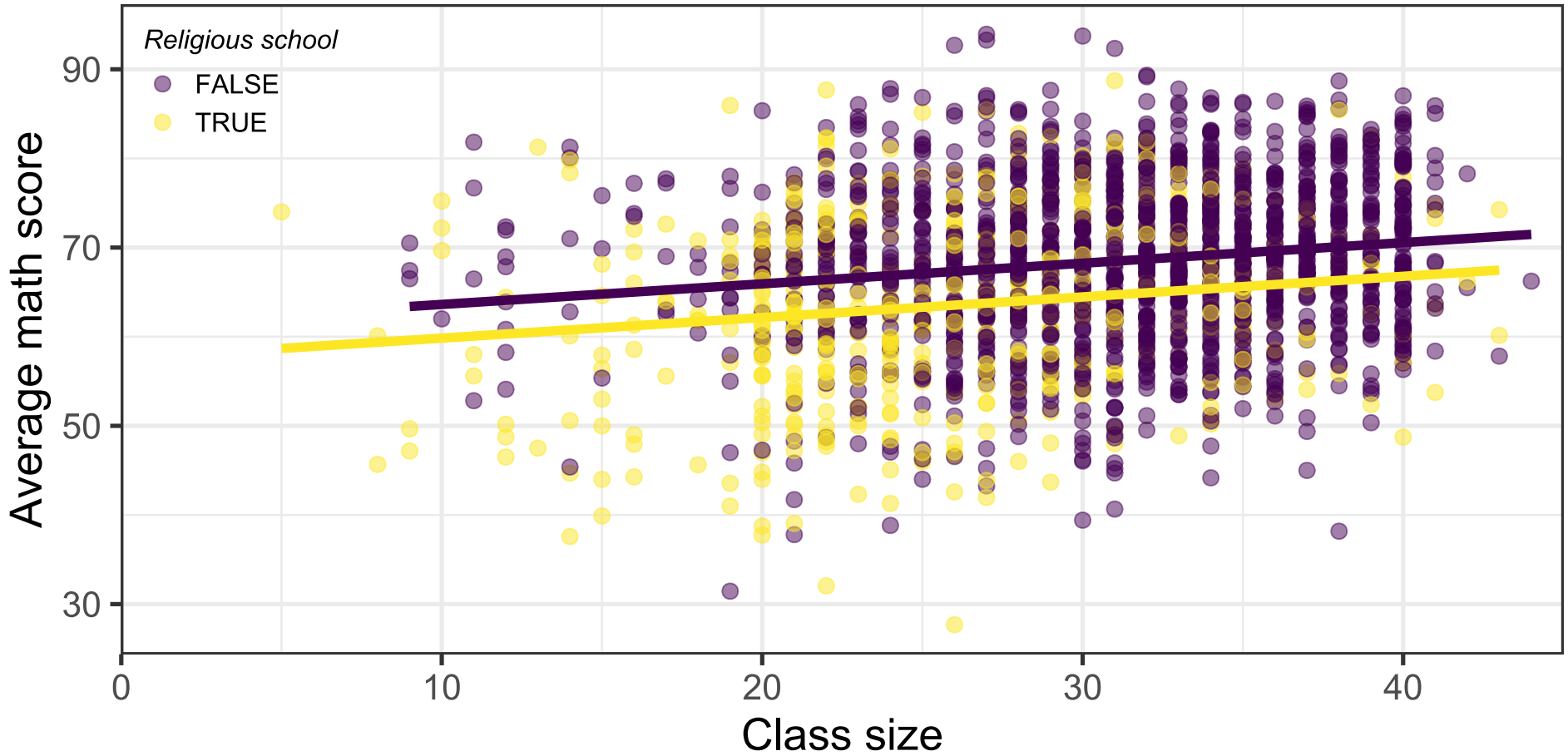
A Numeric and a Dummy Regressor: Visually

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$



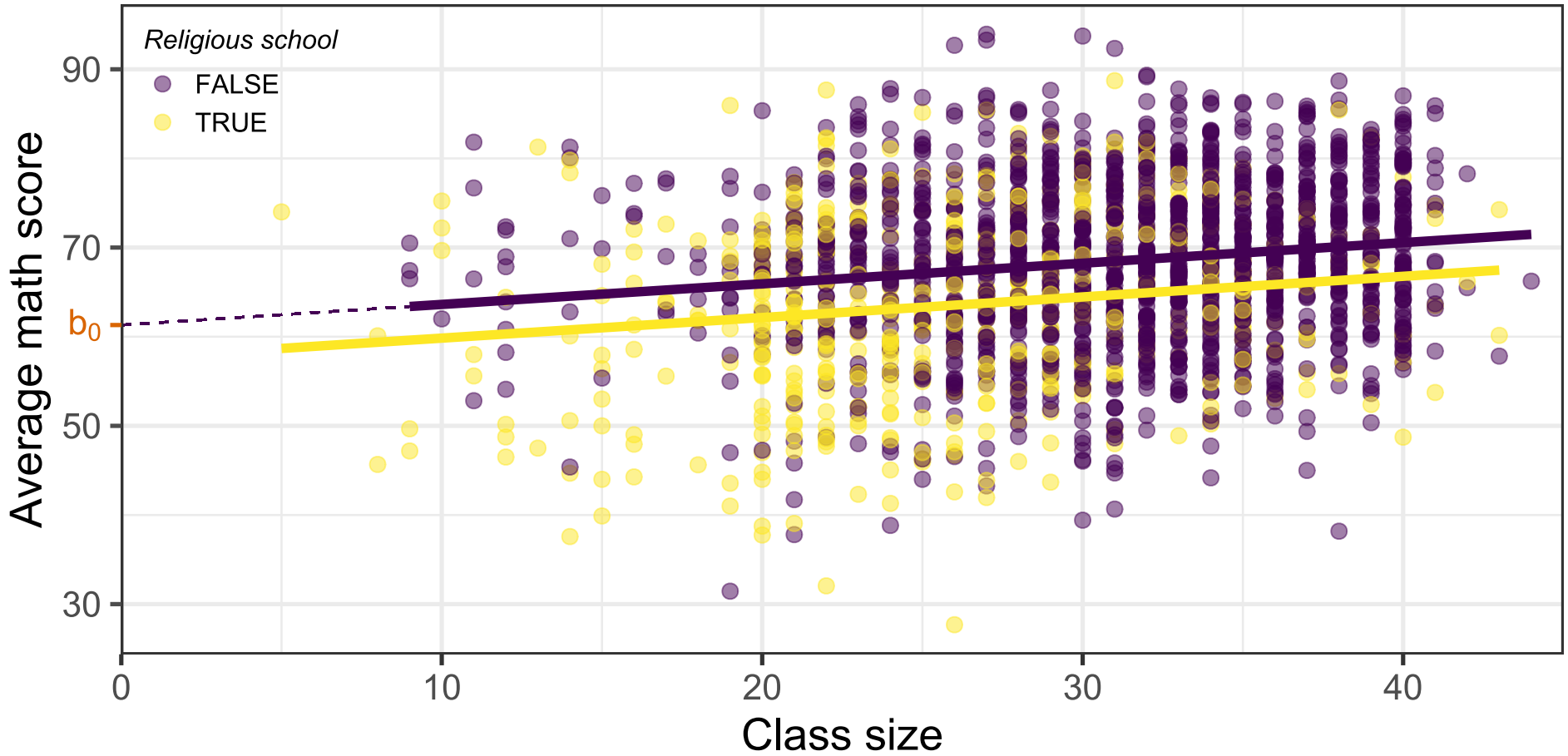
A Numeric and a Dummy Regressor: Visually

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$



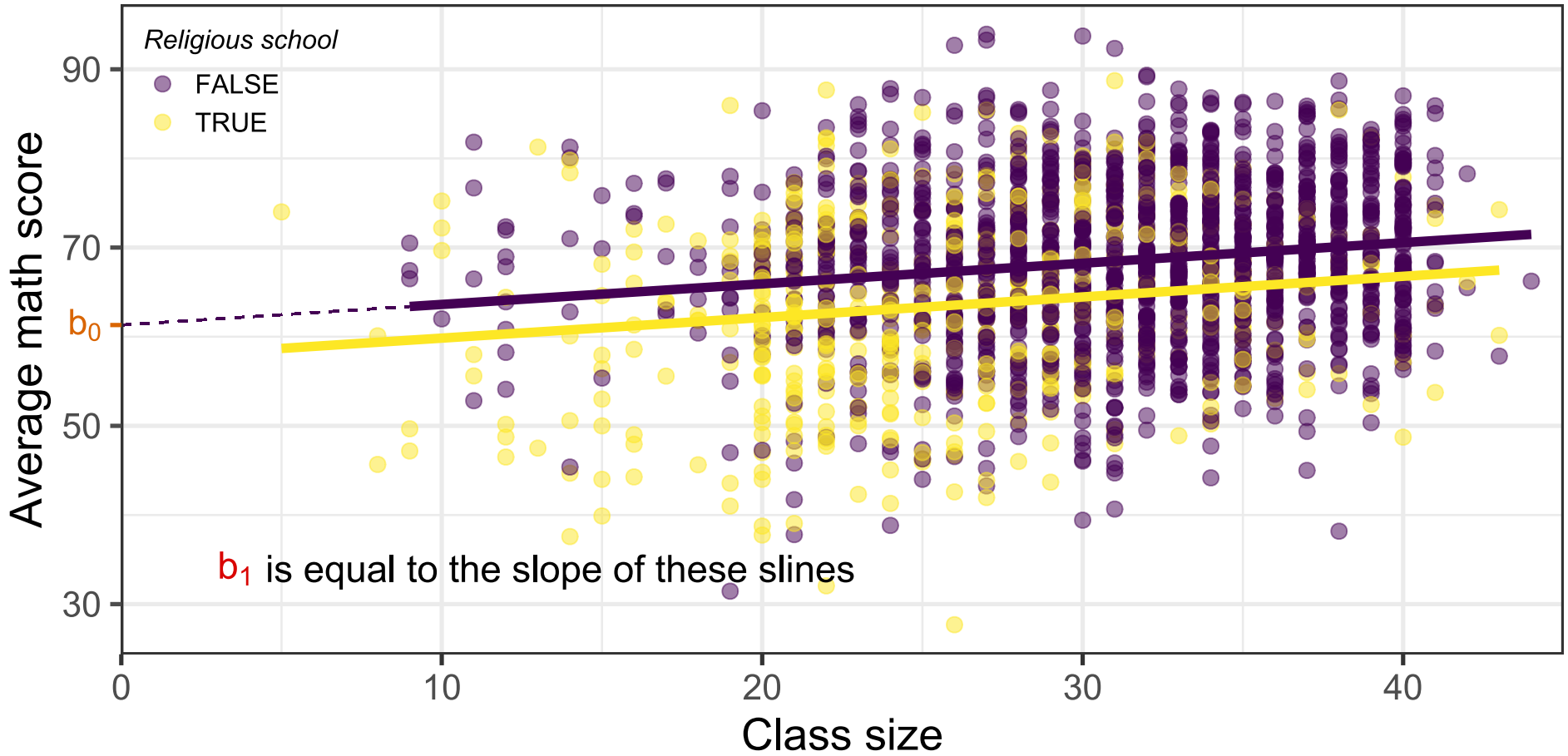
A Numeric and a Dummy Regressor: Visually

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$



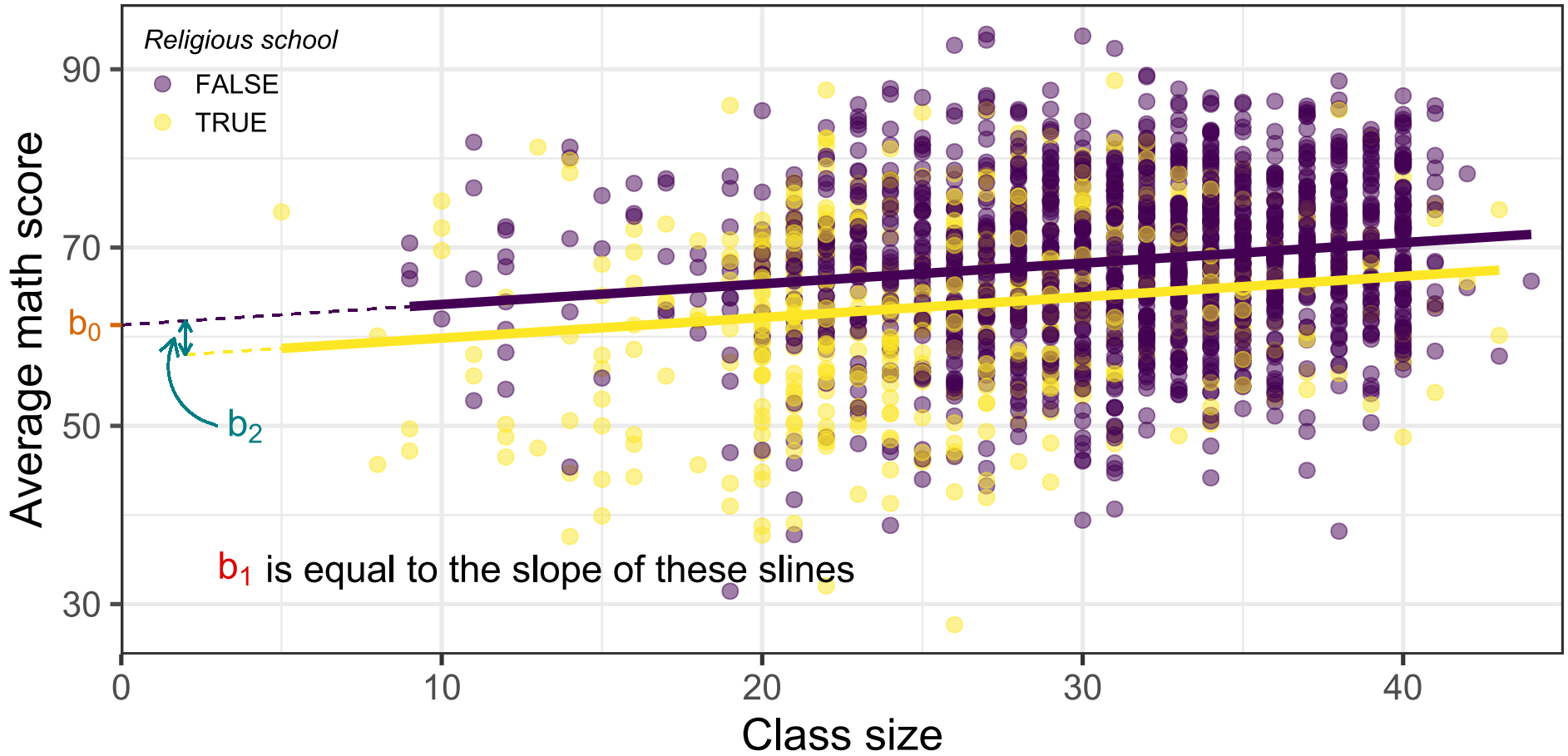
A Numeric and a Dummy Regressor: Visually

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$



A Numeric and a Dummy Regressor: Visually

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$



No Perfect Collinearity

There is one condition to satisfy to add regressors to the model:

> Any additional variable needs to add **at least *some* new information**.

No Perfect Collinearity

There is one condition to satisfy to add regressors to the model:

> Any additional variable needs to add **at least *some* new information**.

In other words, regressors **cannot be perfectly collinear**, i.e. not linear combinations of one another:

$$x_2 \neq ax_1 + b$$

No Perfect Collinearity

There is one condition to satisfy to add regressors to the model:

> Any additional variable needs to add **at least *some* new information**.

In other words, regressors **cannot be perfectly collinear**, i.e. not linear combinations of one another:

$$x_2 \neq ax_1 + b$$

Even if not perfectly correlated, the individual effects of highly correlated regressors are hard to disentangle.

No Perfect Collinearity

There is one condition to satisfy to add regressors to the model:

> Any additional variable needs to add **at least *some* new information**.

In other words, regressors **cannot be perfectly collinear**, i.e. not linear combinations of one another:

$$x_2 \neq ax_1 + b$$

Even if not perfectly correlated, the individual effects of highly correlated regressors are hard to disentangle.

Note that this implies that the number of observations has to be greater than the number of independent variables.

No Perfect Collinearity: Dummy Variable Trap

This condition is particularly relevant for ***categorical variables***, i.e. variables that take a limited of possible "levels" (e.g. gender, seasons, race, education levels, etc.)

No Perfect Collinearity: Dummy Variable Trap

This condition is particularly relevant for ***categorical variables***, i.e. variables that take a limited of possible "levels" (e.g. gender, seasons, race, education levels, etc.)

Let's go back to our `religious` school regression:

```
##
## Call:
## lm(formula = avgmath ~ classize + religious, data = grades)
##
## Coefficients:
## (Intercept)     classize    religious
##     61.3092       0.2311      -3.7800
```

No Perfect Collinearity: Dummy Variable Trap

This condition is particularly relevant for ***categorical variables***, i.e. variables that take a limited of possible "levels" (e.g. gender, seasons, race, education levels, etc.)

Let's go back to our `religious` school regression:

```
##
## Call:
## lm(formula = avgmath ~ classize + religious, data = grades)
##
## Coefficients:
## (Intercept)     classize    religious
##     61.3092       0.2311      -3.7800
```

What if I create a `is_religious` and a `is_notreligious` variable and regress `avgmath` on both (and `classize`)?

No Perfect Collinearity: Dummy Variable Trap

What if I create a `is_religious` and a `is_notreligious` variable and regress `avgmath` on both (and `classize`)?

```
grades <- grades %>%
  mutate(is_religious = (religious == 1),
         is_notreligious = (religious == 0))
lm(avgmath ~ classize + is_religious +
     is_notreligious, grades)
```

```
##
## Call:
## lm(formula = avgmath ~ classize + is_religious + is_notreligious,
##     data = grades)
##
## Coefficients:
##         (Intercept)             classize     is_religiousTRUE  is_notreligiousTRUE
##             61.3092               0.2311              -3.7800                   NA
```

Only one of two has a coefficient! Why?

No Perfect Collinearity: Dummy Variable Trap

What if I create a `is_religious` and a `is_notreligious` variable and regress `avgmath` on both (and `classize`)?

```
grades <- grades %>%
  mutate(is_religious = (religious == 1),
         is_notreligious = (religious == 0))
lm(avgmath ~ classize + is_religious +
     is_notreligious, grades)
```

```
##
## Call:
## lm(formula = avgmath ~ classize + is_religious + is_notreligious,
##     data = grades)
##
## Coefficients:
##         (Intercept)             classize     is_religiousTRUE  is_notreligiousTRUE
##             61.3092               0.2311              -3.7800                   NA
```

Only one of two has a coefficient! Why?

```
grades %>% count(is_religious == 1 - is_notreligious)
```

```
## # A tibble: 1 x 2
##   `is_religious == 1 - is_notreligious`     n
##   <lgl>                                 <int>
## 1 TRUE                                   2019
```

No Perfect Collinearity: Dummy Variable Trap

→ `R` automatically detects perfect collinearity between variables and drops one of them

No Perfect Collinearity: Dummy Variable Trap

$\rightarrow$ `R` automatically detects perfect collinearity between variables and drops one of them

⚠ you have to pay attention to the ***omitted/reference category***: the "baseline" category from which the coefficients are interpreted. Remember:

$$b_2 = \mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) - \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N})$$

No Perfect Collinearity: Dummy Variable Trap

→ `R` automatically detects perfect collinearity between variables and drops one of them

⚠ you have to pay attention to the ***omitted/reference category***: the "baseline" category from which the coefficients are interpreted. Remember:

$$b_2 = \mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) - \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N})$$

This applies to cases when you have more than 2 categories. You don't need to create a dummy variable for each possibility, `R` will detect the categorical variable(s) (as long as they are stored as `character` or `factor`) and do that for you.

No Perfect Collinearity: Dummy Variable Trap

→ `R` automatically detects perfect collinearity between variables and drops one of them

⚠ you have to pay attention to the ***omitted/reference category***: the "baseline" category from which the coefficients are interpreted. Remember:

$$b_2 = \mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) - \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N})$$

This applies to cases when you have more than 2 categories. You don't need to create a dummy variable for each possibility, `R` will detect the categorical variable(s) (as long as they are stored as `character` or `factor`) and do that for you.

But you have to look which category has been chosen as the ***omitted category***.

Task 2: Dummy Variable Trap

10:00

Let's run a regression where there is perfect linear dependence between regressors.

1. Load the *STAR* data from here, using `read.csv`, and assign it to an object called `star_df`. Keep only cases with no `NA`s with the following code:
`star_df <- star_df[complete.cases(star_df),]`

2. Create three dummy variables: (i) `small` equal to `TRUE` if students are in a small class and `FALSE` otherwise; (ii) `regular` equal to `TRUE` if students are in a regular class and `FALSE` otherwise; (iii) `regular_plus` equal to `TRUE` if students are in a regular+aide class and `FALSE` otherwise. (*Hint: To create a dummy, write* `dummy = (variable=="value")`, *inside the appropriate dplyr verb*) Create a last variable, `sum`, equal to the sum of `small`, `regular` and `regular_plus`. What is `sum` equal to? What does this mean?

3. Regress `math` on `regular_plus`. What is the average predicted `math` score of students in a regular+aide class?

4. Regress `math` on `small`, `regular` and `regular_plus`. What do you notice? What's the omitted (reference) category? Does this match the previous question?

5. Regress `math` on `star`. What do you notice? What's the omitted category? Interpret the

Omitted Variable Bias (OVB)

Omitted variable bias: Omitting important control variables from the regression model

This renders the coefficient for your regressor of interest unreliable (*biased*).

Omitted Variable Bias (OVB)

Omitted variable bias: Omitting important control variables from the regression model

This renders the coefficient for your regressor of interest unreliable (*biased*).

Let's note y our outcome variable, x our regressor and z our omitted variable.

Omitted Variable Bias (OVB)

Omitted variable bias: Omitting important control variables from the regression model

This renders the coefficient for your regressor of interest unreliable (*biased*).

Let's note y our outcome variable, x our regressor and z our omitted variable.

We could run regressions for the following models:

Omitted Variable Bias (OVB)

Omitted variable bias: Omitting important control variables from the regression model

This renders the coefficient for your regressor of interest unreliable (*biased*).

Let's note y our outcome variable, x our regressor and z our omitted variable.

We could run regressions for the following models:

1. ***Simple linear model***: $y = b_0 + b_1 x + e$

Omitted Variable Bias (OVB)

Omitted variable bias: Omitting important control variables from the regression model

This renders the coefficient for your regressor of interest unreliable (*biased*).

Let's note y our outcome variable, x our regressor and z our omitted variable.

We could run regressions for the following models:

1. ***Simple linear model***: $y = b_0 + b_1 x + e$
2. ***Multiple linear model***: $y = c_0 + c_1 x + c_2 z + e$

Omitted Variable Bias (OVB)

Omitted variable bias: Omitting important control variables from the regression model

This renders the coefficient for your regressor of interest unreliable (*biased*).

Let's note y our outcome variable, x our regressor and z our omitted variable.

We could run regressions for the following models:

1. ***Simple linear model***: $y = b_0 + b_1 x + e$
2. ***Multiple linear model***: $y = c_0 + c_1 x + c_2 z + e$
3. ***Omitted variable on regressor***: $z = d_0 + d_1 x + e$

Omitted Variable Bias (OVB)

Omitted variable bias: Omitting important control variables from the regression model

This renders the coefficient for your regressor of interest unreliable (*biased*).

Let's note y our outcome variable, x our regressor and z our omitted variable.

We could run regressions for the following models:

1. ***Simple linear model***: $y = b_0 + b_1 x + e$
2. ***Multiple linear model***: $y = c_0 + c_1 x + c_2 z + e$
3. ***Omitted variable on regressor***: $z = d_0 + d_1 x + e$

The formula for the OVB is:

$$\text{OVB} = c_2 \times d_1$$

Omitted Variable Bias (OVB)

Omitted variable bias: Omitting important control variables from the regression model

This renders the coefficient for your regressor of interest unreliable (*biased*).

Let's note y our outcome variable, x our regressor and z our omitted variable.

We could run regressions for the following models:

1. ***Simple linear model***: $y = b_0 + b_1 x + e$
2. ***Multiple linear model***: $y = c_0 + c_1 x + c_2 z + e$
3. ***Omitted variable on regressor***: $z = d_0 + d_1 x + e$

The formula for the OVB is:

$$\text{OVB} = c_2 \times d_1$$

In other words, $b_1 = c_1 + OVB$

Omitted Variable Bias (OVB)

$$\text{OVB} = \underbrace{\text{multiple regression coefficient on omitted variable}}_{c_2} \times \underbrace{\frac{Cov(x,z)}{Var(x)}}_{d_1}$$

From this formula you obtain:

- the OVB's ***magnitude*** (only if you observe z),
- the OVB's ***sign*** (positive/negative): since in practice z is not observed (otherwise you could include it in the regression) this is the most relevant case

Omitted Variable Bias (OVB)

$$\text{OVB} = \underbrace{\text{multiple regression coefficient on omitted variable}}_{c_2} \times \underbrace{\frac{Cov(x,z)}{Var(x)}}_{d_1}$$

From this formula you obtain:

- the OVB's ***magnitude*** (only if you observe z),
- the OVB's ***sign*** (positive/negative): since in practice z is not observed (otherwise you could include it in the regression) this is the most relevant case

Question:

Imagine you want to uncover the relationship between income and years of education. Why might a simple regression of income on years of education not yield reliable estimates? What could be an omitted variable? What's the expected sign of the OVB?

Omitted Variable Bias (OVB): In Practice

Let's go back to our class size and student performance example. We had:

Simple linear model: $\text{average math score} = b_0 + b_1 \text{class size} + e$

```
## (Intercept)    classize
##  57.7939158   0.3174906
```

Omitted Variable Bias (OVB): In Practice

Let's go back to our class size and student performance example. We had:

Simple linear model: $\text{average math score} = b_0 + b_1 \text{class size} + e$

```
## (Intercept)    classize
##  57.7939158   0.3174906
```

Multiple linear model: $\text{average math score} = c_0 + c_1 \text{class size} + c_2 \% \text{ disadvantaged} + e$

```
##   (Intercept)      classize disadvantaged
##   69.94438332    0.07167819   -0.33957877
```

Omitted Variable Bias (OVB): In Practice

Let's go back to our class size and student performance example. We had:

Simple linear model: $\text{average math score} = b_0 + b_1 \text{class size} + e$

```
## (Intercept)    classize
##  57.7939158   0.3174906
```

Multiple linear model: $\text{average math score} = c_0 + c_1 \text{class size} + c_2 \% \text{ disadvantaged} + e$

```
##   (Intercept)      classize disadvantaged
##   69.94438332    0.07167819   -0.33957877
```

Omitted variable on regressor: $\% \text{ disadvantaged} = d_0 + d_1 \text{class size} + e$

```
## (Intercept)    classize
##  35.7809990  -0.7238744
```

Omitted Variable Bias (OVB): In Practice

Let's go back to our class size and student performance example. We had:

Simple linear model: $\text{average math score} = b_0 + b_1 \text{class size} + e$

```
## (Intercept)    classize
##  57.7939158   0.3174906
```

Multiple linear model: $\text{average math score} = c_0 + c_1 \text{class size} + c_2 \% \text{ disadvantaged} + e$

```
##   (Intercept)      classize disadvantaged
##   69.94438332    0.07167819   -0.33957877
```

Omitted variable on regressor: $\% \text{ disadvantaged} = d_0 + d_1 \text{class size} + e$

```
## (Intercept)    classize
##  35.7809990  -0.7238744
```

We obtain:

$$b_1 = 0.317 = \underbrace{0.072}_{c_1} + \underbrace{(-0.34)}_{c_2} \times \underbrace{(-0.724)}_{d_1} = c_1 + OVB$$

Adjusted R^2

- Not of great importance but because it is so widely reported you just need to know it.

Adjusted R^2

- Not of great importance but because it is so widely reported you just need to know it.
- By construction, R^2 will always increase when a new regressor is added to the regression.

Adjusted R^2

- Not of great importance but because it is so widely reported you just need to know it.
- By construction, R^2 will always increase when a new regressor is added to the regression.
- The *adjusted* R^2 imposes a penalty for adding regressors to the model. The details are not crucial as in the vast majority of cases the R^2 and the adjusted R^2 are pretty similar.

Task 3: Recap

10:00

Let's use the *STAR* data (used in the previous task) to review the main concepts covered.

1. Using the filtered data from the previous task, regress `math` on `school` (tabulate the variable to know what it contains). Interpret the coefficients. What's the omitted category? Do you find them surprising? Why? What might be an omitted variable?

2. Compute the share of students qualifying for free lunch (i.e. `lunch` equals "free") by school location category. What do you observe? Add `free` to the previous question's regression. How do the coefficients change?

3. Regress `math` on `star`. After interpreting the coefficients, regress `math` on `star`, `gender`, `ethnicity`, `lunch`, `degree`, `experience` and `school`. Recalling that this is a randomized experiment, does it look like the randomization was well done?

4. What's the adjusted R^2 from the previous multiple regression. How do you interpret it? What might you deduce about the importance of observable individual, teacher and school characteristics in explaining educational outcomes?

5. (Optional) Regress `math` on `gender` and `experience` (the teacher's experience). Interpret the coefficients. How would these regression results look like visually?

On the way to causality

☑ How to manage data? Read it, tidy it, visualise it...

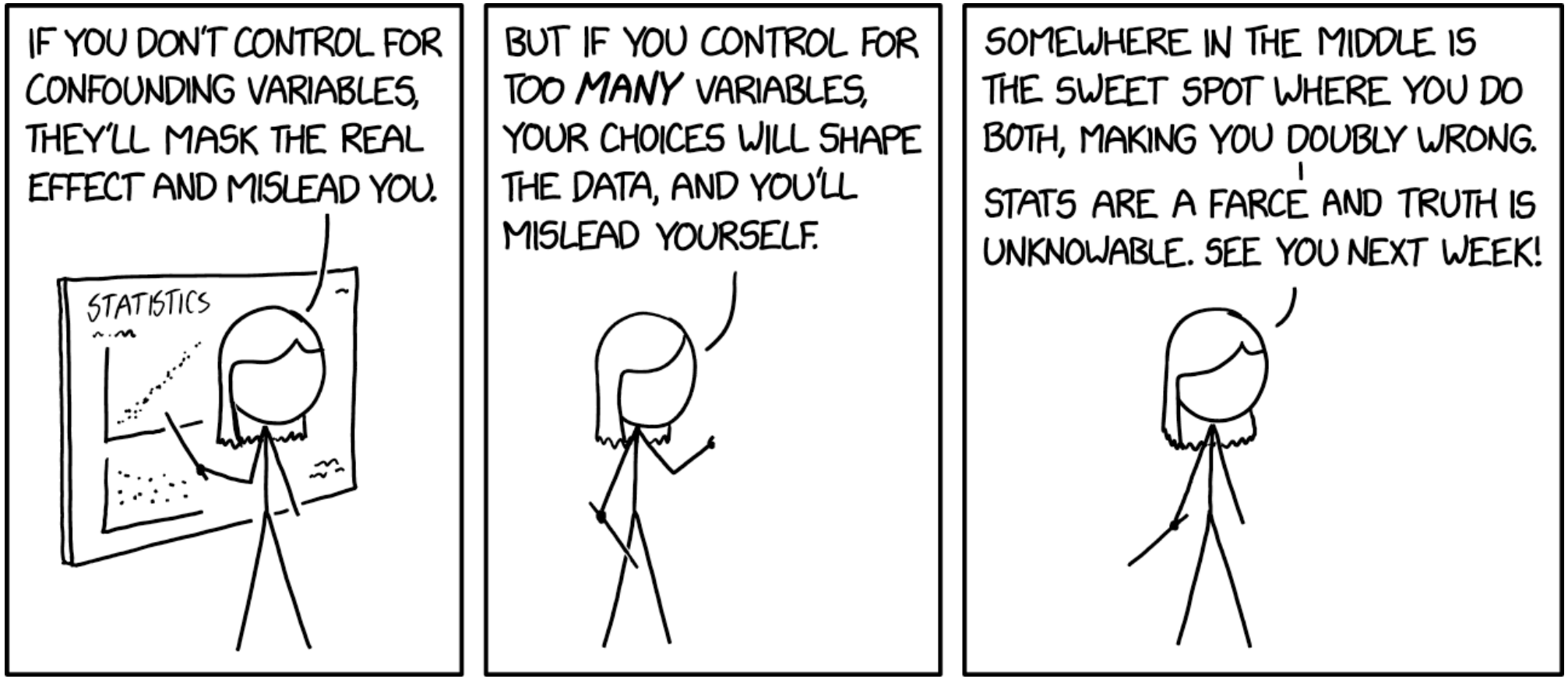
🚧 **How to summarise relationships between variables?** Simple and multiple linear regression... to be continued

☑ What is causality?

❌ What if we don't observe an entire population?

❌ Are our findings just due to randomness?

❌ How to find exogeneity in practice?


IF YOU DON'T CONTROL FOR CONFOUNDING VARIABLES, THEY'LL MASK THE REAL EFFECT AND MISLEAD YOU.
STATISTICS
BUT IF YOU CONTROL FOR TOO MANY VARIABLES, YOUR CHOICES WILL SHAPE THE DATA, AND YOU'LL MISLEAD YOURSELF.
SOMEWHERE IN THE MIDDLE IS THE SWEET SPOT WHERE YOU DO BOTH, MAKING YOU DOUBLY WRONG.
STATS ARE A FARCE AND TRUTH IS UNKNOWABLE. SEE YOU NEXT WEEK!

SciencesPo
DEPARTMENT OF ECONOMICS

SEE YOU NEXT WEEK!

	florian.oswald@sciencespo.fr
	Slides
	Book
	@ScPoEcon
	@ScPoEcon