COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 18: Correlation and Linear Regression

Nov 19, 2025

Reminders

- Final Project Proposal Due tonight
    - 10% of your final project grade!
- HW 7 due Monday (skip Question 4 about the survey)
- Lab 9 this week
    - Last lab assignment of the semester!!
    - Thanksgiving next week, then final project consultations the week after

Last Topic: Prediction

- **Today**: Correlation and Linear Regression
- **Monday, Nov 24**: Least Squares and Residuals ← HW 7 due
- ***Wednesday, Nov 26****: Holiday!*
- **Monday, Dec 1**: Regression Inference ← HW 8 due
  Progress Report due
- **Wednesday, Dec 3**: Special Topics (Data Ethics)
  - Final Project Consultations during Lab
- **Monday, Dec 8**: Special Topics (Data Privacy) ← HW 9 due

Final Project

- Report is a written report
  - Should be in paragraph form… Do not simply list bullet points!
- Progress report we are only checking the written report
  - You do not need to submit your notebooks
- Final Project Report you will submit the completed written report and your notebook
  - Due on the Friday Dec 12 at 11:59pm
  - *No late days for the final report. Submit what you have by the deadline*

Correlation

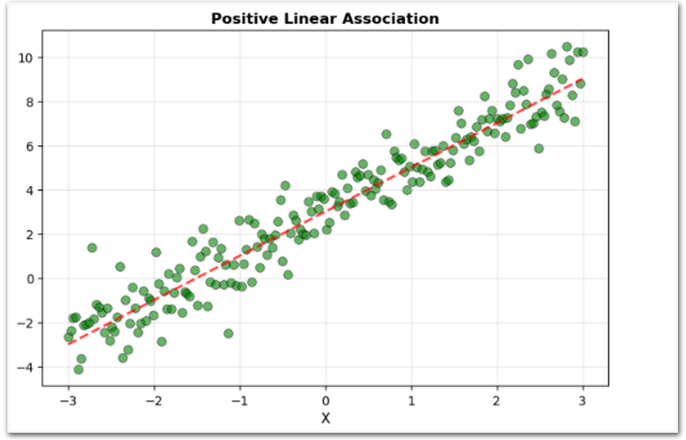
Prediction

- Guessing the future based on data
- To predict the value of a variable:
  - Identify (measurable) attributes that are associated with that variable
  - Describe the relation between the attributes and the variable you want to predict
  - Use the relation to predict the value of a variable

Two Numerical Variables

Trend

- Positive association
- Negative association
- Pattern



Any discernible “shape” in the scatter

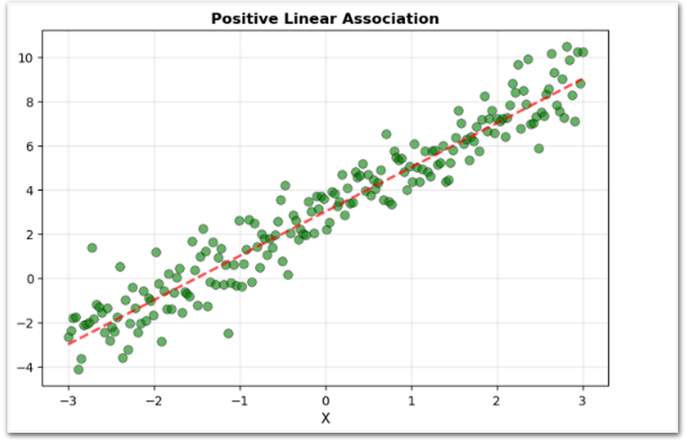
- Linear
- Non-linear

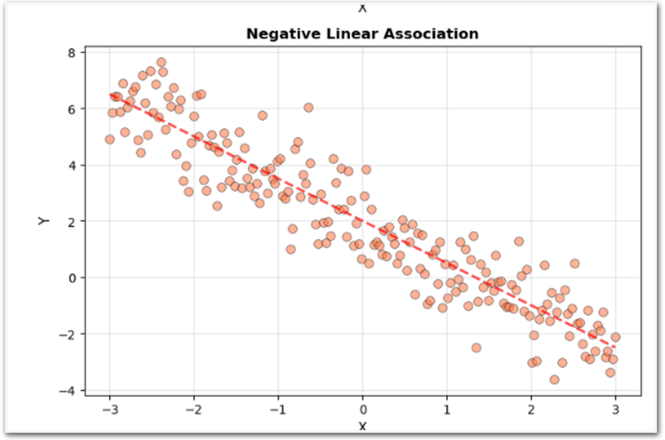
Visualize, then quantify

Two Numerical Variables

Trend

- Positive association
- Negative association
- Pattern





Any discernible “shape” in the scatter

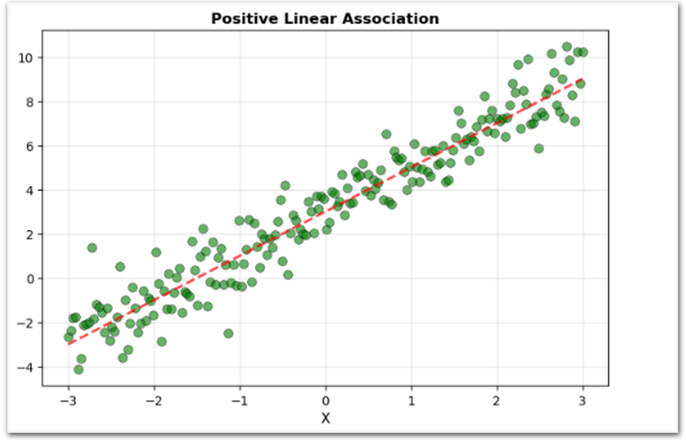
- Linear
- Non-linear

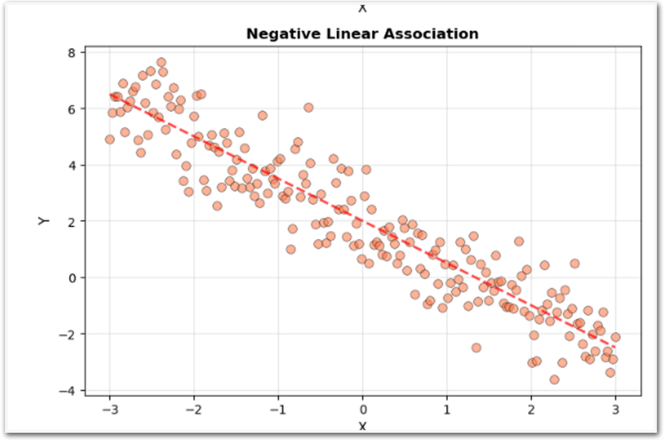
Visualize, then quantify

Two Numerical Variables

Trend

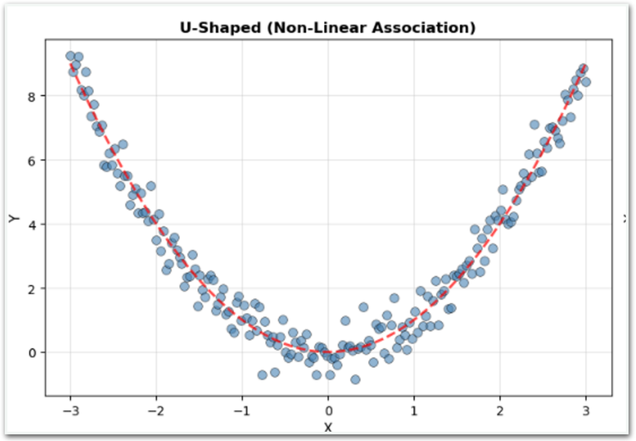
- Positive association
- Negative association
- Pattern





Any discernible “shape” in the scatter

- Linear
- Non-linear



Visualize, then quantify

Two Numerical Variables

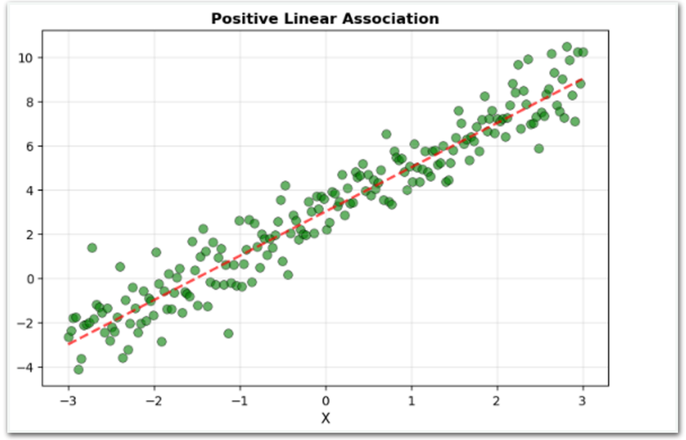
Trend

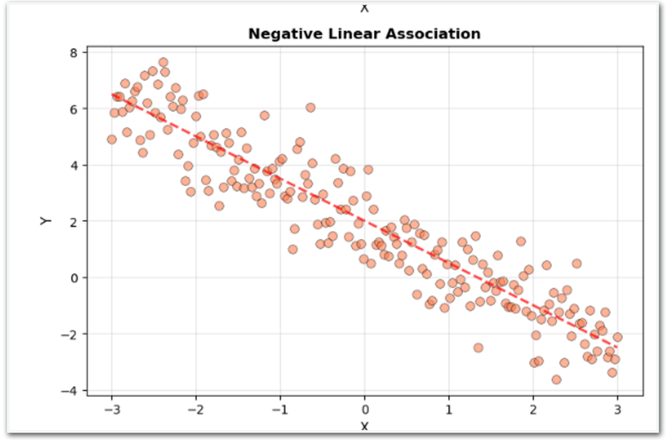
- Positive association
- Negative association
- Pattern

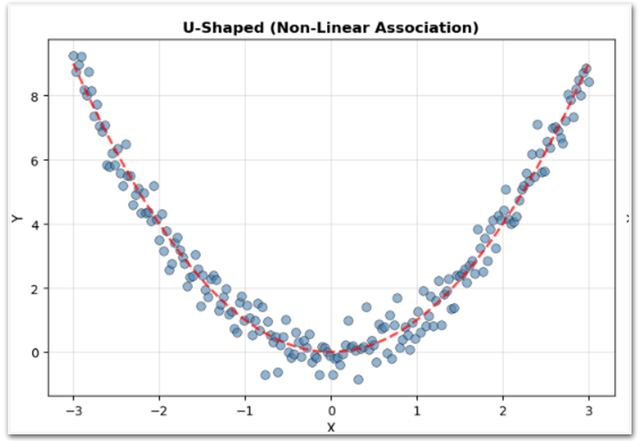
Any discernible “shape” in the scatter

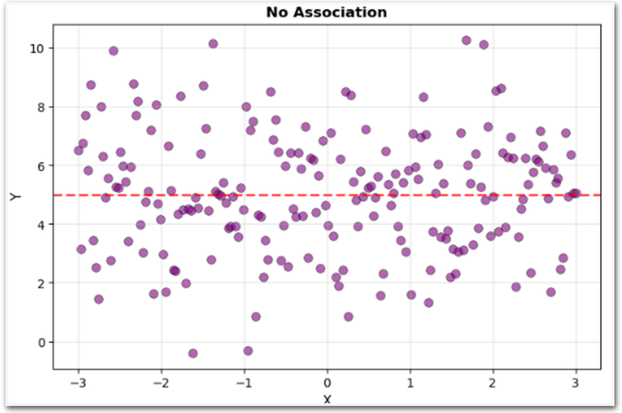
- Linear
- Non-linear

Visualize, then quantify









Correlation Coefficient r

- A metric for the strength of the **linear relationship** between two variables
  - How clustered values are in a scatter plot around a straight line
- r is always between -1 and 1 ($-1 \leq r \leq 1$)
  - $r = 1$: perfect correlation (straight line) sloping upward
  - $r = -1$: perfect correlation (straight line) sloping down
- r = 0: no linear association (*uncorrelated*)

Notebook demo: Correlation Plots

Notes about using r

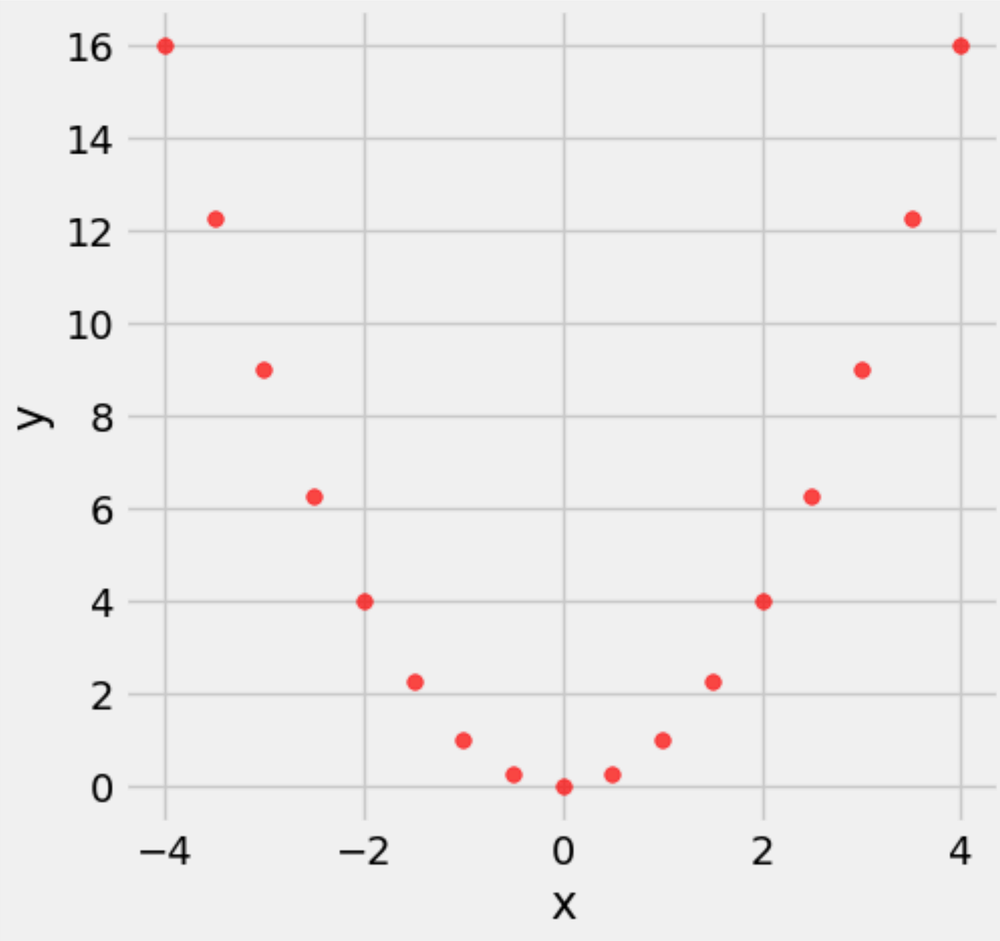
- r measures **linear** association
- Association does *not* imply causation
  - Two variables may be correlated, but one may not cause the other
  - Non-linearity and outliers can affect correlation
- Correlations based on aggregated data (**ecological correlations**) can be misleading
  - Correlations between individuals may be lower than correlation between averages of groups

Questions

True or false?

1. If the correlation coefficient of x and y is 0, then knowing one cannot help us predict the other

Non-linear correlations may have $r \approx 0$



x and y are highly correlated, but not linearly

$r \approx 0$

Questions

True or false?

1. If the correlation coefficient of x and y is 0, then knowing one cannot help us predict the other

2. If x and y have a correlation coefficient of 1, then one must cause the other

Association ≠ causation

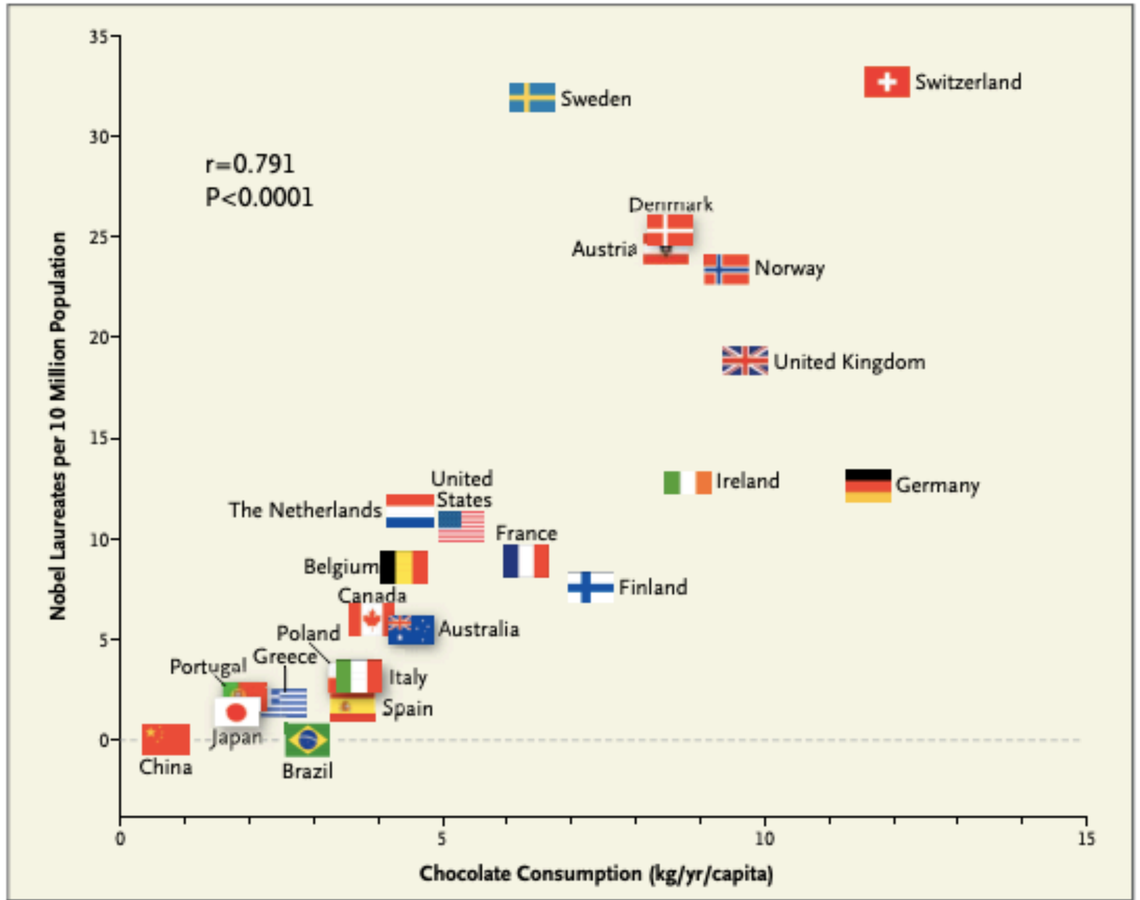


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

https://www.biostat.jhsph.edu/courses/bio621/misc/Chocolate%20consumption%20cognitive%20function%20and%20nobel%20laurates%20(NEJM).pdf

Questions

True or false?

1. If the correlation coefficient of x and y is 0, then knowing one cannot help us predict the other

2. If x and y have a correlation coefficient of 1, then one must cause the other

3. If x and y have a correlation coefficient of -0.8, they have a negative association

Prediction

Guess the future

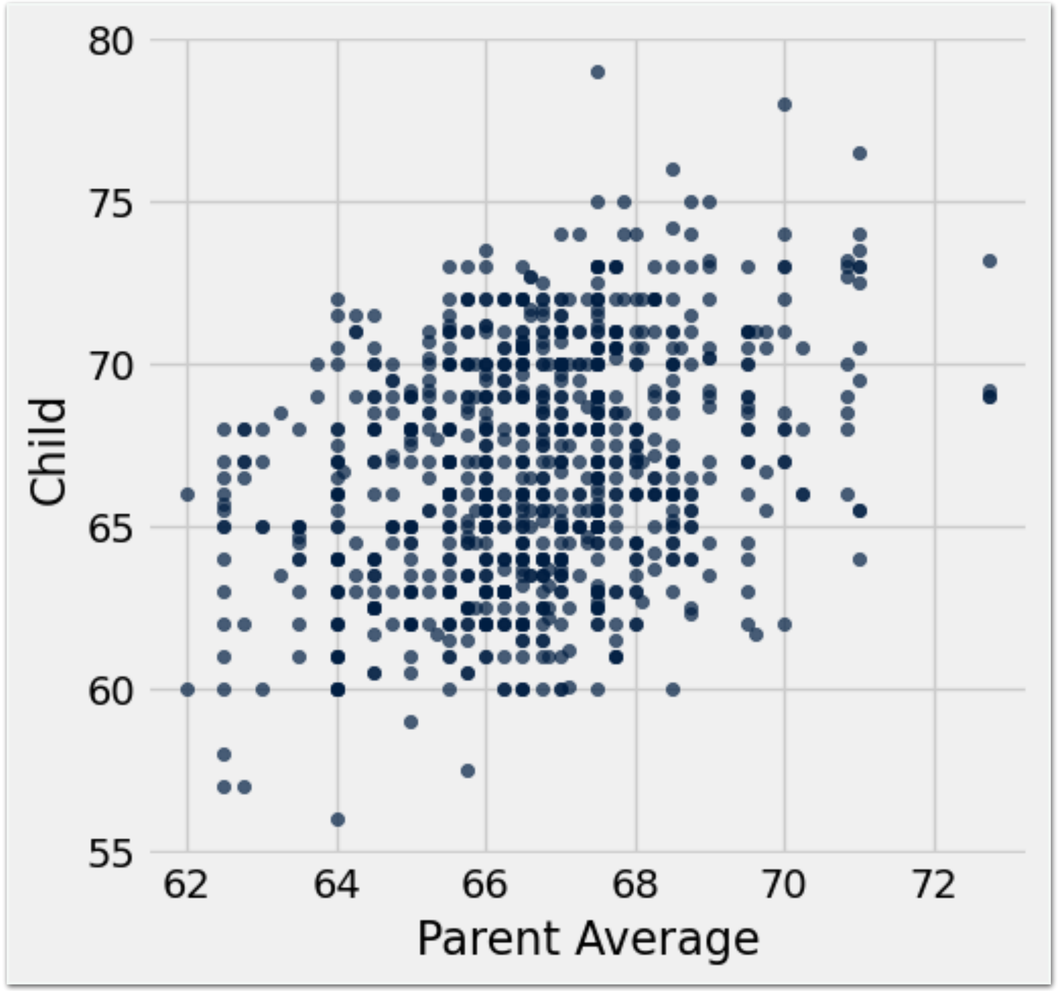
- Based on incomplete information

- One way of making predictions:

  - To predict the outcome for an individual, find others who are like that individual and whose outcomes you know. Use those outcomes as the basis of your prediction

Example: Galton's Heights

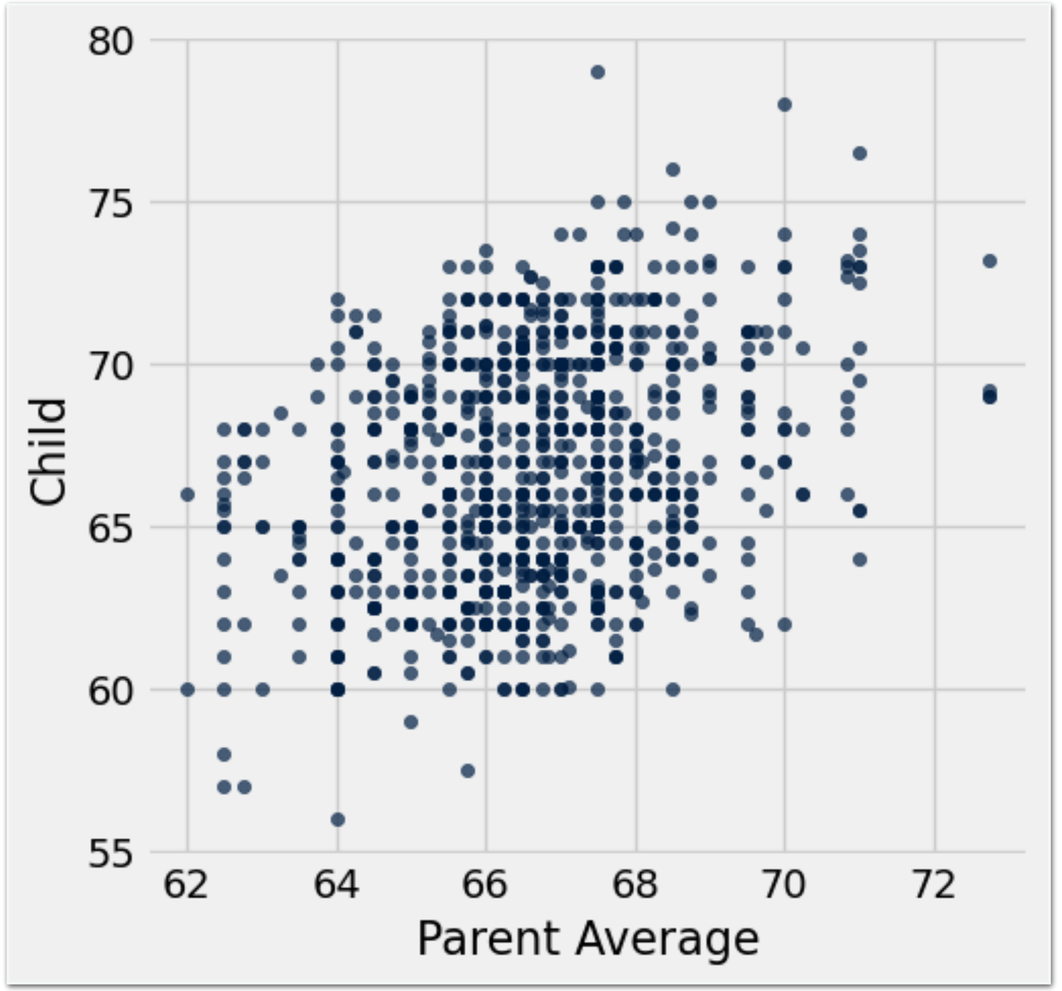
Goal: Predict the height of a new child based on that child's midparent height (average parent height)



Example: Galton's Heights

Goal: Predict the height of a new child based on that child's midparent height (average parent height)

- How could we predict the child's height given a mid parent height of 68 inches?

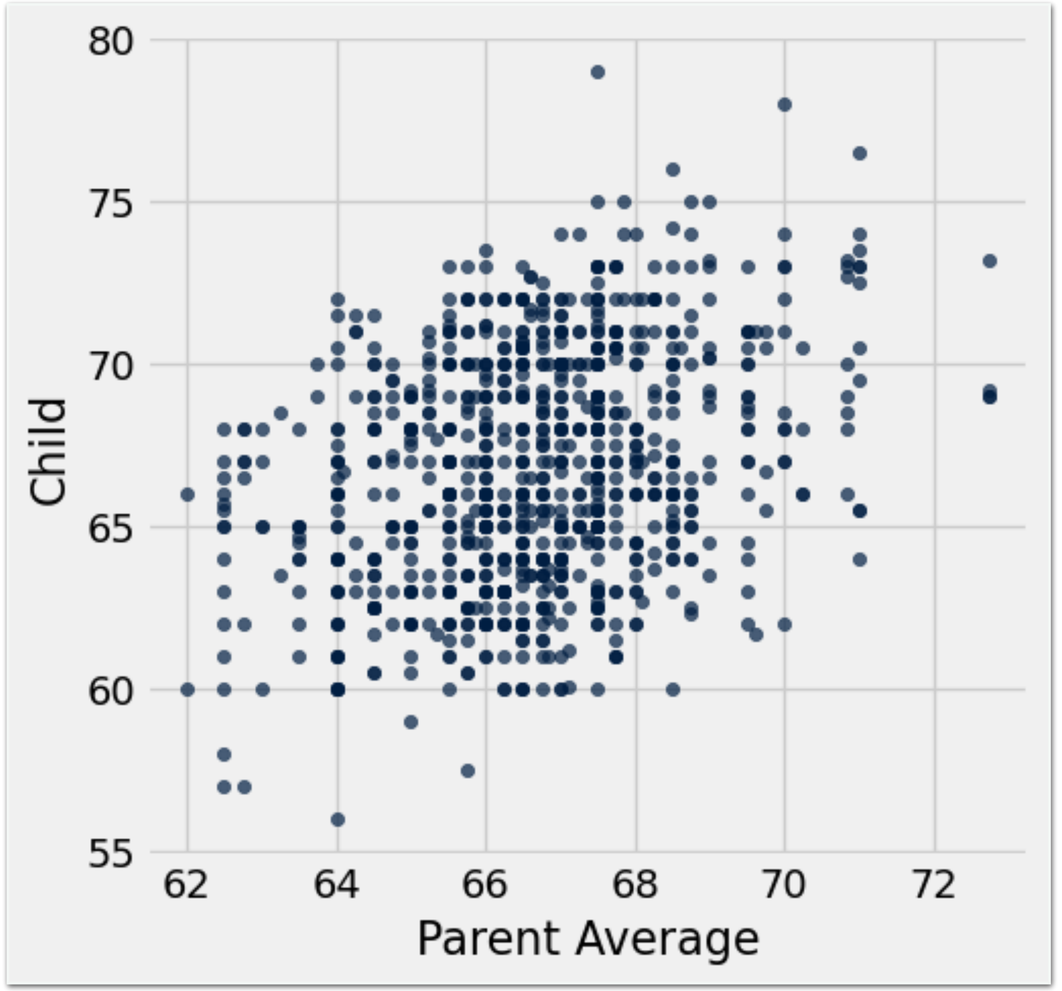


Example: Galton's Heights

Goal: Predict the height of a new child based on that child's midparent height (average parent height)

- How could we predict the child's height given a mid parent height of 68 inches?

- Idea: Use the average height of the children of families whose midparent height is close to 68 inches

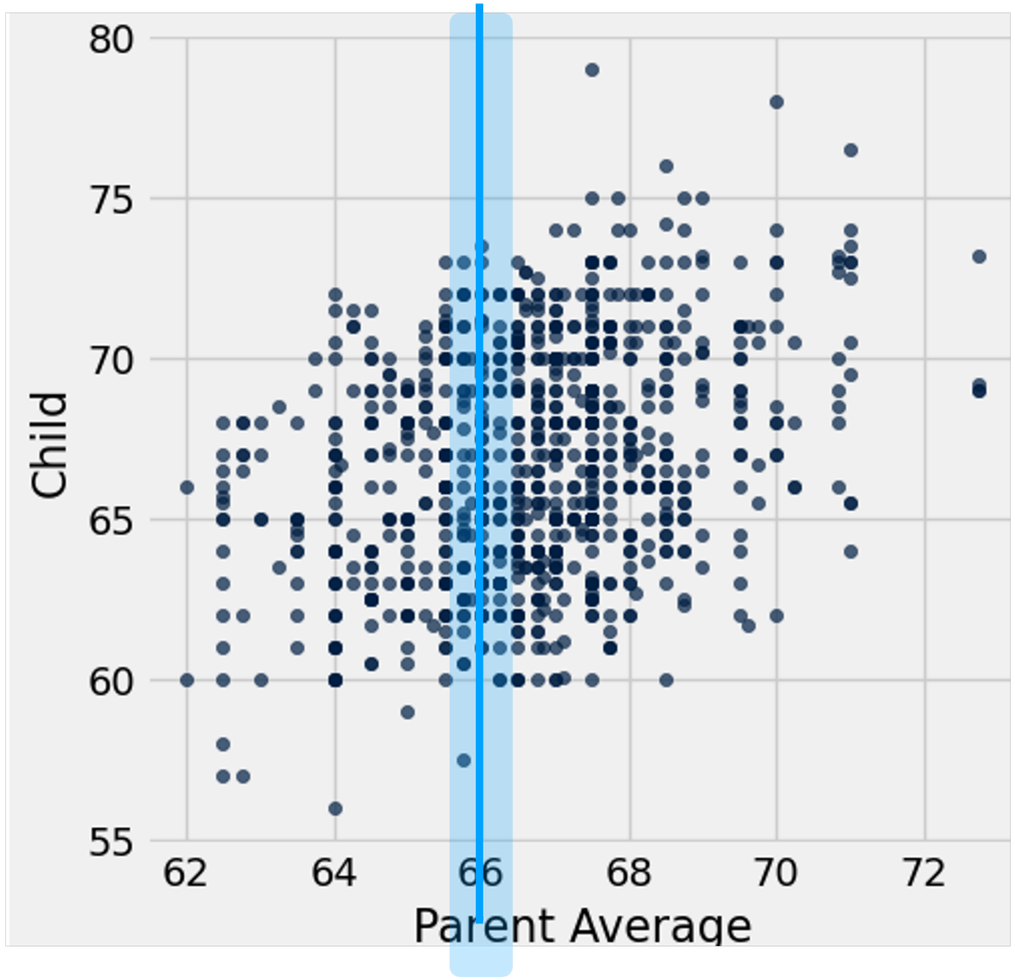


Notebook Demo: Child’s Height

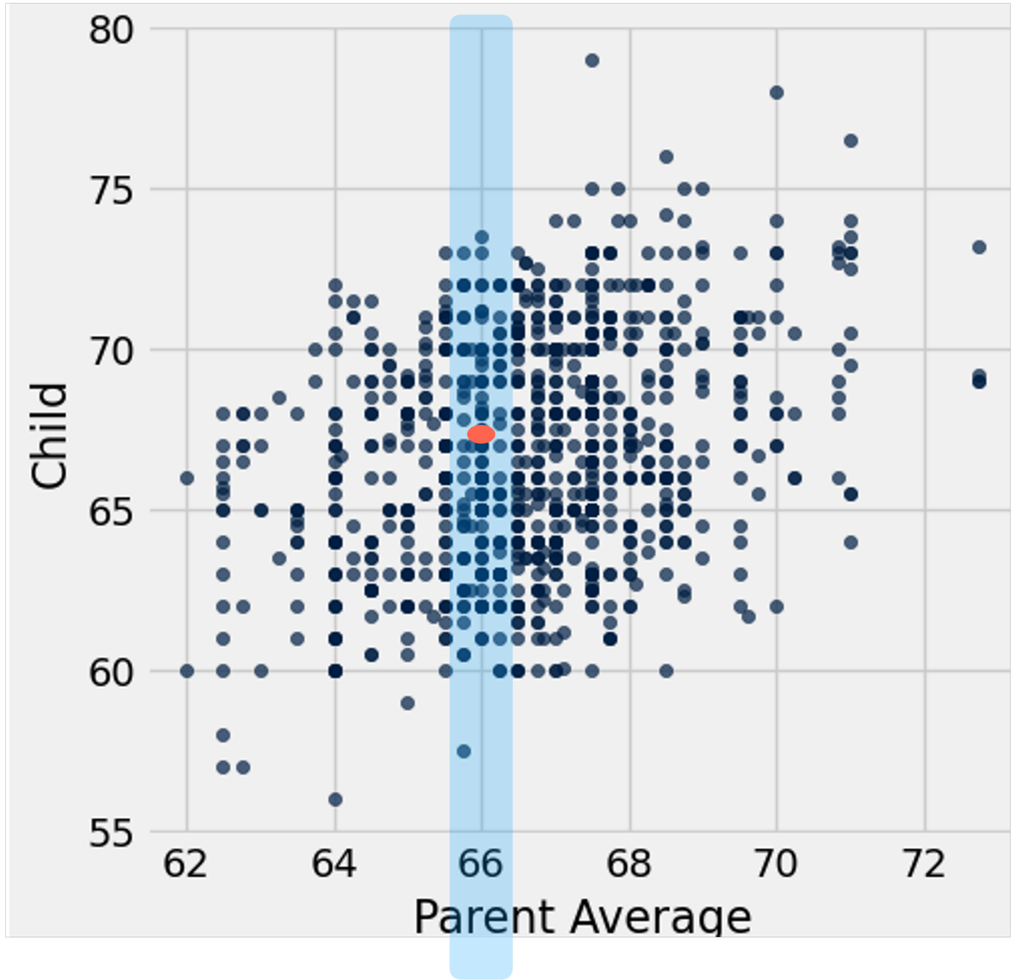
Graph of Average

- For each x value, the prediction is the average of the y values in its nearby group
    - The graph of these predictions is the **graph of averages**
- If the association between x and y is linear, then points in the graph of averages tend to fall on a line
    - This line is called the **regression line**

Graph of Average

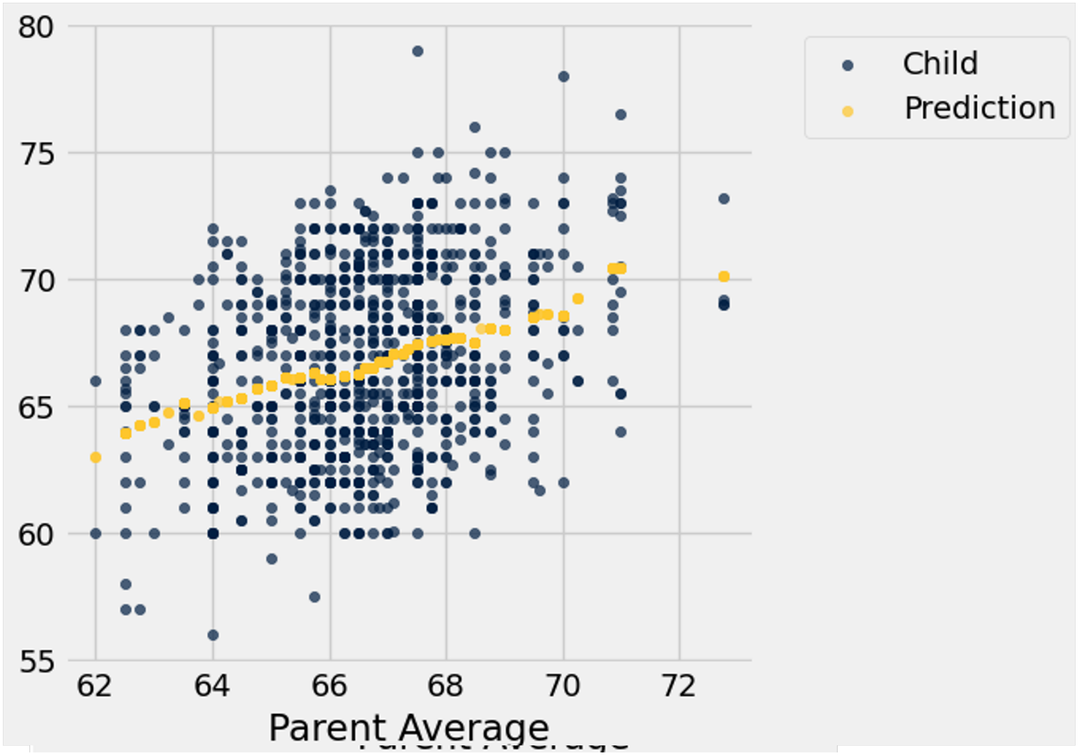


Graph of Average



Nearest Neighbor Regression

- A method for predicting a numerical y given a value of x:
  - Identify the group of points where the values of x are close to the given value
  - The prediction is the average of the y values for the group



Calculating r

Formalula for r

- The correlation coefficient (r) is the average product of x in standard units and y in standard units

  - To determine r, we first convert our values in x & y to standard units

  - We'll denote x_{su} and y_{su}, respectively

$$y_{\text{su}} = r \times x_{\text{su}}$$

This is known as the **linear regression line**

Standard Units

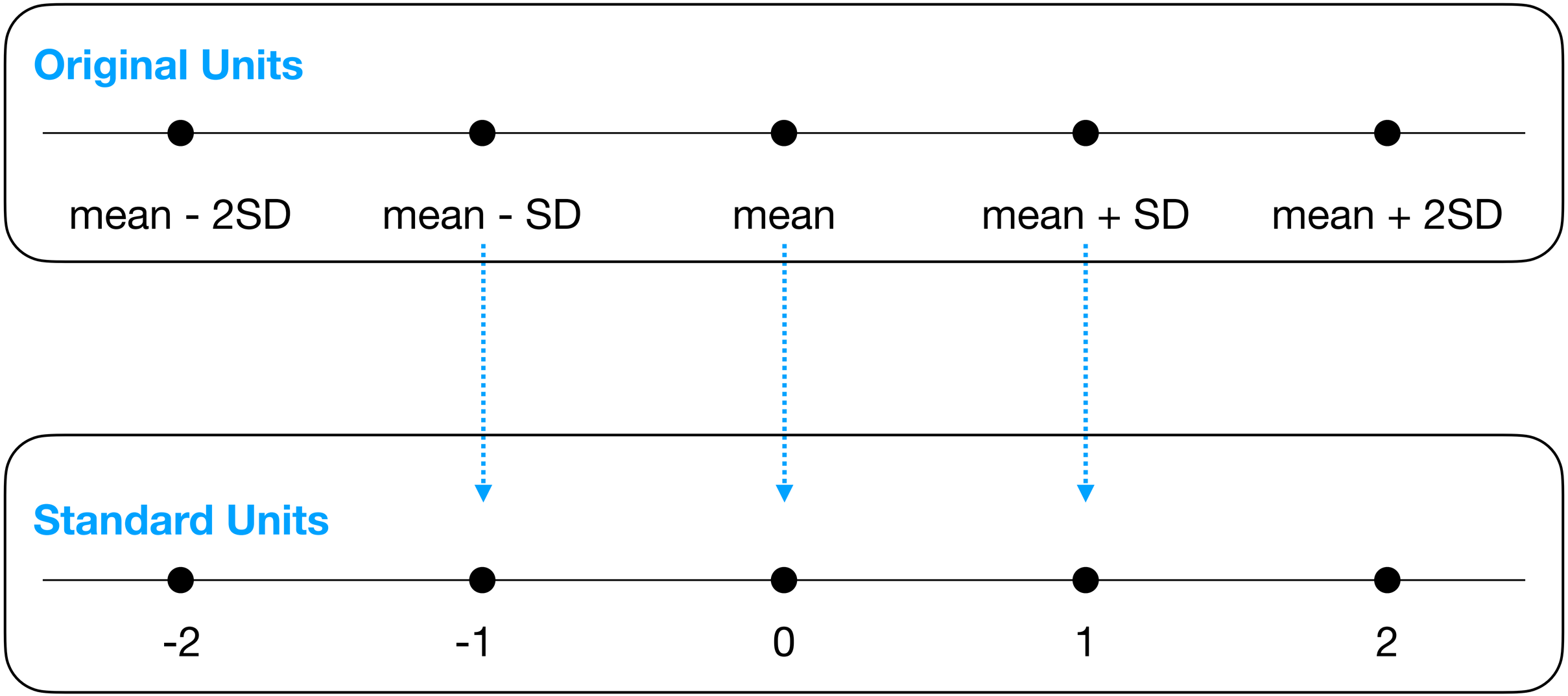
- The quantity z (from "$\text{average} \pm z\ \text{SDs}$" in Chebychev's inequality) measures **standard units**

  - **Standard units** is the number of standard deviations away from the average

- To convert a value (v) to standard units, compare the deviation from the average (μ) with the standard deviation (SD):

$$z = \frac{v - \mu}{\text{SD}}$$

Converting to Standard Units



Correlation Coefficient r

- r is the average of the product of two variables, when both variables are measured in standard units

- What this means for us:

  - r is not affected by changing the units of the measurement of the data

  - r will be the same regardless of which variable is plotted on the x- and y- axes

Regression Line: Slope & Intercept

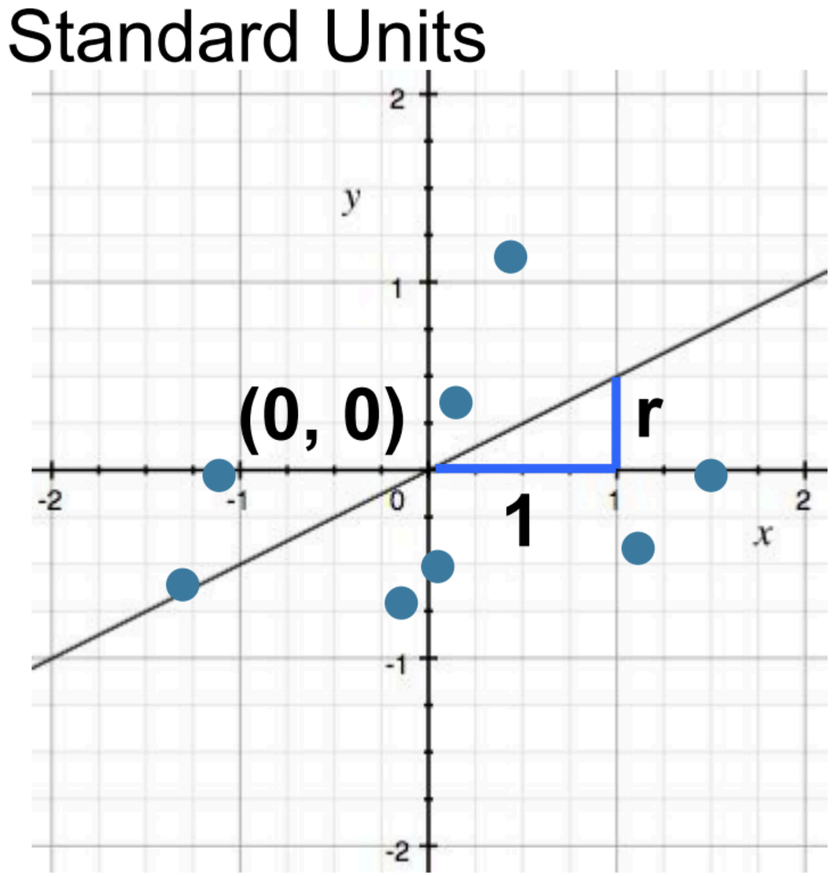
Regression Line

Lines generally can be expressed by

$$y = \text{slope} \times x + \text{intercept}$$

When in standard units, this is equivalent to

$$y_{su} = r \times x_{su}$$



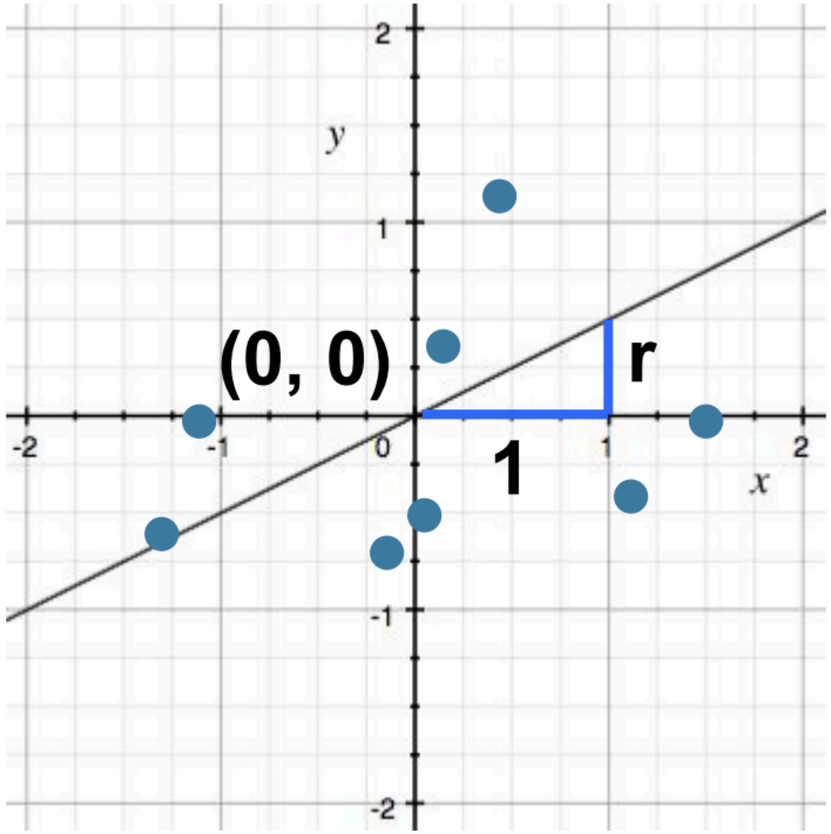
Regression Line: Original Units

$$y_{\text{su}} = r \times x_{\text{su}}$$

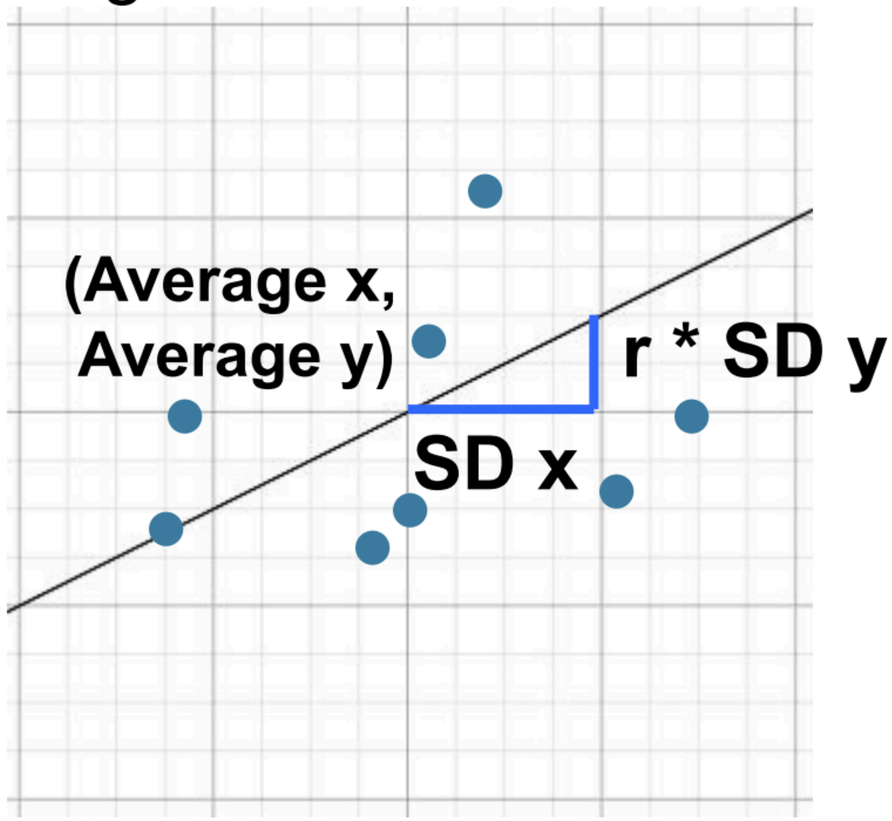
$$\frac{\text{estimate of } y - \text{avg}(y)}{\text{SD of } y} = r \times \frac{x - \text{avg}(x)}{\text{SD of } x}$$

Regression Line: Original Units

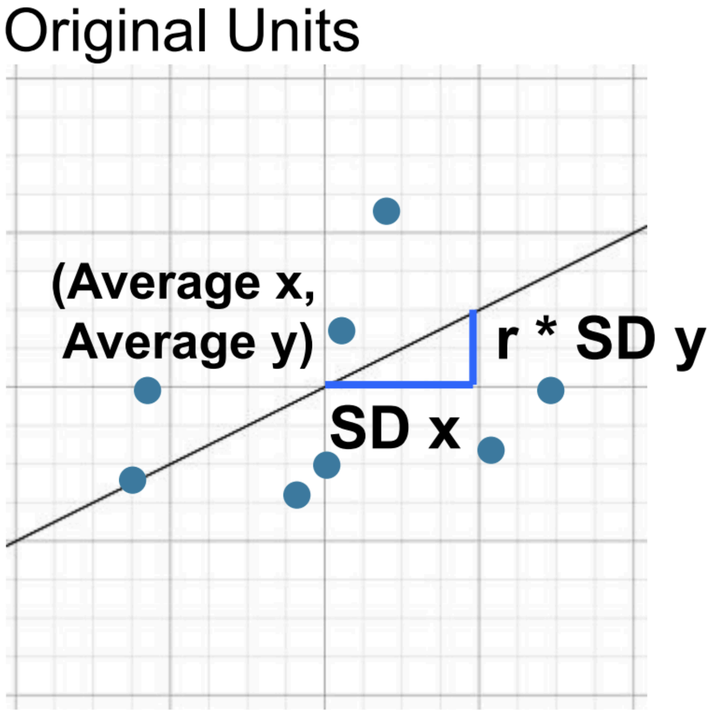
Standard Units



Original Units



Regression Line: Original Units



$$\text{estimate of } y = \text{slope} \times x + \text{intercept}$$

$$\text{slope} = r \times \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept} = \text{avg}(y) - \text{slope} \times \text{avg}(x)$$

Notebook Demo: Regression Line for Height

Next time

- **Today**: Correlation and Linear Regression
- **Monday, Nov 24**: Least Squares and Residuals ← HW 7 due
- ***Wednesday, Nov 26****: Holiday!*
- **Monday, Dec 1**: Regression Inference ← HW 8 due
  Progress Report due
- **Wednesday, Dec 3**: Special Topics (Data Ethics)
  - Final Project Consultations during Lab
- **Monday, Dec 8**: Special Topics (Data Privacy) ← HW 9 due