ECON 3818: Introduction to Statistics with Computer Applications

Kyle Butts

November 10, 2020

Chapter 5: Regression

Review from Last Class

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

- Covariance, σ_{XY} (s_{XY})
    - Direction matters, but magnitude is hard to interpret
- Correlation, ρ_{XY} (r_{XY})
    - Direction and magnitude matter
    - Correlation is always value between $[-1, 1]$

Review from Last Class

Recall:

$$r = \frac{cov(X, Y)}{\sqrt{V(X)} \cdot \sqrt{V(Y)}}$$

- 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

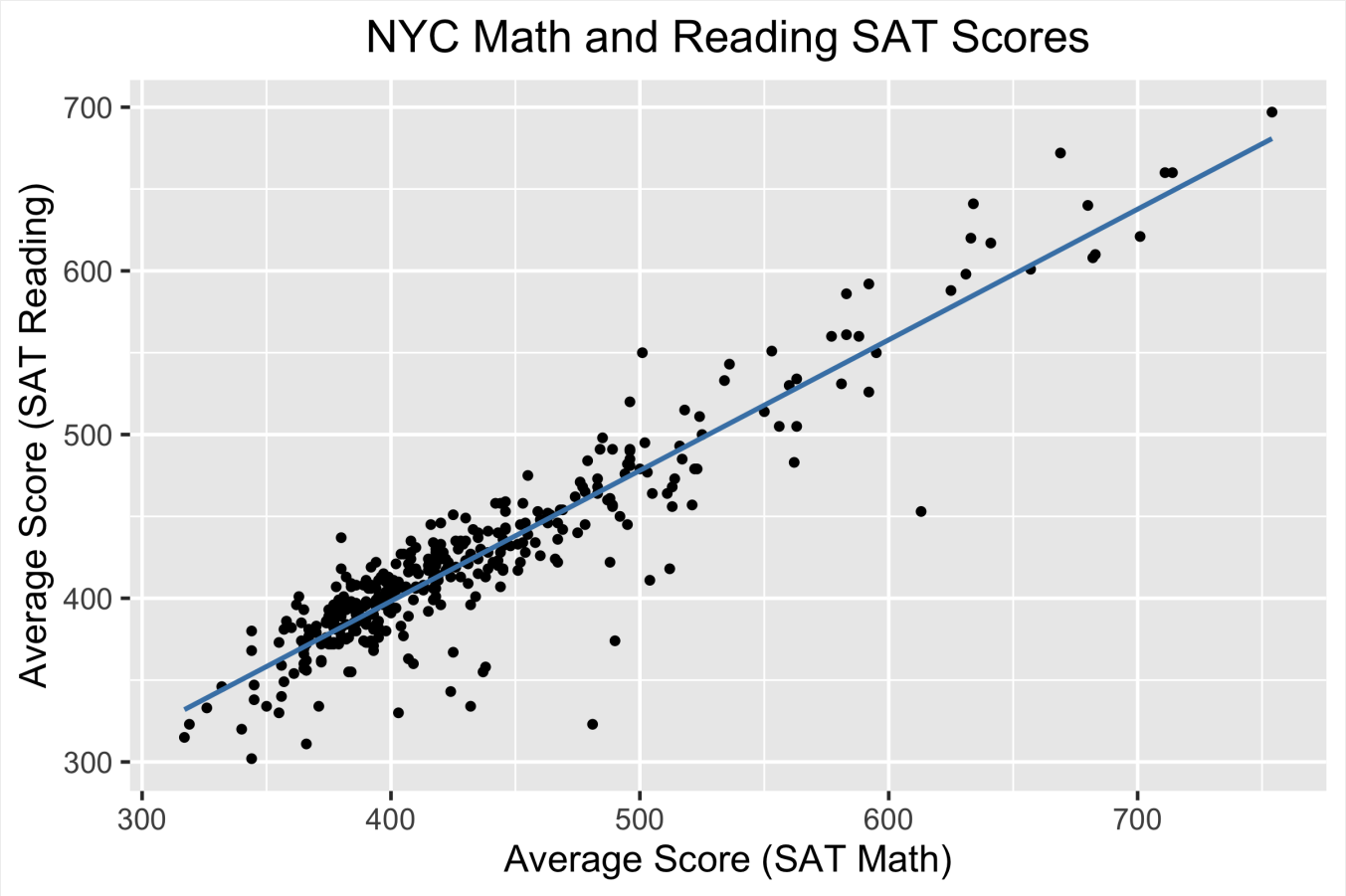
(b) 0.24

(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

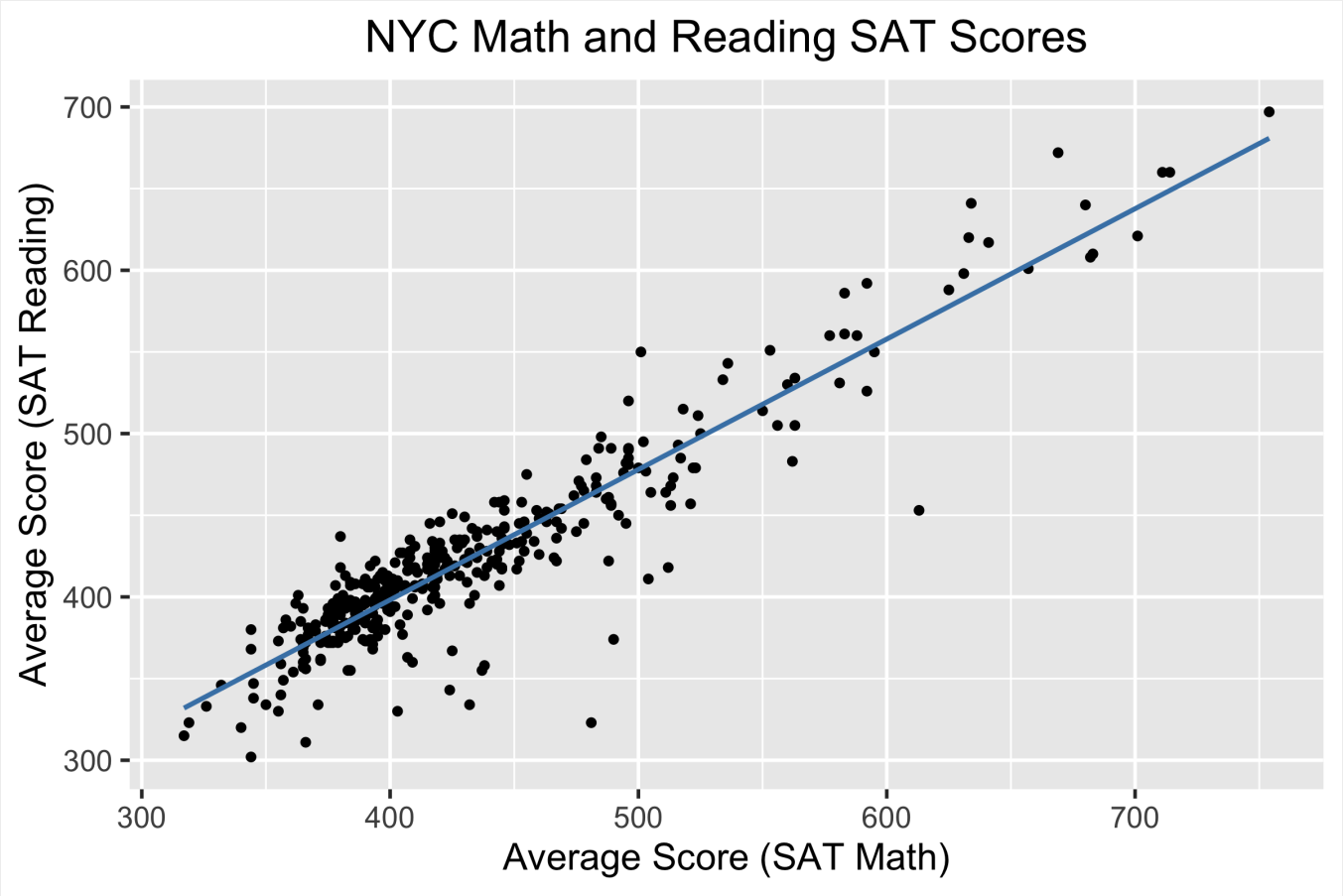
Definition (Regression Line)
Let X and Y be two random variables. A regression line is a straight line that describes how the response variable, Y, changes as the explanatory variable X changes.

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.*

Assuming the relationship is actually linear, how do we find this line of "best fit"?

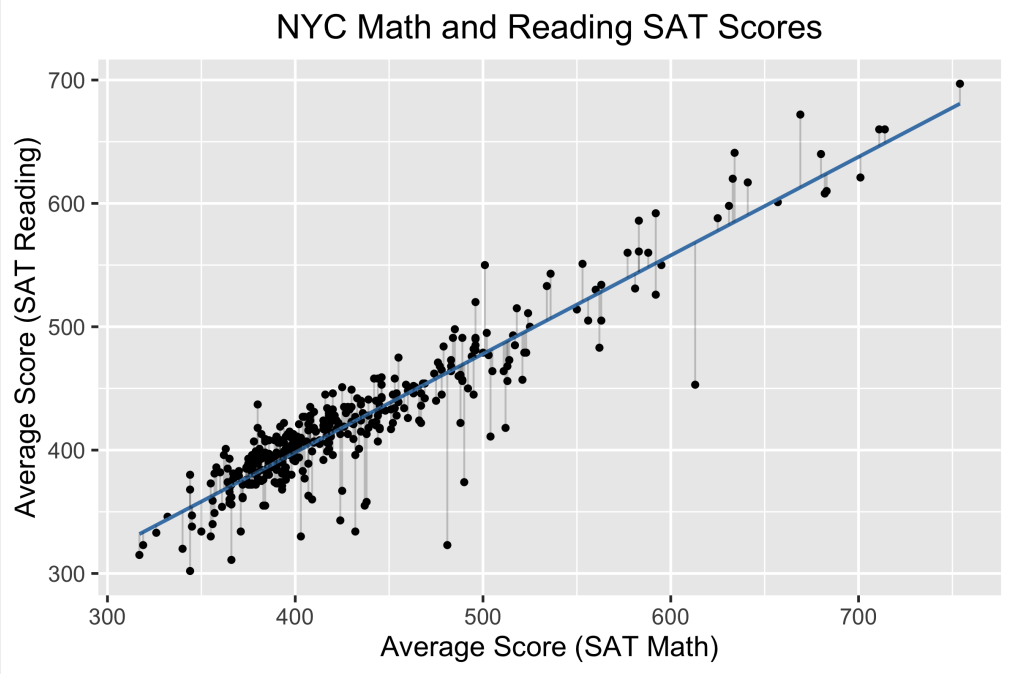
Least Squares Regression Line

What is the best straight line fit for the data? How would you determine it?



Least Squares Regression Line

What is the best straight line fit for the data? How would you determine it?



We want to pick the line that minimizes the total error.

Residual Error

Our goal is to pick a line that produces the least amount of total error. Let's define a residual (error) as the difference between what we observe (Y_i) and what we predict ($\hat{Y}_i$):

$$\varepsilon_i = Y_i - \widehat{Y}_i.$$

We could overestimate or underestimate, so ε_i could be positive or negative. If we add up all the residuals, the positive and negatives will sometimes cancel.

Therefore we want to minimize ε_i^2. Minimizing will produce ε_i^2 as close to 0 as possible.

We call $\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$ the sum of squared residuals.

Least Squares Regression Line

Let's strictly define the least squares regression line.

Definition (Least Squares)
Let $\{X_i, Y_i\}_{i=1}^n$ be a random sample of X and Y. The least squares regression line is the equation $\widehat{Y}_i = a + bX_i$, where a and b solve

$$\min_{a,b} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - (a + bX_i)\right)^2.$$

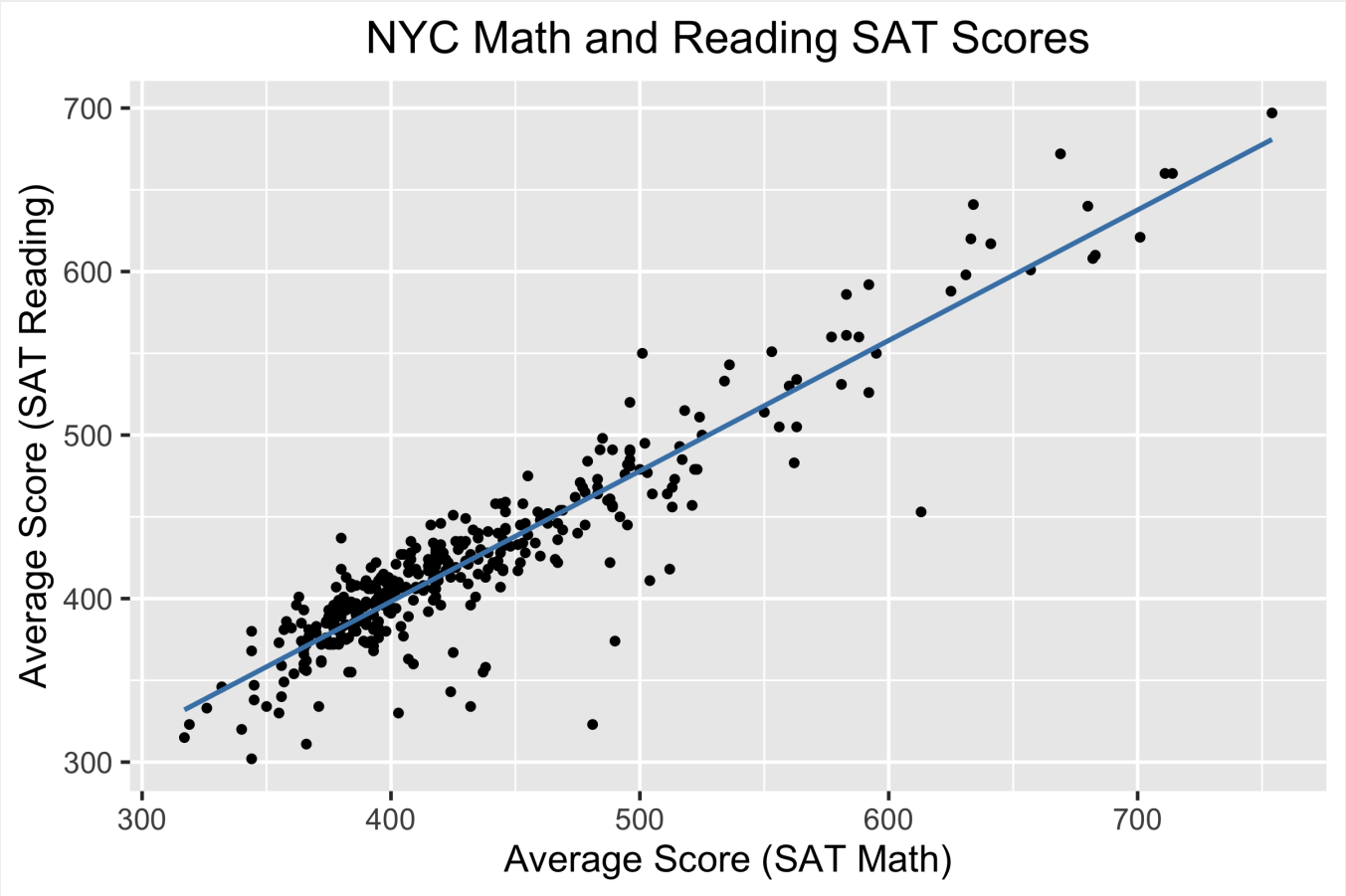
The solution is $b = r_{XY}\frac{s_Y}{s_X}$ and $a = \bar{Y} - b\bar{X}$.

We call $\widehat{Y}_i$ the predicted value of Y_i given X_i.

Example

The least squares regression line for midterm grades is

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



Takeaways of Least Squares Regression

- Calculate regression by minimizing difference between predicted values and actual values
- This leads to the following coefficients:
  - $b = r_{XY} \cdot \frac{s_Y}{s_X}$, the slope
  - $a = \bar{Y} - b\bar{X}$, the intercept

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) $\widehat{Y}_i = 176.5889 + 2.6224 X_i$

(b) $\widehat{Y}_i = 176.5768 + 0.0054 X_i$

(c) $\widehat{Y}_i = 126.9151 + 22.08814 X_i$

(d) $\widehat{Y}_i = 170.6913 + 2.6224 X_i$

Regression Intuition

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:

$$y = a + bx$$

so a is the value of y whenever $x = 0$, and 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
  - The value of $\hat{Y}_i$ whenever X_i=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:

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

- 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,
Final Exam $= 38 + 5.7 *$ 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;

$$\textit{Juuls sold} = 117 - 12.4(\textit{price})$$

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?

Next we define a measure of goodness of fit.

Definition (R^2)
Let $\{X_i, Y_i\}_{i=1}^n$ be a random sample of X and Y and $\widehat{Y}_i$ be the predicted values. We define R-squared as

$$R^2 = \frac{\sum_{i=1}^n \left(\widehat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2} = r_{XY}^2,$$

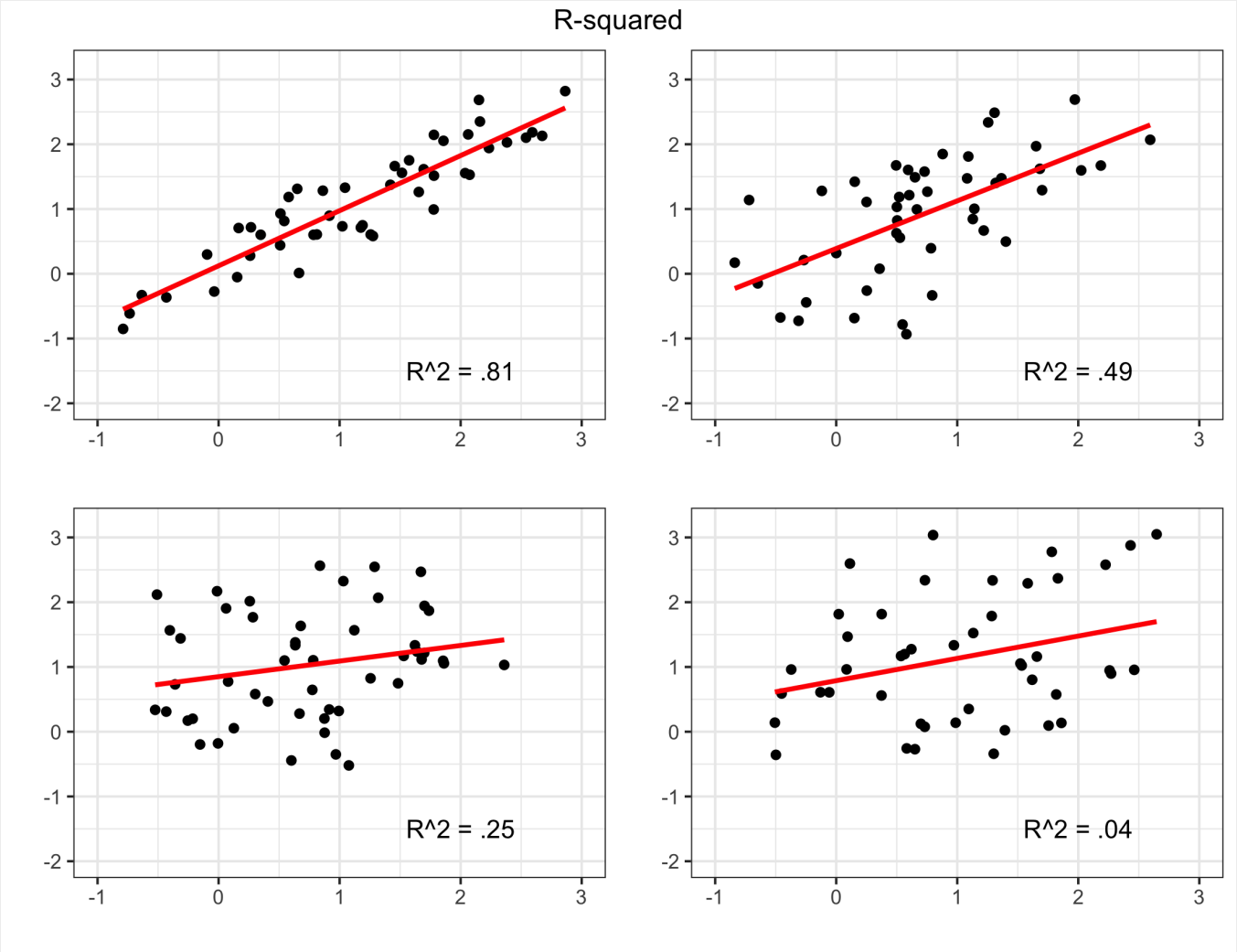
where r_{XY} is the correlation coefficient for X and Y.

Intuition of R^2

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

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

Visualizing R^2



r and R^2

- Correlation, r, describes the strength of a straight-line relationship between two variables
- R^2, or 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 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 fine 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 predicts 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?

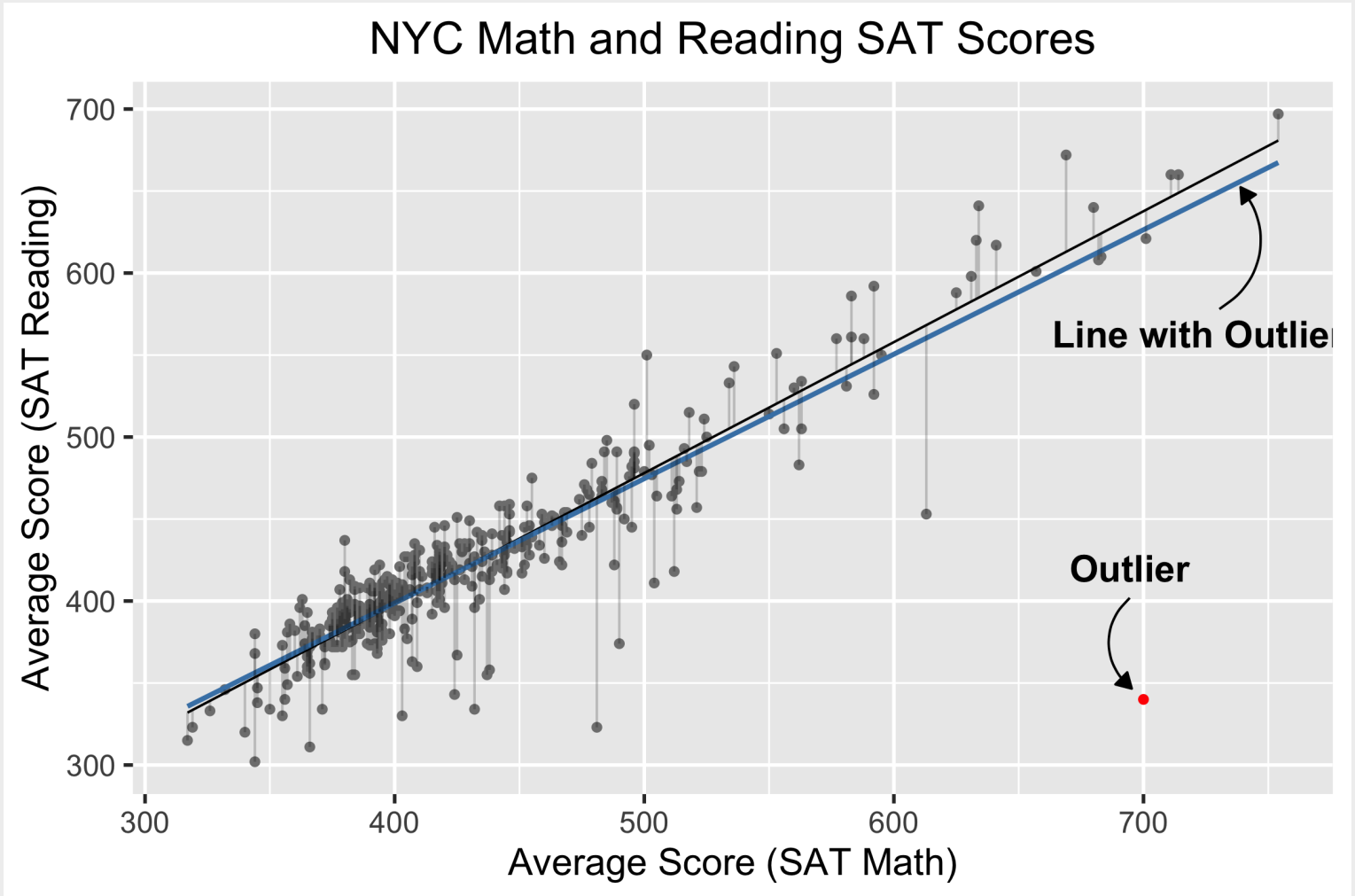
Influential Observations

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

Influential Observations

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

Now we will introduce three other potential issues with these tools

- Ecological correlation
- Extrapolation
- Lurking variables

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)

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

Cautions about Correlation and Regression

With all these limitations and potential issues, it is important to remember:

Correlation does not imply causation

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)

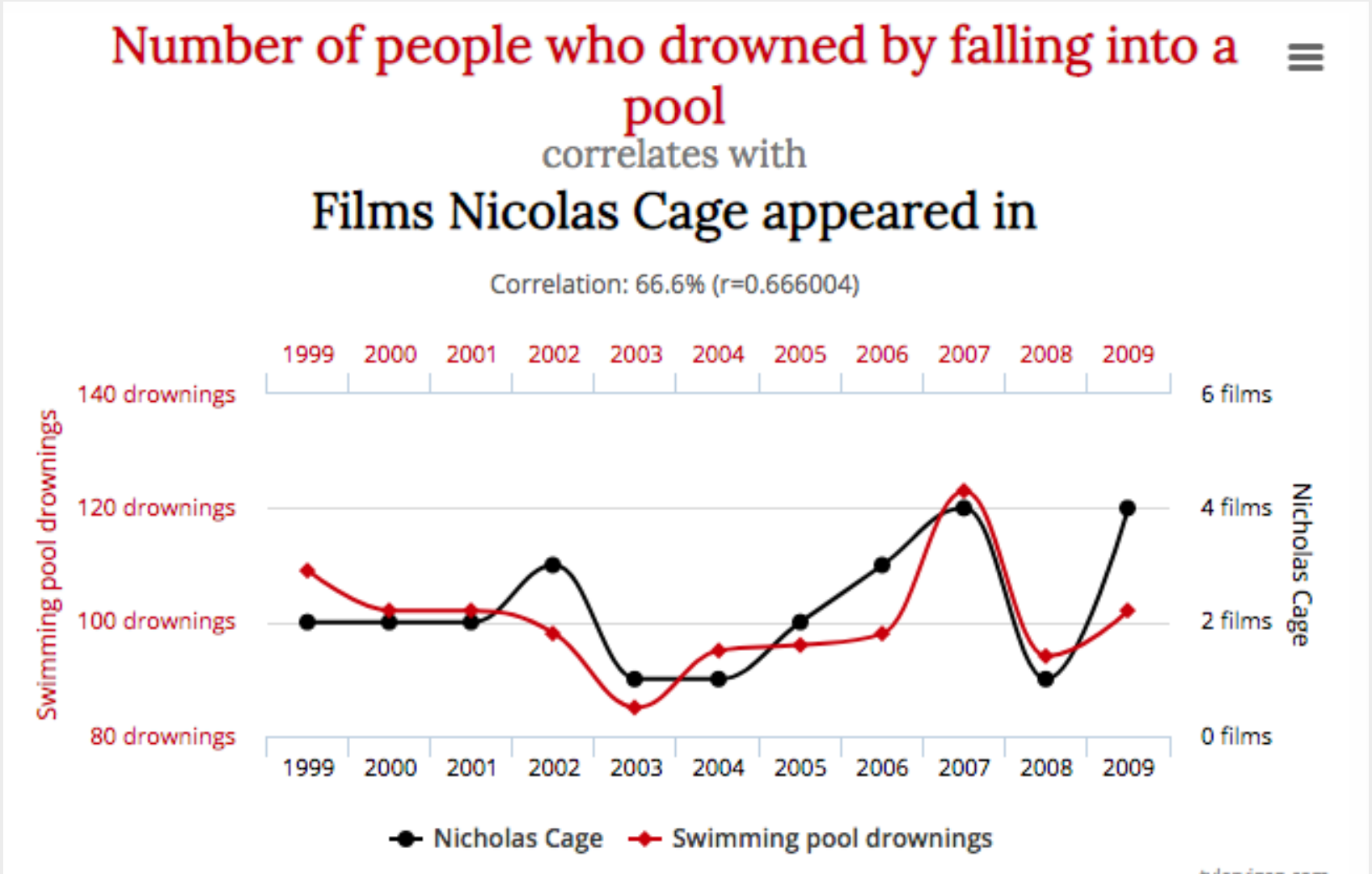
(b) a mistake, because the correlation must be negative

(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)

Spurious Correlations

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



Spurious Correlations