SciencesPo
DEPARTMENT OF ECONOMICS

ScPoEconometrics Advanced

Recap 2

Nikiforos Zampetakis based on Florian Oswald's slides
SciencesPo Paris
2024-02-03

Recap 2

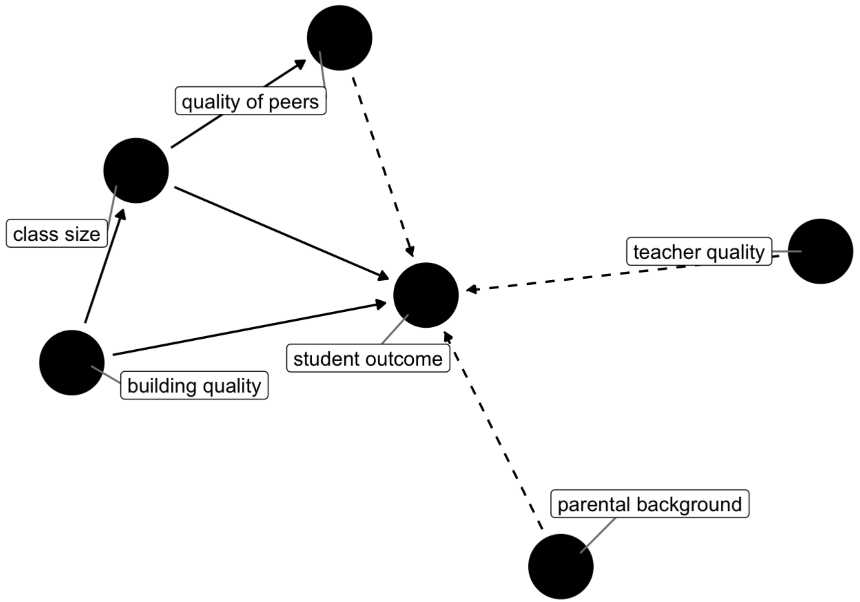
- Last time, we refreshed our basic OLS knowledge
- Today we continue and look at more than one explanatory variable, and associated problems

- But, why *more than one variable*?
- Like, **how many** other variables?
- And, above all: **which ones** ?

We will remember what we meant by **a model**.

Back to the STAR Experiment

- Remember what we learned about the STAR Experiment
- What is the causal impact of class size on test scores?
- $\text{score}_i = \beta_0 + \beta_1 \text{classize}_i + u_i$?
- We use a **model** to order our thoughts about how a causal impact is determined.

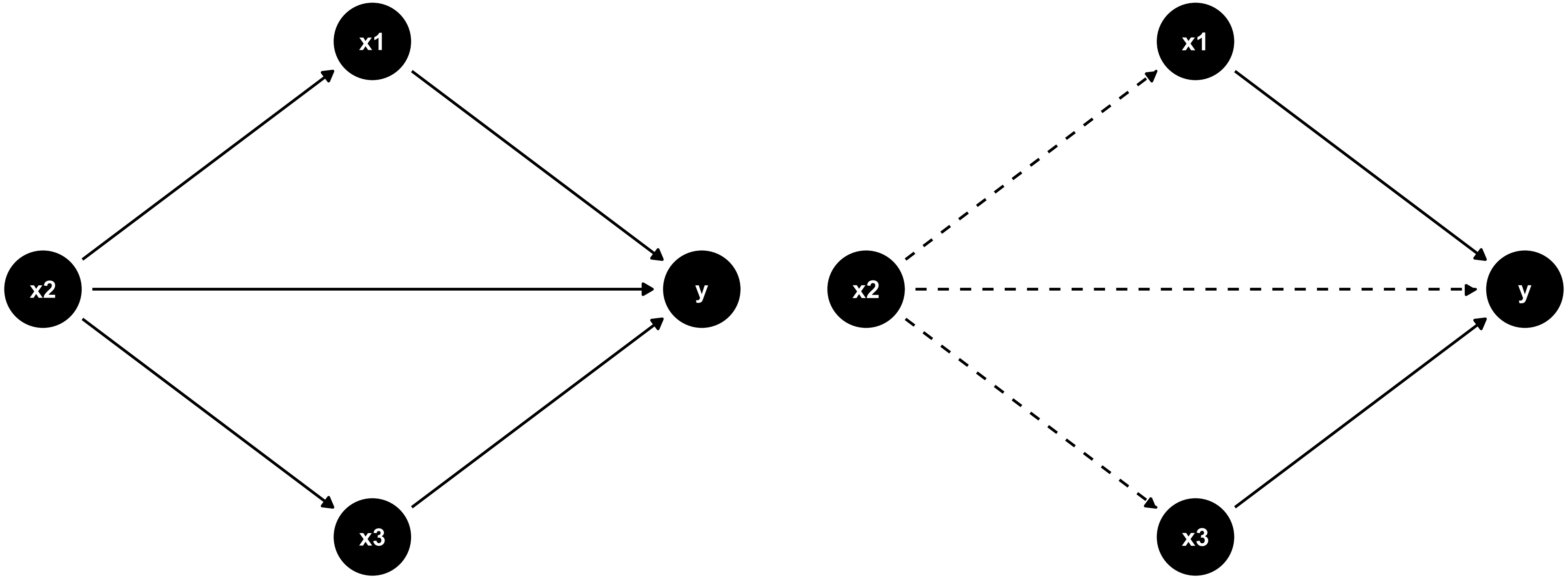


Multiple Variables

Let's augment our model with more variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Spot the Difference 🕵️‍♀️



Omitted-variable bias

Omitted-variable bias

Omitted-variable bias (OVB) arises when we omit a variable that

1. affects our outcome variable y

2. correlates with an explanatory variable x_j

As it's name suggests, this situation leads to bias in our estimate of β_j.

Note: OVB Is not exclusive to multiple linear regression, but it does require multiple variables affect y.

Omitted-variable bias

Example

Let's imagine a simple model for the amount individual i gets paid

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \beta_2 \text{Male}_i + u_i$$

where

- School_i gives i's years of schooling
- Male_i denotes an indicator variable for whether individual i is male.

thus

- β_1: the returns to an additional year of schooling (*ceteris paribus*)
- β_2: the premium for being male (*ceteris paribus*)
  If $\beta_2 > 0$, then there is discrimination against women—receiving less pay based upon gender.

Omitted-variable bias

Example, continued

From our population model

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \beta_2 \text{Male}_i + u_i$$

If a study focuses on the relationship between pay and schooling, *i.e.*,

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + (\beta_2 \text{Male}_i + u_i)$$

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \varepsilon_i$$

where $\varepsilon_i = \beta_2 \text{Male}_i + u_i$.

We used our exogeneity assumption to derive OLS' unbiasedness. But even if $\boldsymbol{E}[u|X] = 0$, it is not true that $\boldsymbol{E}[\varepsilon|X] = 0$ so long as $\beta_2 \neq 0$.

Specifically, $\boldsymbol{E}[\varepsilon|\text{Male} = 1] = \beta_2 + \boldsymbol{E}[u|\text{Male} = 1] \neq 0$. **Now OLS is biased.**

Omitted-variable bias

Example, continued

Let's try to see this result graphically.

The population model:

$$\text{Pay}_i = 20 + 0.5 \times \text{School}_i + 10 \times \text{Male}_i + u_i$$

Our regression model that suffers from omitted-variable bias:

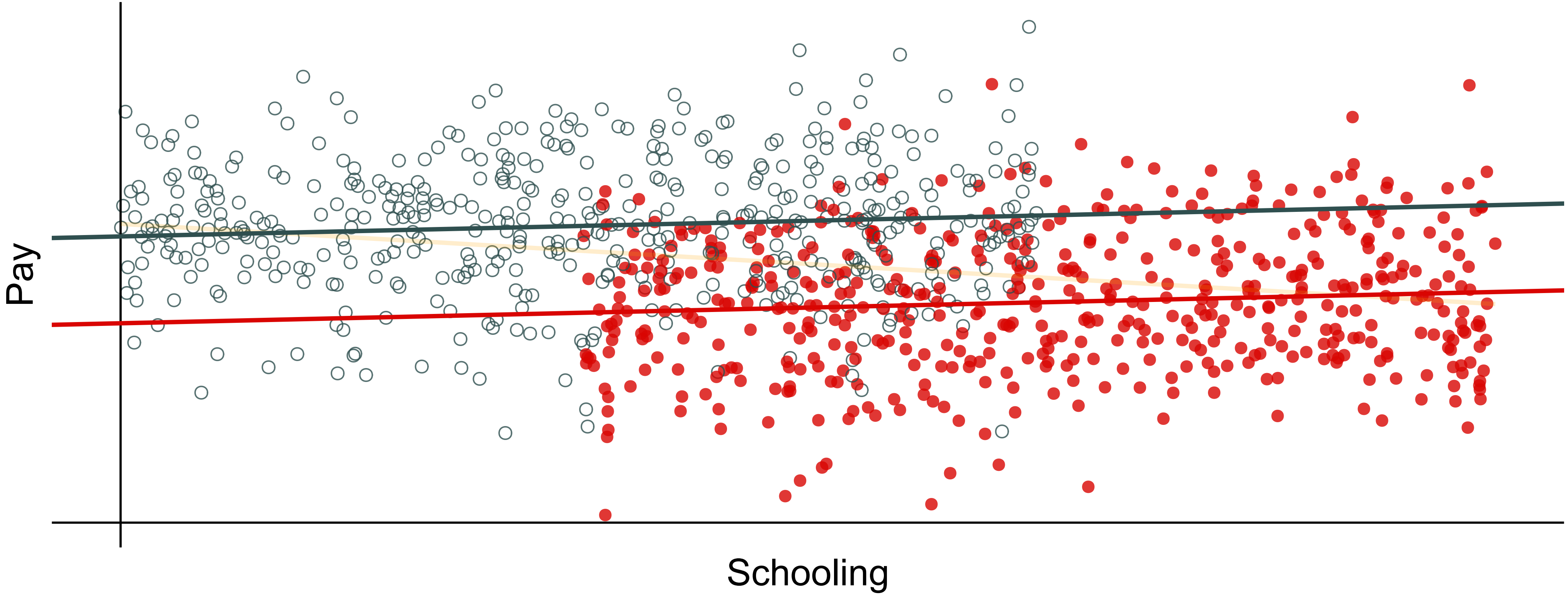
$$\text{Pay}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{School}_i + e_i$$

Finally, imagine that women, on average, receive more schooling than men.

Omitted-variable bias

Example, continued: $\text{Pay}_i = 20 + 0.5 \times \text{School}_i + 10 \times \text{Male}_i + u_i$

Unbiased regression estimate: $\widehat{\text{Pay}}_i = 20.9 + 0.4 \times \text{School}_i + 9.1 \times \text{Male}_i$



Omitted-variable bias

Solutions

1. Don't omit variables

2. Instrumental variables and two-stage least squares (coming soon): If we could find something that **only** affects x_1 but *not* the omitted variable, we can make progress!

3. Use multiple observations for the same unit i: panel data.

Warning: There are situations in which neither solution is possible.

1. Proceed with caution (sometimes you can sign the bias).

2. The key is to have a mental map of what *should* belong to the model.

Interpreting coefficients

Interpreting coefficients

Continuous variables

Consider the relationship

$$\text{Pay}_i = \beta_0 + \beta_1 \,\text{School}_i + u_i$$

where

- Pay_i is a continuous variable measuring an individual's pay
- School_i is a continuous variable that measures years of education

Interpreting coefficients

Interpretations

- β_0: the y-intercept, *i.e.*, Pay when $\text{School} = 0$
- β_1: the expected increase in Pay for a one-unit increase in School

Interpreting coefficients

Continuous variables

Consider the model

$$y = \beta_0 + \beta_1 x + u$$

Differentiate the model:

$$\frac{dy}{dx} = \beta_1$$

Task 1: Interpretation (4 minutes)

1. Load the `wage1` dataset from the `wooldridge` package. you may have to install this first.
2. Run `skimr::skim` on the dataset to get an overview. what is the fraction of nonwhite in the data?
3. Regressing wage on education and tenure, what is the interpretation of the `tenure` coefficient? You may need to consult `?wage1` here.

Task 1: Solution

1.Load the data

```
library(tidyverse)
library(wooldridge)

data("wage1")
```

2.Run `skimr::skim` on the dataset to get an overview. what is the fraction of nonwhite in the data?

```
library(skimr)

wage1 %>% skim()
```

3.Regressing wage on education and tenure, what is the interpretation of the tenure coefficient? You may need to consult ?wage1 here.

```
summary(lm(wage ~ educ + tenure, data = wage1))
```

Interpreting coefficients

Categorical variables

Consider the relationship

$$\text{Pay}_i = \beta_0 + \beta_1 \, \text{Female}_i + u_i$$

where

- Pay_i is a continuous variable measuring an individual's pay
- Female_i is a binary/indicator variable taking 1 when i is female

Interpreting coefficients

Interpretations

- β_0: the expected Pay for males (*i.e.*, when $\text{Female} = 0$)
- β_1: the expected difference in Pay between females and males
- $\beta_0 + \beta_1$: the expected Pay for females

Interpreting coefficients

Categorical variables

Derivations

$$
\begin{aligned}
\boldsymbol{E}[\text{Pay}|\text{Male}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 0 + u_i] \\
&= \boldsymbol{E}[\beta_0 + 0 + u_i] \\
&= \beta_0
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{E}[\text{Pay}|\text{Female}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 1 + u_i] \\
&= \boldsymbol{E}[\beta_0 + \beta_1 + u_i] \\
&= \beta_0 + \beta_1
\end{aligned}
$$

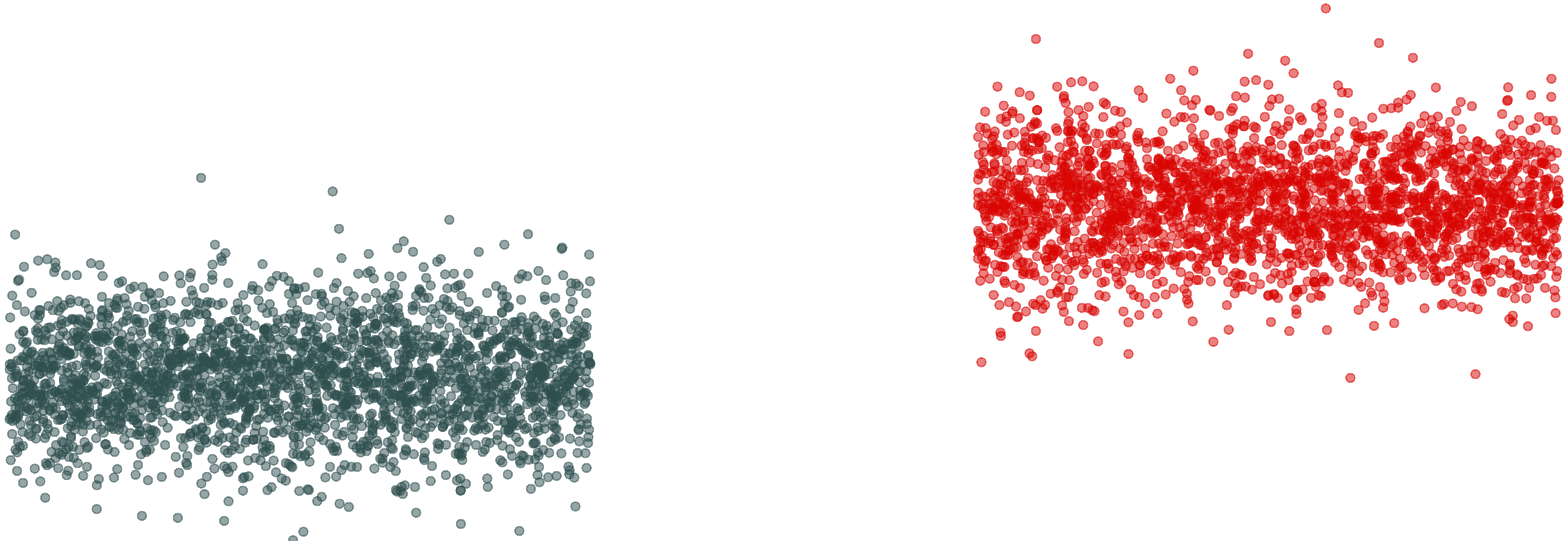
Note: If there are no other variables to condition on, then $\hat{\beta}_1$ equals the difference in group means, *e.g.*, $\overline{x}_{\text{Female}} - \overline{x}_{\text{Male}}$.

Note$_2$: The *holding all other variables constant* interpretation also applies for categorical variables in multiple regression settings.

Interpreting coefficients

Categorical variables

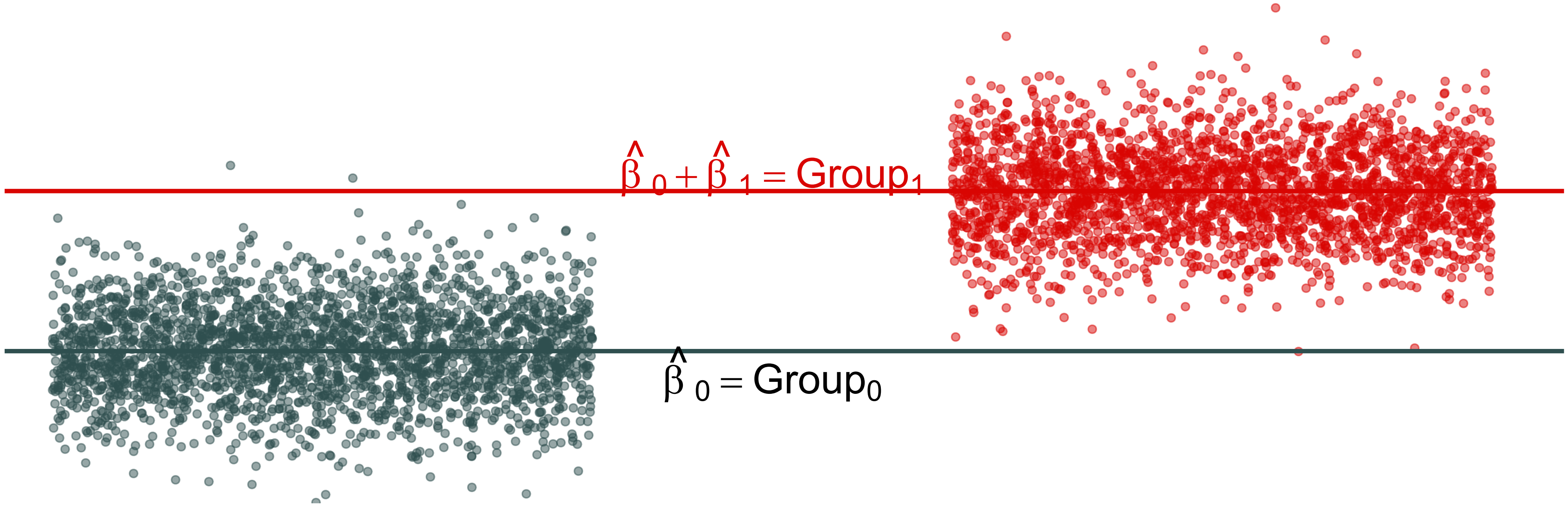
$y_i = \beta_0 + \beta_1 x_i + u_i$ for binary variable $x_i = \{0, 1\}$



Interpreting coefficients

Categorical variables

$y_i = \beta_0 + \beta_1 x_i + u_i$ for binary variable $x_i = \{0,\ 1\}$



Task 2: Categorical Variables (3 Minutes)

- Continue with the `wage1` dataset.
- Now regress `wage` on `female`. What is $E[wage|\text{male}]$?
- Add `married` to the regression. Now what is $E[wage|\text{female,not married}]$?

Task 2: Solution

- Now regress `wage` on `female`. What is $E[wage|\text{male}]$?

```
summary(lm(wage ~ female, data = wage1))
```

- Add `married` to the regression. Now what is $E[wage|\text{female,not married}]$?

```
summary(lm(wage ~ female + married, data = wage1))
```

Interpreting coefficients

Interactions

Interactions allow the effect of one variable to change based upon the level of another variable.

Examples

1. Does the effect of schooling on pay change by gender?
2. Does the effect of gender on pay change by race?
3. Does the effect of schooling on pay change by experience?

Interpreting coefficients

Interactions

Previously, we considered a model that allowed women and men to have different wages, but the model assumed the effect of school on pay was the same for everyone:

$$\text{Pay}_i = \beta_0 + \beta_1 \, \text{School}_i + \beta_2 \, \text{Female}_i + u_i$$

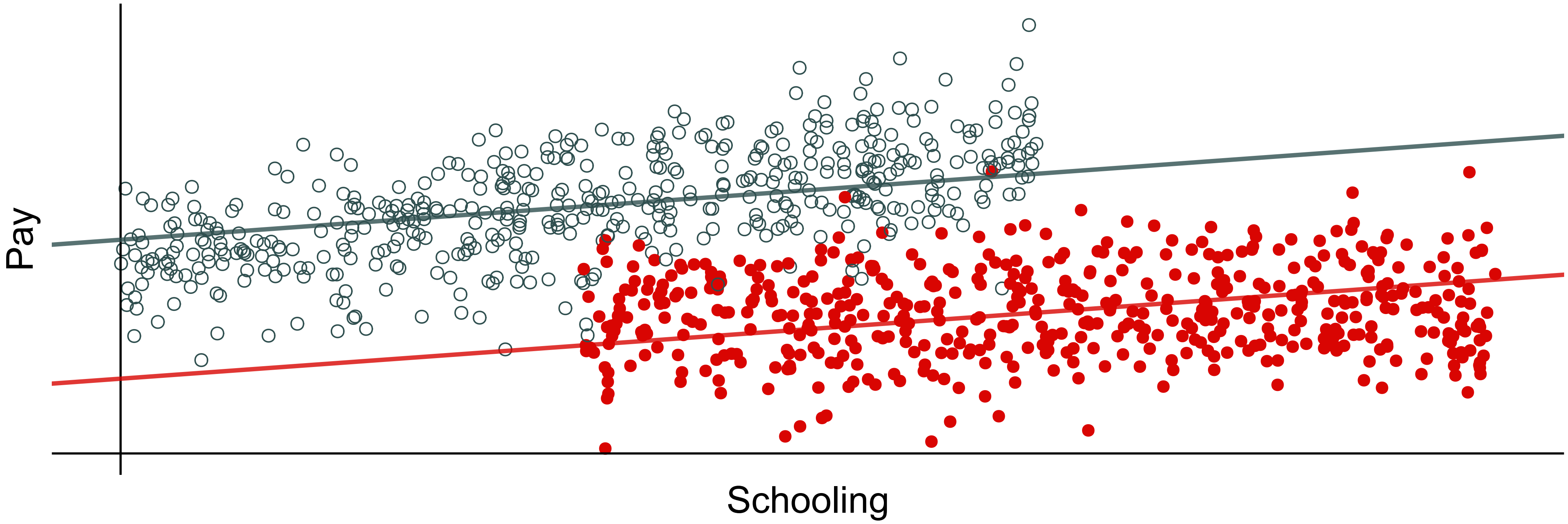
but we can also allow the effect of school to vary by gender:

$$\text{Pay}_i = \beta_0 + \beta_1 \, \text{School}_i + \beta_2 \, \text{Female}_i + \beta_3 \, \text{School}_i \times \text{Female}_i + u_i$$

Interpreting coefficients

Interactions

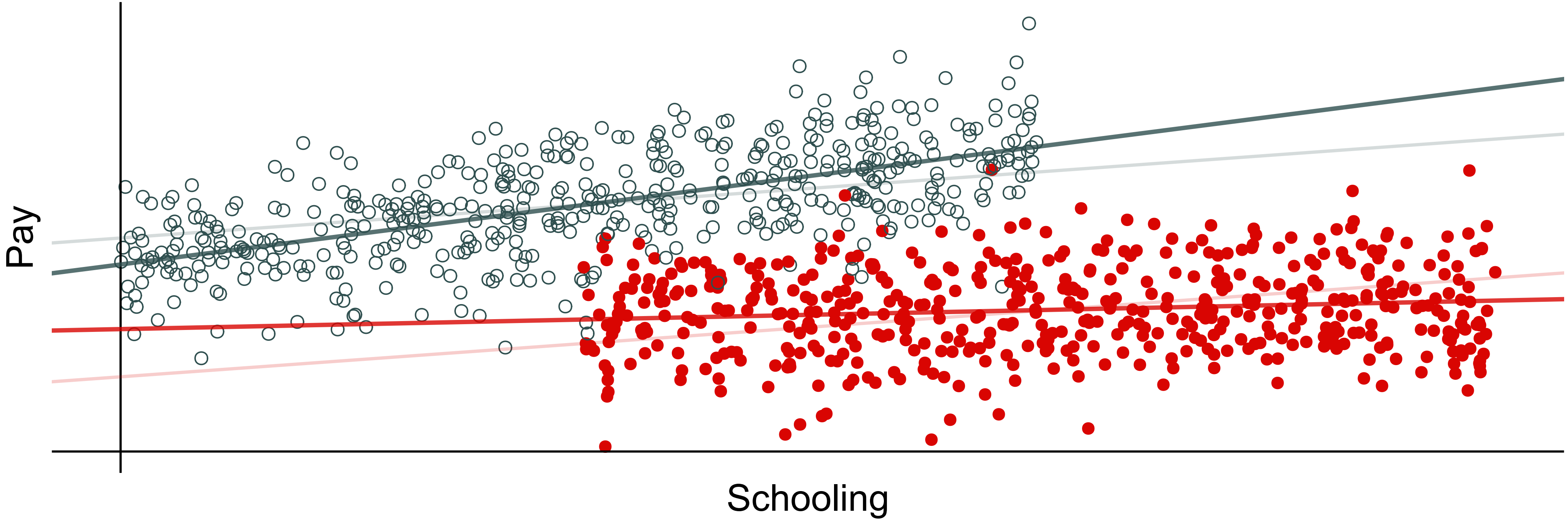
The model where schooling has the same effect for everyone (**F** and **M**):



Interpreting coefficients

Interactions

The model where schooling's effect can differ by gender (**F** and **M**):



Interpreting coefficients

Interactions

Interpreting coefficients can be a little tricky with interactions, but the key[†] is to carefully work through the math.

$$\text{Pay}_i = \beta_0 + \beta_1 \, \text{School}_i + \beta_2 \, \text{Female}_i + \beta_3 \, \text{School}_i \times \text{Female}_i + u_i$$

Expected returns for an additional year of schooling for women:

$$\begin{aligned} \boldsymbol{E}[\text{Pay}_i | \text{Female} \wedge \text{School} = \ell + 1] - \boldsymbol{E}[\text{Pay}_i | \text{Female} \wedge \text{School} = \ell] &= \\ \boldsymbol{E}[\beta_0 + \beta_1(\ell + 1) + \beta_2 + \beta_3(\ell + 1) + u_i] - \boldsymbol{E}[\beta_0 + \beta_1 \ell + \beta_2 + \beta_3 \ell + u_i] &= \\ \beta_1 + \beta_3 & \end{aligned}$$

Similarly, β_1 gives the expected return to an additional year of schooling for men. Thus, β_3 gives the **difference in the returns to schooling** for women and men.

† As is often the case with econometrics.

Task 3: Interactions (4 minutes)

- Same dataset!
- Regress wage on experience, female indicator and their interaction. What is the interpretation of all the coefficients here? Can you distinguish them from zero?
- What is the expected wage for a male with 5 years of experience?

Task 3: Solution

- Regress wage on experience, female indicator and their interaction. What is the interpretation of all the coefficients here? Can you distinguish them from zero?

```
lm1 = lm(wage ~ female*exper, data = wage1)

summary(lm1)

summary(lm(wage ~ exper + female +female:exper, data = wage1))
```

- What is the expected wage for a male with 5 years of experience?

```
male5 = tibble(female = 0, exper = 5)

predict(lm1, male5)
```

Interpreting coefficients

Log-linear specification

In economics, you will frequently see logged outcome variables with linear (non-logged) explanatory variables, *e.g.*,

$$\log(\text{price}_i) = \beta_0 + \beta_1 \text{bdrms}_i + u_i$$

This specification changes our interpretation of the slope coefficients.

```r
data(hprice1,package = "wooldridge")
lm(log(price) ~ bdrms, data = hprice1) %>% tidy()
```

```
#> # A tibble: 2 × 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    5.04     0.126      39.9  3.13e-57
#> 2 bdrms          0.167    0.0345      4.85 5.43e- 6
```

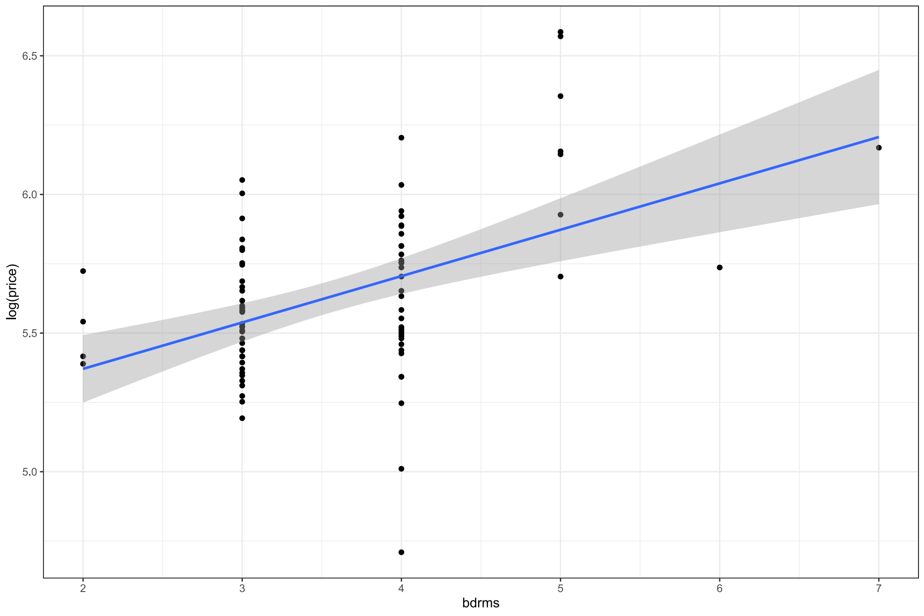
Interpreting coefficients

Log-linear specification

Interpretation

- A one-unit increase in our explanatory variable increases the outcome variable by *approximately* $\beta_1 \times 100$ percent.

- *Example:* An additional bedroom increases sales prices of a house by *approximately* 16 percent (for $\beta_1 = 0.16$).



Interpreting coefficients

Log-linear specification

Consider the log-linear model

$$\log(y) = \beta_0 + \beta_1\, x + u$$

and differentiate

$$\frac{dy}{y} = \beta_1 dx$$

So a marginal change in x (*i.e.*, dx) leads to a $\beta_1 dx$ **percentage change** in y.

Interpreting coefficients

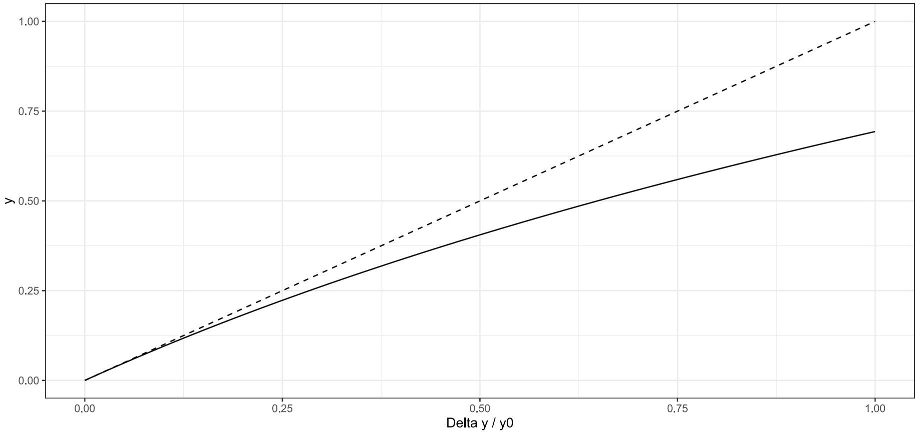
Log-linear specification

What about that **approximation** part?

> An additional bedroom increases sales prices of a house by *approximately* 16 percent (for $\beta_1 = 0.16$).

- $\%\Delta y \approx 0.16 \times 100 = 16\%$.
- Good approximation as long as Δy is not too big.
- We approximate

$$\log\left(\frac{\Delta y}{y_0} + 1\right) \approx \frac{\Delta y}{y_0}$$



Interpreting coefficients

Log-linear specification

What about that **approximation** part?

> An additional bedroom increases sales prices of a house by *approximately* 16 percent (for $\beta_1 = 0.16$).

- $\%\Delta y \approx 0.16 \times 100 = 16\%$.
- Good approximation as long as Δy is not too big.
- We approximate

$$\log\left(\frac{\Delta y}{y_0} + 1\right) \approx \frac{\Delta y}{y_0}$$

- The **exact** formula is

$$\%\Delta y = 100 \times (\exp(\Delta x \beta) - 1)$$

- In our case:

$$\%\Delta y = 100 \times (\exp(0.16) - 1) = 17.3$$

Task 4

- same Dataset!
- Now regress *log wage* on education and tenure. How does the interpretation of the coefficient on education change?

Task 4: Solution

- Now regress *log wage* on education and tenure. How does the interpretation of the coefficient on education change?

```
summary(lm(lwage ~ educ + tenure, data = wage1))
```

Interpreting coefficients

Log-log specification

Similarly, econometricians frequently employ log-log models, in which the outcome variable is logged *and* at least one explanatory variable is logged

$$\log(\text{price}_i) = \beta_0 + \beta_1 \, \log(\text{sqrft}_i) + u_i$$

Interpretation:

- A one-percent increase in x will lead to a β_1 percent change in y.
- Often interpreted as an elasticity.

Interpreting coefficients

Log-log specification

Consider the log-log model

$$\log(y) = \beta_0 + \beta_1 \log(x) + u$$

and differentiate

$$\frac{dy}{y} = \beta_1 \frac{dx}{x}$$

which says that for a one-percent increase in x, we will see a β_1 percent increase in y. As an elasticity:

$$\frac{dy}{dx}\frac{x}{y} = \beta_1$$

Task 5

- Load the `hprice1` dataset from the `wooldridge` package.
- Regress log price on log sqrft. What is the interpretation on `log(sqrft)`?
- What is the $E[\text{price}|\text{sqrft} = 115]$ (Caution! not `log price`!)

Task 5

- Load the `hprice1` dataset from the `wooldridge` package.

```
data("hprice1")
```

- Regress log price on log sqrft. What is the interpretation on `log(sqrft)`?

```
summary(lm(lprice ~ lsqrft, data = hprice1))
```

Interpreting coefficients

Log-log specification

```
lm(log(price) ~ log(sqrft), data = hprice1) %>% tic
```

```
#> # A tibble: 2 × 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   -0.975    0.641      -1.52 1.32e- 1
#> 2 log(sqrft)     0.873    0.0846     10.3  1.05e-16
```

- a 1% increase in square footage of the house leads to a 0.873% increase in sales price.
- Notice the absence of *units* here (it's all in **percent** terms of both variables involved).

Interpreting coefficients

Log-linear with a binary variable

Note: If you have a log-linear model with a binary indicator variable, the interpretation for the coefficient on that variable changes.

Consider again

$$\log(y_i) = \beta_0 + \beta_1 x_1 + u_i$$

for binary variable x_1.

The *approximate* interpretation of β_1 is as before:

> When x_1 changes from 0 to 1, y will change by $100 \times \beta_1$ percent.

```
#>
#> Call:
#> lm(formula = log(price) ~ log(lotsize) + log(sqrft) + bdrms +
#>     colonial, data = hprice1)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -0.69479 -0.09750 -0.01619  0.09151  0.70228
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  -1.34959    0.65104  -2.073   0.0413 *
#> log(lotsize)  0.16782    0.03818   4.395 3.25e-05 ***
#> log(sqrft)    0.70719    0.09280   7.620 3.69e-11 ***
#> bdrms         0.02683    0.02872   0.934   0.3530
#> colonial      0.05380    0.04477   1.202   0.2330
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.1841 on 83 degrees of freedom
#> Multiple R-squared:  0.6491,    Adjusted R-squared:  0.6322
#> F-statistic: 38.38 on 4 and 83 DF,  p-value: < 2.2e-16
```

Approximate

- When *colonial* changes from 0 to 1 (i.e. house *becomes* colonial), y will change by $100 \times \beta_1 = 5.37$ percent.

Exact

- When *colonial* changes from 1 to 0, y will change by $100 \times \left(e^{\beta_1} - 1\right) = 5.52$ percent.

Uncertainty and inference

Uncertainty and inference

Is there more?

Up to this point, we know OLS has some nice properties, and we know how to estimate an intercept and slope coefficient via OLS.

Our current workflow:

- Get data (points with x and y values)
- Regress y on x
- Plot the OLS line (*i.e.*, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$)
- Done?

But how do we actually **learn** something from this exercise?

Uncertainty and inference

Linkup with Intro Course

This is related to *Intro Course material*:

1. Sampling
2. Hypothesis Testing
3. Regression Inference

Uncertainty and inference

There is more

But how do we actually **learn** something from this exercise?

- Based upon our value of $\hat{\beta}_1$, can we rule out previously hypothesized values?
- How confident should we be in the precision of our estimates?
- How well does our model explain the variation we observe in y?

We need to be able to deal with uncertainty. Enter: **Inference.**

Uncertainty and inference

Learning from our errors

As our previous simulation pointed out, our problem with **uncertainty** is that we don't know whether our sample estimate is *close* or *far* from the unknown population parameter.[†]

However, all is not lost. We can use the errors $\left(e_i = y_i - \hat{y}_i\right)$ to get a sense of how well our model explains the observed variation in y.

When our model appears to be doing a "nice" job, we might be a little more confident in using it to learn about the relationship between y and x.

Now we just need to formalize what a "nice job" actually means.

†: Except when we run the simulation ourselves—which is why we like simulations.

Uncertainty and inference

Learning from our errors

First off, we will estimate the variance of u_i (recall: $\text{Var}(u_i) = \sigma^2$) using our squared errors, *i.e.*,

$$s^2 = \frac{\sum_i e_i^2}{n - k}$$

where k gives the number of slope terms and intercepts that we estimate (*e.g.*, β_0 and β_1 would give $k = 2$).

s^2 is an unbiased estimator of σ^2.

Uncertainty and inference

Learning from our errors

We know that the variance of $\hat{\beta}_1$ (for simple linear regression) is

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{s^2}{\sum_i \left(x_i - \overline{x}\right)^2}$$

which shows that the variance of our slope estimator

1. increases as our disturbances become noisier
2. decreases as the variance of x increases

Uncertainty and inference

Learning from our errors

More common: The **standard error** of $\hat{\beta}_1$

$$\hat{\text{SE}}\left(\hat{\beta}_1\right) = \sqrt{\frac{s^2}{\sum_i \left(x_i - \overline{x}\right)^2}}$$

Recall: The standard error of an estimator is the standard deviation of the estimator's distribution.

Uncertainty and inference

Learning from our errors

Standard error output is standard in `R`'s `lm`:

```
#> [[1]]
#>  [1] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
#>  [7] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics" 
#> [13] "grDevices" "utils"     "datasets"  "methods"   "base"      "broom"    
#> [19] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
#> [25] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics" 
#> [31] "grDevices" "utils"     "datasets"  "methods"   "base"     
#> 
#> [[2]]
#>  [1] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
#>  [7] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics" 
#> [13] "grDevices" "utils"     "datasets"  "methods"   "base"      "broom"    
#> [19] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
#> [25] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics" 
#> [31] "grDevices" "utils"     "datasets"  "methods"   "base"     
#> 
#> [[3]]
#>  [1] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
#>  [7] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics" 
#> [13] "grDevices" "utils"     "datasets"  "methods"   "base"      "broom"    
#> [19] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
#> [25] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics" 
#> [31] "grDevices" "utils"     "datasets"  "methods"   "base"     
```

Uncertainty and inference

Learning from our errors

We use the standard error of $\hat{\beta}_1$, along with $\hat{\beta}_1$ itself, to learn about the parameter β_1.

After deriving the distribution of $\hat{\beta}_1$,[†] we have two (related) options for formal statistical inference (learning) about our unknown parameter β_1:

- **Confidence intervals:** Use the estimate and its standard error to create an interval that, when repeated, will generally[††] contain the true parameter.
- **Hypothesis tests:** Determine whether there is statistically significant evidence to reject a hypothesized value or range of values.

†: *Hint:* it's normal with the mean and variance we've derived/discussed above)
††: *E.g.*, Similarly constructed 95% confidence intervals will contain the true parameter 95% of the time.

Uncertainty and inference

Confidence intervals

We construct $(1-\alpha)$-level confidence intervals for β_1

$$\hat{\beta}_1 \pm t_{\alpha/2,\text{df}}\ \hat{\text{SE}}\left(\hat{\beta}_1\right)$$

$t_{\alpha/2,\text{df}}$ denotes the $\alpha/2$ quantile of a t dist. with $n-k$ degrees of freedom.

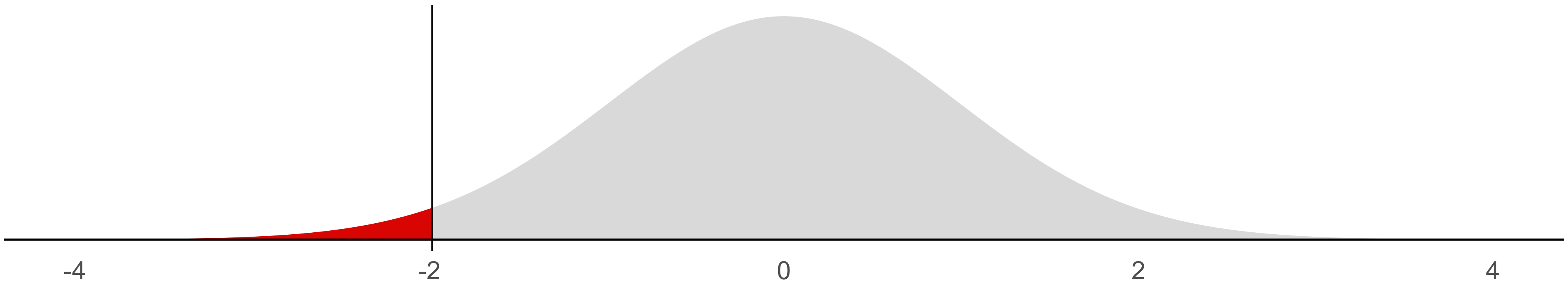
Uncertainty and inference

Confidence intervals

We construct $(1-\alpha)$-level confidence intervals for β_1

$$\hat{\beta}_1 \pm t_{\alpha/2,\text{df}} \, \hat{\text{SE}}\left(\hat{\beta}_1\right)$$

For example, 100 obs., two coefficients (*i.e.*, $\hat{\beta}_0$ and $\hat{\beta}_1 \implies k=2$), and $\alpha = 0.05$ (for a 95% confidence interval) gives us $t_{0.025,\,98} = -1.98$



Uncertainty and inference

Confidence intervals

We construct $(1-\alpha)$-level confidence intervals for β_1

$$\hat{\beta}_1 \pm t_{\alpha/2,\text{df}}\ \hat{\text{SE}}\left(\hat{\beta}_1\right)$$

Example:

```
lm(y ~ x, data = pop_df) %>% tidy(conf.int = TRUE)
```

```
#> # A tibble: 2 × 7
#>   term        estimate std.error statistic  p.value conf.low conf.high
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
#> 1 (Intercept)    2.53     0.422       6.00 3.38e- 8    1.69      3.37 
#> 2 x              0.567    0.0793      7.15 1.59e-10    0.410     0.724
```

Our 95% confidence interval is thus $0.567 \pm 1.98 \times 0.0793 = [0.410,\ 0.724]$

Uncertainty and inference

Confidence intervals

So we have a confidence interval for β_1, *i.e.*, $[0.410,\ 0.724]$.

What does it mean?

Informally: The confidence interval gives us a region (interval) in which we can place some trust (confidence) for containing the parameter.

More formally: If repeatedly sample from our population and construct confidence intervals for each of these samples, $(1-\alpha)$ percent of our intervals (*e.g.*, 95%) will contain the population parameter *somewhere in the interval*.

Now back to our simulation...

Uncertainty and inference

Confidence intervals

We drew 10,000 samples (each of size $n = 30$) from our population and estimated our regression model for each of these simulations:

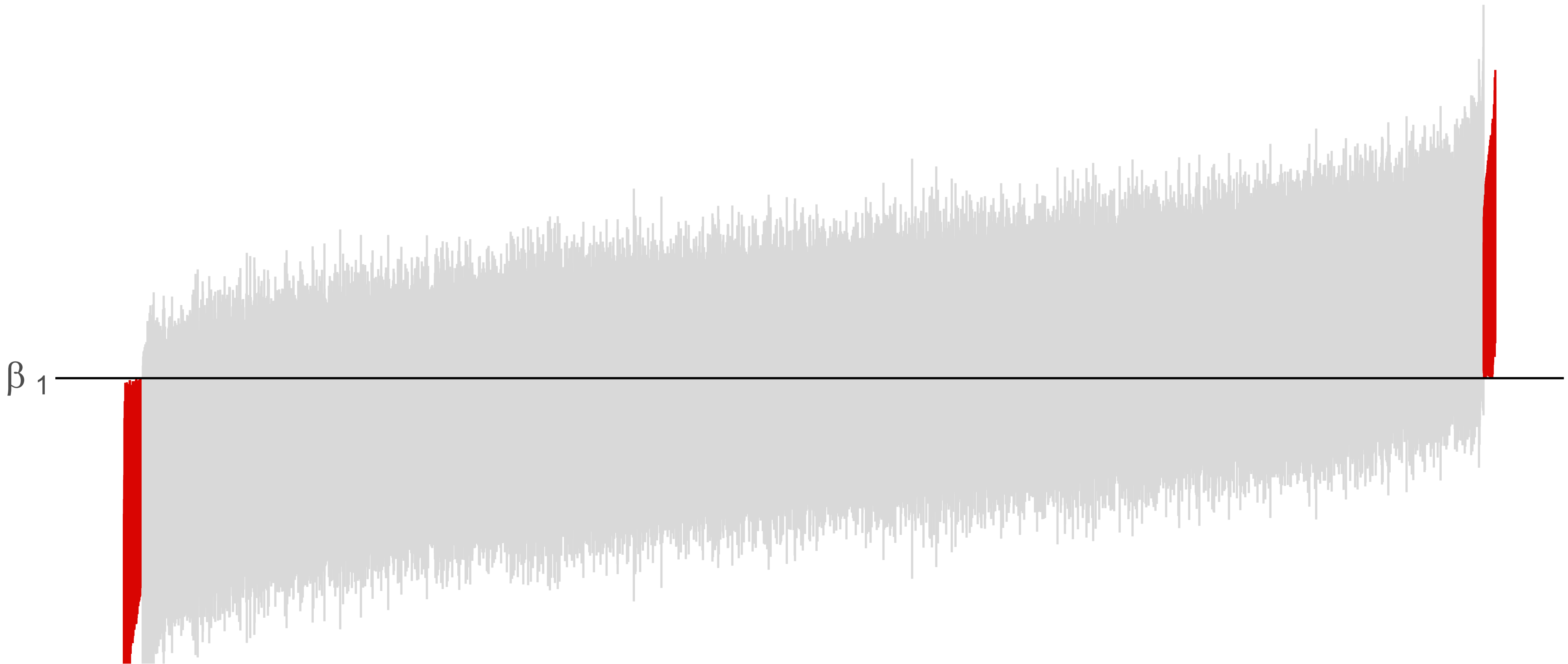
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

(repeated 10,000 times)

Now, let's estimate 95% confidence intervals for each of these intervals…

Uncertainty and inference

Confidence intervals



Uncertainty and inference

Hypothesis testing

In many applications, we want to know more than a point estimate or a range of values. We want to know what our statistical evidence says about existing theories.

We want to test hypotheses posed by officials, politicians, economists, scientists, friends, weird neighbors, *etc.*

Examples

- Does increasing police presence **reduce crime**?
- Does building a giant wall **reduce crime**?
- Does shutting down a government **adversely affect the economy**?
- Does legal cannabis **reduce drunk driving** or **reduce opiod use**?
- Do air quality standards **increase health** and/or **reduce jobs**?

Uncertainty and inference

Hypothesis testing

Hypothesis testing relies upon very similar results and intuition.

While uncertainty certainly exists, we can still build *reliable* statistical tests (rejecting or failing to reject a posited hypothesis).

OLS t test Our (null) hypothesis states that β_1 equals a value c, *i.e.*, $H_o:\ \beta_1 = c$

From OLS's properties, we can show that the test statistic

$$t_{\text{stat}} = \frac{\hat{\beta}_1 - c}{\hat{\text{SE}}\left(\hat{\beta}_1\right)}$$

follows the t distribution with $n-k$ degrees of freedom.

Uncertainty and inference

Hypothesis testing

For an α-level, **two-sided** test, we reject the null hypothesis (and conclude with the alternative hypothesis) when

$$\left|t_{\text{stat}}\right| > \left|t_{1-\alpha/2,\, df}\right|$$

meaning that our **test statistic is more extreme than the critical value**.

Alternatively, we can calculate the **p-value** that accompanies our test statistic, which effectively gives us the probability of seeing our test statistic *or a more extreme test statistic* if the null hypothesis were true.

Very small p-values (generally < 0.05) mean that it would be unlikely to see our results if the null hyopthesis were really true—we tend to reject the null for p-values below 0.05.

Uncertainty and inference

Hypothesis testing

`R` and `stata`s default to testing hypotheses against the value zero.

```
lm(y ~ x, data = pop_df) %>% tidy()
```

```
#> # A tibble: 2 × 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    2.53     0.422       6.00 3.38e- 8
#> 2 x              0.567    0.0793      7.15 1.59e-10
```

Ho: $\beta_1 = 0$ *vs.* Ha: $\beta_1 \neq 0$

$t_\text{stat} = 7.15$ and $t_{0.975,\,28} = 2.05$ which implies p-value < 0.05

Therefore, we reject Ho.

Uncertainty and inference

F tests

You will sometimes see F tests in econometrics.

We use F tests to test hypotheses that involve multiple parameters
(*e.g.*, $\beta_1 = \beta_2$ or $\beta_3 + \beta_4 = 1$),

rather than a single simple hypothesis
(*e.g.*, $\beta_1 = 0$, for which we would just use a t test).

Uncertainty and inference

F tests

Example

Economists love to say "Money is fungible."

Imagine that we might want to test whether money received as income actually has the same effect on consumption as money received from tax rebates/returns.

$$\text{Consumption}_i = \beta_0 + \beta_1 \text{Income}_i + \beta_2 \text{Rebate}_i + u_i$$

Uncertainty and inference

F tests

Example, continued

We can write our null hypothesis as

$$H_o:\ \beta_1 = \beta_2 \iff H_o:\ \beta_1 - \beta_2 = 0$$

Imposing this null hypothesis gives us the **restricted model**

$$\text{Consumption}_i = \beta_0 + \beta_1 \text{Income}_i + \beta_1 \text{Rebate}_i + u_i$$

$$\text{Consumption}_i = \beta_0 + \beta_1 \left(\text{Income}_i + \text{Rebate}_i\right) + u_i$$

Uncertainty and inference

F tests

Example, continued

To this the null hypothesis $H_o:\ \beta_1 = \beta_2$ against $H_a:\ \beta_1 \neq \beta_2$,
we use the F statistic

$$F_{q,\, n-k-1} = \frac{\left(\text{SSE}_r - \text{SSE}_u\right)/q}{\text{SSE}_u/(n-k-1)}$$

which (as its name suggests) follows the F distribution with q numerator degrees of freedom and $n-k-1$ denominator degrees of freedom.

Here, q is the number of restrictions we impose via H_o.

Uncertainty and inference

F tests

Example, continued

The term SSE_r is the sum of squared errors (SSE) from our **restricted model**

$$\text{Consumption}_i = \beta_0 + \beta_1 \left(\text{Income}_i + \text{Rebate}_i\right) + u_i$$

and SSE_u is the sum of squared errors (SSE) from our **unrestricted model**

$$\text{Consumption}_i = \beta_0 + \beta_1 \text{Income}_i + \beta_2 \text{Rebate}_i + u_i$$

SciencesPo
DEPARTMENT OF ECONOMICS

END

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