

ECON 3818

Chapter 5

Kyle Butts

29 November 2021

Chapter 5: Regression

Review from Last Class

Recall the ways we discussed relationships between two random variables X and Y :

Covariance, σ_{XY} (sample analogue: s_{XY})

- Direction matters, but magnitude is hard to interpret

Correlation, ρ_{XY} (sample analogue: r_{XY})

- Direction and magnitude matter
- Correlation is always value between $[-1, 1]$

Review from Last Class

Recall:

$$r = \frac{Cov(X, Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}$$

- Correlation is a function of covariance, just normalizes the magnitudes so we can interpret.

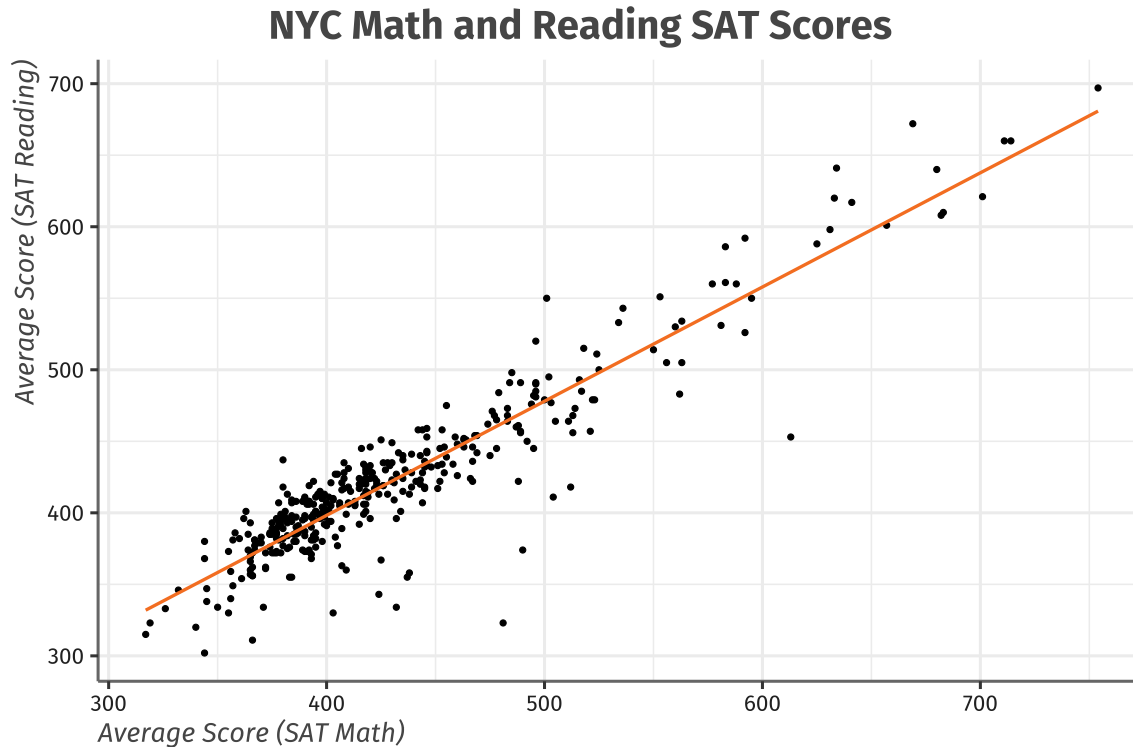
Clicker Question

Suppose you calculate the sample covariance, $s_{XY} = 1.2$, and the sample standard deviations $s_X = 2$ and $s_Y = 2.5$. What is the sample correlation, r_{XY} ?

- a. 0.0576
- b. 0.24
- c. 0.048
- d. 4.17

Relationship between X and Y

We often summarize the relationship between X and Y using a straight line:



This is called the line of best fit, or the **regression line**.

Regression Line Example

In the previous example, the regression line is given by

$$\text{Average } \widehat{\text{SAT Reading}} = 78.87 + 0.7983 * \text{Average SAT Math}$$

The $\hat{}$ symbol means that we are *predicting* average SAT reading score.

Example: If a school has an average SAT math score of 600, we would predict their SAT reading score would be

$$\text{Average } \widehat{\text{SAT Reading}} = 78.87 + 0.7983 * 600 = 557.85.$$

Regression Line

Let X and Y be two random variables. A **regression line** is a straight line that describes how the response variable, Y , changes as the explanatory variable X changes.

We often use a regression line to predict the value of Y for a given value of X , *when we believe the relationship between X and Y is linear.*

Since we are predicting Y with a line, our regression line will be given by:

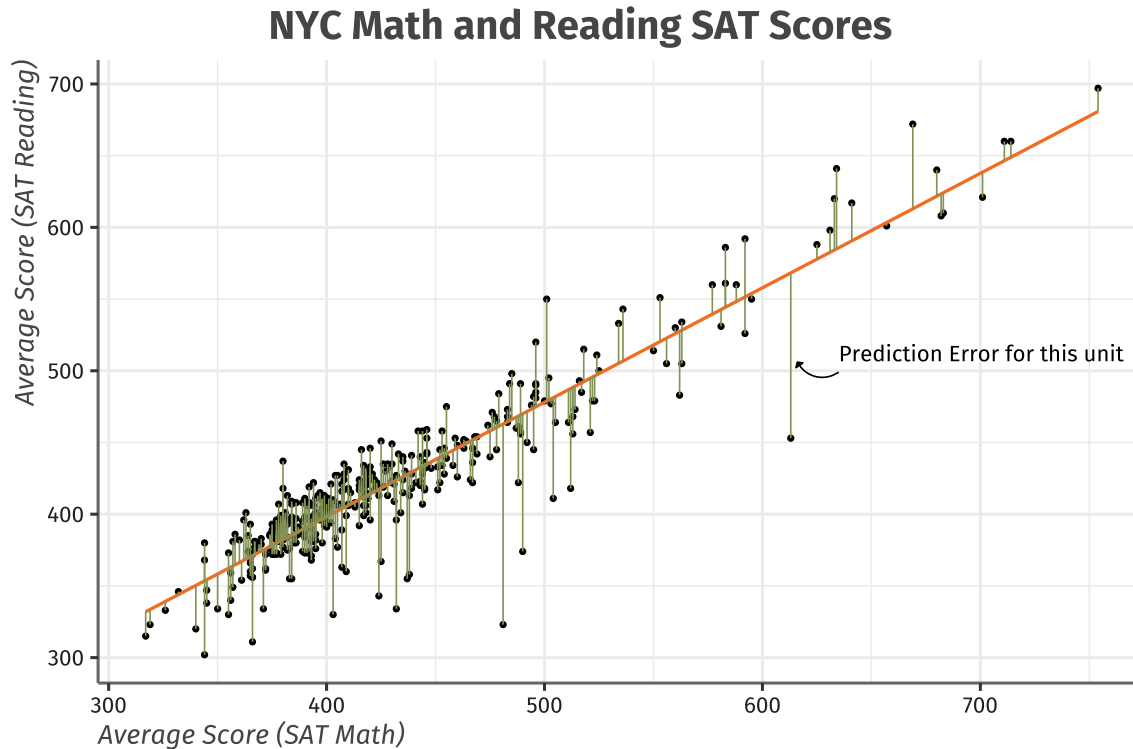
$$\underbrace{\hat{Y}_i}_{\text{Prediction for } Y_i} = \underbrace{\hat{a}}_{\text{Intercept}} + \underbrace{\hat{b}}_{\text{Slope}} * X_i$$

\hat{Y}_i is called the **predicted value**

Residual Errors

However, \hat{Y} will not be equal to Y . Therefore our **prediction error** for unit i is given by

$$\varepsilon_i = Y_i - \hat{Y}_i$$



Being the "best"

Question: How do we estimate the intercept \hat{a} and the slope \hat{b} line of "best fit"?

Answer:

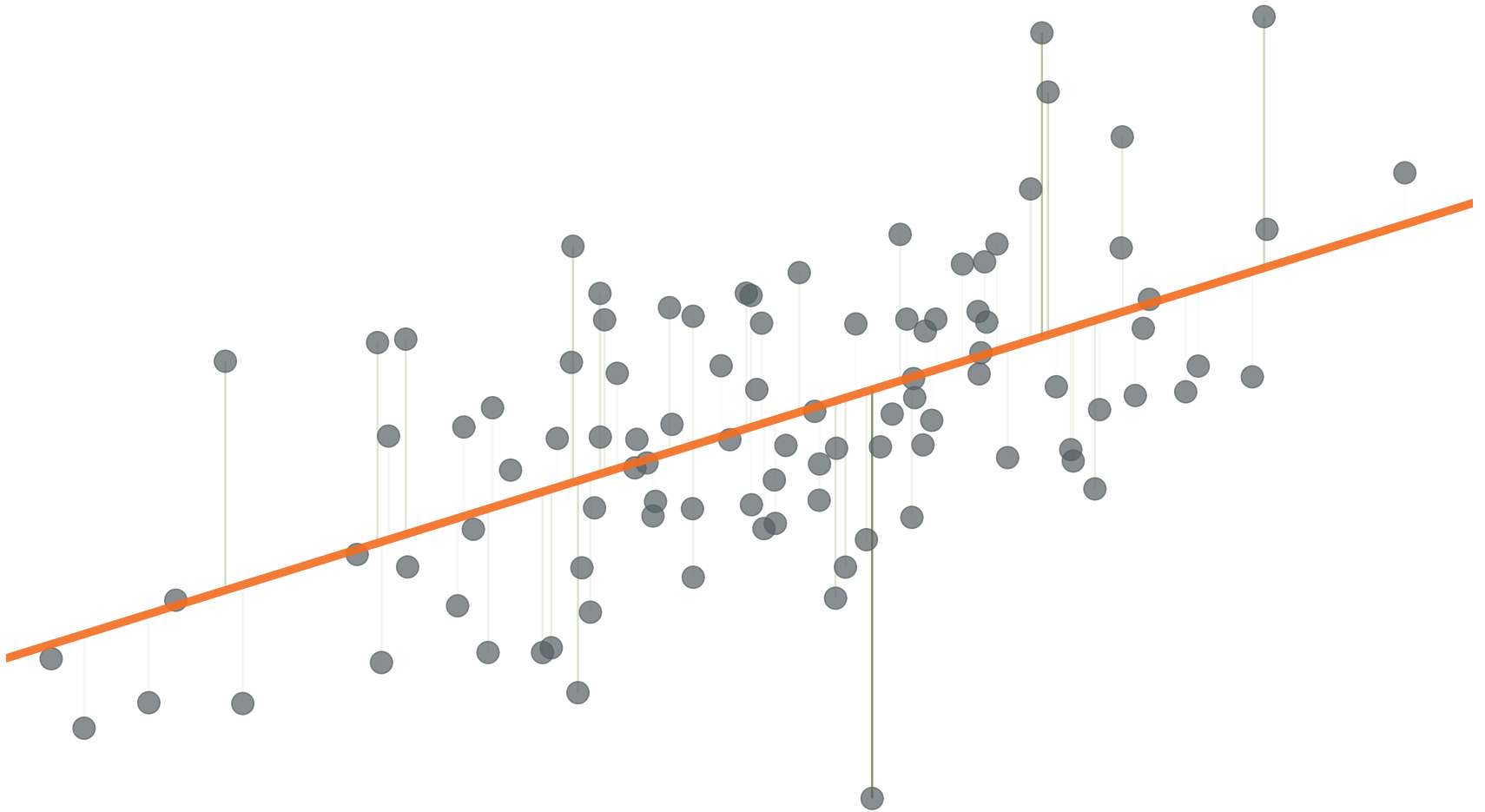
1. What about minimizing of $\sum_i \varepsilon_i$, i.e. *lower total prediction error*?
 - ε_i could be positive or negative if we over/under-estimate \hat{Y}_i . If we add up all the residuals, the positive and negatives will sometimes cancel out (same reasoning as variance). This isn't good because we want to be 'penalized' for all prediction errors.
2. Therefore we minimize ε_i^2 because each ε_i^2 will be positive.
 - In general, *the line of best-fit* means the line that minimizes the **sum of squared errors** (SSE):

Sum of Squared Errors

$$SSE = \sum_{i=1}^n (y_i - \hat{a} - \hat{b}X_i)^2$$

OLS vs. other lines/estimators

The OLS estimate is the combination of \hat{a} and \hat{b} that minimize SSE.



Least Squares Regression Line

Let $\{X_i, Y_i\}_{i=1}^n$ be a *simple random sample* of X and Y .

The **least squares regression line** is the equation $\hat{Y}_i = \hat{a} + \hat{b}X_i$, where \hat{a} and \hat{b} minimize

$$\sum_{i=1}^n (Y_i - a - b * X_i)^2$$

The solution is:

$$\hat{b} = r_{XY} \frac{s_Y}{s_X}$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

Clicker Question

Consider the NHIS dataset. Let Y be a person's weight in pounds, and X be the number of drinks per day they consume (on average). You calculate the following:

$$\bar{Y} = 176.5889, \bar{X} = 2.2489, s_Y = 39.86577, s_X = 1.804856, r_{XY} = 0.1187268$$

What is the regression line you fit to the data?

- a. $\hat{Y}_i = 176.5889 + 2.6224X_i$
- b. $\hat{Y}_i = 176.5768 + 0.0054X_i$
- c. $\hat{Y}_i = 126.9151 + 22.08814X_i$
- d. $\hat{Y}_i = 170.6913 + 2.6224X_i$

Interpreting a Regression

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes

We often use a regression line to predict the value of y for a given value of x , when we believe the relationship is linear

A linear relationship is of the form:

$$y = a + bx$$

- a is the value of y whenever $x = 0$
- b is the amount y changes when x increases by one

Interpreting a Regression

Lets go back to our clicker question, we calculated the line of best fit to be:

$$\hat{Y}_i = 170.69 + 2.62X_i$$

where Y is a person's weight in pounds, and X is the number of drinks per day they consume (on average)

- Our intercept, 170.69 is the predicted weight for someone who doesn't consume any alcohol
 - Our best guess at Y when $X = 0$,
- Our slope, 2.62 is the amount predicted weight increases when number of drinks per day increases by 1
 - The amount \hat{Y}_i changes when X_i increases by 1

Interpreting a Regression

Say we calculate the following regression line from hours studied and final exam grades:

$$\widehat{\text{Final Exam}} = 38 + 5.7 * \text{Hours of Studying}$$

- 38 is predicted score with no studying
- Each hour of studying increases predicted final exam score by 5.7 points

Clicker Question

Given that same regression line, $\text{Final Exam} = 38 + 5.7 * \text{Hours of Studying}$, what is the predicted final exam score if you study 8 hours?

- a. 83.6
- b. 45.6
- c. 96.3

Clicker Question

A store in Boulder calculates a least squares line that describes how price (in dollars) of juuls affects the quantity sold;

$$\widehat{\text{Juuls sold}} = 117 - 12.4 * \text{price}$$

If price *decreases* by 1 dollar, what happens to number of juuls sold?

- a. Quantity decreases by 12.4
- b. Quantity increases 12.4
- c. Quantity decreases by 117
- d. Quantity increases by 117

Properties of Regression Residuals

The slope, b , and intercept, a , of the regression line are found by minimizing $\sum_{i=1}^n \varepsilon_i^2$.

This forces $\bar{\varepsilon} = 0$. Hence, an **assumption** of regression is that $E(\varepsilon) = 0$.

Intuitively, this assumption means that the error in your prediction is due entirely to randomness.

Overview of Regression Analysis

A researcher is studying the relationship between high school students' SAT scores and their GPA during their freshman year of college. The data has a linear correlation coefficient of 0.503. Additional sample statistics are summarized in the table below:

Variable	Description	Sample Mean	Sample Std. Dev
X	SAT score	$\bar{X} = 1501.72$	$s_X = 104.14$
Y	GPA	$\bar{Y} = 3.3$	$s_Y = 0.45$

- What is the slope and intercept of this regression line? Write the linear regression using the $Y = a + bX$ format.
- Interpret the slope and intercept coefficients
- What is the predicted GPA if the student got a 1600 on the SAT?

$$R^2$$

Next we define a measure to evaluate how well the regression line fits:

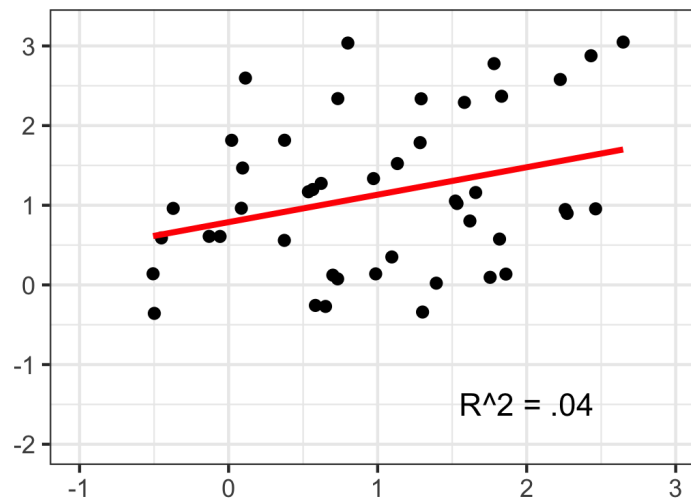
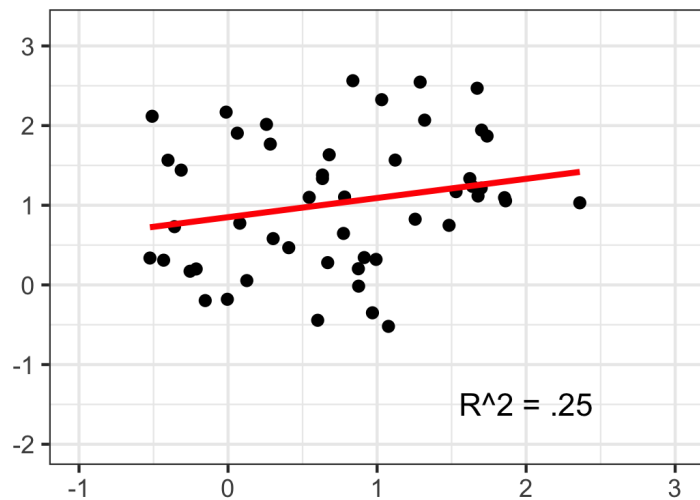
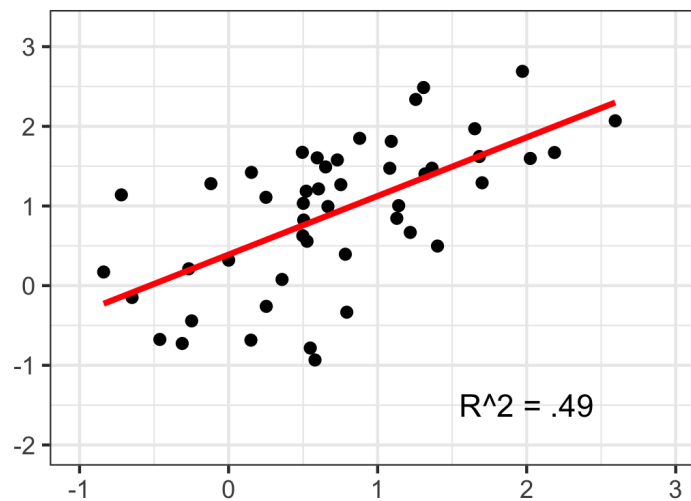
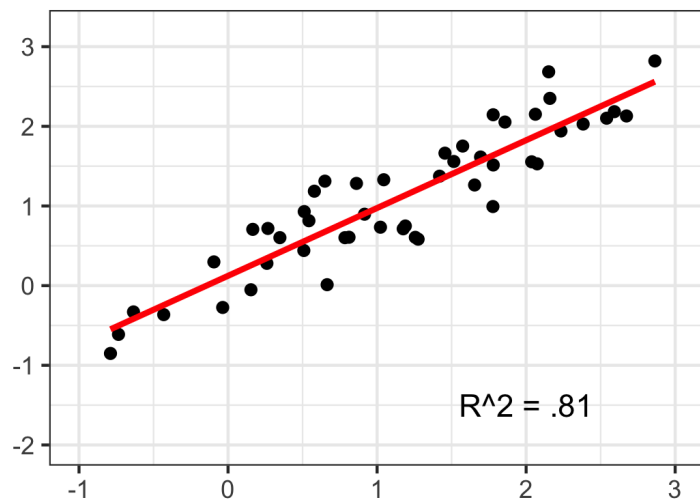
$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Intuition of R^2

Intuitively, R^2 measures the percent of variation in Y explained by the model.

$$R^2 = \frac{\text{variation in } \hat{y} \text{ along the regression line as } x \text{ varies}}{\text{total variation in observed values of } y}$$

R-squared



r and R^2

Correlation, r , describes the strength of a straight-line relationship between two variables

R^2 , is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x

$$R^2 = r^2$$

r and R^2

Lets say we have $r = -0.7786$ and $R^2 = (-0.7786)^2 = 0.6062$ between exercise and fat gain.

- $r = -0.7786$, there is a strong negative linear relationship between time exercised and amount of weight gained
- $R^2 = 0.6062$, about 61% of the variation in fat gained is accounted for by the linear relationship between fat gain and exercise. This means about 39% of the change in fat gained is not explained by this relationship

Clicker Question

Say we run a regression on the temperature and the amount of gas used to heat a home. We find that the $r = -0.99$ and $R^2 = 0.98$. This suggests that:

- a. although temperature and gas used are very correlated, the temperature does not make very good predictions of the amount of gas used
- b. gas used increases by 0.99 cubic feet for each additional degree colder it is outside
- c. prediction of gas used based off temperature will be quite accurate

R^2 Sidebar

A small R^2 does not mean the result is uninteresting. All it means is that the x variable alone does not explain a large portion of the variation in y.

Example: You find significant relationship between exercise and income, but it has a small R^2 .

We know income is determined by a variety of variables -- parent's income, education, innate ability, experience, etc.

- Your result isn't uninteresting, it just means there is a lot of variation in income *not due* to exercise, which is exactly what we'd expect.

R^2 Example

Recall from our previous example that a researcher calculated a correlation coefficient 0.503 between SAT scores and college freshman GPA. This implies an R^2 of 0.253.

What does this R^2 mean?

Does this make sense? What other things could explain the variation in freshman year GPA?

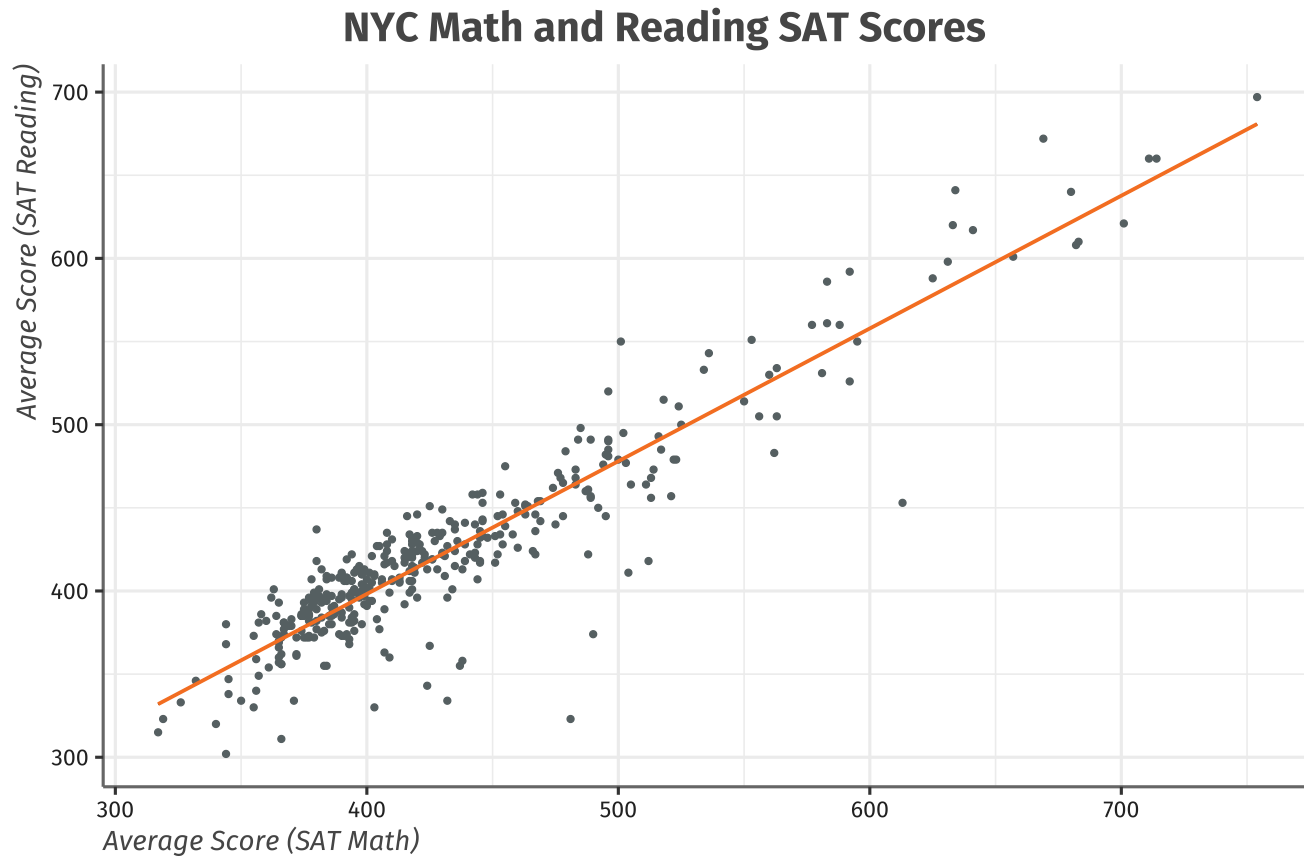
Influential Observations

Our regression line is sensitive to outliers, either in the x or y dimension

- We say an outlier is **influential** if deleting it changes our regression line substantially
- The amount by which the line changes is called the **leverage** an influential observation has

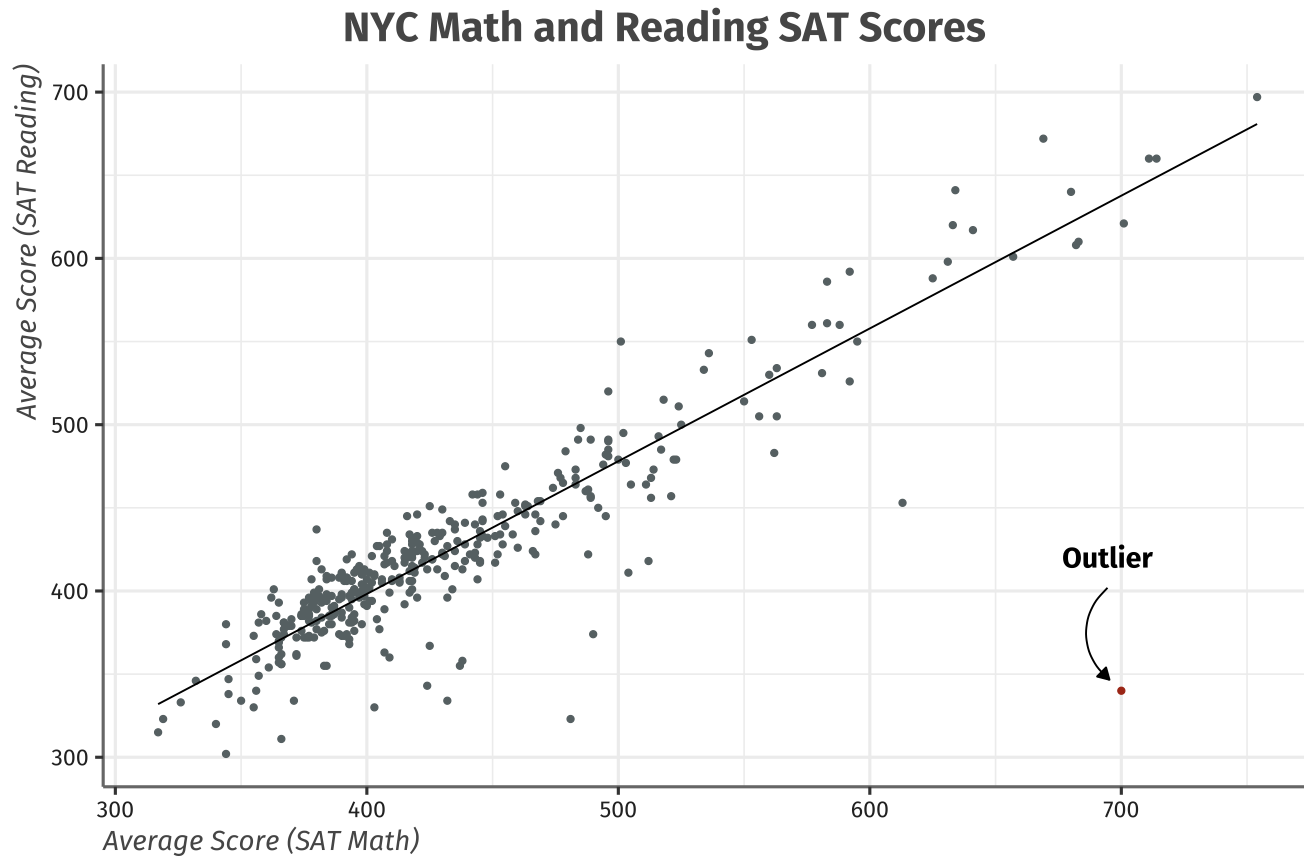
Example of Influential Observation

Let's revisit the NYC Math and Reading SAT Scores example.



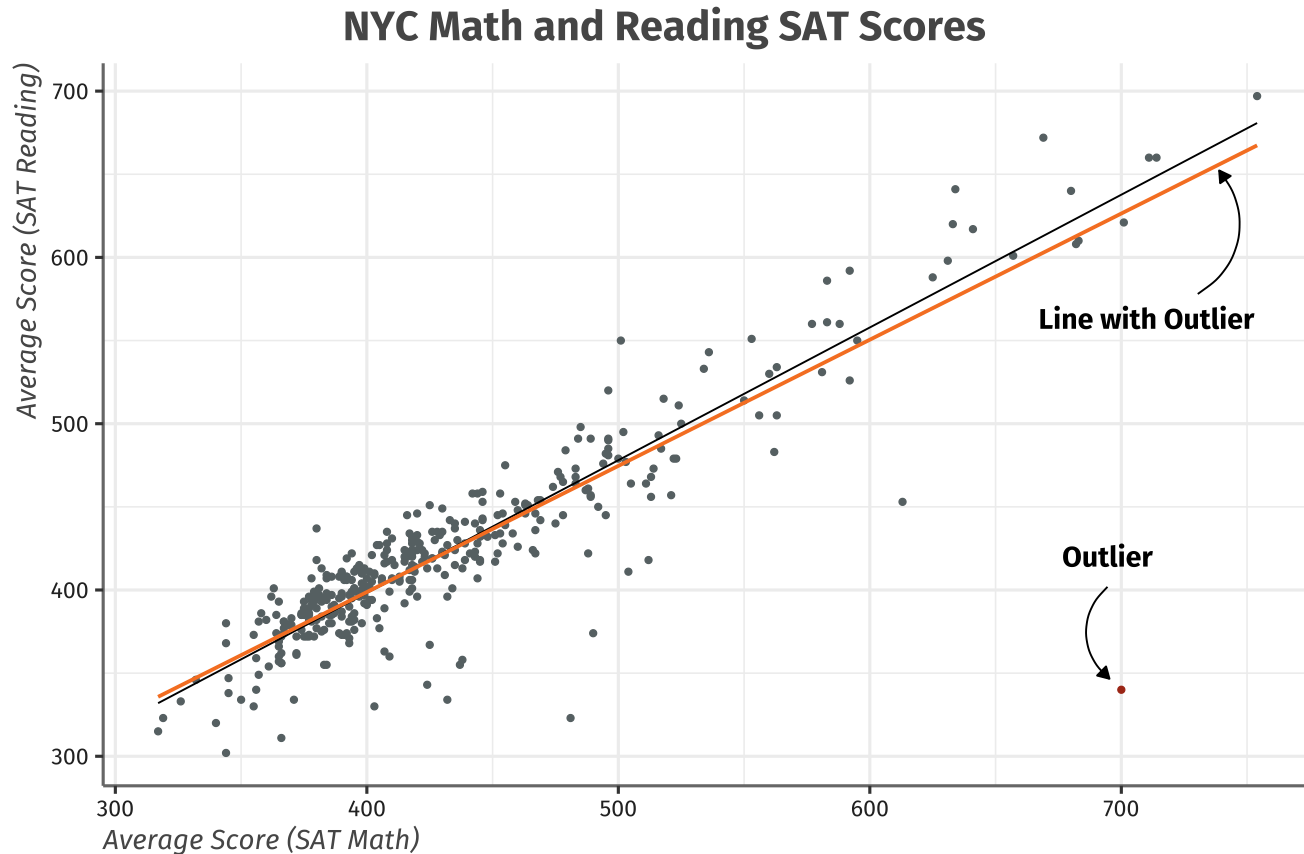
Example of Influential Observation

Suppose we had an outlier of a Math score of 700 and a Reading score of 340.



Example of Influential Observation

Suppose we had an outlier of a Math score of 700 and a Reading score of 340. That data point has quite a bit of leverage because it is an extreme outlier.



Cautions about Correlation and Regression

Correlation and regression are powerful tools for describing the relationship between two variables, but they have their limitations

- These tools only describe linear relationships
- They are not resistant to outliers

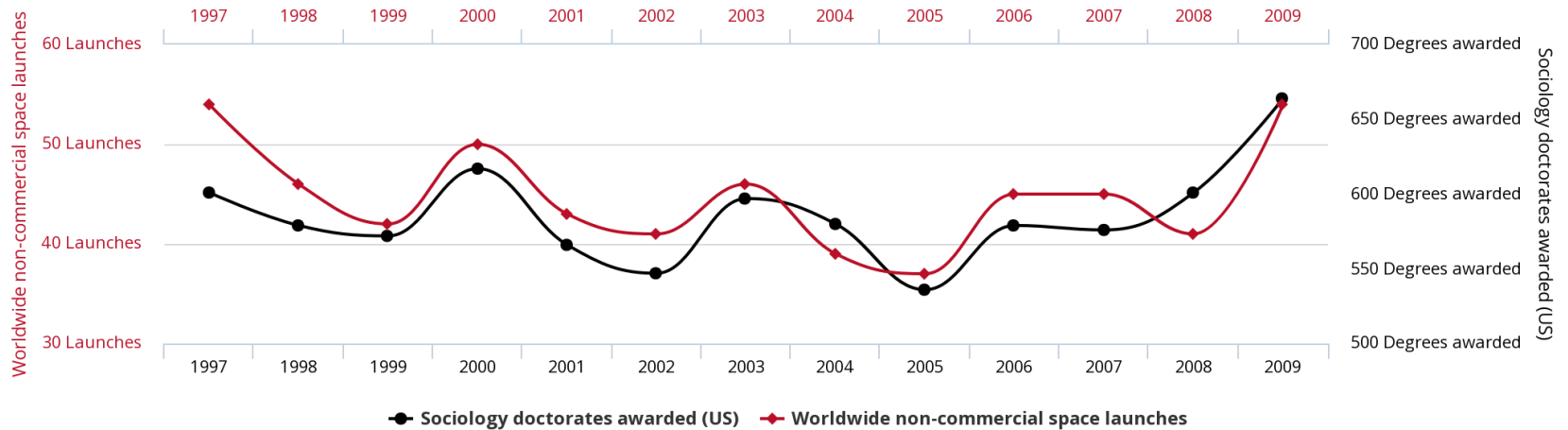
There is one more major issue with discussing correlations:

Correlation does not imply causation

Spurious Correlations

Spurious Correlation is things move in the same direction, even if they are completely unrelated.

Worldwide non-commercial space launches
correlates with
Sociology doctorates awarded (US)

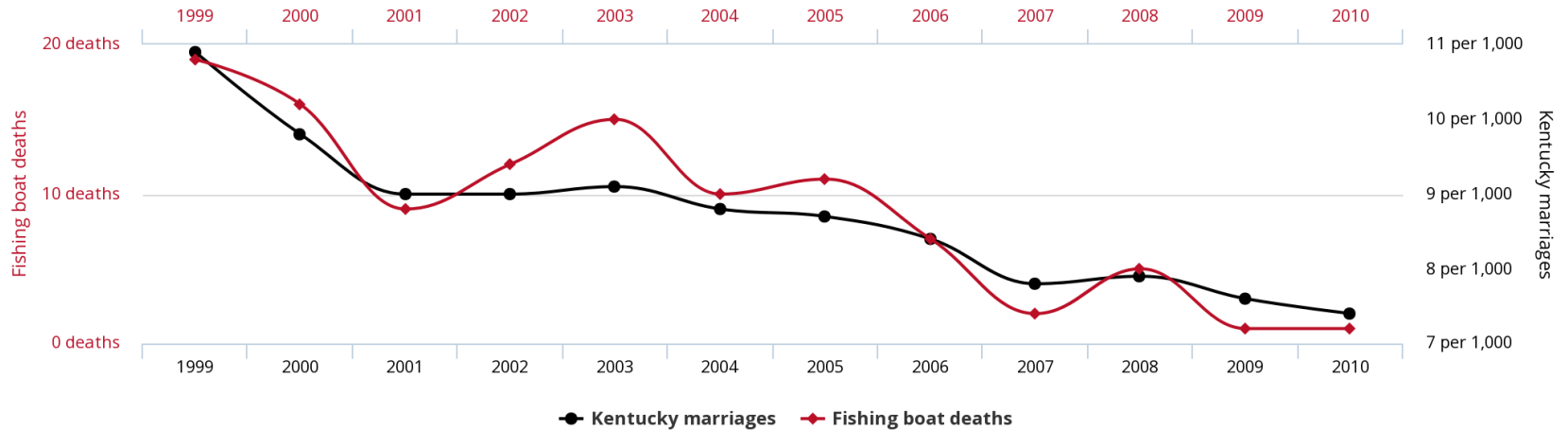


tylervigen.com

Spurious Correlations

Spurious Correlation is things move in the same direction, even if they are completely unrelated.

People who drowned after falling out of a fishing boat
correlates with
Marriage rate in Kentucky



tylervigen.com

Lurking Variable

A **lurking variable** is a variable that is not among the explanatory or response in a study and yet may influence the interpretation of relationships among those variables. Also known as **omitted variable bias**.

- Experience in music and test scores -- family background is a lurking variable
- Ice cream sales and number of violent police reports -- both of these things increase when the weather is warm

Ecological Correlation

A correlation based on averages rather than on individuals is called an **ecological correlation**. The idea being that the correlation between averages may be stronger than the correlation at the individual level

Examples:

- Number of years of education and average income level
- Hours of weekly exercise and body mass index

These relationships are very strong when we look at everyone, but may not be as strong when we analyze at the individual level

Extrapolation

Extrapolation is the use of a regression line for a prediction far outside the range of values of the explanatory variable x that you used to obtain the line

The idea here is that not many relationships are linear for *all values of x*

Example:

- Age and height, eventually you stop growing
- Experience and pay, eventually your salary levels off (usually)

Clicker Question

A study of elementary school children, ages 6 to 11, finds a high positive correlation between shoe size and score on a test of reading comprehension. The observed correlation is most likely due to:

- a. cause and effect (larger shoe size causes higher reading comprehension)
- b. a mistake, because the correlation must be negative
- c. the effect of a lurking variable, such as age or years of reading experience
- d. reverse cause and effect (higher reading comprehension causes larger shoe size)