Data Science for Everyone

Week 12: Correlation & Regression Concept Review

Angela Lai

April 24, 2020

New York University

Outline

- Logistics
- Concept review
- Demo

Logistics

- Homework 3/4 due at 8 p.m. ET on Monday, April 27
- Project due at 8 p.m. ET on Monday, May 4

Logistics

- Labs 9 and 10 are **optional**!
- Your two lowest lab grades out of 11 will be dropped, meaning that these can only help your grade if you choose to complete them
- OPTIONAL Lab 9 out, due at 8 p.m. ET on Wednesday, April 29

Project Tips

Common questions:

- How should I format my project report?
- What should I be careful about?

Project Tips

- Clearly indicate every question you answer
- Best practices: create a Jupyter Notebook like the homework assignments. Fill that out with your code and analyses and, in the end, print to PDF. Make a cell (or cells) for every question.
- Be very careful about how you interpret the regression coefficients, p-values, and confidence intervals

Review

Mean: the average of a set of numbers.

The **standard deviation** roughly measures how far, on average, numbers are from their mean.

We can use these to convert our data into standard units.

Regression Review

Linear regression: a linear approach to modeling the relationship between a dependent variable and one or more independent variables.

The **correlation coefficient** measures linear association between two variables. We write it as r.

Regression Review

Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Formula for standard deviation:

$$s_x = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

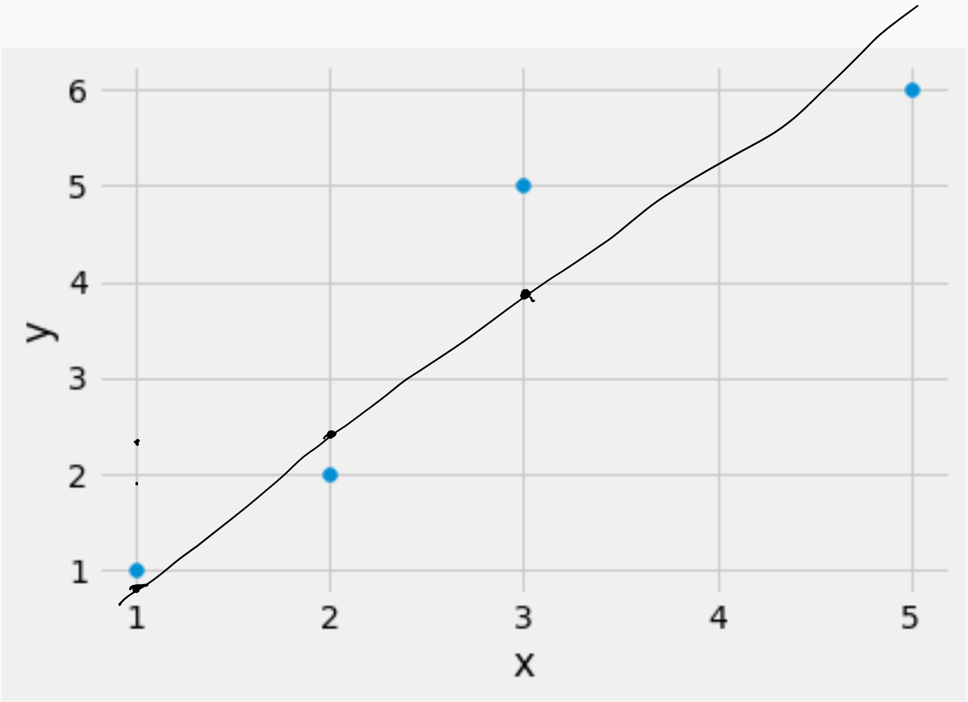
(Note that some formulas have $n - 1$ on the bottom instead of n).

Correlation coefficient:

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Regression Review

We want to find a linear model for the relationship between x and y. That is, we want to find slope b and intercept a for $y = bx + a$.



Regression Review

Recall:

$$\text{intercept} = a = \bar{y} - b\bar{x}$$

$$\text{slope} = b = r \cdot \left(\frac{s_y}{s_x}\right)$$

Our data: {(1,1), (2,2), (3,5), (5, 6)}

Let's work this out by hand (mostly).

x	y
1	1
2	2
3	5
5	6

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n=4} x_i = \frac{1+2+3+5}{4} = 2.75$$

$$\bar{y} = \frac{1+2+5+6}{4} = 3.5$$

$$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}} = \sqrt{\frac{1}{4}\left((1-2.75)^2 + (2-2.75)^2 + (3-2.75)^2 + (5-2.75)^2\right)}$$

$$= 1.479$$

$$s_y = \sqrt{\frac{1}{4}\left((1-3.5)^2 + (2-3.5)^2 + (5-3.5)^2 - (6-3.5)^2\right)}$$

$$= 2.062$$

$$r_{xy} = r_{yx} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$= \frac{1}{n \cdot s_x \cdot s_y}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{4 \cdot 1.479 \cdot 2.062}\left[(1-2.75)(1-3.5) + \ldots\right.$$

$$\approx 0.943$$

$y = \underline{b}x + \underline{a}$

$$b = r\left(\frac{s_y}{s_x}\right) = 0.943\left(\frac{2.062}{1.479}\right) \approx 1.315$$

$$a = \bar{y} - b\bar{x} = 3.5 - 1.315 \times 2.75 \approx -.116$$

Regression Review

What do we get after working this out by hand?

$$\bar{x} = 2.75, \bar{y} = 3.5$$

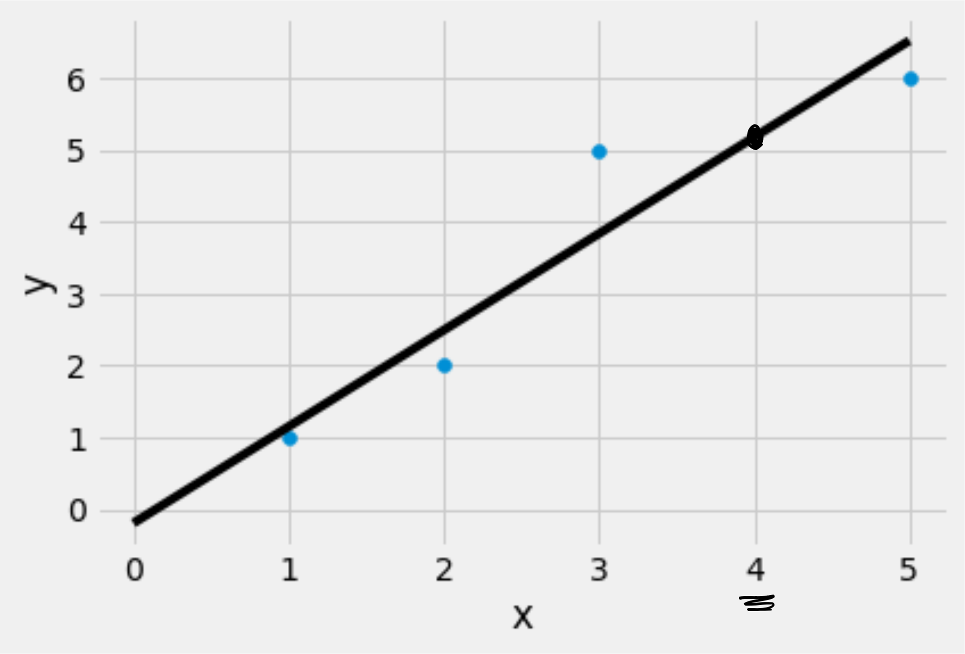
$$s_x \approx 1.479, s_y \approx 2.062$$

$$r_{xy} \approx 0.943, R^2 \approx 0.889$$

$$b = 1.315, a = -0.115$$

Let's plot the linear model we get with this a and b!

Regression Review



Demo

Let’s see how we can do this with code!