SciencesPo
DEPARTMENT OF ECONOMICS

Applied Data Analysis for Public Policy Studies

Multiple Regression Model

Michele Fioretti
SciencesPo Paris
2020-10-19

Recap from last week

- Causality versus correlation
- The Potential Outcome Framework a.k.a. Rubin's Causal Model
- Randomized controlled trials (RCTs)

Recap from last week

- Causality versus correlation
- The Potential Outcome Framework a.k.a. Rubin's Causal Model
- Randomized controlled trials (RCTs)

Today - Multiple Regression Model

- **Multiple Regression Model**
- *Extensions:* standardized regression; log models; interacting variables
- Empirical applications:
  - *Class size* and *student performance*
  - *Wage*, *education* and *gender*

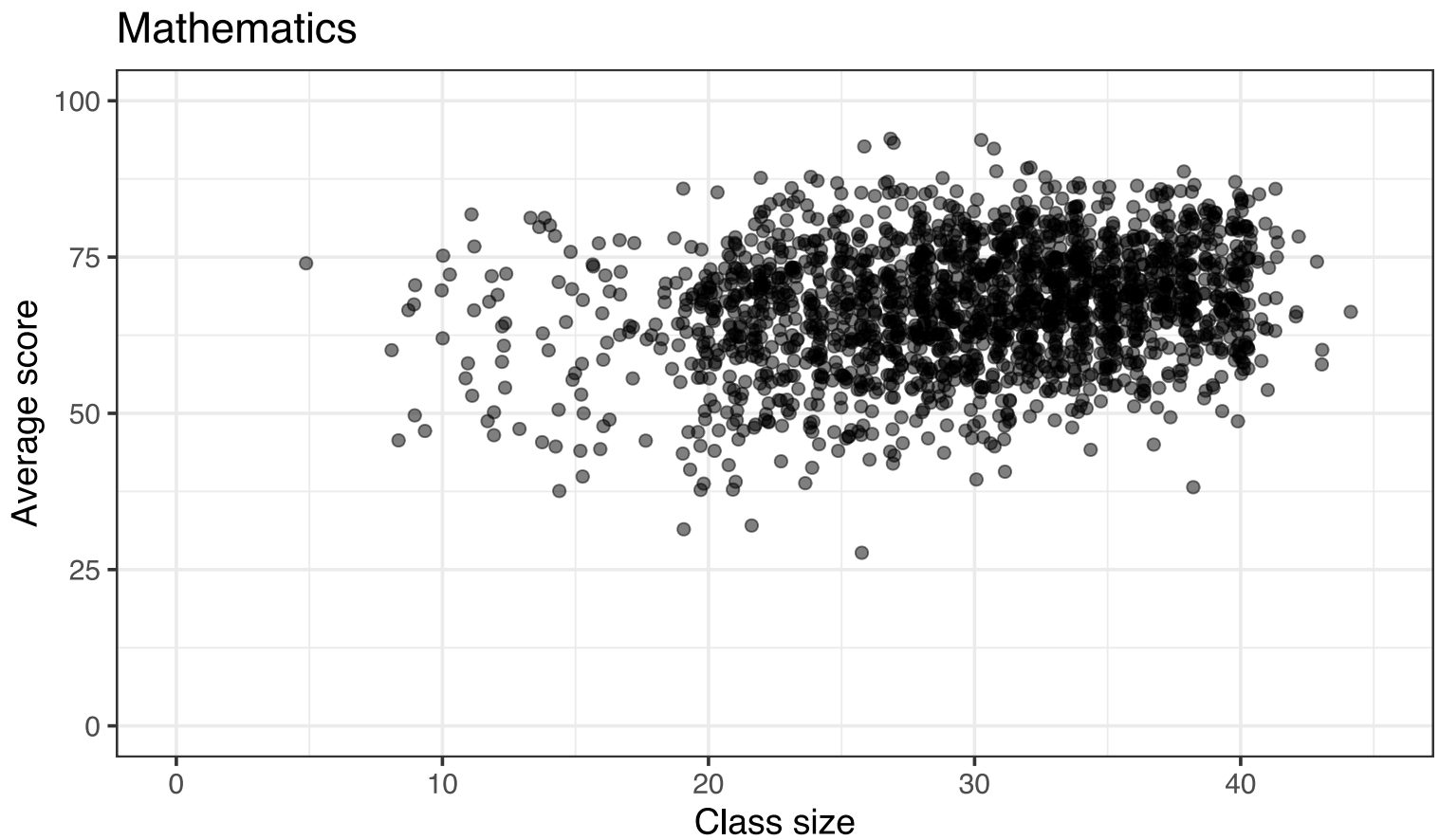
Class size and student performance

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

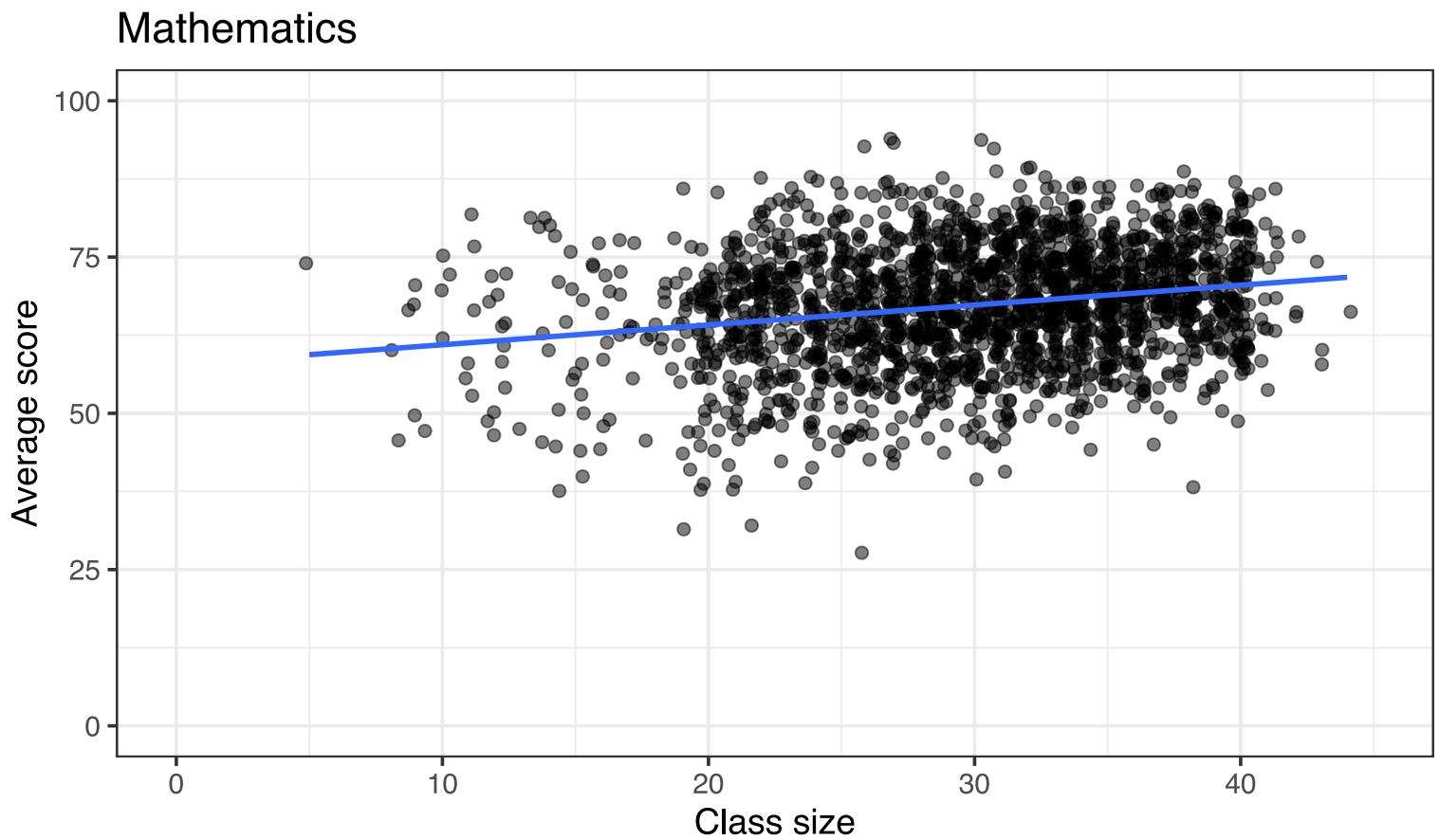
Class size and student performance

- 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.

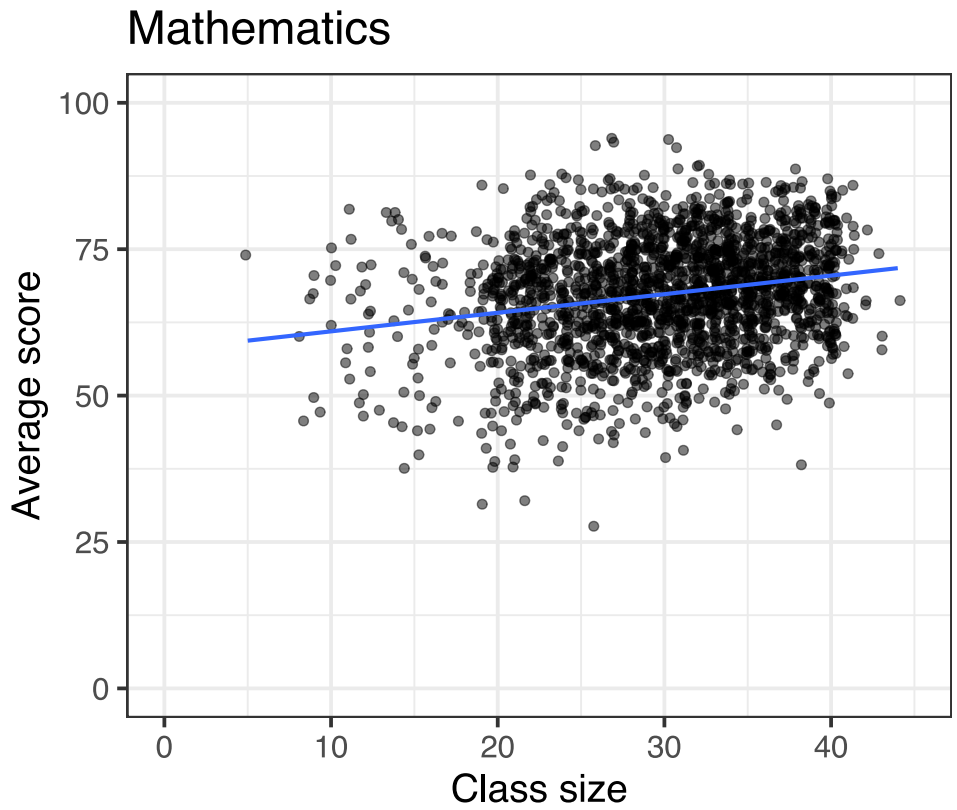
Class size and student performance: Raw relationship



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
```

- *Interpretation:* ***On average***, a 1-student increase in class size is ***associated*** to a 0.32 points increase in average math score.

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

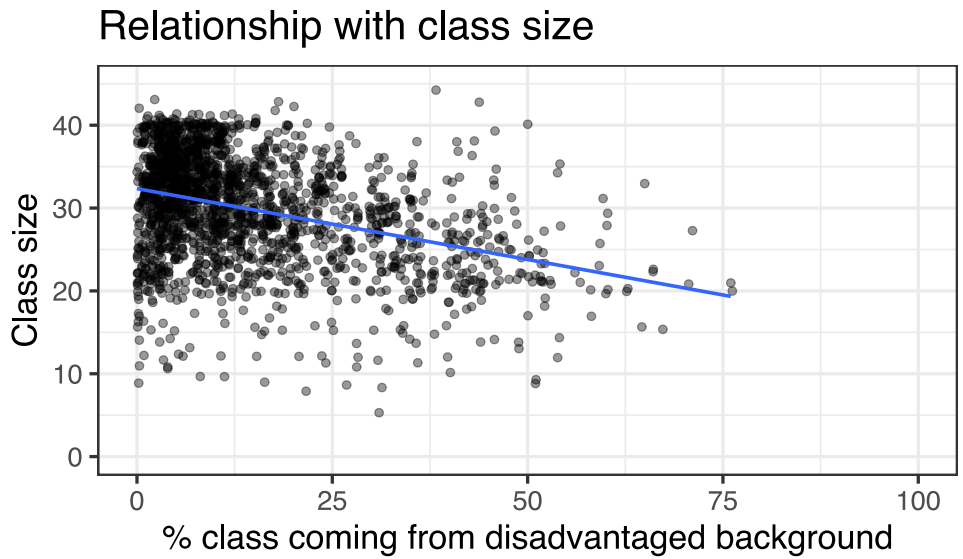
- Let's go back to Angrist and Lavy's (1999)'s analysis of the effect of class size on student performance in Israel.
- With a **simple linear regression**, we found that class size was positively ***associated*** with students' scores in maths and reading.
- This is intuitively unexpected, and contrasts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

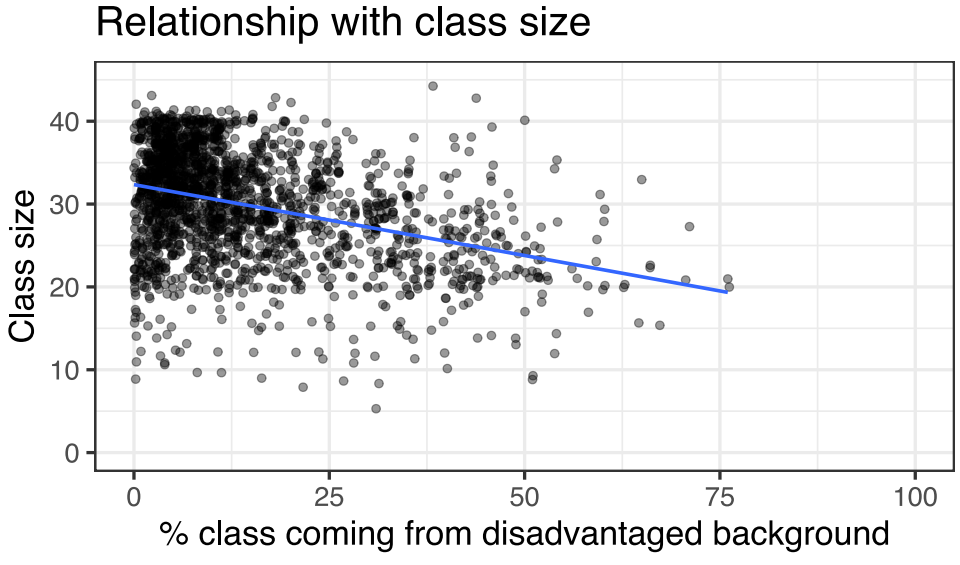
Link between **class size** and **the share of student who come from** `disadvantaged` **background** in the class.



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

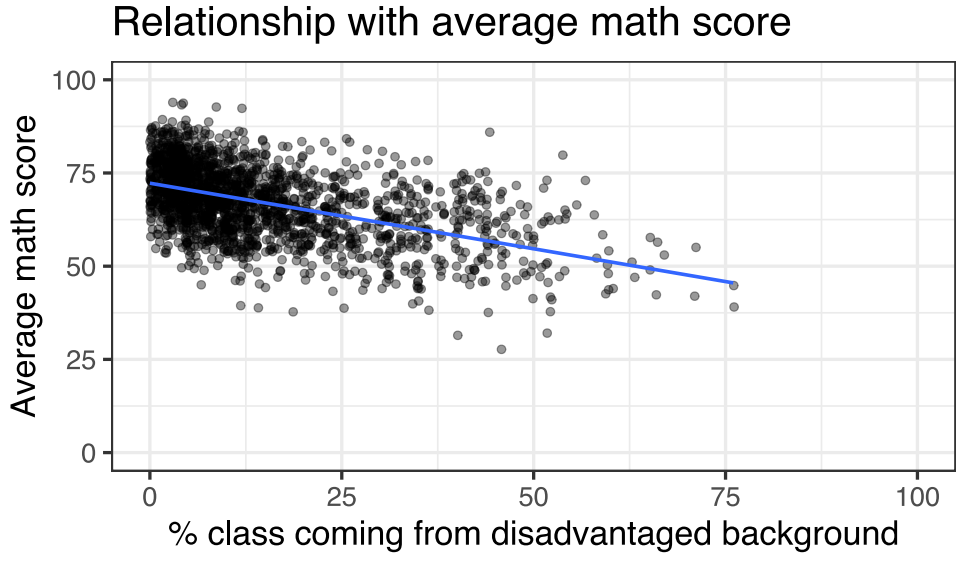
Class size and student performance: Confounders

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



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

Link between **average math score** and **the share of student who come from** `disadvantaged` **background** 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:

$$\text{mathscore}_i = b_0 + b_1 \text{classize}_i + b_2 \text{disadvantaged}_i + e_i$$

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:

$$\text{mathscore}_i = b_0 + b_1 \text{classize}_i + b_2 \text{disadvantaged}_i + e_i$$

- 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***.

⚠ 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***.

⚠ 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.

Mutiple 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.

Mutiple 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 indidivudal 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 indidivudal 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:
$\text{mathscore}_i = b_0 + b_1 \text{classize}_i + b_2 \text{disadvantaged}_i + e_i$

```
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 = 69.9: When `class size` and `disadvantaged` are set to 0, the *predicted* value of the average math score is 69.9.

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 = 69.9: When `class size` and `disadvantaged` are set to 0, the *predicted* value of the average math score is 69.9.
- b_1 = 0.07: Keeping the share 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 = 69.9: When `class size` and `disadvantaged` are set to 0, the *predicted* value of the average math score is 69.9.
- b_1 = 0.07: Keeping the share 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

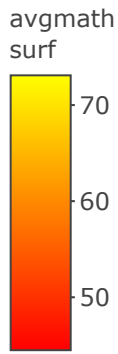
```
##
## Call:
## lm(formula = avgmath ~ classize + disadvantaged, data = grades)
##
## Coefficients:
##   (Intercept)       classize  disadvantaged
##      69.94438        0.07168       -0.33958
```

Answers

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

- 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.

Multiple Regression in 3D



Task 1 (10 minutes)

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`. Assign the output to a new object `reg`.

3. Look at the coefficients of this regression by running `reg$coefficients`. How do they compare with the math score regression coefficients?

4. 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.

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 disantangle.

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 disantangle.

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

Task 2 (7 minutes)

Still need to think about this perfect collinearity issue? Let's run a regression where there is perfect linear dependence between regressors.

1. Load the *STAR* data from here 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),]`
   Keep only second graders and small and regular class groups.

2. In the same dataset, add a dummy variable `star_df$small` equal to `TRUE` if students are in a small class and `FALSE` if they are in a regular class. In the same way, add a dummy variable `star_df$regular` equal to `TRUE` if students are NOT in a small class and to `FALSE` if they are in a small class. Clearly these two variables are perfectly collinear.

- Regress `small` on `math`. What is the average predicted `math` score of students in a small class?

- Regress `regular` on `math`. What is the average predicted `math` score of students in a regular class?

- Regress `small` and `regular` on `math`. What do you notice? Can you explain why?

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.

Ommitted variable bias (OVB)

- ***ommitted variable bias***: Omitting important control variables from the regression model
- This renders the coefficient for your regressor of interest unreliable.

Ommitted variable bias (OVB)

- ***ommitted variable bias***: Omitting important control variables from the regression model
- This renders the coefficient for your regressor of interest unreliable.
- Recall our previous example of class size and disadvantaged students. Not taking into account the `disadvantaged` variable led to a OVB for our class size estimate.

Ommitted variable bias (OVB)

- ***ommitted variable bias***: Omitting important control variables from the regression model
- This renders the coefficient for your regressor of interest unreliable.
- Recall our previous example of class size and disadvantaged students. Not taking into account the `disadvantaged` variable led to a OVB for our class size estimate.
- The formula for OVB is:

$$\begin{aligned}\text{OVB} = &\{\text{Relationship between } disadvantaged_i \text{ and } classize_i\} \\ &* \{\text{Effect of } disadvantaged_i \text{ in multiple regression}\}\end{aligned}$$

Ommitted variable bias (OVB)

- ***ommitted variable bias***: Omitting important control variables from the regression model
- This renders the coefficient for your regressor of interest unreliable.
- Recall our previous example of class size and disadvantaged students. Not taking into account the `disadvantaged` variable led to a OVB for our class size estimate.
- The formula for OVB is:

$$\begin{aligned}\text{OVB} = &\{\text{Relationship between } disadvantaged_i \text{ and } classize_i\} \\ &* \{\text{Effect of } disadvantaged_i \text{ in multiple regression}\}\end{aligned}$$

- From this formula you obtain both the *mangitude* of OVB and its *sign* (positive/negative).

Ommitted variable bias (OVB)

- ***ommitted variable bias***: Omitting important control variables from the regression model
- This renders the coefficient for your regressor of interest unreliable.
- Recall our previous example of class size and disadvantaged students. Not taking into account the `disadvantaged` variable led to a OVB for our class size estimate.
- The formula for OVB is:

$$\begin{aligned}\text{OVB} = &\{\text{Relationship between } disadvantaged_i \text{ and } classize_i\} \\ &* \{\text{Effect of } disadvantaged_i \text{ in multiple regression}\}\end{aligned}$$

- From this formula you obtain both the *mangitude* of OVB and its *sign* (positive/negative).
- The relationship between `disadvantaged` and `classize` is negative, and the effect of `disadvantaged` in the multiple regression is negative, therefore the OVB is positive. In other words, ommitting `disadvantaged` from the regression leads to an inflated coefficient for class size.

Variations from the baseline model

Depending on the dataset and the relationships between the variables of interest, you may need to move away from the baseline model.

- We will focus on 3 important variations:
  - *Standardized regressions*
  - *Log models*
  - *Interactions between regressors*

Variations from the baseline model

Depending on the dataset and the relationships between the variables of interest, you may need to move away from the baseline model.

- We will focus on 3 important variations:
  - *Standardized regressions*
  - *Log models*
  - *Interactions between regressors*
- In each case, the way we estimate these coefficients does not change (i.e OLS).

Variations from the baseline model

Depending on the dataset and the relationships between the variables of interest, you may need to move away from the baseline model.

- We will focus on 3 important variations:
  - *Standardized regressions*
  - *Log models*
  - *Interactions between regressors*
- In each case, the way we estimate these coefficients does not change (i.e OLS).
- However, the way we interpret our coefficients $(b_1, b_2, \ldots, b_k)$ does change!

Standardized Regression

Let's define what *standardizing* a variable means.

> ***Standardizing*** a variable z means to *demean* the variable and to divide the demeaned value by its own standard deviation:

$$z_i^{stand} = \frac{z_i - \bar{z}}{\sigma(z)}$$

where $\bar{z}$ is the mean of z and $\sigma(z)$ is the standard deviation of z, i.e. $\sigma(z) = \sqrt{\text{VAR}(z)}$.

- Why would we do that in the first place?

Standardized Regression

Let's define what *standardizing* a variable means.

> ***Standardizing*** a variable z means to *demean* the variable and to divide the demeaned value by its own standard deviation:

$$z_i^{stand} = \frac{z_i - \bar{z}}{\sigma(z)}$$

where $\bar{z}$ is the mean of z and $\sigma(z)$ is the standard deviation of z, i.e. $\sigma(z) = \sqrt{\text{VAR}(z)}$.

- Why would we do that in the first place?
- Intuitively, standardizing **puts variables on the same scale** so we can compare them.
- In our class size and student performance example, it will help to interpret:
  - The **magnitude** of the effects,
  - The **relative importance of each variable**.

Standardizing Regression: Interpretation

- If the dependent variable y is standardized:

Standardizing Regression: Interpretation

- If the dependent variable y is standardized:
  - By definition, b_k measures the predicted change in y^{stand} associated with a one unit increase in x_k.
  - If y^{stand} increases by one, it means that y increases by one standard deviation. So b_k measures the change in y **as a share of y's standard deviation**.

Standardizing Regression: Interpretation

- If the dependent variable y is standardized:
  - By definition, b_k measures the predicted change in y^{stand} associated with a one unit increase in x_k.
  - If y^{stand} increases by one, it means that y increases by one standard deviation. So b_k measures the change in y **as a share of y's standard deviation**.
- If the regressor x_k is standardized:

Standardizing Regression: Interpretation

- If the dependent variable y is standardized:
  - By definition, b_k measures the predicted change in y^{stand} associated with a one unit increase in x_k.
  - If y^{stand} increases by one, it means that y increases by one standard deviation. So b_k measures the change in y **as a share of y's standard deviation**.
- If the regressor x_k is standardized:
  - By definition, b_k measures the predicted change in y associated with a one unit increase in x_k^{stand} .
  - If x_k^{stand} increases by one unit, it means that x_k increases by one standard deviation. So b_k measures the predicted change in y **associated with an increase in x_k by one standard deviation**.

Task 3: To do at home (7 minutes)

Let's go back our `grades` dataset. These are the estimates we got from regressing the math test score on the full set of regressors.

```
##       (Intercept)          classize     disadvantaged school_enrollment
##       70.31886504        0.02524502        -0.34728316         0.01834627
##            female         religious
##       -1.20818830        1.15754786
```

1. Create a new variable `avgmath_stand` equal to the the standardized math score. You can use the `standardize()` function from the `jtools` package or do it by hand with base `R`.

2. Run the full regression using the standardized math test score as the dependent variable. Interpret the coefficients and their magnitude.

3. If you were to pick the most influential variable on the math score, what would it be?

4. Add the standardized variables for each *continuous* regressor as `<regressor>_stand`.

    - Would it make sense to standardize the `religious` variable?

5. Regress `avgmath_stand` on the full set of standardized regressors and `religious`. Discuss the relative influence of the regressors.

Log Models

- The models we have seen so far can be called ***level-level*** specifications. Both the dependent and the independent variables have been measured in level.

Log Models

- The models we have seen so far can be called ***level-level*** specifications. Both the dependent and the independent variables have been measured in level.
    - This *level* can be: euros, years, number of students,... and even percentage.

Log Models

- The models we have seen so far can be called ***level-level*** specifications. Both the dependent and the independent variables have been measured in level.
  - This *level* can be: euros, years, number of students,... and even percentage.
- Taking the log of the dependent and/or the independent variable(s) leads us to define 3 other types of regressions:
  - ***Log - level***: $log(y_i) = b_0 + b_1 x_{1,i} + \ldots + e_i$
  - ***Level - log***: $\mathrm{y}_i = b_0 + b_1 log(x_{1,i}) + \ldots + e_i$
  - ***Log - log***: $log(y_i) = b_0 + b_1 log(x_{1,i}) + \ldots + e_i$

Log Models: Interpretation

Here is a table that summarise how to interpret regressor coefficients in each case:

Model	Equation	Interpretation of b_1
Level - Level	$y = b_0 + b_1 x_1 + e$	**One unit** increase in x_1 is associated, on average, with b_1 **unit change** in y
Log - Level	$log(y) = b_0 + b_1 x_1 + e$	**One unit** increase in x_1 is associated, on average, with b_1 **percent change** in y
Level - Log	$y = b_0 + b_1 log(x_1) + e$	**One percent** increase in x_1 is associated, on average, with $b_1/100$ **unit change** in y
Log - Log	$log(y) = b_0 + b_1 log(x_1) + e$	**One percent** increase in x_1 is associated, on average, with b_1 **percent change** in y

- It looks like cooking recipes but of course it can be derived with some calculus.

When do we use log models?

There are several reasons why we would want to use log models:

When do we use log models?

There are several reasons why we would want to use log models:

- Account for ***non linearity*** in the relationship between y and x.

When do we use log models?

There are several reasons why we would want to use log models:

- Account for ***non linearity*** in the relationship between y and x.
- Interpret coefficients as ***elasticities*** which play a central role in economic theory.

When do we use log models?

There are several reasons why we would want to use log models:

- Account for ***non linearity*** in the relationship between y and x.
- Interpret coefficients as ***elasticities*** which play a central role in economic theory.
- Limit the influence of ***outliers*** (due to the concavity of the *log* function).

Interacting Regressors

- We interact two regressors when we believe the effect of one depends on the value of the other.
    - *Example:* The returns to education on wage vary by gender.

Interacting Regressors

- We interact two regressors when we believe the effect of one depends on the value of the other.
    - *Example:* The returns to education on wage vary by gender.
- In practice, if we interact x_1 and x_2, we would write our model like this :

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + b_3 x_{1,i} * x_{2,i} + \ldots + e_i$$

Interacting Regressors

- We interact two regressors when we believe the effect of one depends on the value of the other.
    - *Example:* The returns to education on wage vary by gender.
- In practice, if we interact x_1 and x_2, we would write our model like this :

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + b_3 x_{1,i} * x_{2,i} + \ldots + e_i$$

- The interpretation of b_1, b_2, and b_3 will depend on the type of x_1 and x_2.

Interacting Regressors

- We interact two regressors when we believe the effect of one depends on the value of the other.
    - *Example:* The returns to education on wage vary by gender.
- In practice, if we interact x_1 and x_2, we would write our model like this :

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + b_3 x_{1,i} * x_{2,i} + \ldots + e_i$$

- The interpretation of b_1, b_2, and b_3 will depend on the type of x_1 and x_2.
- Let's focus on the cases where one regressor is a dummy/categorical variable and the other is continuous.
- It will give you the intuition for the other cases:
    - Both regresors are dummies/categorical variables,
    - Both regresors are continuous variables.

Interacting Regressors

- Let's go back to the *STAR* experiment data.

Interacting Regressors

- Let's go back to the *STAR* experiment data.
- How does the effect of being in a small vs regular class vary with the experience of the teacher?

Interacting Regressors

- Let's go back to the *STAR* experiment data.
- How does the effect of being in a small vs regular class vary with the experience of the teacher?
- Our regression model becomes:

$$\text{score}_i = b_0 + b_1 \text{small}_i + b_2 \text{experience}_i + b_3 \text{small}_i * \text{experience}_i + e_i$$

Interacting Regressors

- Let's go back to the *STAR* experiment data.
- How does the effect of being in a small vs regular class vary with the experience of the teacher?
- Our regression model becomes:

$$\text{score}_i = b_0 + b_1 \text{small}_i + b_2 \text{experience}_i + b_3 \text{small}_i * \text{experience}_i + e_i$$

Effect of small class with teacher with 10 years of experience?

Interacting Regressors

- Let's go back to the *STAR* experiment data.
- How does the effect of being in a small vs regular class vary with the experience of the teacher?
- Our regression model becomes:

$$\text{score}_i = b_0 + b_1\text{small}_i + b_2\text{experience}_i + b_3\text{small}_i * \text{experience}_i + e_i$$

Effect of small class with teacher with 10 years of experience?

$$\mathbb{E}[\text{score}_i|\text{small}_i = 1 \ \& \ \text{experience}_i = 10] = b_0 + b_1 + b_2 * 10 + b_3 * 10$$

Interacting Regressors

- Let's go back to the *STAR* experiment data.
- How does the effect of being in a small vs regular class vary with the experience of the teacher?
- Our regression model becomes:

$$\text{score}_i = b_0 + b_1\text{small}_i + b_2\text{experience}_i + b_3\text{small}_i * \text{experience}_i + e_i$$

Effect of small class with teacher with 10 years of experience?

$$\mathbb{E}[\text{score}_i|\text{small}_i = 1 \ \& \ \text{experience}_i = 10] = b_0 + b_1 + b_2 * 10 + b_3 * 10$$

$$\mathbb{E}[\text{score}_i|\text{small}_i = 0 \ \& \ \text{experience}_i = 10] = b_0 + b_2 * 10$$

Interacting Regressors

- Let's go back to the *STAR* experiment data.
- How does the effect of being in a small vs regular class vary with the experience of the teacher?
- Our regression model becomes:

$$\text{score}_i = b_0 + b_1\text{small}_i + b_2\text{experience}_i + b_3\text{small}_i * \text{experience}_i + e_i$$

Effect of small class with teacher with 10 years of experience?

$$\mathbb{E}[\text{score}_i|\text{small}_i = 1 \ \& \ \text{experience}_i = 10] = b_0 + b_1 + b_2 * 10 + b_3 * 10$$

$$\mathbb{E}[\text{score}_i|\text{small}_i = 0 \ \& \ \text{experience}_i = 10] = b_0 + b_2 * 10$$

$$\begin{aligned}\mathbb{E}[\text{score}_i|\text{small}_i = 1 \ \& \ \text{experience}_i = 10] - \mathbb{E}[\text{score}_i|\text{small}_i = 0 \ \& \ \text{experience}_i = 10] \\ = b_0 + b_1 + b_2 * 10 + b_3 * 10 - (b_0 + b_2 * 10) \\ = b_1 + b_3 * 10\end{aligned}$$

Interacting Regressors

Running the regression for the `math` score (for all grades), we obtain:

```
lm(math ~ small+ experience + small*experience, star_df)
```

```
##
## Call:
## lm(formula = math ~ small + experience + small * experience,
##     data = star_df)
##
## Coefficients:
##          (Intercept)             smallTRUE            experience
##              534.1919                15.8906                 1.3305
## smallTRUE:experience
##              -0.3034
```

Interpretation:

Interacting Regressors

Running the regression for the `math` score (for all grades), we obtain:

```
lm(math ~ small+ experience + small*experience, star_df)
```

```
##
## Call:
## lm(formula = math ~ small + experience + small * experience,
##     data = star_df)
##
## Coefficients:
##          (Intercept)             smallTRUE            experience
##              534.1919                15.8906                 1.3305
## smallTRUE:experience
##               -0.3034
```

Interpretation:

- The interaction term allows the impact of being in a small class to vary with the experience of the teacher.

Interacting Regressors

Running the regression for the `math` score (for all grades), we obtain:

```
lm(math ~ small+ experience + small*experience, star_df)
```

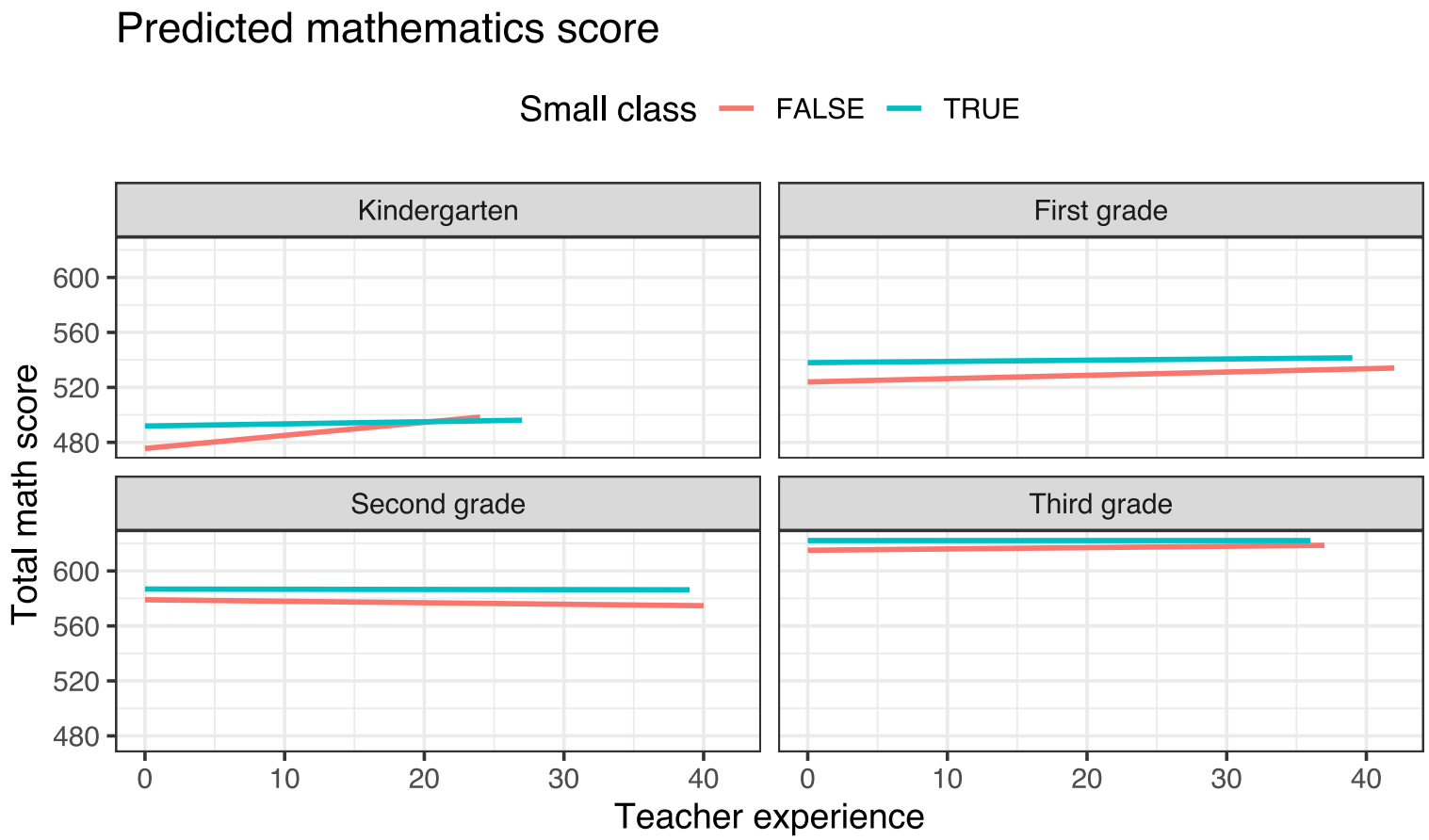
```
##
## Call:
## lm(formula = math ~ small + experience + small * experience,
##     data = star_df)
##
## Coefficients:
##          (Intercept)             smallTRUE            experience
##              534.1919                15.8906                 1.3305
## smallTRUE:experience
##               -0.3034
```

Interpretation:

- The interaction term allows the impact of being in a small class to vary with the experience of the teacher.
- In particular, we still observe a positive impact of being in a small class on math score,
- but this effect is decreasing in the experience of the teacher.

Interacting Regressors: Visually

Interacting Regressors: Visually by grade



Wages, Education and Gender

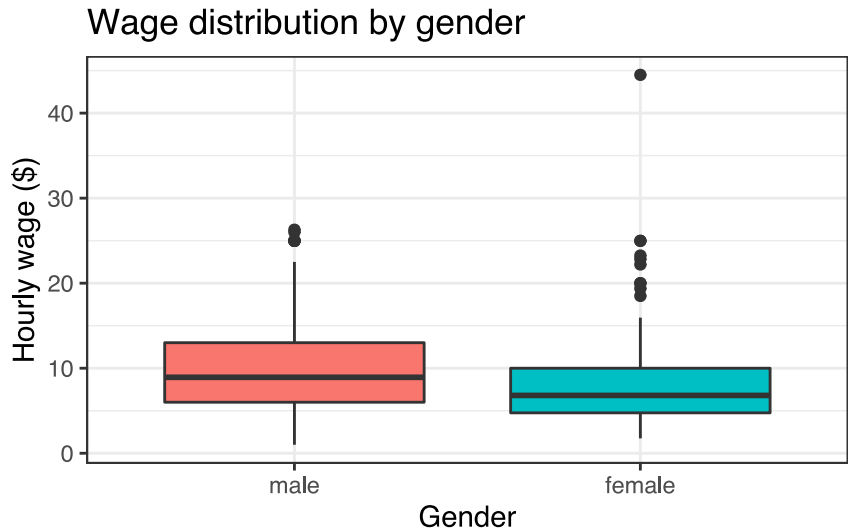
- Let's use these new *tools* to investigate the relationship between wages, education and gender.

Wages, Education and Gender

- Let's use these new *tools* to investigate the relationship between wages, education and gender.
- We will use data from the **C**urrent **P**opulation **S**urvey (CPS), the U.S. government's monthly survey of unemployment and labor force participation.
- We'll use a sample of 534 individuals from the May 1985 CPS available in the `AER` package.

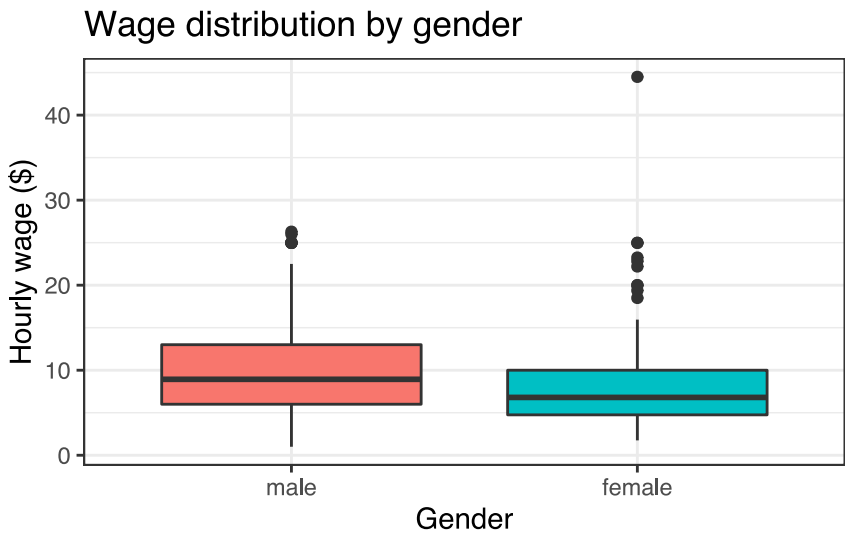
Wages, Education and Gender

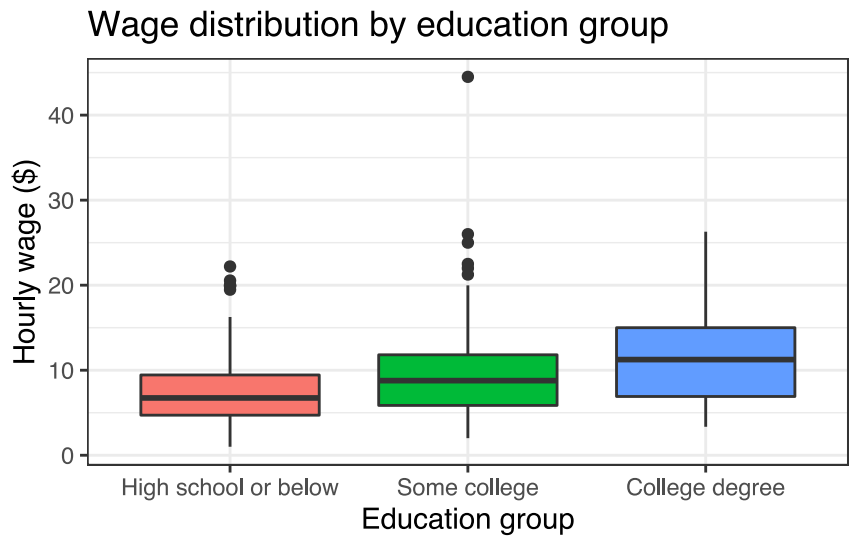
- Let's use these new *tools* to investigate the relationship between wages, education and gender.
- We will use data from the **C**urrent **P**opulation **S**urvey (CPS), the U.S. government's monthly survey of unemployment and labor force participation.
- We'll use a sample of 534 individuals from the May 1985 CPS available in the `AER` package.



Wages, Education and Gender

- Let's use these new *tools* to investigate the relationship between wages, education and gender.
- We will use data from the **C**urrent **P**opulation **S**urvey (CPS), the U.S. government's monthly survey of unemployment and labor force participation.
- We'll use a sample of 534 individuals from the May 1985 CPS available in the `AER` package.





Wages, Education and Gender

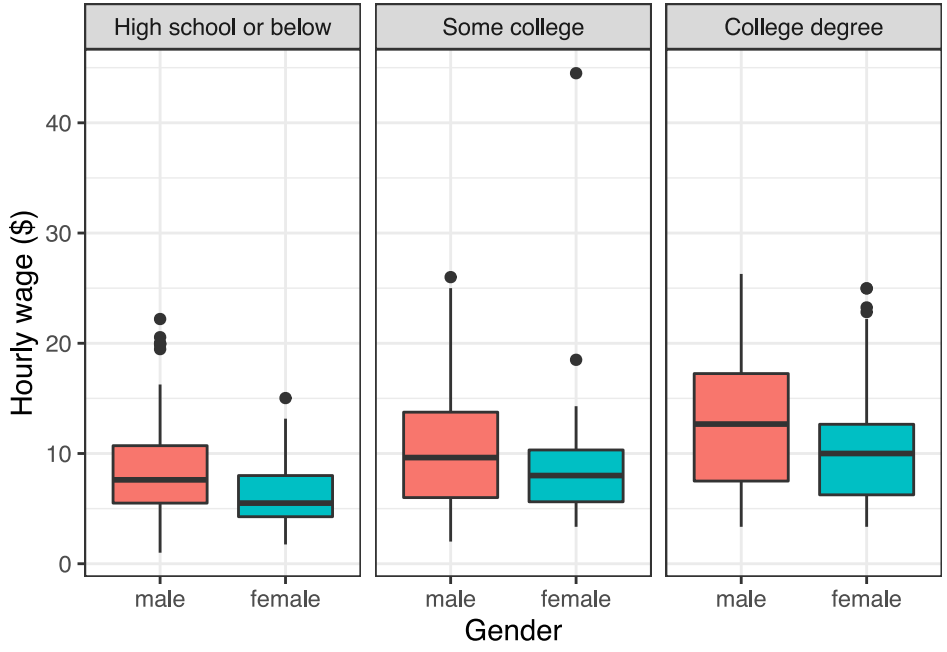
- So, wages are higher for men and higher for educated people, well, nothing really new.

Wages, Education and Gender

- So, wages are higher for men and higher for educated people, well, nothing really new.
- But, are the returns to education the same for women and men?
- In other words, how does the gender gap evolve for different levels of education?

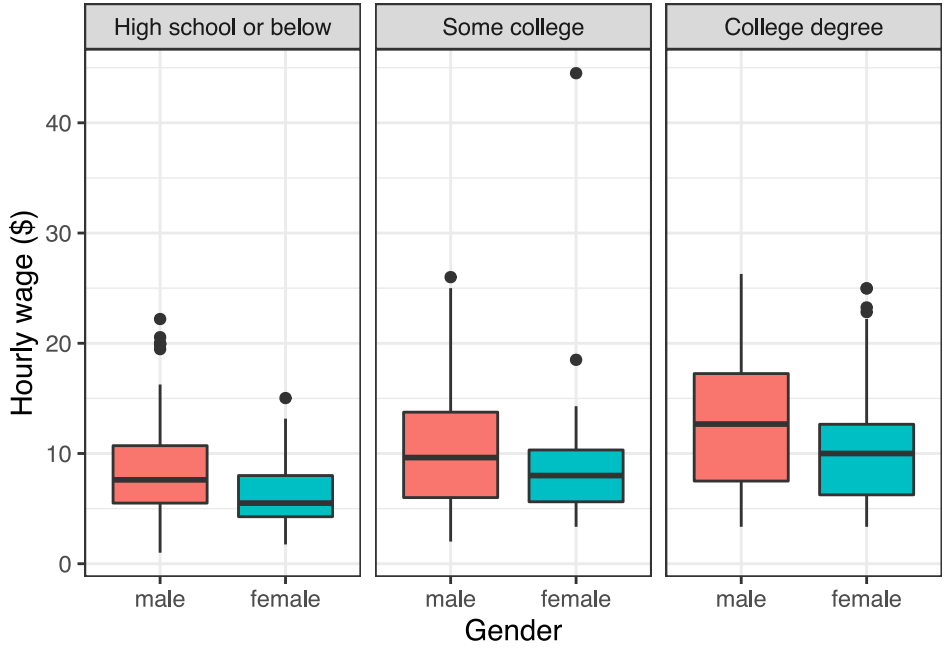
Wages, Education and Gender

- So, wages are higher for men and higher for educated people, well, nothing really new.
- But, are the returns to education the same for women and men?
- In other words, how does the gender gap evolve for different levels of education?



Wages, Education and Gender

- So, wages are higher for men and higher for educated people, well, nothing really new.
- But, are the returns to education the same for women and men?
- In other words, how does the gender gap evolve for different levels of education?

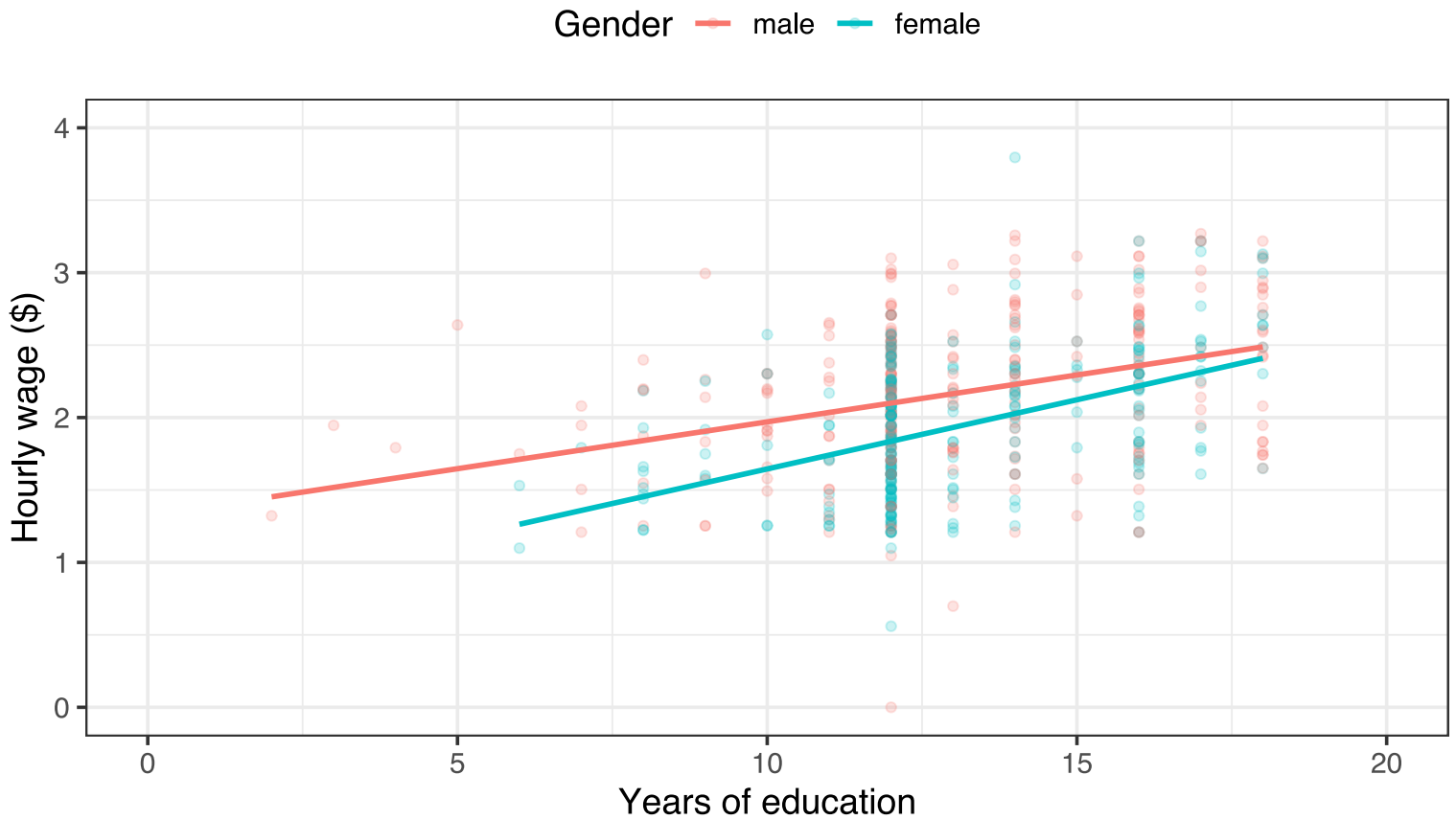


- We do observe a gender gap for each education group
- But it's not really clear if/how it varies with education
- Let's test it with a regression!

Task 4 (10 minutes)

1. Load the data `CPS1985` from the `AER` package.
2. Look at the `help` to get the definition of each variable: `?CPS1985`
3. We don't know if people are working part-time or full-time, does it matter here?
4. Create the `log_wage` variable equal to the log of `wage`.
5. Regress `log_wage` on `gender` and `education`, and save it as `reg1`. Interpret each coefficient.
6. Regress the `log_wage` on `gender`, `education` and their interaction `gender*education`, save it as `reg2`. Interpret each coefficient. Does the gender wage gap decrease with education?
7. Add all of other regressors to `reg2` and save it as `reg3`. Do our coefficients of interest change?

Wage, Gender and Education - Visually (Simple Model)



Teaser for the Next 3 Lectures

- You may have noticed that since the beginning we always work with **samples** drawn from the overall population.

Teaser for the Next 3 Lectures

- You may have noticed that since the beginning we always work with **samples** drawn from the overall population.
- Each time, imagine we could draw another sample from population:
  - Would we obtain the same results?
  - In other words, how confident can we be that our estimates (sign, magnitude) are not just driven by hazard?

Teaser for the Next 3 Lectures

- You may have noticed that since the beginning we always work with **samples** drawn from the overall population.
- Each time, imagine we could draw another sample from population:
  - Would we obtain the same results?
  - In other words, how confident can we be that our estimates (sign, magnitude) are not just driven by hazard?
- The next chapters will answwer those kind of questions:
  - We'll present the notion of **sampling**, and
  - Understand what **statistical inference** is and how to do it.

SciencesPo
DEPARTMENT OF ECONOMICS

SEE YOU NEXT WEEK!

	michele.fioretti@sciencespo.fr
	Slides
	Book
	@ScPoEcon
	@ScPoEcon