SciencesPo
DEPARTMENT OF ECONOMICS

Applied Data Analysis for Public Policy Studies

Chategorical Variables

Michele Fioretti
SciencesPo Paris
2020-09-11

Categorical Variables

- We have seen different data types in the 1st session.

Categorical Variables

- We have seen different data types in the 1st session.
- One of them was `factor`, representing **categorical** data:

Categorical Variables

- We have seen different data types in the 1st session.
- One of them was `factor`, representing **categorical** data:
- A person is *male* or *female*

Categorical Variables

- We have seen different data types in the 1st session.
- One of them was `factor`, representing **categorical** data:
- A person is *male* or *female*
- A plane is *passenger*, *cargo* or *military*

Categorical Variables

- We have seen different data types in the 1st session.
- One of them was `factor`, representing **categorical** data:
- A person is *male* or *female*
- A plane is *passenger*, *cargo* or *military*
- Some good is produced in *Spain*, *France*, *China* or *UK*.

Binary/Boolean/Dummy

- A *dummy* variable is either `TRUE` or `FALSE` (or `0` or `1`).
- We use dummies to mark **category membership**: if member, then `TRUE`.
- for example,

$$\text{is.male}_i = \begin{cases} 1 & \text{if } i \text{ is male} \\ 0 & \text{if } i \text{ is not male.} \end{cases}$$

- Notice that whether `0` corresponds to `TRUE` or `FALSE` is up to you. Just be consistent!

Dummy Variables

- We defined `is.male`...
- ... Similarly, for females,

$$\text{is.female}_i = \begin{cases} 1 & \text{if } i \text{ is female} \\ 0 & \text{if } i \text{ is not female.} \end{cases}$$

Dummy Variables

- We defined `is.male`...
- ... Similarly, for females,

$$\text{is.female}_i = \begin{cases} 1 & \text{if } i \text{ is female} \\ 0 & \text{if } i \text{ is not female.} \end{cases}$$

- Let's all create this dataset:

```
df1 = data.frame(income=c(3000,5000,7500,3500),
                 sex=c("male","female","male","female"))
```

Falling into The Dummy Variable Trap

- Let's run regression
  $y = b_0 + b_1 is.male + b_2 is.female$
- First, we create those dummy variables:

```
df1$is.male = df1$sex == "male"
df1$is.female = df1$sex == "female"
```

Falling into The Dummy Variable Trap

- Let's run regression $y = b_0 + b_1 is.male + b_2 is.female$
- First, we create those dummy variables:

```
df1$is.male = df1$sex == "male"
df1$is.female = df1$sex == "female"
```

- and then let's run this:

```
lm(income ~ is.male + is.female,df1)
```

- What do you see? 🤔

Falling into The Dummy Variable Trap

- Let's run regression $y = b_0 + b_1 is.male + b_2 is.female$
- First, we create those dummy variables:

```
df1$is.male = df1$sex == "male"
df1$is.female = df1$sex == "female"
```

- and then let's run this:

```
lm(income ~ is.male + is.female,df1)
```

- What do you see? 🤔

```
lm(income ~ is.male + is.female,df1)
```

```
##
## Call:
## lm(formula = income ~ is.male + is.female, data = df1)
##
## Coefficients:
##   (Intercept)    is.maleTRUE  is.femaleTRUE
##          4250           1000             NA
```

The Trap: Multicolinearity

```
df1$linear_comb = df1$is.male + df1$is.female
df1
```

```
##   income    sex is.male is.female linear_comb
## 1   3000   male    TRUE     FALSE           1
## 2   5000 female   FALSE      TRUE           1
## 3   7500   male    TRUE     FALSE           1
## 4   3500 female   FALSE      TRUE           1
```

The Trap: Multicolinearity

```r
df1$linear_comb = df1$is.male + df1$is.female
df1
```

```
##   income    sex is.male is.female linear_comb
## 1   3000   male    TRUE     FALSE           1
## 2   5000 female   FALSE      TRUE           1
## 3   7500   male    TRUE     FALSE           1
## 4   3500 female   FALSE      TRUE           1
```

- Oops. `is.male + is.female` is **always** equal `1`!
- In other words, `is.male = 1 - is.female`. A **perfect collinearity**!
- Multiple regression fails. 😠

Drop One Category

- Notice: Inclusion of both dummies doesn't add anything
- If someone is `male` they are *not* `female`.
- So we **drop one of the categories**. Only do $y = b_0 + b_1 is.female$

```
lm1 = lm(income ~ is.female,df1)
lm1
```

```
##
## Call:
## lm(formula = income ~ is.female, data = df1)
##
## Coefficients:
##   (Intercept)  is.femaleTRUE
##          5250          -1000
```

Drop One Category

- Notice: Inclusion of both dummies doesn't add anything

- If someone is `male` they are *not* `female`.

- So we **drop one of the categories**. Only do $y = b_0 + b_1 is.female$

```
lm1 = lm(income ~ is.female,df1)
lm1
```

```
##
## Call:
## lm(formula = income ~ is.female, data = df1)
##
## Coefficients:
##   (Intercept)  is.femaleTRUE
##          5250          -1000
```

- Would we get a ***different slope and intercept*** if we were to $y = b_0 + b_1 is.male$ instead?

Drop One Category

- Notice: Inclusion of both dummies doesn't add anything

- If someone is `male` they are *not* `female`.

- So we **drop one of the categories**. Only do $y = b_0 + b_1 is.female$

```
lm1 = lm(income ~ is.female,df1)
lm1
```

```
##
## Call:
## lm(formula = income ~ is.female, data = df1)
##
## Coefficients:
##   (Intercept)  is.femaleTRUE
##          5250          -1000
```

- Would we get a ***different slope and intercept*** if we were to $y = b_0 + b_1 is.male$ instead?

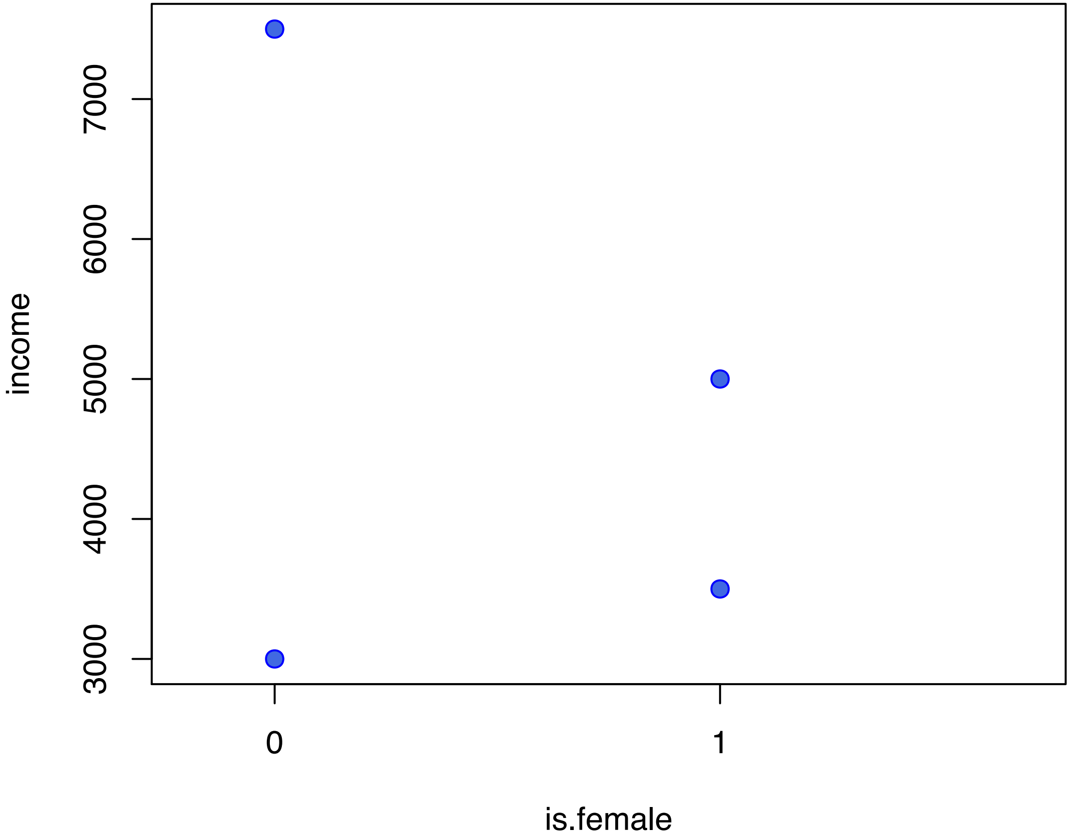
```
lm2 = lm(income ~ is.female,df1)
lm2
```

```
##
## Call:
## lm(formula = income ~ is.female, data = df1)
##
## Coefficients:
##   (Intercept)  is.femaleTRUE
##          5250          -1000
```

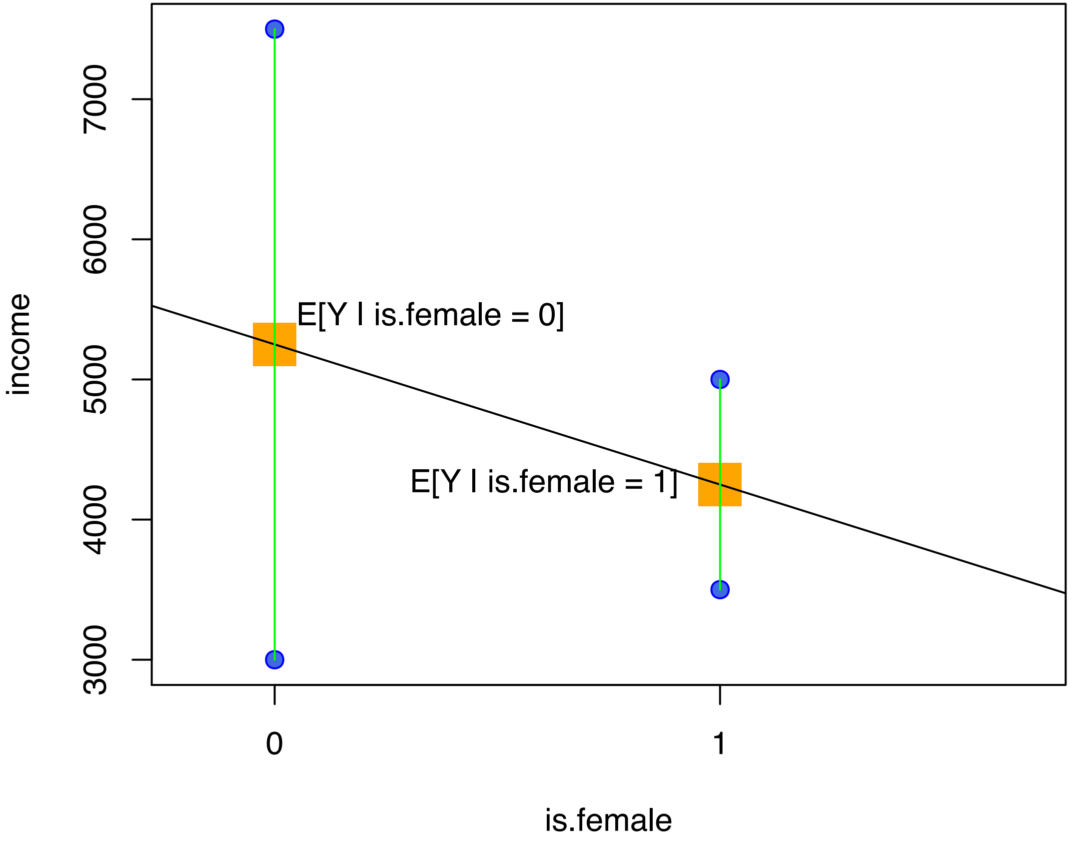
Interpretation of Dummies

- Let's go back to the case where we excluded the variable `is.male`.
- So what's the effect of being `male` now?
    - Well, *male* means `is.female = 0`. So `male` is **subsumed in the intercept**!
    - At `is.female = 0`, i.e. $\hat{y} = b_0 + b_1 \cdot 0 = 5250$
- Coefficient on `is.female` is $b_1 =$ -1000. It measures the *difference in intercept from being female.*
    - That means: $\hat{y} = b_0 + b_1 \cdot 1 = 4250$

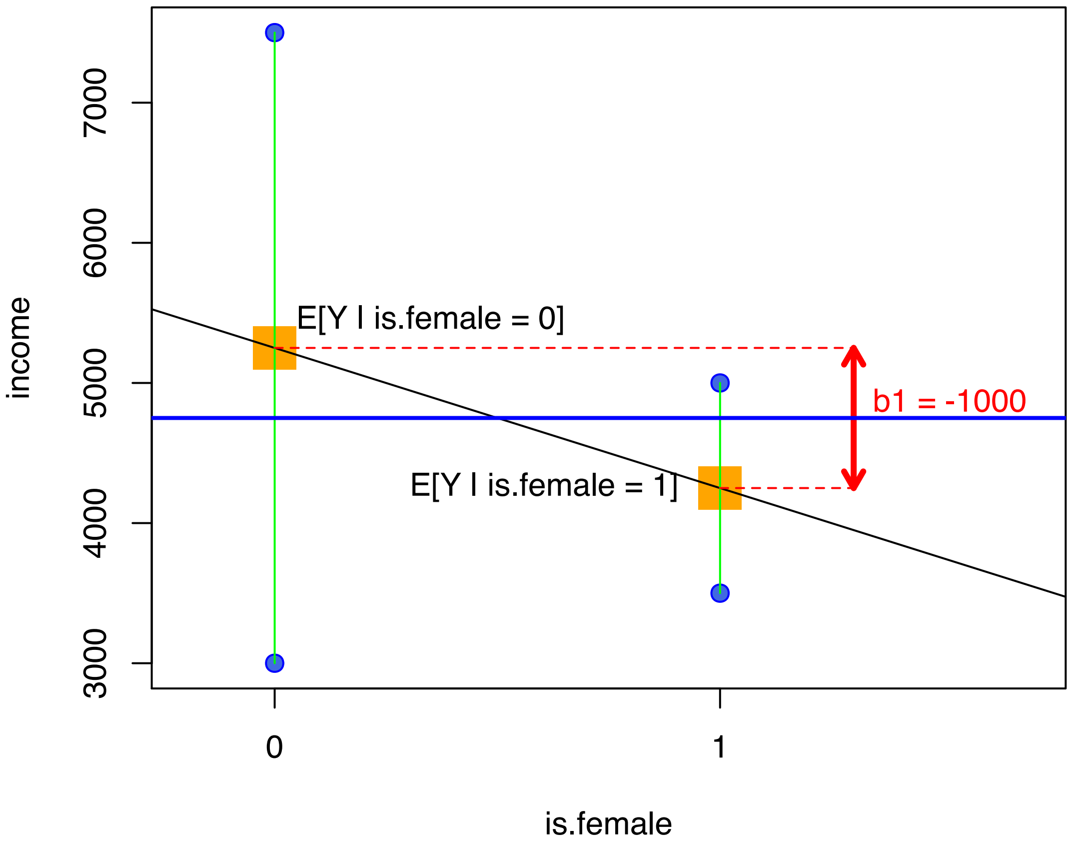
Our Dataset in a Picture



Regression connects Conditional Means!



b_1 is *Difference* in Conditional Means



Interpretation of Dummy Coefficient b_1

- So, we have seen that

$$b_1 = E[Y|\text{is.female} = 1] - E[Y|\text{is.female} = 0]$$

- This was the meaning of the red arrow.

App!

- Time for you to play around with the Binary Regression!
- Try to find the best line again!

```r
library(ScPoApps)
launchApp("reg_dummy")
```

Dummy *and* X

- What if we added $\text{exper}_i \in \mathbb{N}$ to that regression?

$$y_i = b_0 + b_1 \text{is.female}_i + b_2 \text{exper}_i + e_i$$

- As before, dummy acts as intercept shifter. We have

$$y_i = \begin{cases} b_0 + b_1 + b_2 \text{exper}_i + e_i & \text{if is.female=1} \\ b_0 + \quad\quad + b_2 \text{exper}_i + e_i & \text{if is.female=0} \end{cases}$$

- intercept is $b_0 + b_1$ for women but b_0 for men
- Slope b_2 **identical** for both!

App!

```
library(ScPoApps)
launchApp("reg_dummy_example")
```

More than Two Levels: `factor`

- Sometimes two categories are not enough.
- The `R` data type `factor` can represent more than just `0` and `1` in terms of categories.
- Function `factor` takes a numeric vector `x` and a vector of `labels`. Each value of `x` is associated to a `label`:

```
factor(x = c(1,1,2,4,3,4),labels = c("HS","someCol","BA","MSc"))
```

```
## [1] HS      HS      someCol MSc     BA      MSc    
## Levels: HS someCol BA MSc
```

- `factor` in an `lm` object automatically chooses an omitted/reference category!

Log Wages and Dummies {#factors}

- Let us illustrate the simplest use of `factors` in `R`.

- Going back to our wage example, let's say that a worker's wage depends on their education as well as their gender:

$$\ln w_i = b_0 + b_1 educ_i + b_2 female_i + e_i$$

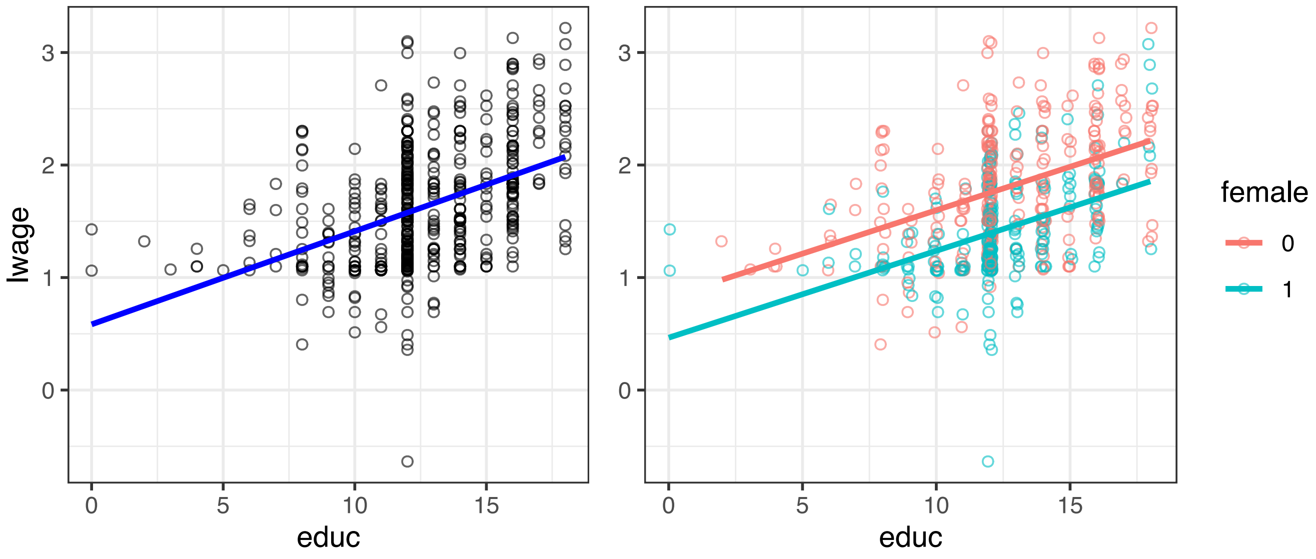
```r
data("wage1", package = "wooldridge")
wage1$female = as.factor(wage1$female)  # convert 0-1 to factor
lm_w = lm(lwage ~ educ, data = wage1)
lm_w_sex = lm(lwage ~ educ + female, data = wage1)
```

Let's Plot the Outcomes

	Dependent variable:	
	lwage	
	(1)	(2)
educ	0.083***	0.077***
	(0.008)	(0.007)
female1		-0.361***
		(0.039)
Constant	0.584***	0.826***
	(0.097)	(0.094)
Observations	526	526
R^2	0.186	0.300

Interpretation

- `R` chooses a *reference category* (by default the first of all levels by order of appearance), which is excluded - here this is `female==0`.
- The interpretation is that b_2 measures the effect of being female *relative* to being male.
- `R` automatically creates a dummy variable for each potential level, excluding the first category.



Interactions

- It can be useful to let the slope of a certain variable vary with ***another*** regressor.
- For instance, what if women with higher education had better wages than similar men?

Interactions

- It can be useful to let the slope of a certain variable vary with ***another*** regressor.
- For instance, what if women with higher education had better wages than similar men?

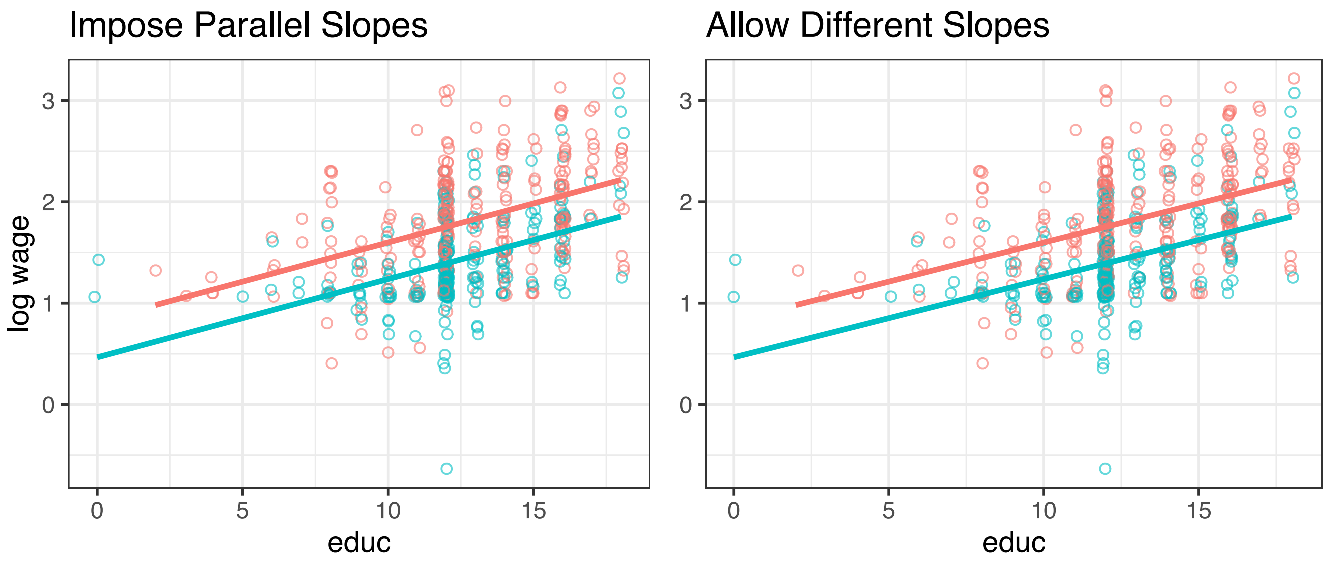
$$\ln w = b_0 + b_1 \text{female} + b_2 \text{educ} + b_3 (\text{female} \times \text{educ}) + e$$

- `female` is a factor with levels `0` and `1`: i.e. the interaction term b_3 will be zero for all men.

```r
# No need to write all variables, R expands to full interactions model!!
lm_w_interact <- lm(lwage ~ educ * female , data = wage1)
lm_w_interact
```

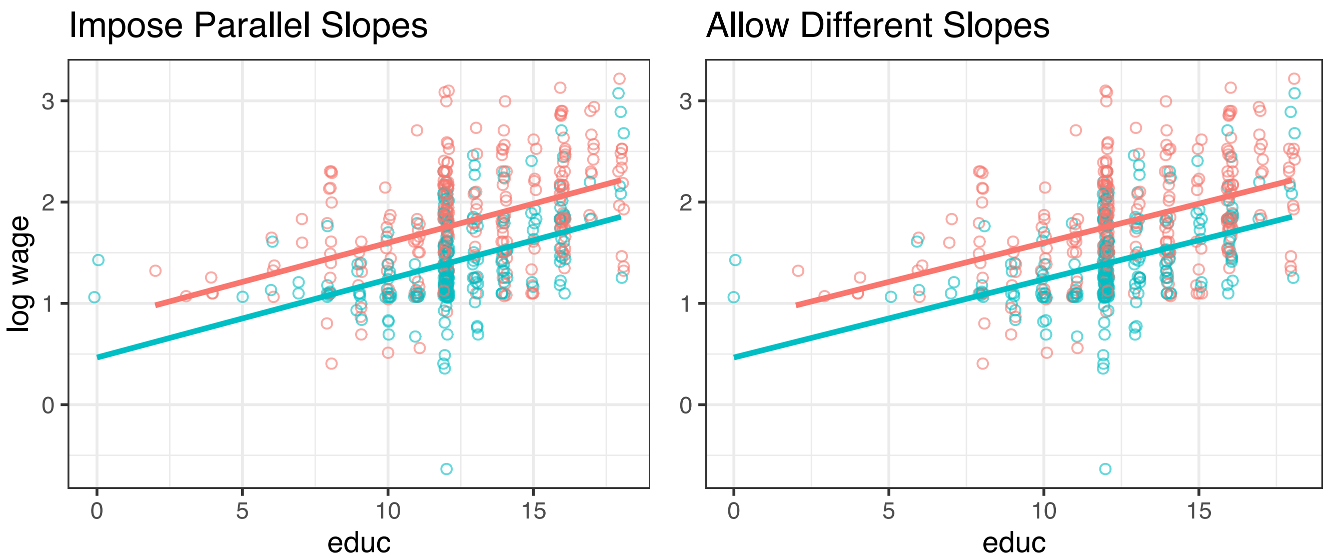
```
##
## Call:
## lm(formula = lwage ~ educ * female, data = wage1)
##
## Coefficients:
##  (Intercept)          educ        female1   educ:female1
##    8.260e-01     7.723e-02     -3.601e-01     -6.408e-05
```

Let's Plot Our Results



- Are the slope different?

Let's Plot Our Results



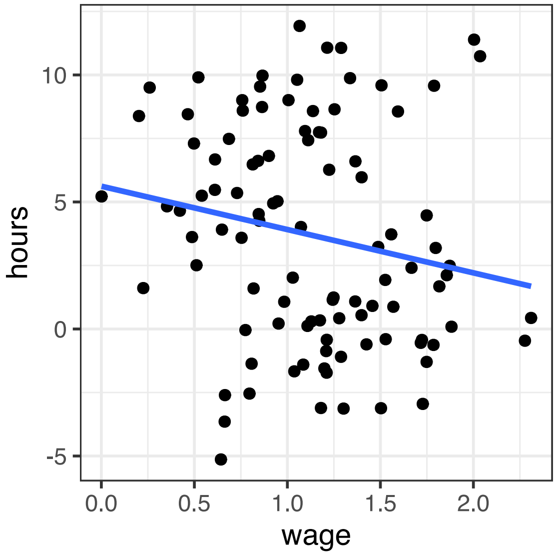
- Are the slope different?
- Right panel allows slopes to be different - turns out they are not!
- **Next session:** how can we ***test*** whether slopes are different?

Last but not Least: Individual Heterogeneity

1. Suppose we have data on hourly wage and a the number of hours worked by workers
2. We want to study the labour supply of those workers: regression `hours_worked ~ wage`.
3. We expect a positive coefficient on `wage`: higher wage => more hours worked.
4. Additional info: workers are either in group `g=0` or `g=1`.

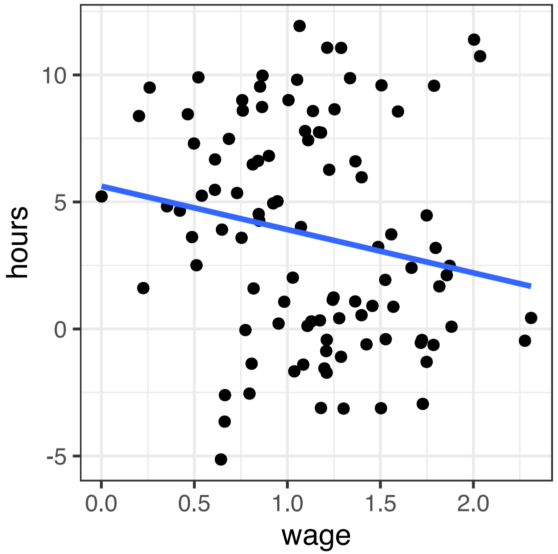
Last but not Least: Individual Heterogeneity

1. Suppose we have data on hourly wage and a the number of hours worked by workers
2. We want to study the labour supply of those workers: regression `hours_worked ~ wage`.
3. We expect a positive coefficient on `wage`: higher wage => more hours worked.
4. Additional info: workers are either in group `g=0` or `g=1`.



Last but not Least: Individual Heterogeneity

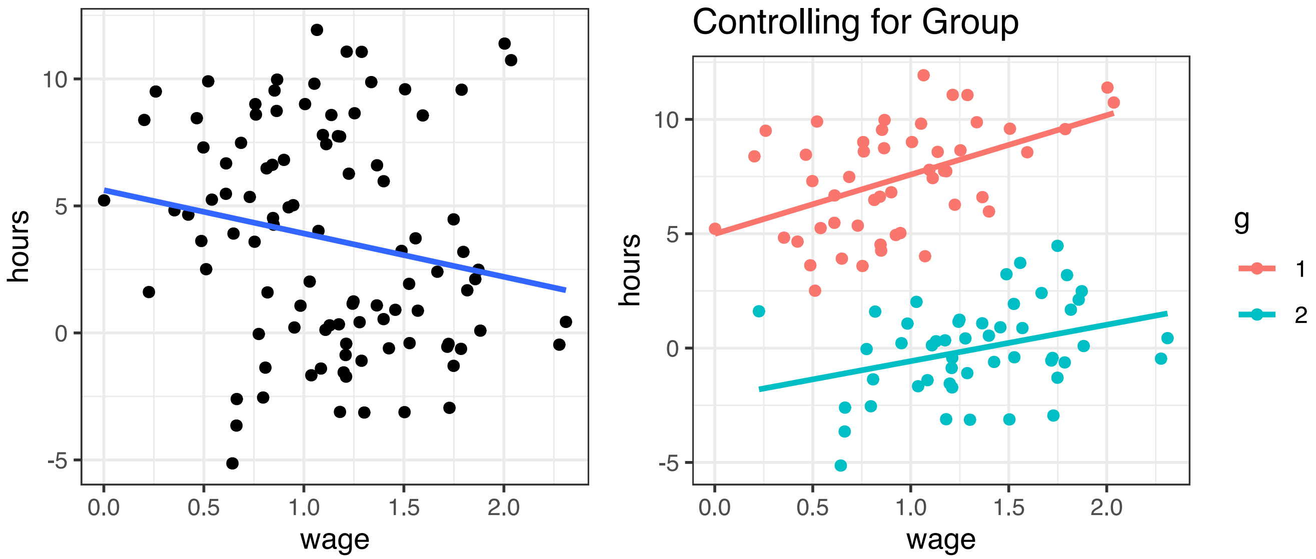
1. Suppose we have data on hourly wage and a the number of hours worked by workers
2. We want to study the labour supply of those workers: regression `hours_worked ~ wage`.
3. We expect a positive coefficient on `wage`: higher wage => more hours worked.
4. Additional info: workers are either in group `g=0` or `g=1`.



- … *a negative relation*?

Are We Missing Something?

- Let's run the same analysis *controlling* by group:



- This is an artificial example; yet it shows the importance of group-specific effects.

- What if groups are ***unobserved***? - we will need advanced methods that are beyond the scope of this course to infer the groups.

- If ***known***, you should include a group dummy so as to control for group effects.

SciencesPo
DEPARTMENT OF ECONOMICS

END

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