SciencesPo
DEPARTMENT OF ECONOMICS

ScPoEconometrics

Linear Regression Extensions

Florian Oswald, Gustave Kenedi and Pierre Villedieu
SciencesPo Paris
2021-03-22

Quick "Quiz" on Last Week's Material

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

   OR

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

Today - Linear Regression Extensions

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

Today - Linear Regression Extensions

Depending on the data 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:

1. *Standardized* regression

2. *Non-linear relationships*: log and polynomial models

3. *Interactions* between variables

Today - Linear Regression Extensions

Depending on the data 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:

1. *Standardized* regression

2. *Non-linear relationships*: log and polynomial models

3. *Interactions* between variables

In each case, the way we estimate these coefficients does not change (i.e OLS).

Today - Linear Regression Extensions

Depending on the data 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:

1. ***Standardized*** regression
2. ***Non-linear relationships***: log and polynomial models
3. ***Interactions*** between variables

In each case, the way we estimate these coefficients does not change (i.e OLS).

Empirical applications:

(i) *class size* and *student performance*, (ii) *college tuition* and *earnings potential*, (iii) *wage*, *education* and *gender*

Standardized Regression

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. $\sigma(z) = \sqrt{\text{Var}(z)}$.

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. $\sigma(z) = \sqrt{\text{Var}(z)}$.

z^{stand} now has mean 0 and standard deviation 1, i.e. $\overline{z^{stand}} = 0$ and $\sigma(z^{stand}) = 1$

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. $\sigma(z) = \sqrt{\text{Var}(z)}$.

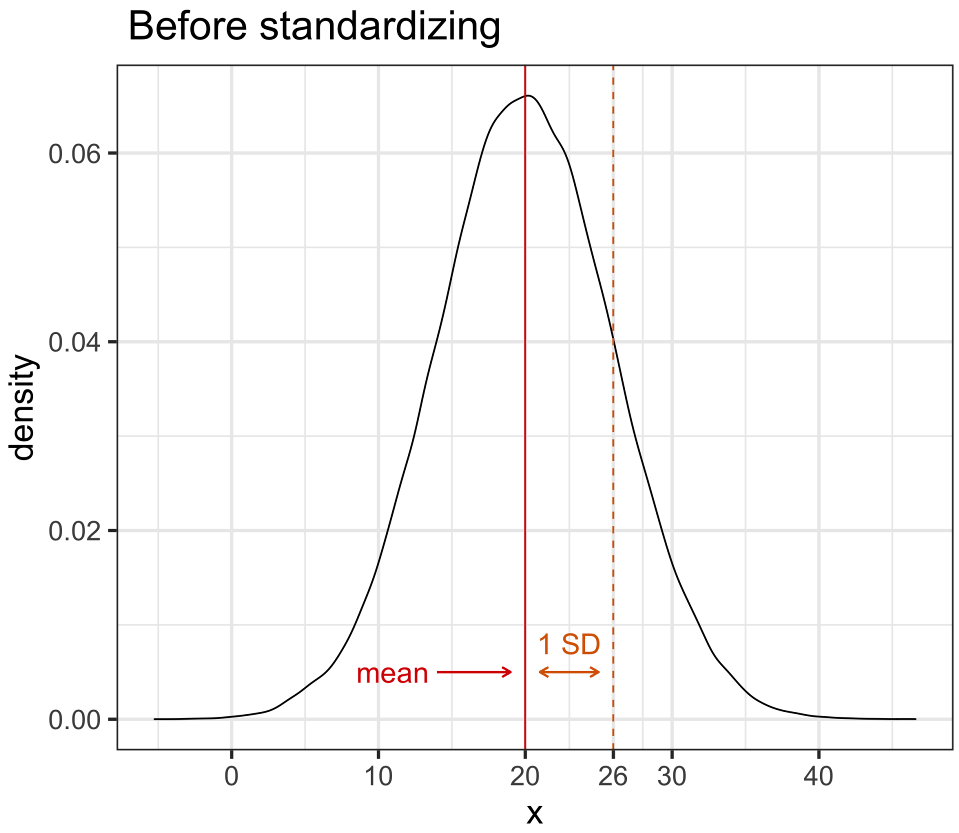
z^{stand} now has mean 0 and standard deviation 1, i.e. $\overline{z^{stand}} = 0$ and $\sigma(z^{stand}) = 1$

Intuitively, standardizing ***puts variables on the same scale*** so we can compare them.

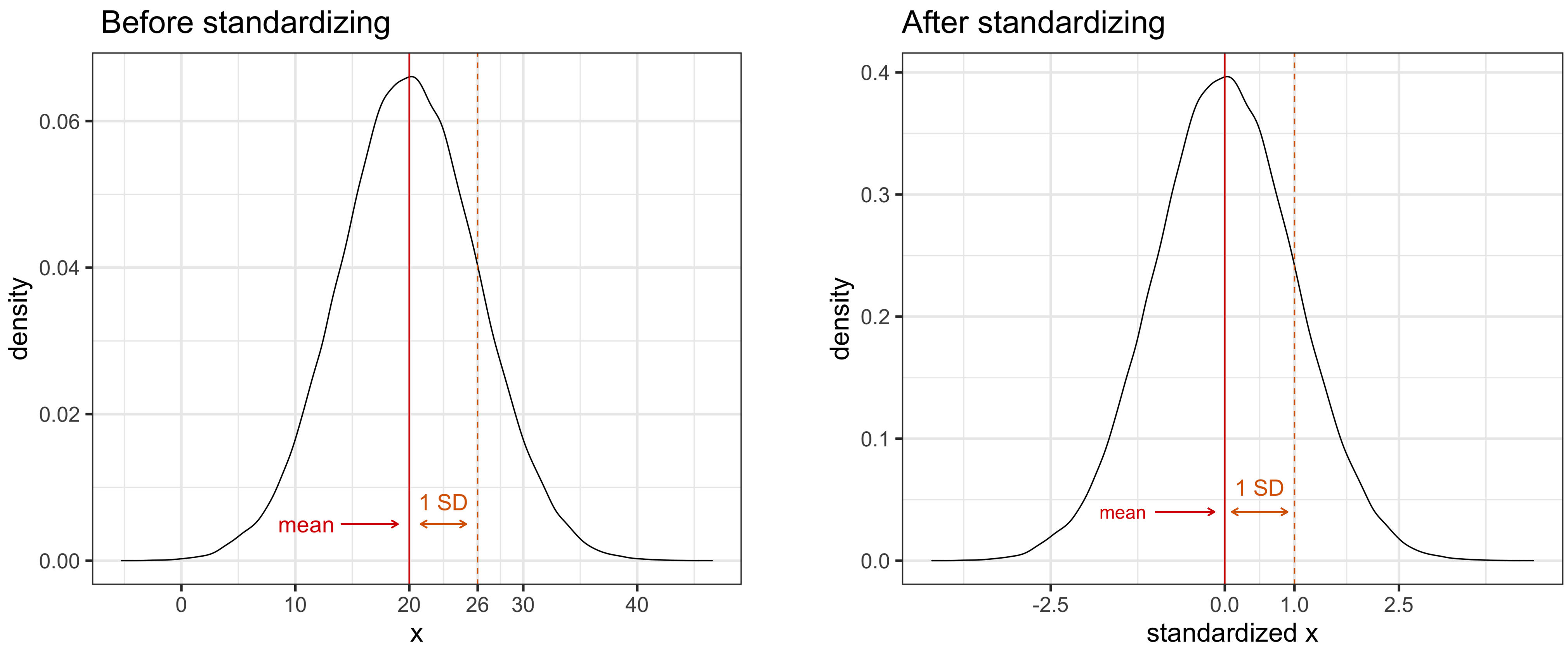
In our class size and student performance example, it will help to interpret:

- The **magnitude** of the effects,
- The **relative importance of each variable**.

Standardized Regression: Graphically



Standardized Regression: Graphically



Standardized Regression: Interpretation

If the ***dependent*** variable y is standardized, i.e. the model is $y^{stand} = b_0 + \sum_{k=1}^{K} b_k x_k + e$:

Standardized Regression: Interpretation

If the ***dependent*** variable y is standardized, i.e. the model is $y^{stand} = b_0 + \sum_{k=1}^{K} b_k x_k + e$:

- 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**.

Standardized Regression: Interpretation

If the ***dependent*** variable y is standardized, i.e. the model is $y^{stand} = b_0 + \sum_{k=1}^{K} b_k x_k + e$:

- 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, i.e. the model is $y = b_0 + \sum_{k=1}^{K} b_k x_k^{stand} + e$:

Standardized Regression: Interpretation

If the ***dependent*** variable y is standardized, i.e. the model is $y^{stand} = b_0 + \sum_{k=1}^{K} b_k x_k + e$:

- 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, i.e. the model is $y = b_0 + \sum_{k=1}^{K} b_k x_k^{stand} + e$:

- 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 1: Standardized regression

07:00

Let's go back our grades dataset. Remember that to load the data you need to use the `read_dta()` function from the `haven` package. These are the estimates we got from regressing average math test scores on the full set of regressors.

```
##       (Intercept)          classize      disadvantaged school_enrollment            female         religious
##       78.560725298       0.003320773      -0.389333008       0.000758258       0.923710499       2.876146701
```

1. Create a new variable `avgmath_stand` equal to the standardized math score. You can use the `scale()` function (combined with `mutate`) 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. Create the standardized variables for each *continuous* regressor as `<regressor>_stand`.

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

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

Non-Linear Relationships

Accounting for Non-Linear Relationships

There are two main "methods":

Accounting for Non-Linear Relationships

There are two main "methods":

1. ***Log*** models

Accounting for Non-Linear Relationships

There are two main "methods":

1. *Log* models
2. *Polynomial* models

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.

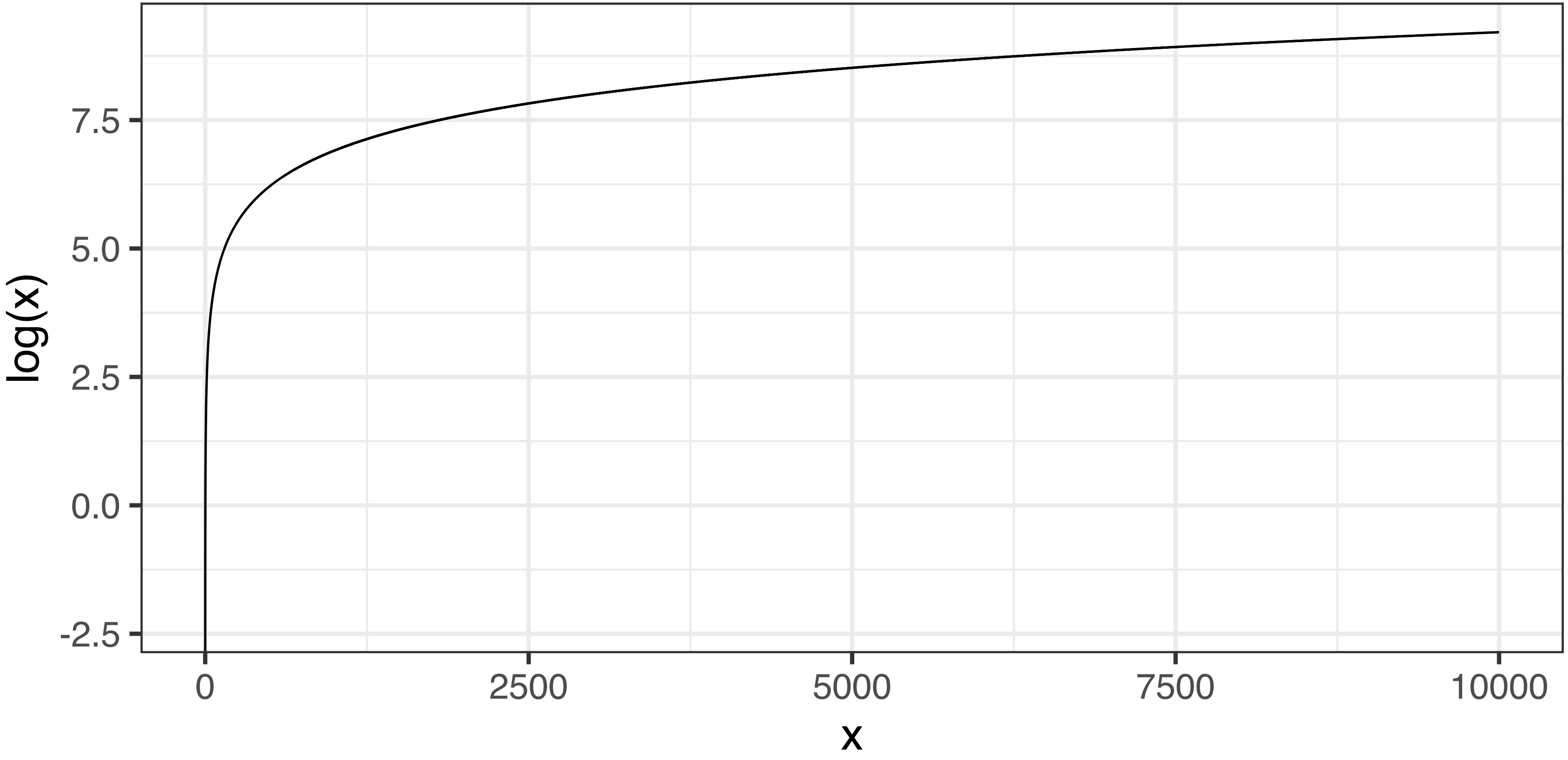
Log Models

- 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 *natural* 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***: $log(y_i) = b_0 + b_1 log(x_{1,i}) + \ldots + e_i$

The (natural) log Function: A Primer 😉



The (natural) log Function: A Primer 😉

The natural log function is the inverse function of the exponential function. , i.e.
$log(e^x) = x$

The (natural) log Function: A Primer 😉

The natural log function is the inverse function of the exponential function. , i.e. $log(e^x) = x$

$\rightarrow$ since for all x, $e^x > 0 \implies$ natural log function is only defined for ***strictly positive values***! (It is not defined in 0!)

The (natural) log Function: A Primer 😉

The natural log function is the inverse function of the exponential function. , i.e. $log(e^x) = x$

$\rightarrow$ since for all x, $e^x > 0 \implies$ natural log function is only defined for ***strictly positive values***! (It is not defined in 0!)

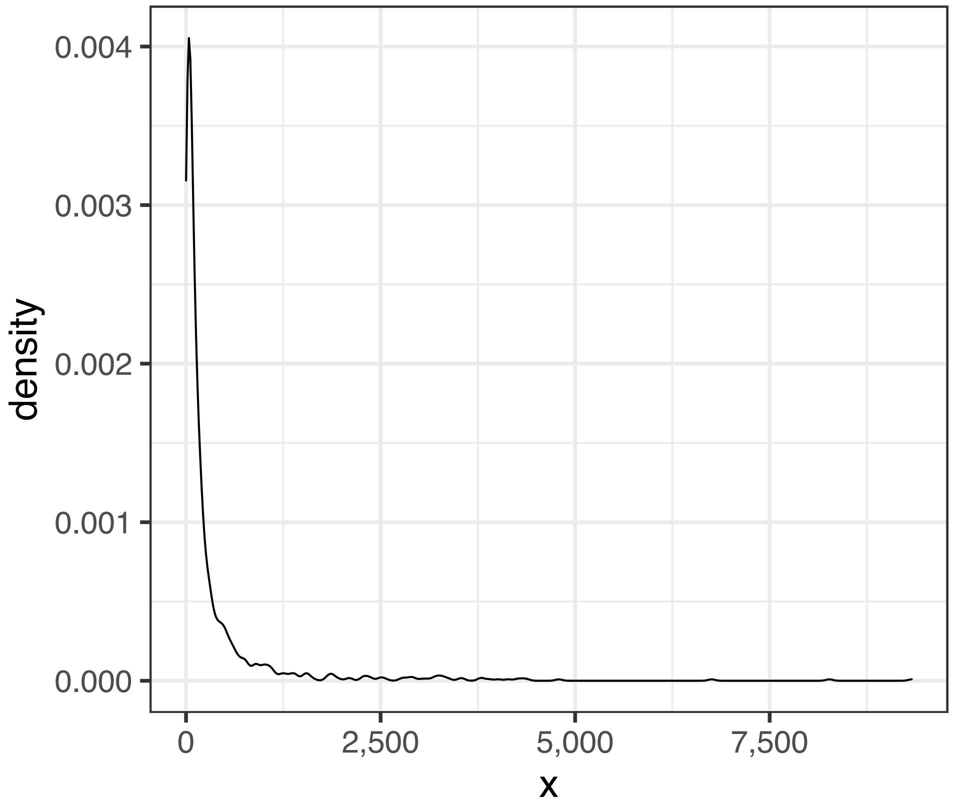
⚠️ You can only log your variables if they don't take 0 or negative values! Always think about this when taking the log of your dependent or independent variable(s)

The (natural) log Function: A Primer 😉

If you have very ***skewed distributions*** taking the log will render it more ***normally distributed***

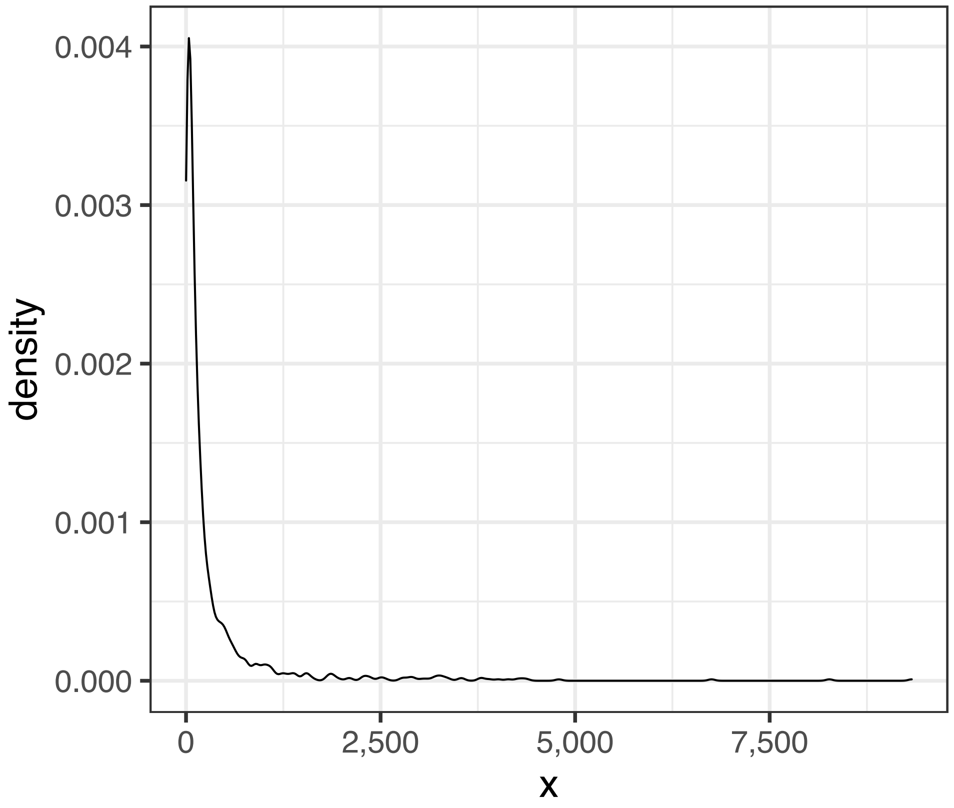
The (natural) log Function: A Primer 😉

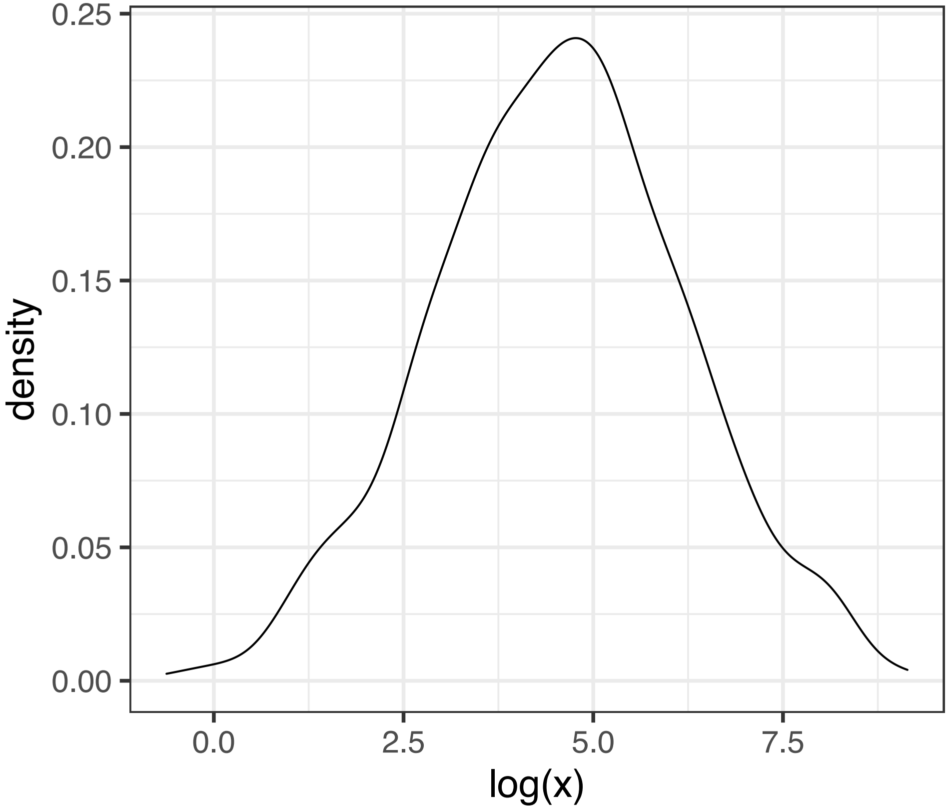
If you have very ***skewed distributions*** taking the log will render it more ***normally distributed***



The (natural) log Function: A Primer 😉

If you have very ***skewed distributions*** taking the log will render it more ***normally distributed***





Log Models: Simplified Interpretations

Specification	Model	Interpretation of b_1
Level - Level	$y = b_0 + b_1 x + e$	A **one unit** increase in x is associated, on average, with a b_1 **unit change** in y
Log - Level	$log(y) = b_0 + b_1 x + e$	A **one unit** increase in x is associated, on average, with a $b_1 \times 100$ **percent change** in y
Level - Log	$y = b_0 + b_1 log(x) + e$	A **one percent** increase in x is associated, on average, with a $b_1 / 100$ **unit change** in y
Log - Log	$log(y) = b_0 + b_1 log(x) + e$	A **one percent** increase in x is associated, on average, with a b_1 **percent change** in y

Log Models: Simplified Interpretations

Specification	Model	Interpretation of b_1
Level - Level	$y = b_0 + b_1 x + e$	A **one unit** increase in x is associated, on average, with a b_1 **unit change** in y
Log - Level	$log(y) = b_0 + b_1 x + e$	A **one unit** increase in x is associated, on average, with a $b_1 \times 100$ **percent change** in y
Level - Log	$y = b_0 + b_1 log(x) + e$	A **one percent** increase in x is associated, on average, with a $b_1 / 100$ **unit change** in y
Log - Log	$log(y) = b_0 + b_1 log(x) + e$	A **one percent** increase in x is associated, on average, with a b_1 **percent change** in y

- This may look like cooking recipes but of course it can be derived with some relatively simple maths.

Log Models: Simplified Interpretations

Specification	**Model**	**Interpretation of b_1**
Level - Level	$y = b_0 + b_1 x + e$	A **one unit** increase in x is associated, on average, with a b_1 **unit change** in y
Log - Level	$log(y) = b_0 + b_1 x + e$	A **one unit** increase in x is associated, on average, with a $b_1 \times 100$ **percent change** in y
Level - Log	$y = b_0 + b_1 log(x) + e$	A **one percent** increase in x is associated, on average, with a $b_1 / 100$ **unit change** in y
Log - Log	$log(y) = b_0 + b_1 log(x) + e$	A **one percent** increase in x is associated, on average, with a b_1 **percent change** in y

- This may look like cooking recipes but of course it can be derived with some relatively simple maths.
- ⚠️ these interpretations are only true for ***small*** changes in x and small/or b_1. What happens if we want to know the change in y for big changes in x or when b_1 is large?

Log Models: General Interpretations

For ***any increase in x, Δx, and any b_1*** $(\Delta x = 5\% = 0.05 \implies 1 + \Delta x = 1.05)$:

Specification	**Model**	**Interpretation of b_1**
Level - Level	$y = b_0 + b_1 x + e$	A **one unit** increase in x is associated, on average, with a b_1 **unit change** in y
Log - Level	$log(y) = b_0 + b_1 x + e$	A **one unit** increase in x is associated, on average, with a $(e^{b_1} - 1) \times 100$ **percent change** in y
Level - Log	$y = b_0 + b_1 log(x) + e$	A Δx **percent** increase in x is associated, on average, with a $b_1 \times log(1 + \Delta x)$ **unit change** in y
Log - Log	$log(y) = b_0 + b_1 log(x) + e$	A Δx **percent** increase in x is associated, on average, with a $((1 + \Delta x)^{b_1} - 1) \times 100$ **percent change** in y

Log Models: Approximations

Why are the approximations shown previously true?

Log Models: Approximations

Why are the approximations shown previously true?

Log-Level

General interpretation: A **one unit** increase in x is associated, on average, with a $(e^{b_1} - 1) \times 100$ **percent change** in y.

Simplified interpretation: A **one unit** increase in x is associated, on average, with a $b_1 \times 100$ **percent change** in y.

Log Models: Approximations

Why are the approximations shown previously true?

Log-Level

General interpretation: A **one unit** increase in x is associated, on average, with a $(e^{b_1} - 1) \times 100$ **percent change** in y.

Simplified interpretation: A **one unit** increase in x is associated, on average, with a $b_1 \times 100$ **percent change** in y.

This is because, for small b_1, $e^{b_1} \approx 1 + b_1 \iff b_1 \approx e^{b_1} - 1$

Log Models: Approximations

Why are the approximations shown previously true?

Log-Level

General interpretation: A **one unit** increase in x is associated, on average, with a $(e^{b_1} - 1) \times 100$ **percent change** in y.

Simplified interpretation: A **one unit** increase in x is associated, on average, with a $b_1 \times 100$ **percent change** in y.

This is because, for small b_1, $e^{b_1} \approx 1 + b_1 \iff b_1 \approx e^{b_1} - 1$

$\rightarrow$ for $b_1 = 0.04$, $e^{b_1} - 1 = e^{0.04} - 1 = 0.0408$

Log Models: Approximations

Why are the approximations shown previously true?

Log-Level

General interpretation: A **one unit** increase in x is associated, on average, with a $(e^{b_1} - 1) \times 100$ **percent change** in y.

Simplified interpretation: A **one unit** increase in x is associated, on average, with a $b_1 \times 100$ **percent change** in y.

This is because, for small b_1, $e^{b_1} \approx 1 + b_1 \iff b_1 \approx e^{b_1} - 1$

$\rightarrow$ for $b_1 = 0.04$, $e^{b_1} - 1 = e^{0.04} - 1 = 0.0408$

$\rightarrow$ for $b_1 = 0.5$, $e^{b_1} - 1 = e^{0.5} - 1 = 0.6487$

Log Models: Approximations

Why are the approximations shown previously true?

Level-Log

General interpretation: A Δx **percent** increase in x is associated, on average, with a $b_1 \times log(1 + \Delta x)$ **unit change** in y.

Simplified interpretation: A **one percent** increase in x is associated, on average, with a $b_1/100$ **unit change** in y.

Log Models: Approximations

Why are the approximations shown previously true?

Level-Log

General interpretation: A Δx **percent** increase in x is associated, on average, with a $b_1 \times log(1 + \Delta x)$ **unit change** in y.

Simplified interpretation: A **one percent** increase in x is associated, on average, with a $b_1/100$ **unit change** in y.

This is because for small Δx, $log(1 + \Delta x) \approx \Delta x$

Log Models: Approximations

Why are the approximations shown previously true?

Level-Log

General interpretation: A Δx **percent** increase in x is associated, on average, with a $b_1 \times log(1 + \Delta x)$ **unit change** in y.

Simplified interpretation: A **one percent** increase in x is associated, on average, with a $b_1/100$ **unit change** in y.

This is because for small Δx, $log(1 + \Delta x) \approx \Delta x$

$\rightarrow$ for $\Delta x = 1\% = 0.01$, $log(1 + \Delta x) = log(1.01) = 0.01$ (hence the $/100$ in the simplified interpretation)

Log Models: Approximations

Why are the approximations shown previously true?

Level-Log

General interpretation: A Δx **percent** increase in x is associated, on average, with a $b_1 \times log(1 + \Delta x)$ **unit change** in y.

Simplified interpretation: A **one percent** increase in x is associated, on average, with a $b_1/100$ **unit change** in y.

This is because for small Δx, $log(1 + \Delta x) \approx \Delta x$

$\rightarrow$ for $\Delta x = 1\% = 0.01$, $log(1 + \Delta x) = log(1.01) = 0.01$ (hence the $/100$ in the simplified interpretation)

$\rightarrow$ for $\Delta x = 20\% = 0.20$, $log(1 + \Delta x) = log(1.20) = 0.18$

Log Models: Approximations

Why are the approximations shown previously true?

Log-Log

General interpretation: A Δx **percent** increase in x is associated, on average, with a $((1 + \Delta x)^{b_1} - 1) \times 100$ **percent change** in y.

Simplified interpretation: A **one percent** increase in x is associated, on average, with a b_1 **percent change** in y.

Log Models: Approximations

Why are the approximations shown previously true?

Log-Log

General interpretation: A Δx **percent** increase in x is associated, on average, with a $((1+\Delta x)^{b_1} - 1) \times 100$ **percent change** in y.

Simplified interpretation: A **one percent** increase in x is associated, on average, with a b_1 **percent change** in y.

This is because for small $|b_1| \times \Delta x$,
$(1+\Delta x)^{b_1} \approx 1 + b_1 \times \Delta x \iff b_1 \times \Delta x \times 100 \approx ((1+\Delta x)^{b_1} - 1) \times 100$

Log Models: Approximations

Why are the approximations shown previously true?

Log-Log

General interpretation: A Δx **percent** increase in x is associated, on average, with a $((1+\Delta x)^{b_1}-1)\times 100$ **percent change** in y.

Simplified interpretation: A **one percent** increase in x is associated, on average, with a b_1 **percent change** in y.

This is because for small $|b_1| \times \Delta x$,
$(1+\Delta x)^{b_1} \approx 1 + b_1 \times \Delta x \iff b_1 \times \Delta x \times 100 \approx ((1+\Delta x)^{b_1}-1)\times 100$

$\rightarrow$ for $\Delta x = 1\% = 0.01$ and $b_1 = 0.5$,
$((1+\Delta x)^{b_1}-1)\times 100 = (1.01^{0.5}-1)\times 100 = 0.5$

Log Models: Approximations

Why are the approximations shown previously true?

Log-Log

General interpretation: A **Δx percent** increase in x is associated, on average, with a $((1+\Delta x)^{b_1}-1)\times 100$ **percent change** in y.

Simplified interpretation: A **one percent** increase in x is associated, on average, with a b_1 **percent change** in y.

This is because for small $|b_1|\times\Delta x$,
$(1+\Delta x)^{b_1}\approx 1+b_1\times\Delta x \iff b_1\times\Delta x\times 100\approx((1+\Delta x)^{b_1}-1)\times 100$

$\rightarrow$ for $\Delta x = 1\% = 0.01$ and $b_1 = 0.5$,
$((1+\Delta x)^{b_1}-1)\times 100=(1.01^{0.5}-1)\times 100=0.5$

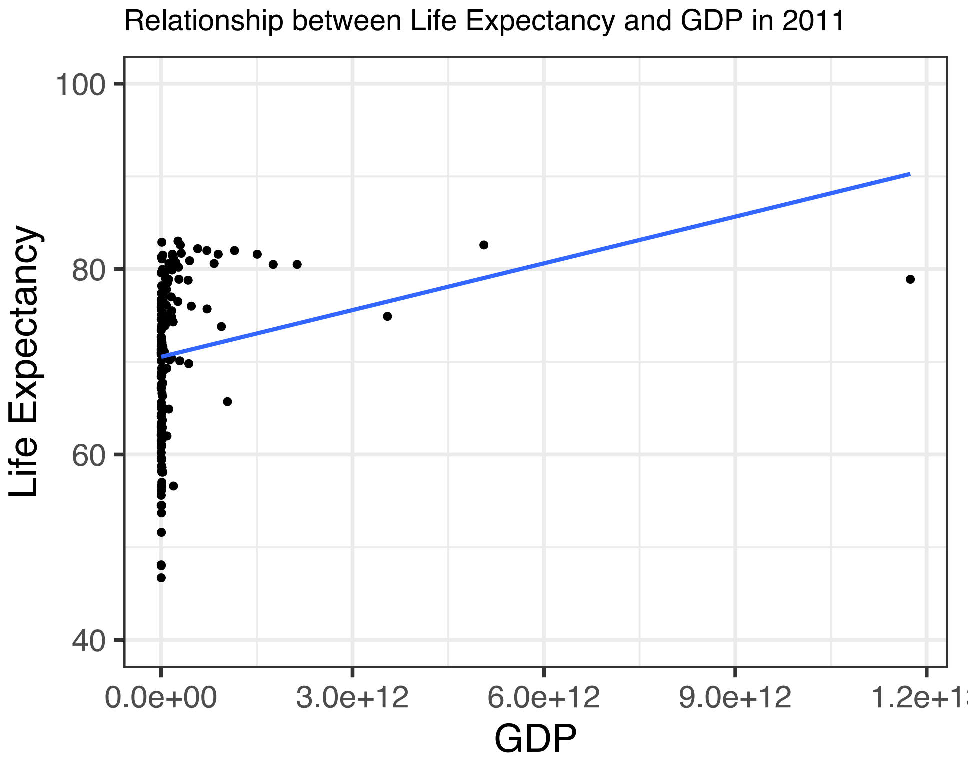
$\rightarrow$ for $\Delta x = 10\% = 0.10$ and $b_1 = 10$,
$((1+\Delta x)^{b_1}-1)\times 100=(1.1^{10}-1)\times 100=159.37$

When Should You Use log Models?

1. If the relationship betwen x and y looks like a log or exponential function.

When Should You Use log Models?

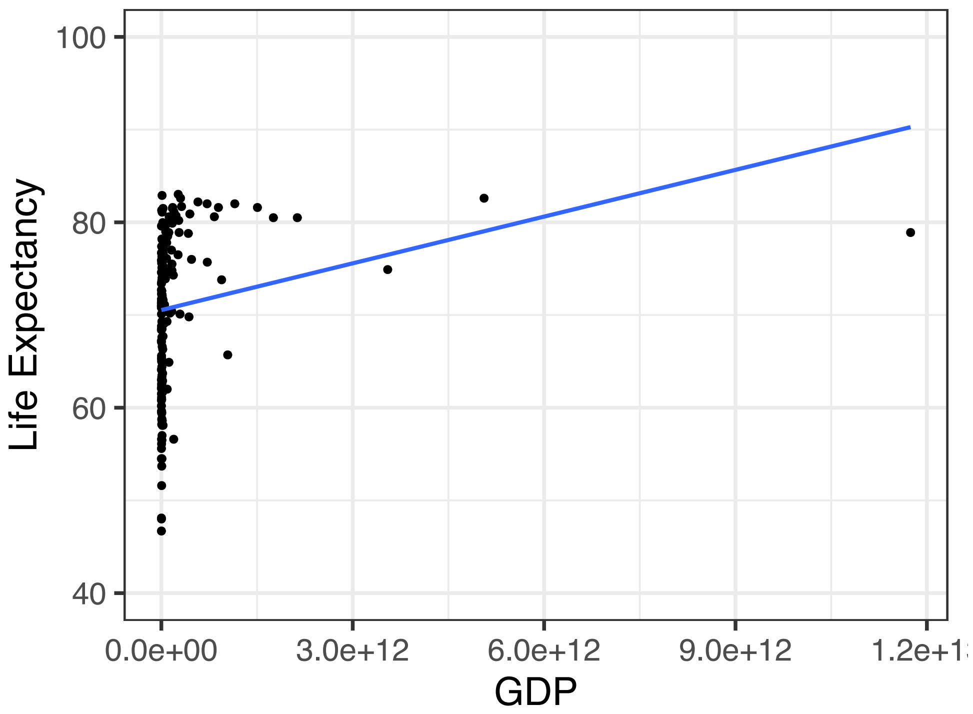
1. If the relationship betwen x and y looks like a log or exponential function.



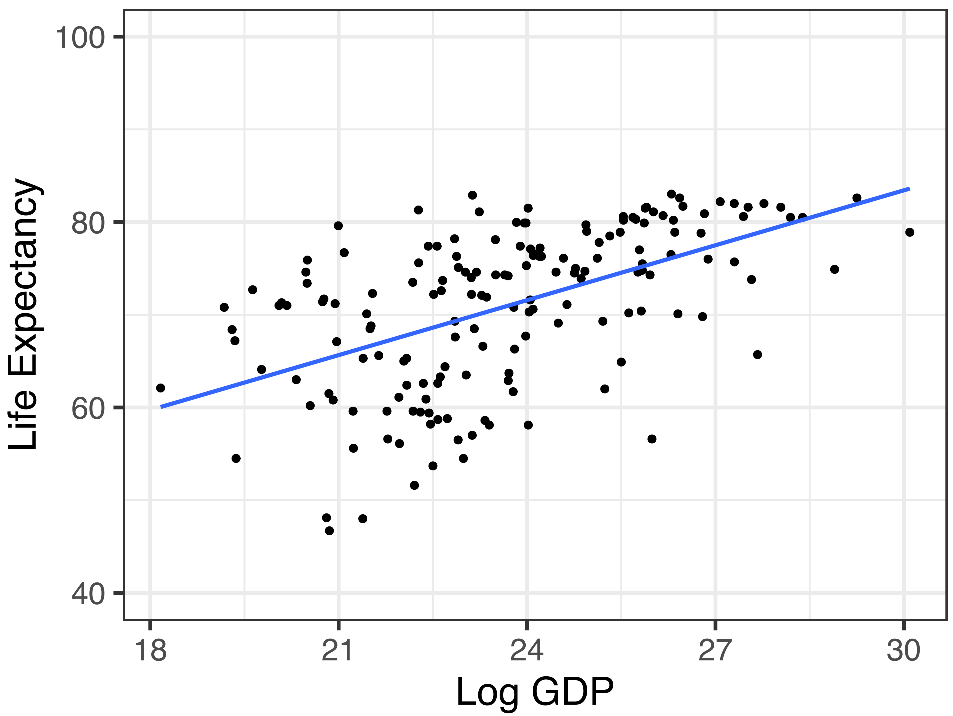
Data from gapminder data in dslabs package.

When Should You Use log Models?

1. If the relationship betwen x and y looks like a log or exponential function.



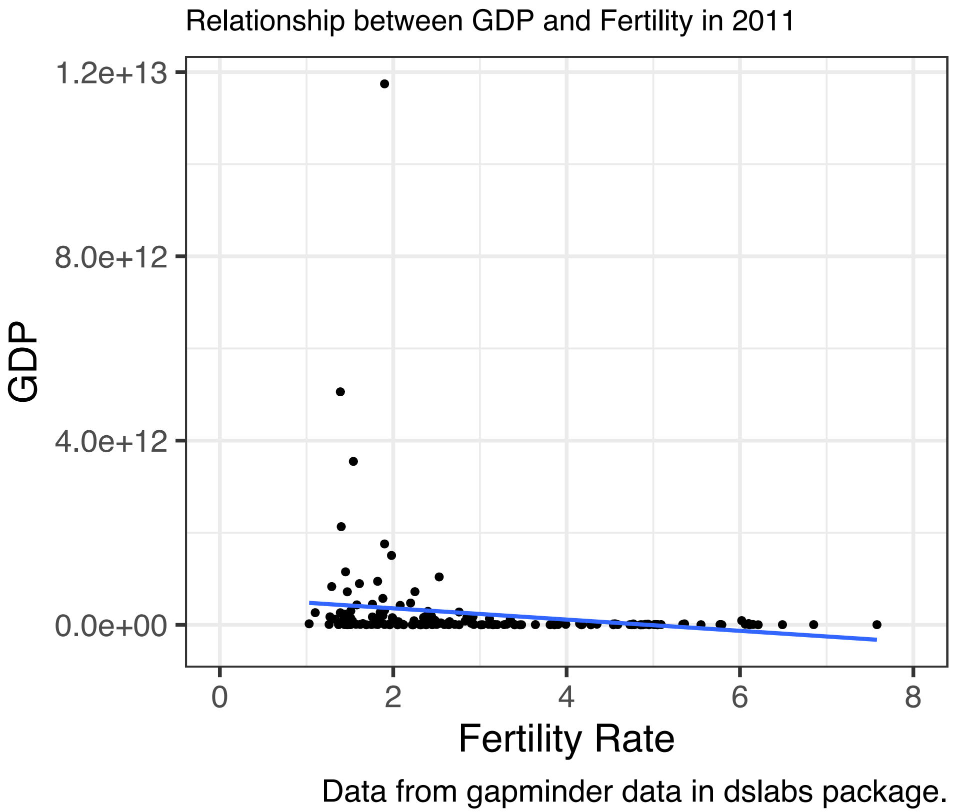
Data from gapminder data in dslabs package.



Data from gapminder data in dslabs package.

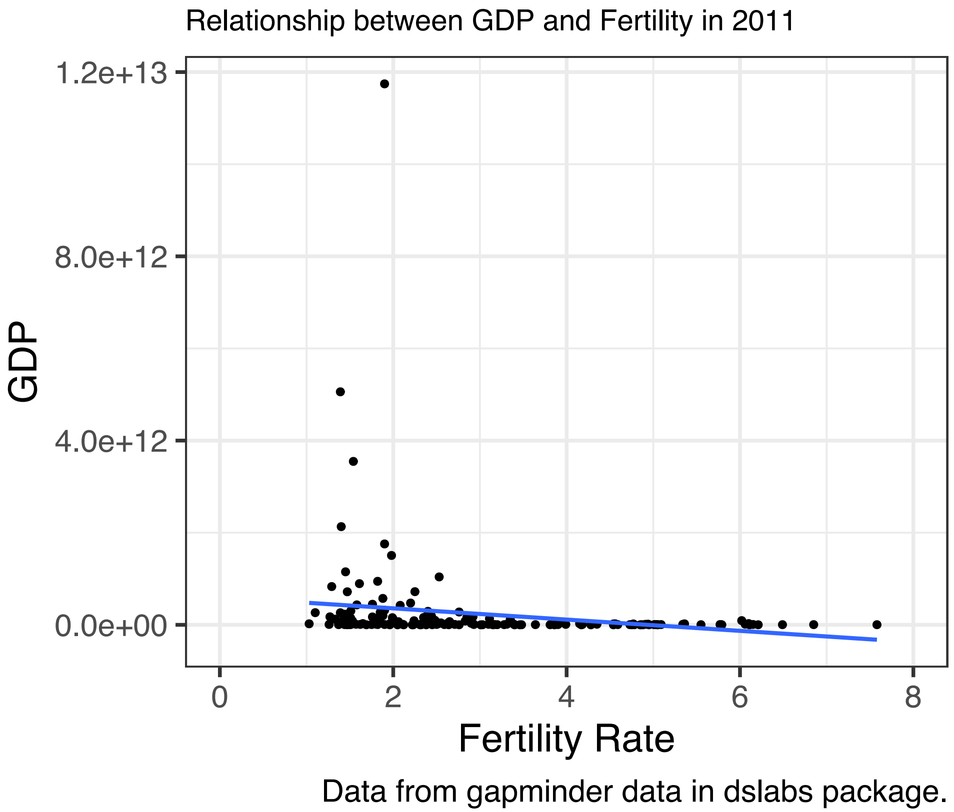
When Should You Use log Models?

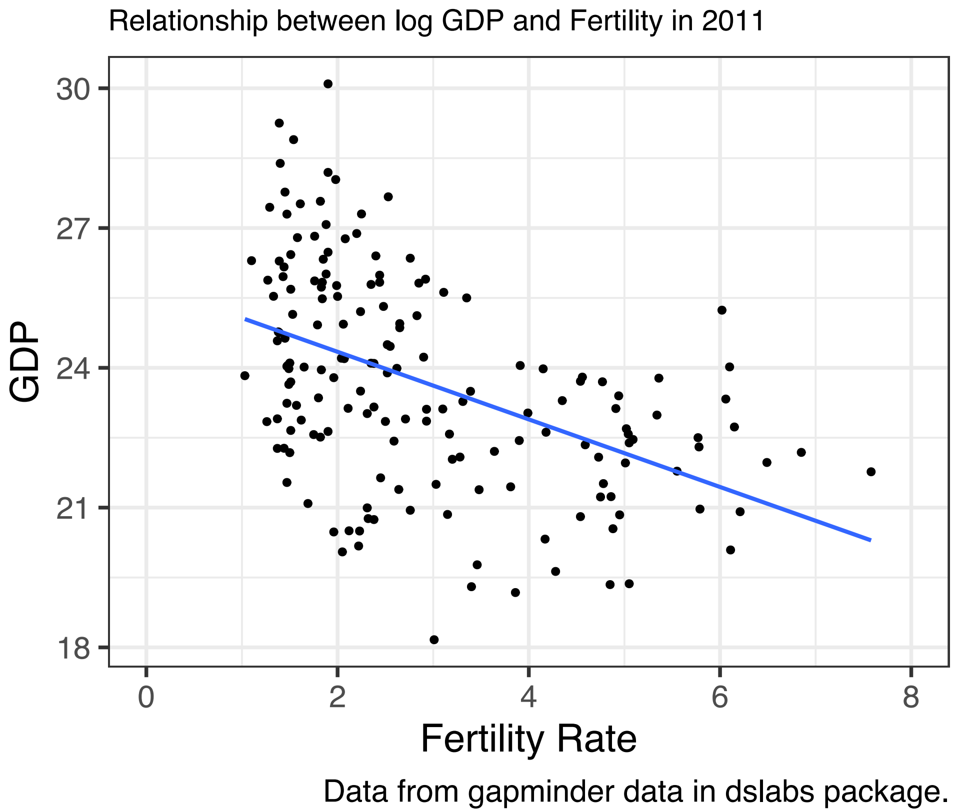
1. If the relationship betwen x and y looks like a log or exponential function.



When Should You Use log Models?

1. If the relationship betwen x and y looks like a log or exponential function.





When Should You Use log Models?

1. If the relationship betwen x and y looks like a log or exponential function.

2. To easily interpret coefficients as ***elasticities*** which play a central role in economic theory.

Elasticity of y with respect to x: percent change in y following a 1% increase in x.

Accounting for Other Types Non-Linear Relationships

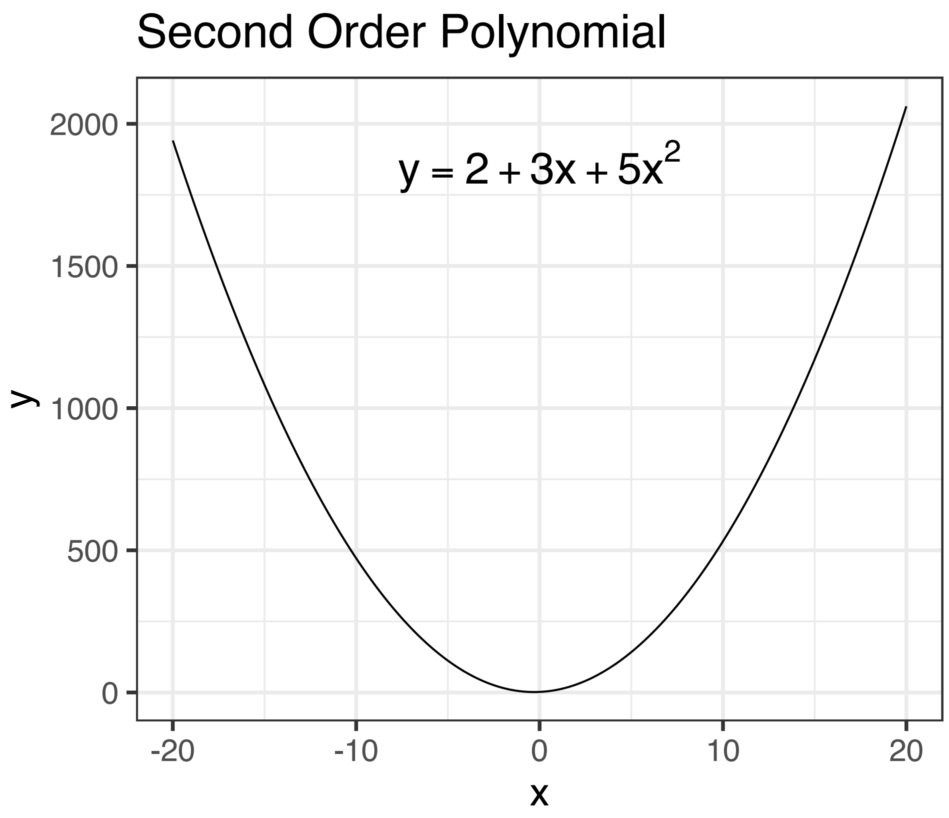
What if the relationship between x and y is not exponential/log?

Accounting for Other Types Non-Linear Relationships

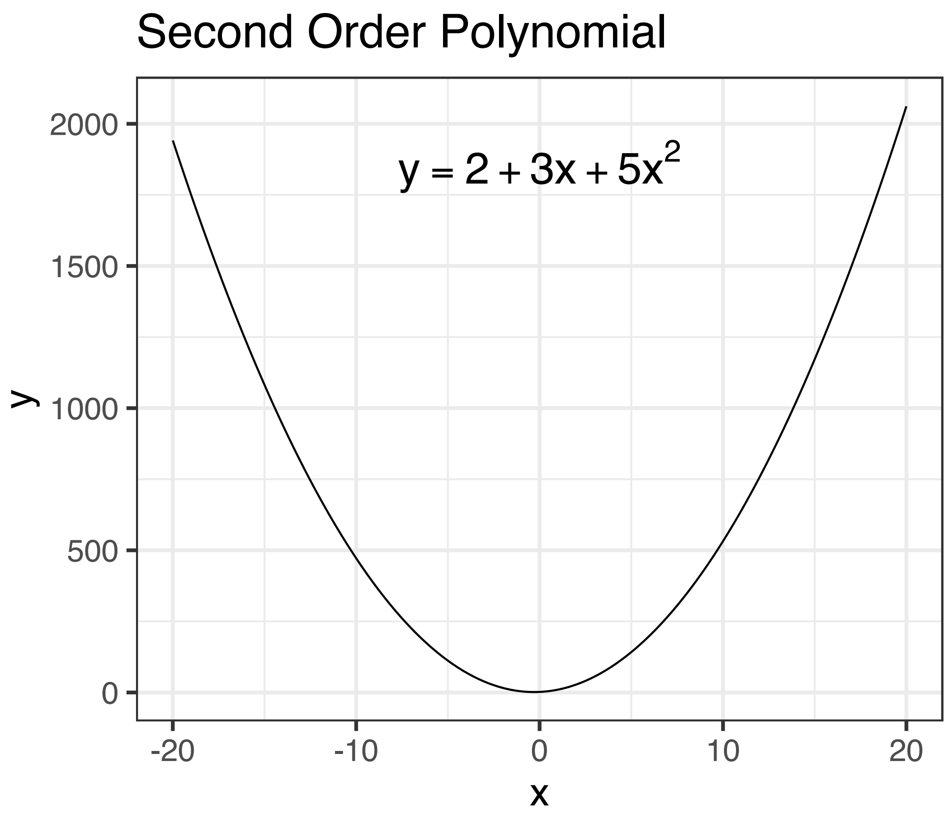
What if the relationship between x and y is not exponential/log?

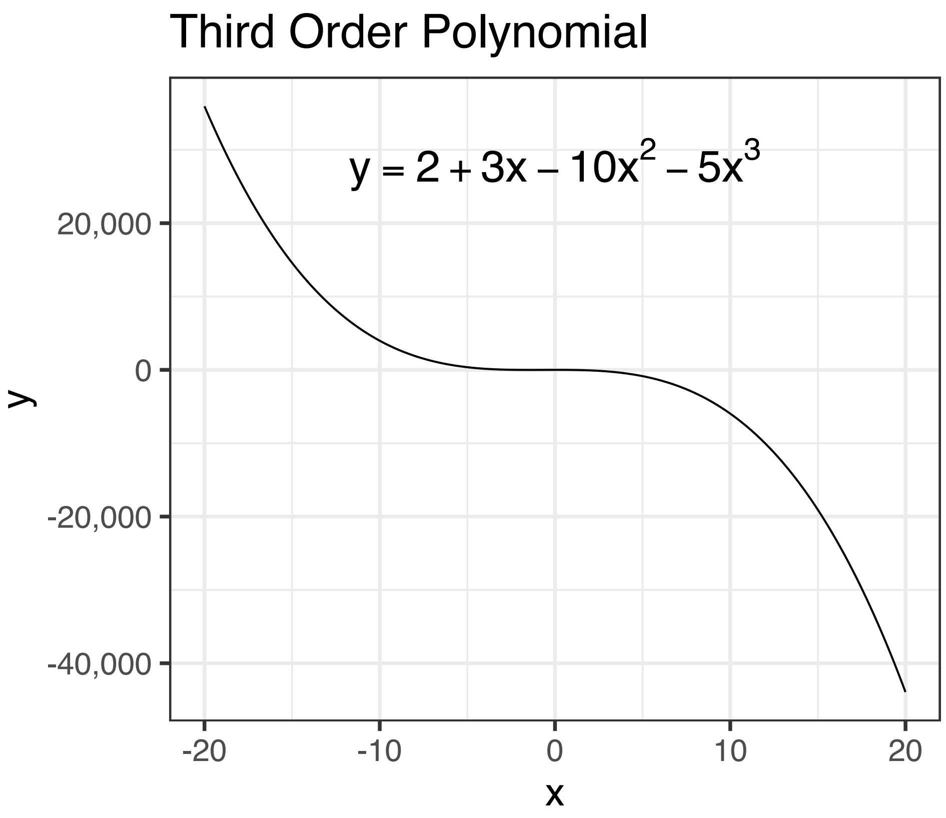
$\rightarrow$ ***polynomial*** regressions: just take a polynomial function of the regressor!

Polynomial Wut? 😟



Polynomial Wut? 😟





Polynomial Regressions

What does this mean in practice?

Polynomial Regressions

What does this mean in practice?

$\rightarrow$ add a higher order of the regressor to the regression, depending on the visual (or expected) relationship

Polynomial Regressions

What does this mean in practice?

$\longrightarrow$ add a higher order of the regressor to the regression, depending on the visual (or expected) relationship

Several ways of doing this in `R`:

```
lm(y ~ x + I(x^2) + I(x^3), data)
```

Polynomial Regressions

What does this mean in practice?

$\longrightarrow$ add a higher order of the regressor to the regression, depending on the visual (or expected) relationship

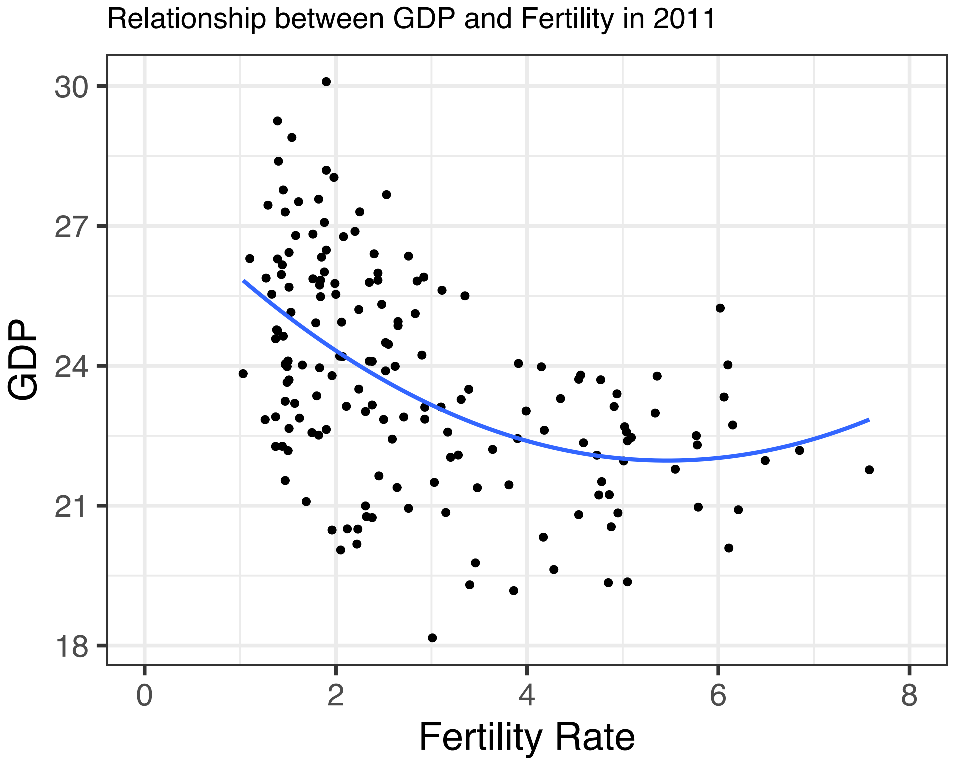
Several ways of doing this in `R`:

```
lm(y ~ x + I(x^2) + I(x^3), data)
```

```
lm(y ~ poly(x, 3, raw = TRUE), data)
```

Polynomial Regressions

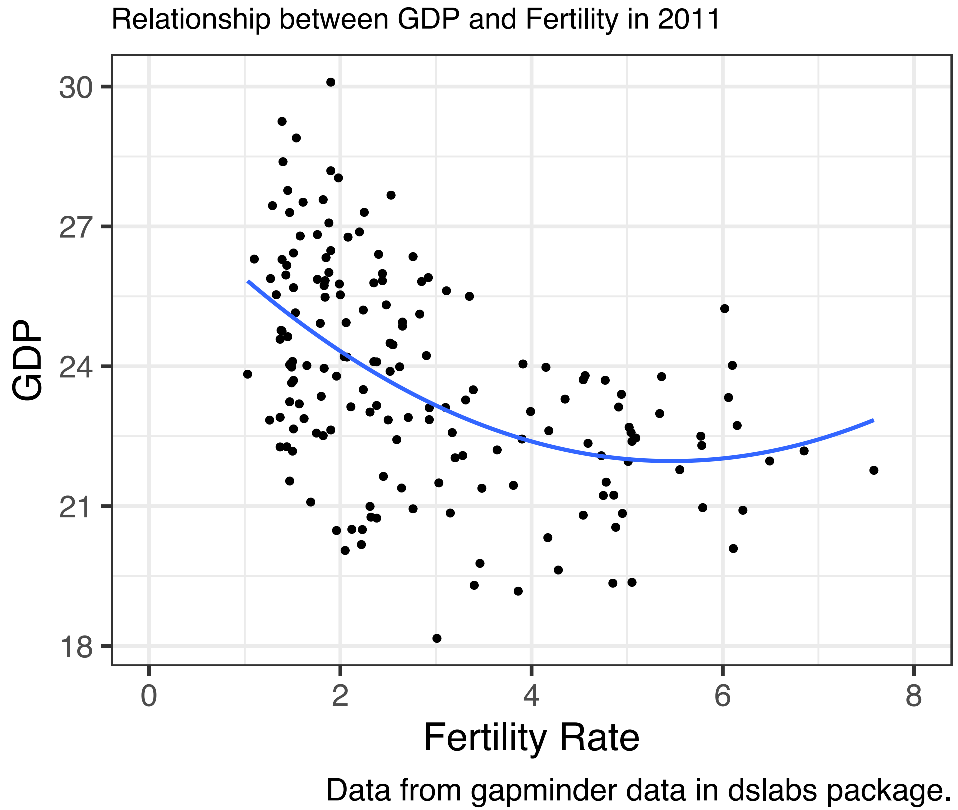
2nd order:



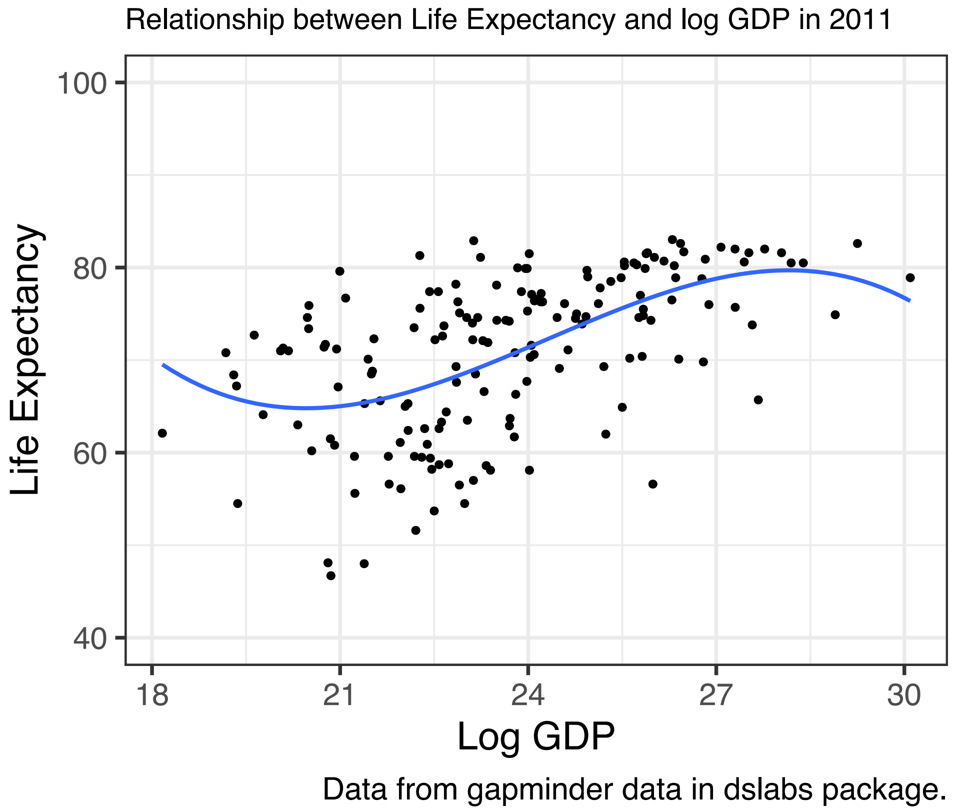
Data from gapminder data in dslabs package.

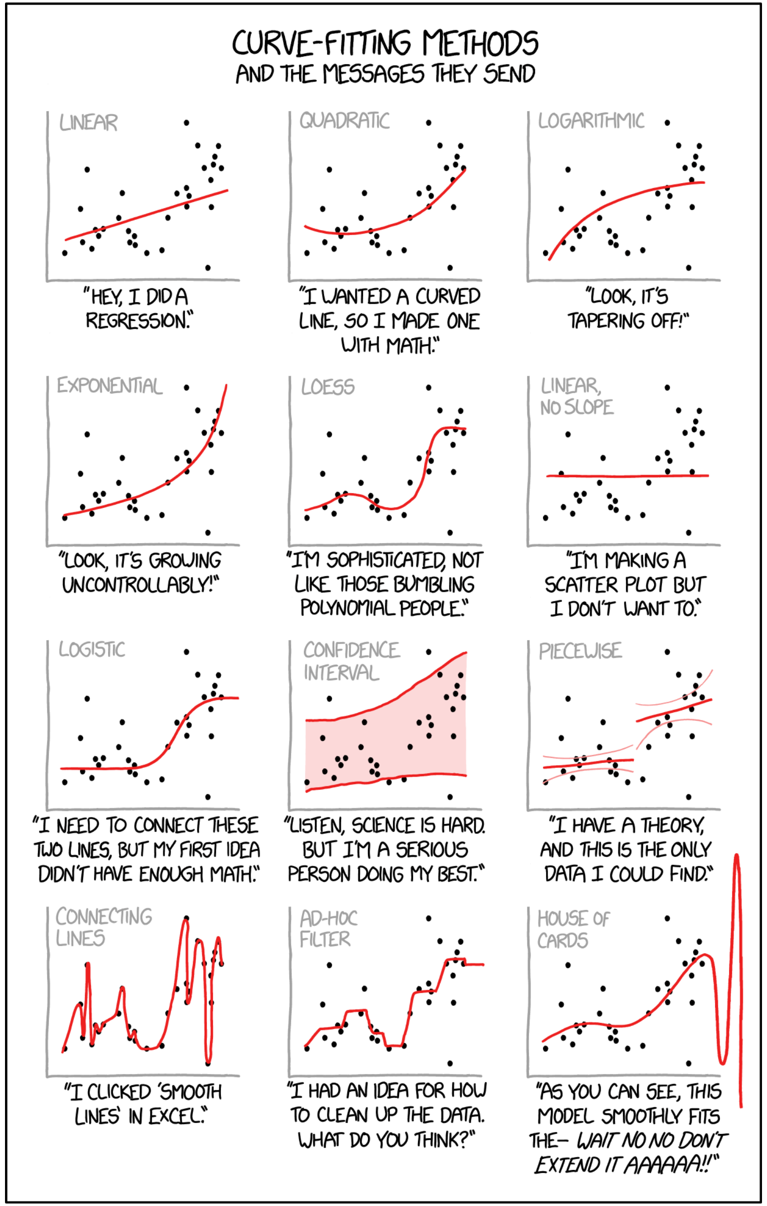
Polynomial Regressions

2nd order:



3rd order:




CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Task 2: Non-linear relationships

10:00

1. Load the data here. This dataset contains information about tuition and estimated incomes of graduates for universities in the US. More details can be found here.

2. Create a scatter plot of estimated mid career pay (`mid_career_pay`) $(y - axis)$ as a function of out of state tuition (`out_of_state_tuition`) $(x - axis)$. Would you say the relationship is broadly linear or rather non-linear? Use `geom_smooth(method = "lm", se = F) + geom_smooth(method = "lm", se = F, formula = y ~ poly(x, 2, raw= T))` to fit both a linear and 2nd order regression line. This time which seems most appropriate?

3. Create a variable equal to out of state tuition divided by 1000. Regress mid career pay on out of state tuition divided by 1000. Interpret the coefficient.

4. Regress mid career pay on out of state tuition divided by 1000 and its square. *Hint:* you can use either `poly(x, 2, raw = T)` or `x + I(x^2)`, where x is your regressor. What does the positive sign on the squared term imply?

Interaction Terms

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   smallTRUE:experience
##             534.1919                15.8906                 1.3305                -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  smallTRUE:experience
##             534.1919               15.8906                1.3305               -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)
```

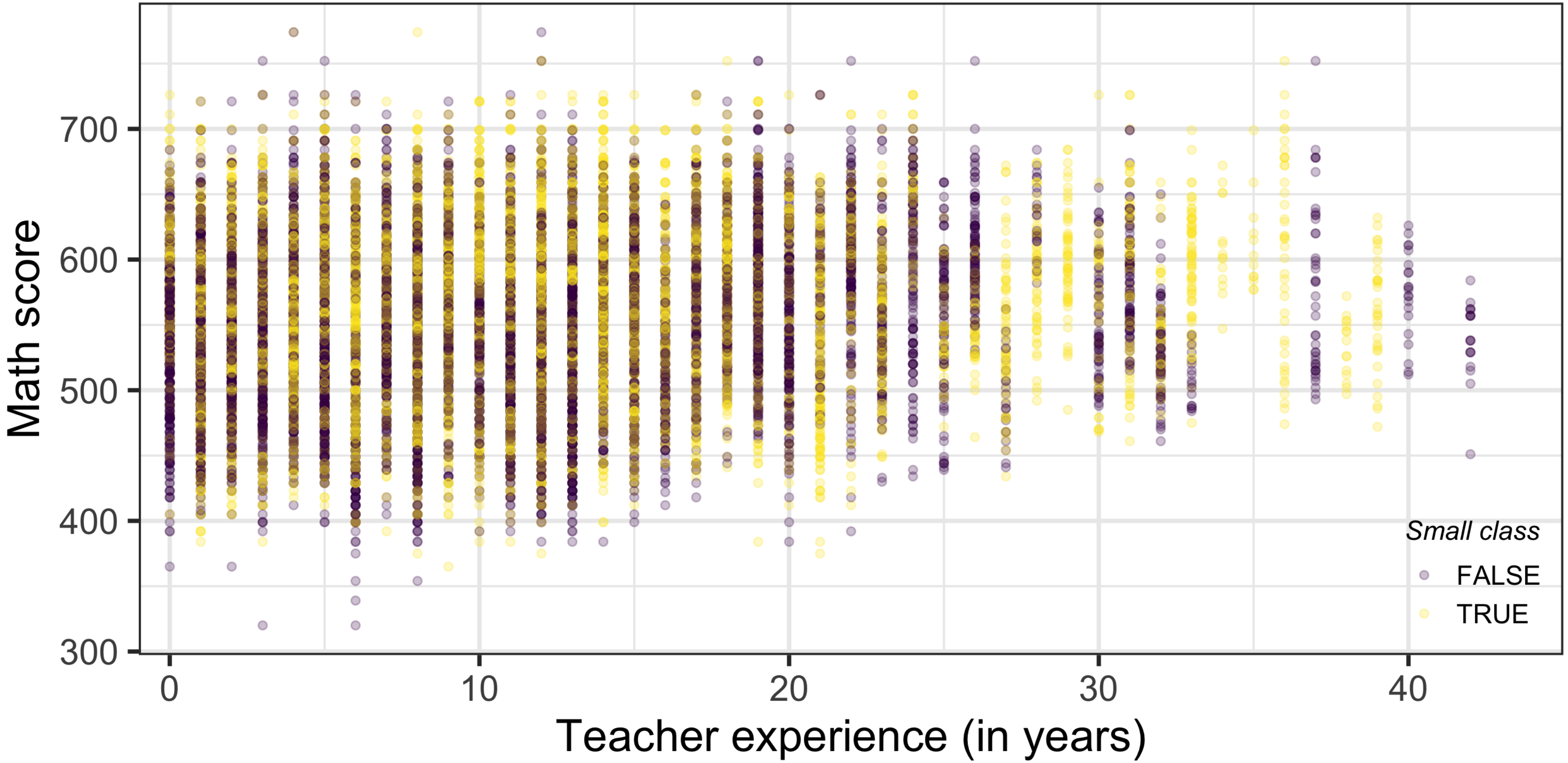
```
##
## Call:
## lm(formula = math ~ small + experience + small * experience,
##     data = star_df)
##
## Coefficients:
##          (Intercept)                smallTRUE               experience  smallTRUE:experience
##             534.1919                  15.8906                   1.3305        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

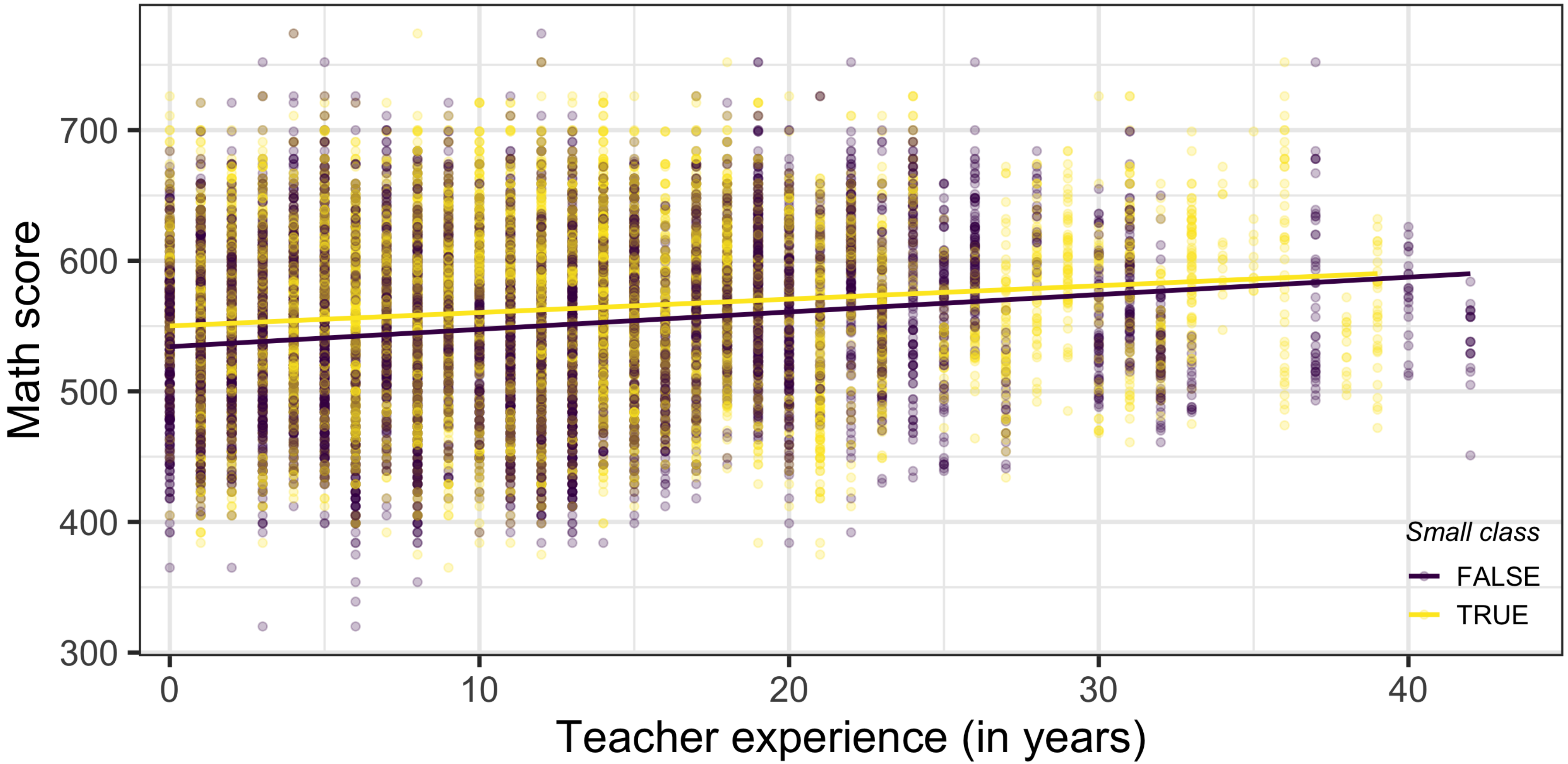
Interacting Regressors: Visually

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



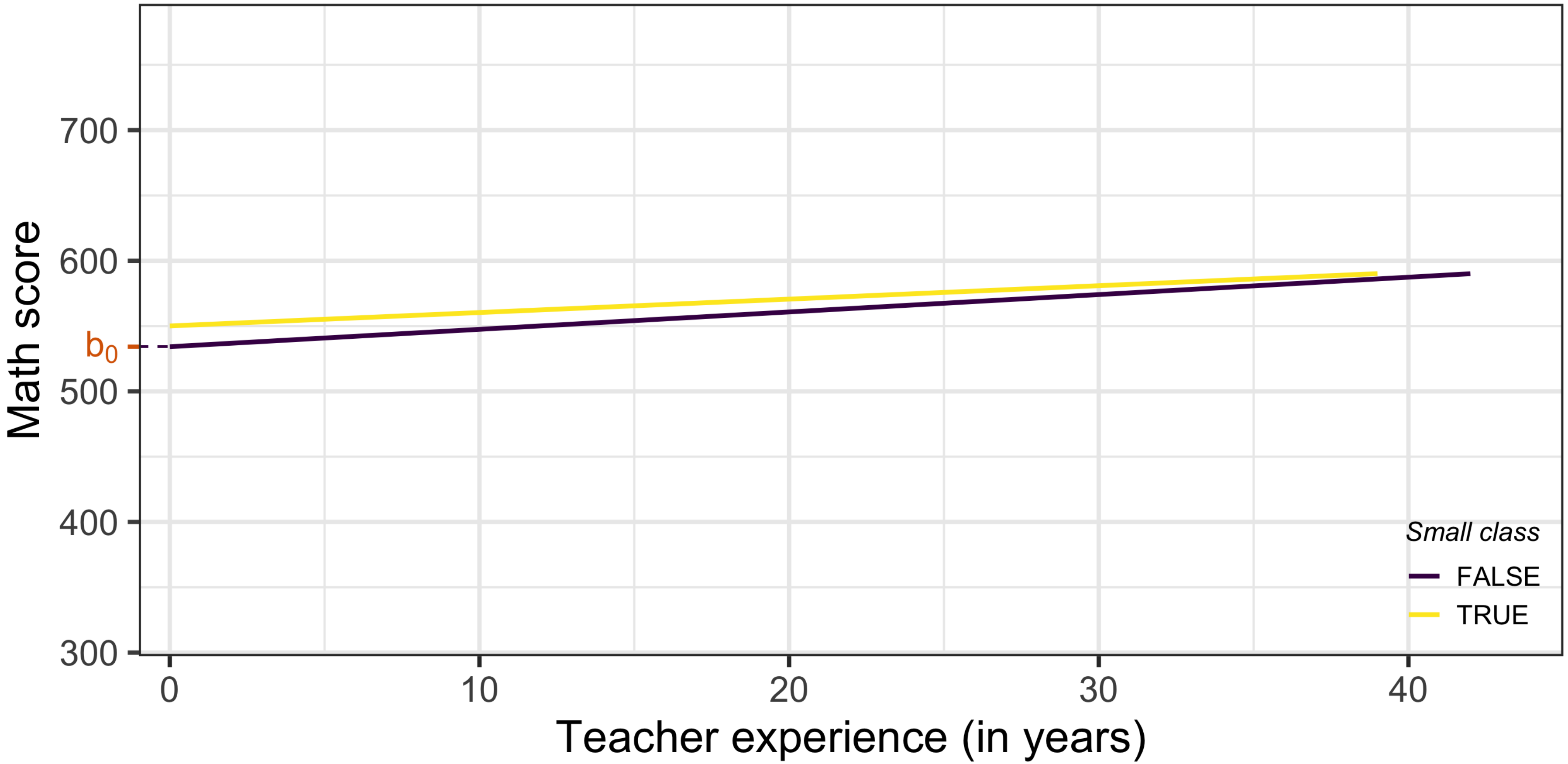
Interacting Regressors: Visually

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



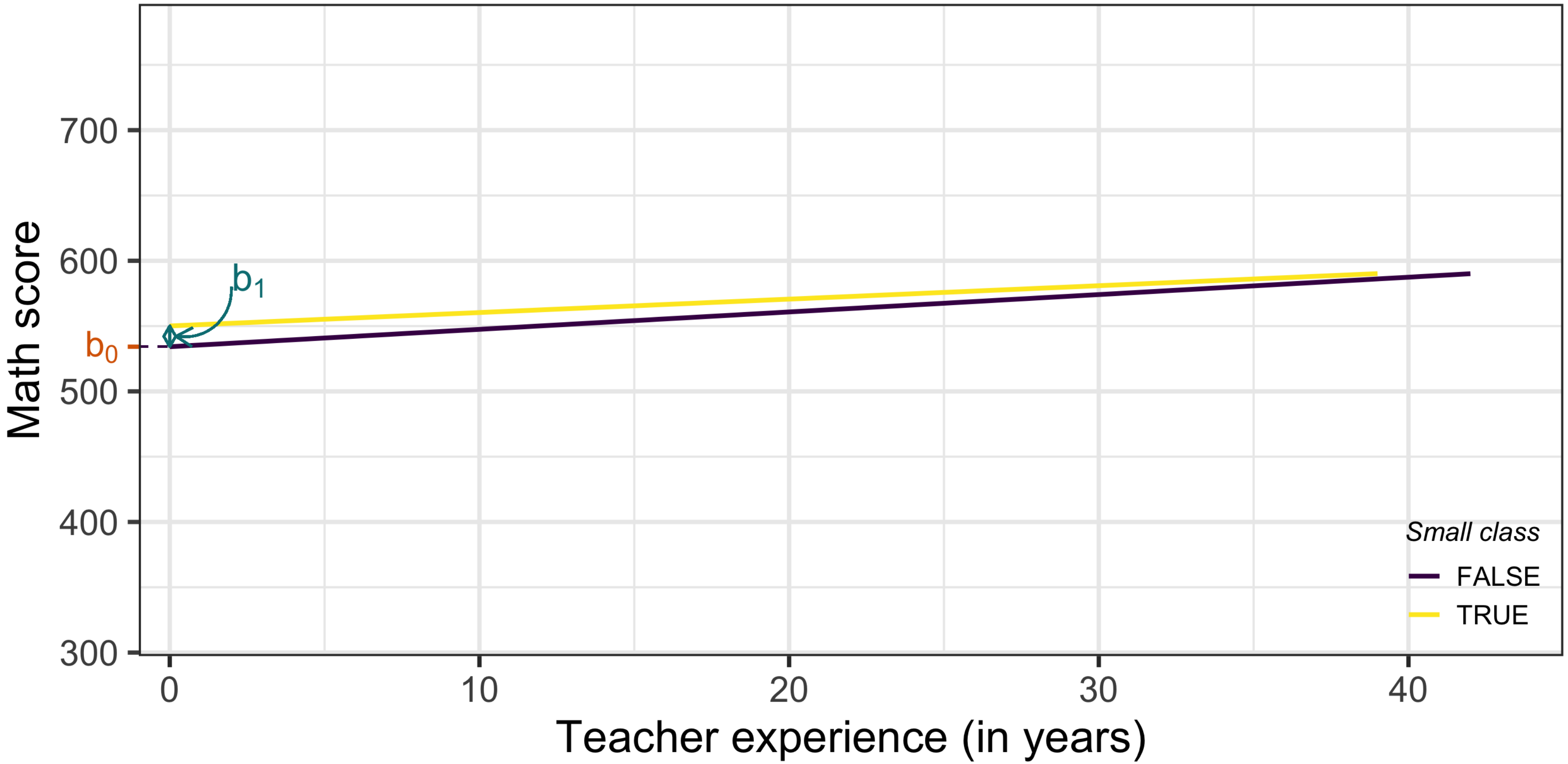
Interacting Regressors: Visually

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



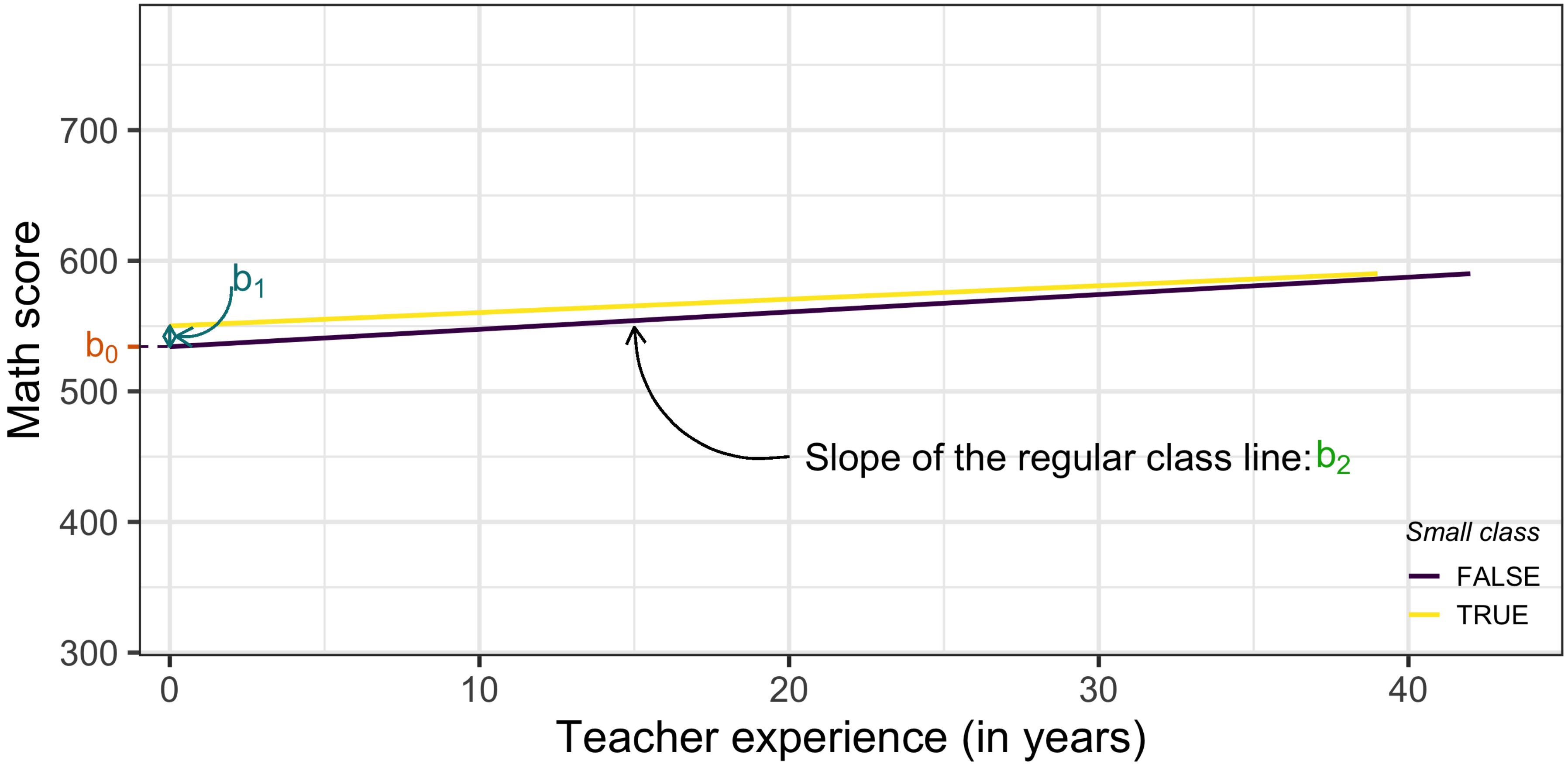
Interacting Regressors: Visually

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



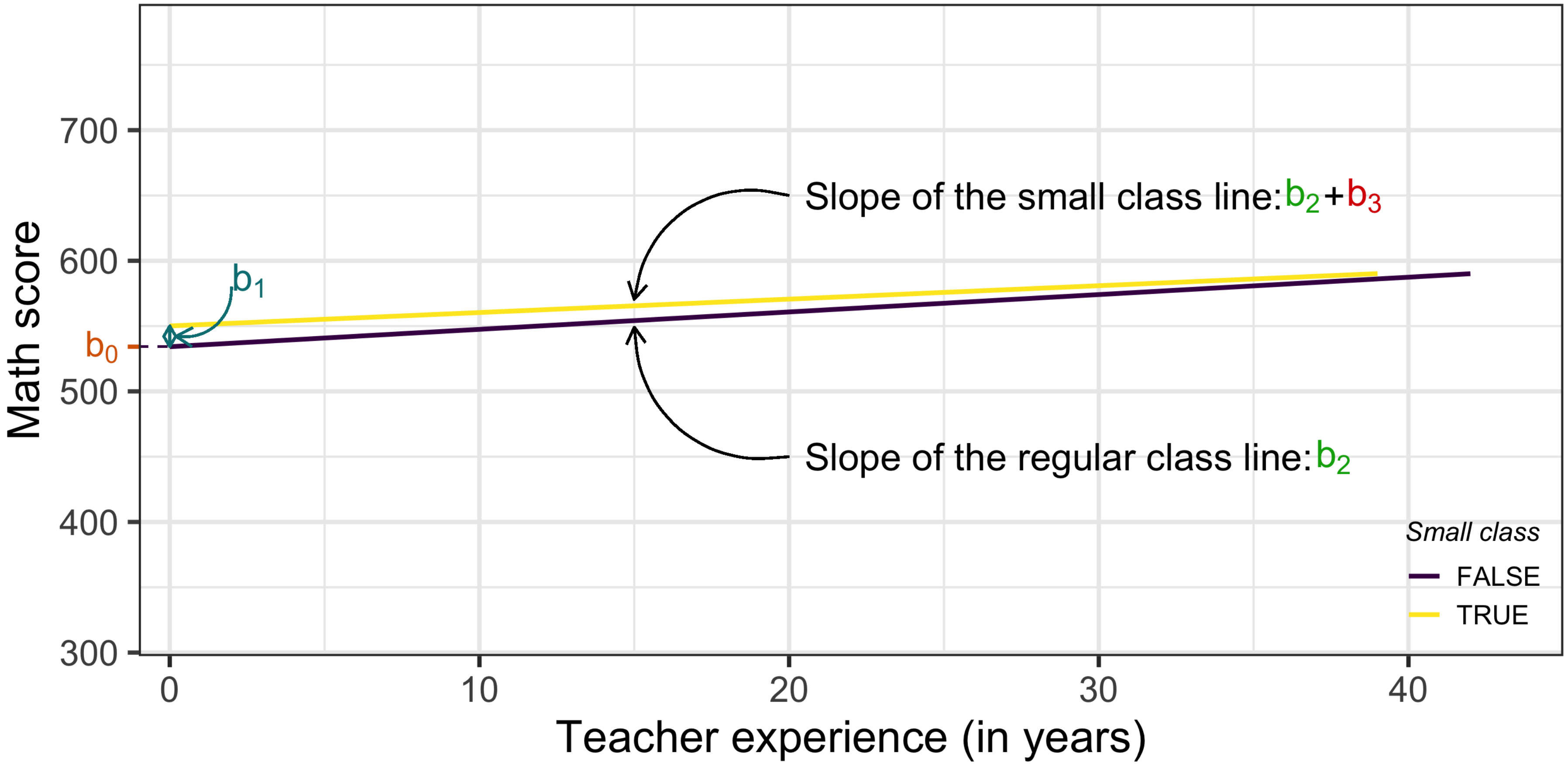
Interacting Regressors: Visually

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



Interacting Regressors: Visually

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



Task 3: Wages, education and gender in 1985

10:00

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. Create a plot showing this interaction. (*Hint:* use the `color = gender` argument in `aes` and `geom_smooth(method = "lm", se = F)` to obtain a regression line per gender.)

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 randomness?

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 randomness?
- We will answer 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!

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