COMS BC1016

Introduction to Computational Thinking and Data Science

Lecture 20: Correlation and Linear Regression

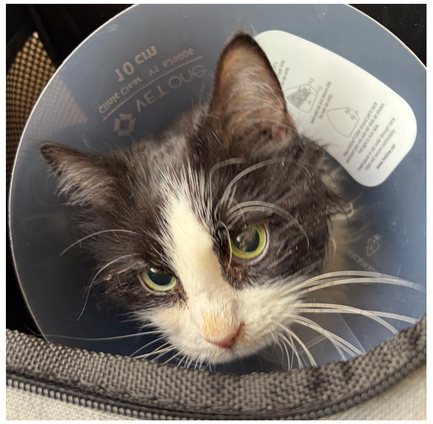
April 13, 2026

Logistics

- **Extra credit opportunity**: Attend Olive's talk Thursday, April 16 at 12pm in Milstein 402
  - Olive Franzese-McLaughlin will be giving a talk titled "Cryptographically Verified AI/ML Audits"
  - Worth 5% on the lowest non-dropped homework (comes out to 0.25% of the final grade)
- Submit your Final Project Proposals by Friday on Gradescope as a group
  - **All of your final project components should be submitted as a Python notebook (.ipynb)**

Upcoming office hours

- My office hours this week will be on Wednesday 3pm-5pm
  - Subject to no more Gertrude emergencies
- I will not have office hours next week while I am traveling
- TAs and computing fellows are expected to have office hours as normal
  - See EdStem for any updates



Gertrude’s surgery went well!

11	4/06	Inference		HW 6 - Confidence Intervals
	4/08		Lab 8 - Normal Distribution and Variability of Sample Means	HW 5 Due 11:59PM
	4/13	Prediction		HW 7 - Sample Sizes and Confidence Intervals
12	4/15		Lab 9 - Regression	HW 6 Due 11:59PM **Final Project Proposals** Due Friday, 4/17
13	4/20	Special Topics: Data Ethics (Computing Fellows Workshop)		HW 8 - Linear Regression
	4/22	Special Topics: Bias in AI (Guest lecture by Murad Megjhani)	Lab: Final Project Work Time	HW 7 Due 11:59PM
14	4/27	Prediction		**Final Project Progress Reports** HW 9 - Regression Inference (Optional, Due 5/4)
	4/29	(Either special topics or prediction)	Lab: Final Project Consultations	HW 8 Due 11:59PM
15	5/04	Wrap up!		HW 9 Due 11:59PM (Optional) **Final Project Report** Due Friday 5/8

Last Time: Central Limit Theorem

Central Limit Theorem

Definition:

If a sample is large and drawn at random with replacement

Then regardless of the distribution,

The **probability distribution of the sample average** is roughly normal

Central Limit Theorem for Sample Mean

Definition:

If you draw a large random sample with replacement from a population, then, regardless of the distribution of the population,

the **probability distribution of the sample mean** is roughly normal,

centered at the population mean, with

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Central Limit Theorem for Sample Mean

Definition:

roughly how far off sample means are from the population mean

sample with replacem

bution of the p

Notice this side only depends on sample size and population SD…
Population size does not affect sample mean!

the **probab y distribution of the sampl**

centered at th population mean, with

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Central Limit Theorem for Sample Mean

Definition:

If you [illegible] replacement from a population,
[illegible]tion,
[illegible] is roughly normal,
centered at the population mean,

Population SD is a **constant**.

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Central Limit Theorem for Sample Mean

Definition:

If you [illegible] replacement from a population,
[illegible]tion,
[illegible] is roughly normal,
centered at the population mean,

Population SD is a **constant**.
The only thing affecting sample mean SD is sample size

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Central Limit Theorem for Sample Mean

Definition:

If you draw a large random sample with replacement from a population, then, regardless of the distribution of the population,

the **probability distribution of the sample mean** is roughly normal,

centered at the population mean, with

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Central Limit Theorem for Sample Mean

Def

We have better bounds on proportion of data that fall within $\pm z$ SDs for normal distributions

ent from a population,

ation,

the **probability distri on of the sample mean** is roughly normal,

centered at the populatio mean, with

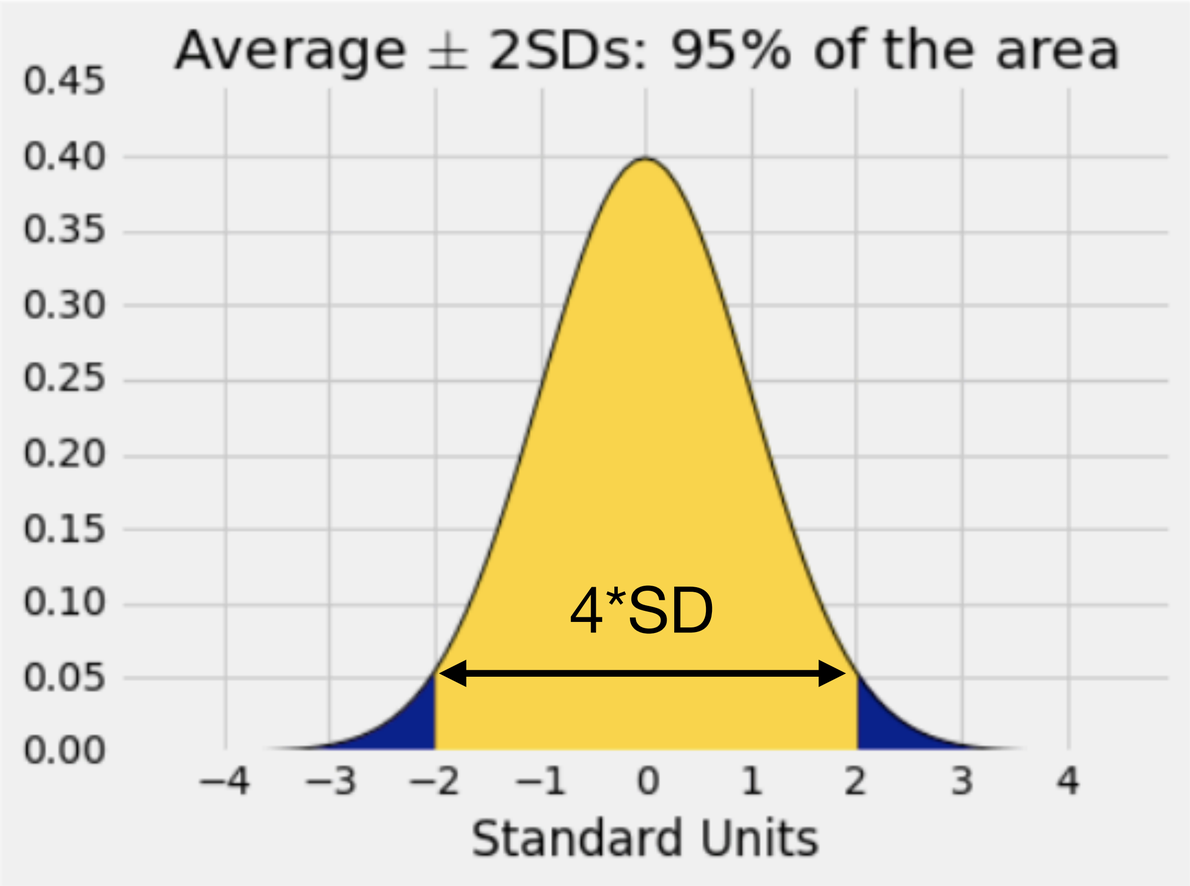
$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Normal vs All Distributions

Range	All Distributions (Chebyshev's)	Normal Distribution
mean $\pm$ 1 SDs	At least 0%	At least 68%
mean $\pm$ 2 SDs	At least 75%	At least 95%
mean $\pm$ 3 SDs	At least 89%	At least 99%

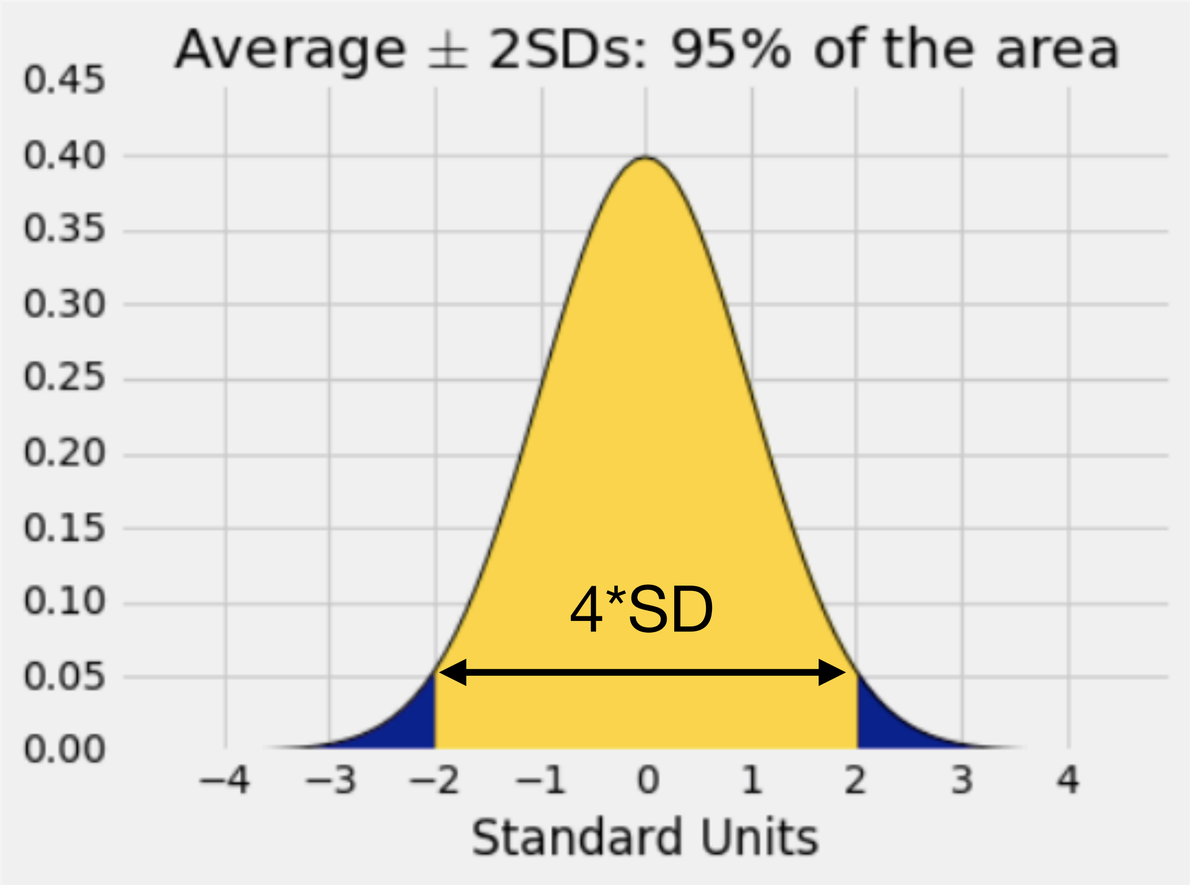
Connecting to confidence intervals

For a normal distribution, 95% of the data is in the range average + 2SDs



Connecting to confidence intervals

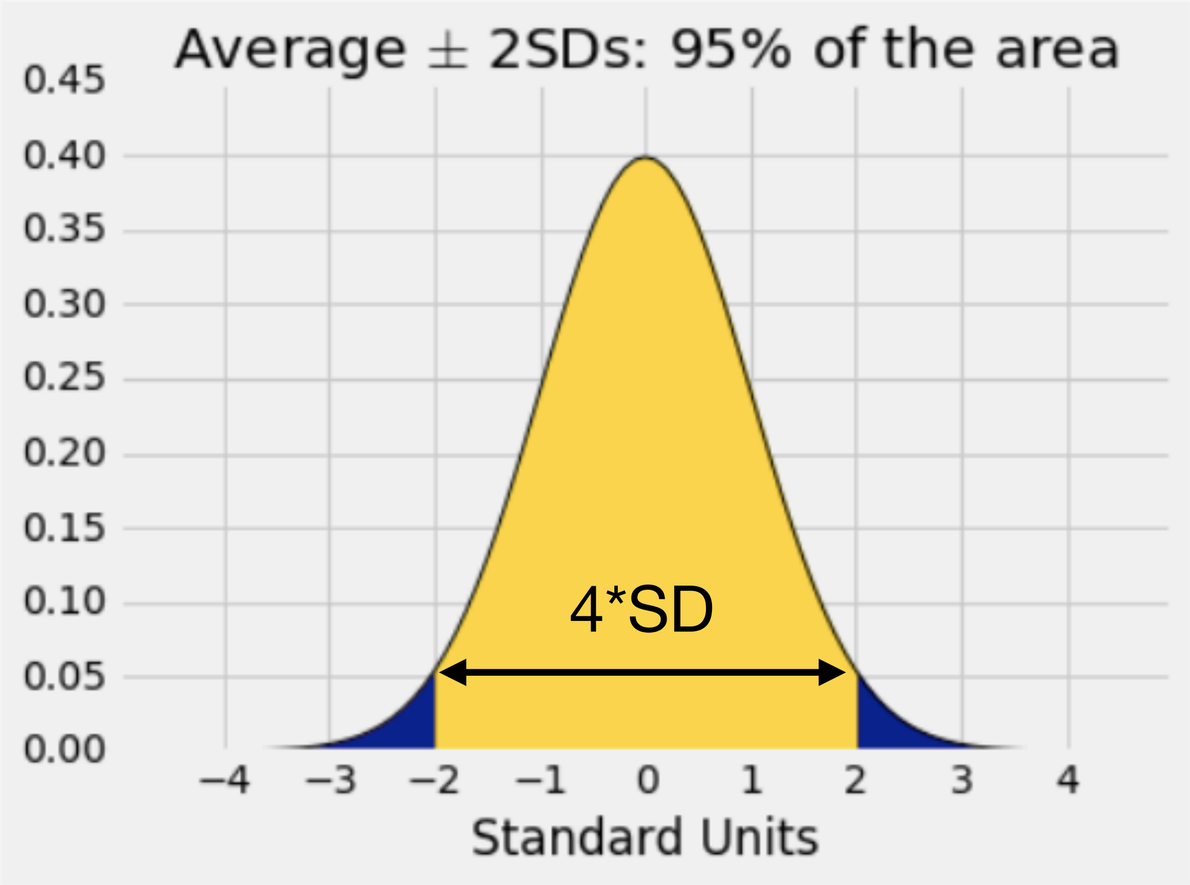
For a normal distribution, 95% of the data is in the range average + 2SDs



$$\text{sample means SD} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Connecting to confidence intervals

For a normal distribution, 95% of the data is in the range average + 2SDs



$$\text{sample means SD} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

$$95\,\%\ \ \text{CI} = 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Bounding the Width of a CI?

- Suppose we want to use bootstrap to construct a 95% confidence interval but we want this interval to be fairly narrow.

- How large should our sample be?

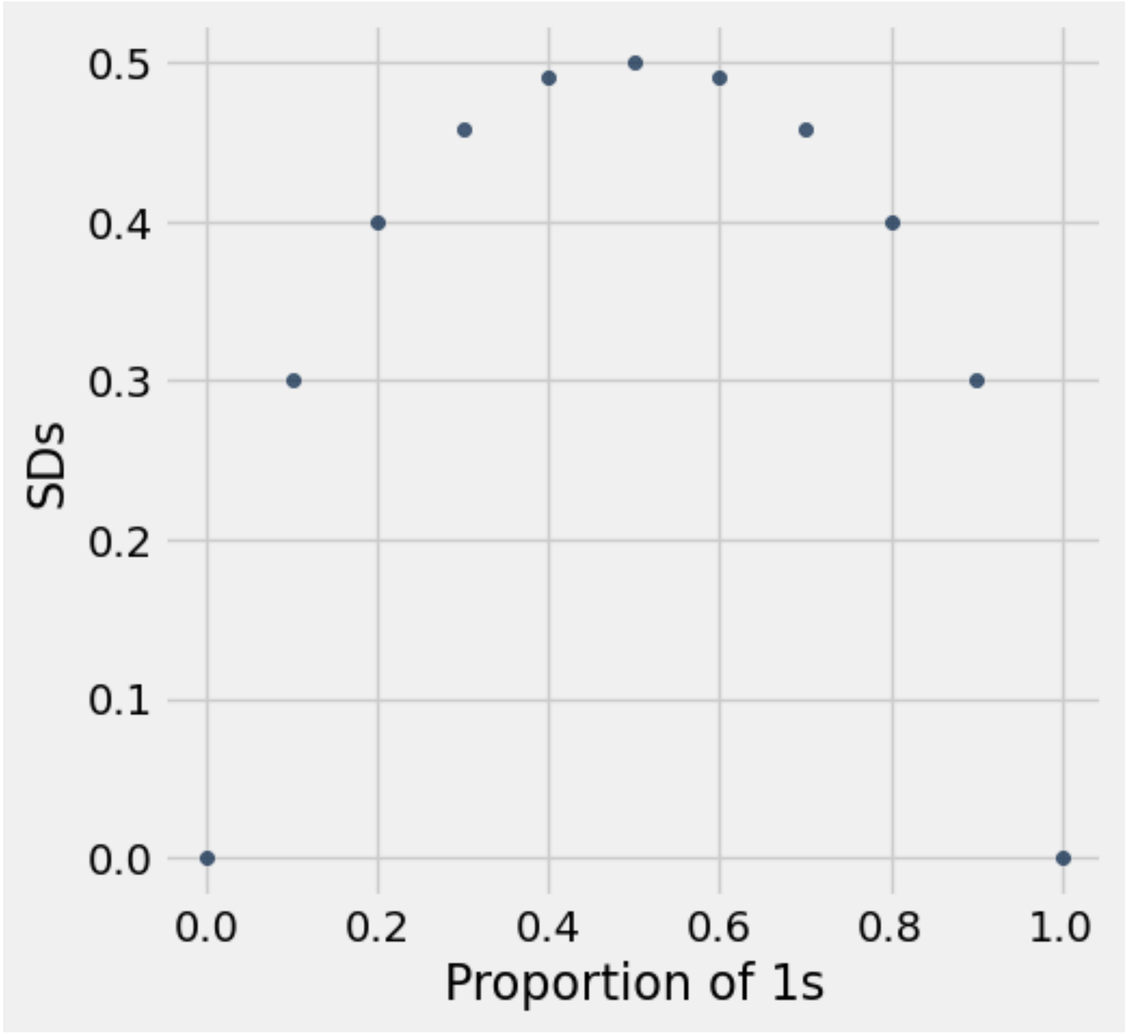
$$\text{sample means SD} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

$$95\,\%\ \text{CI} = 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Population SD for Situations with 2 Outcomes

- For situations with only 2 outcomes, the SD ranges from 0 to 0.5, with a max value of 0.5

- Thus, to estimate worst case scenario (most conservative sample size needed), you can use the maximum SD=0.5



Bounding the Width of a CI?

- Suppose we want to use bootstrap to construct a 95% confidence interval but we want this interval to be fairly narrow.

- How large should our sample be?

- Population SD is at most 0.5

  - Can bound the sample size

$$95\%\ \text{CI} = 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Bounding the Width of a CI?

- Suppose we want to use bootstrap to construct a 95% confidence interval but we want this interval to be fairly narrow.

- How large should our sample be?

- Population SD is at most 0.5

  - Can bound the sample size

$$95\%\ \text{CI} = 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

$$95\%\ \text{CI} \geq 4 * \frac{0.5}{\sqrt{\text{sample size}}}$$

Bounding the Width of a CI?

- Suppose we want to use bootstrap to construct a 95% confidence interval but we want this interval to be fairly narrow.

- How large should our sample be?

- Population SD is at most 0.5

  - Can bound the sample size

$$95\%\ \text{CI} = 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

$$95\%\ \text{CI} \geq 4 * \frac{0.5}{\sqrt{\text{sample size}}}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{95\%\ \text{CI}}$$

Bounding the Width of a CI?

- Suppose we want to use bootstrap to construct a 95% confidence interval but we want this interval to be fairly narrow.

- How large should our sample be?

- Population SD is at most 0.5

  - Can bound the sample size

$$95\%\ \text{CI} = 4 * \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

$$95\%\ \text{CI} \geq 4 * \frac{0.5}{\sqrt{\text{sample size}}}$$

$$\sqrt{\text{sample size}} \geq 4 * \frac{0.5}{95\%\ \text{CI}}$$

$$\text{sample size} \geq \left(4 * \frac{0.5}{95\%\ \text{CI}}\right)^2$$

Today’s Lecture

- Correlation
  - Correlation coefficient
- Linear Regression

Correlation

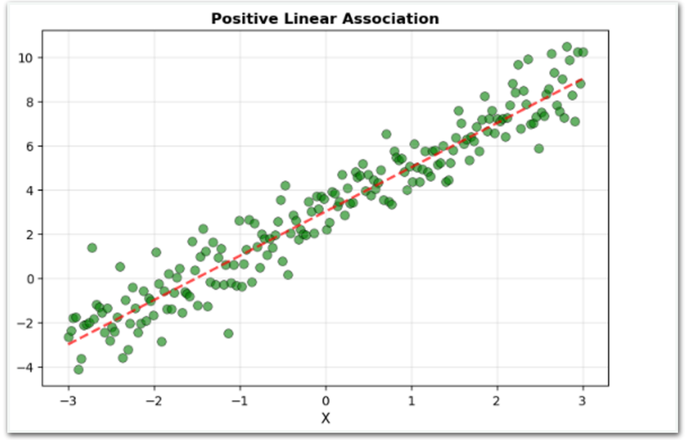
Prediction

- 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

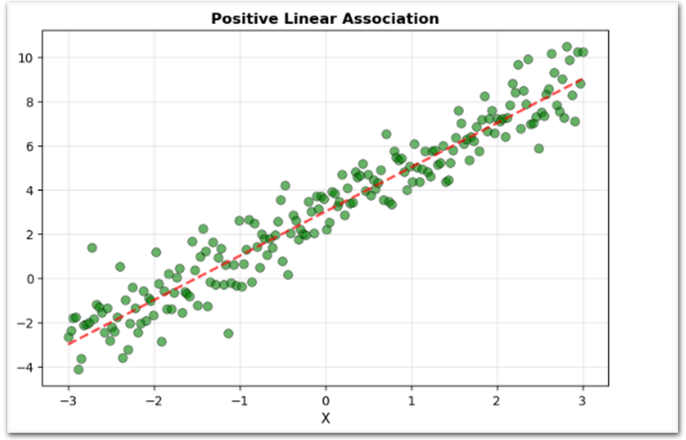
- Linear
- Non-linear

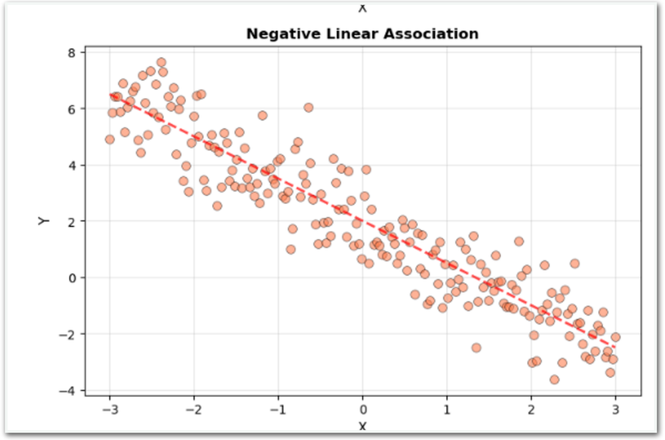
Visualize, then quantify

Two Numerical Variables

Trend

- Positive association
- Negative association
- Pattern





Any discernible “shape” in the scatter

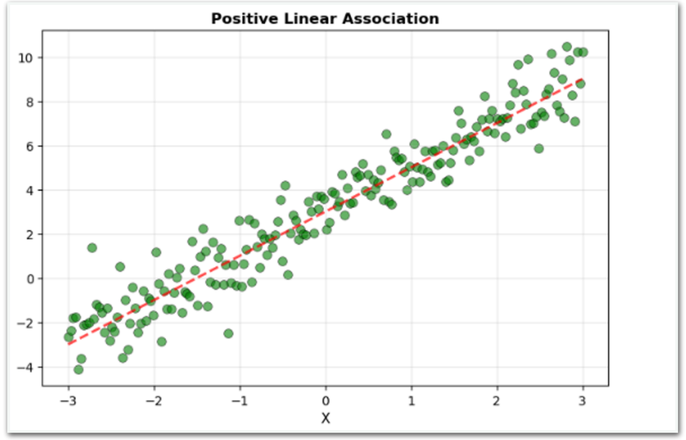
- Linear
- Non-linear

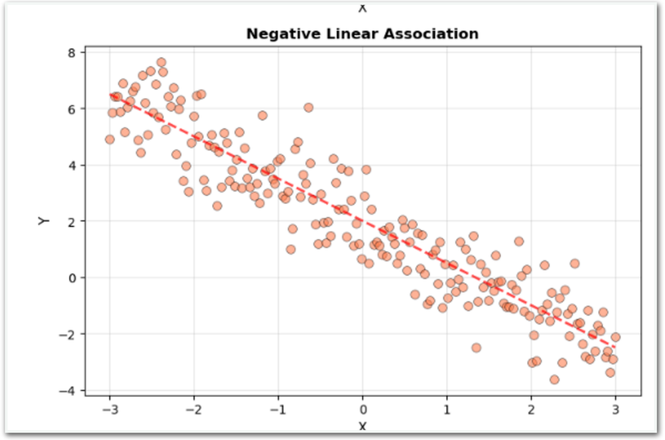
Visualize, then quantify

Two Numerical Variables

Trend

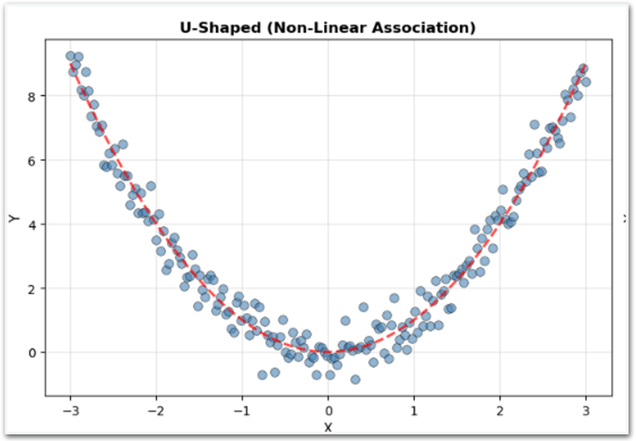
- Positive association
- Negative association
- Pattern





Any discernible “shape” in the scatter

- Linear
- Non-linear



Visualize, then quantify

Two Numerical Variables

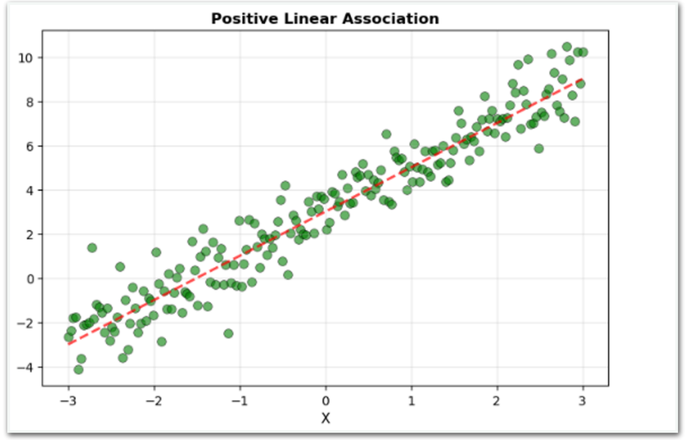
Trend

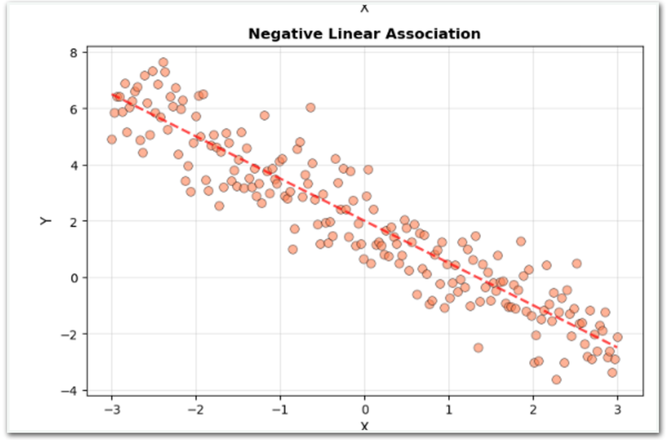
- Positive association
- Negative association
- Pattern

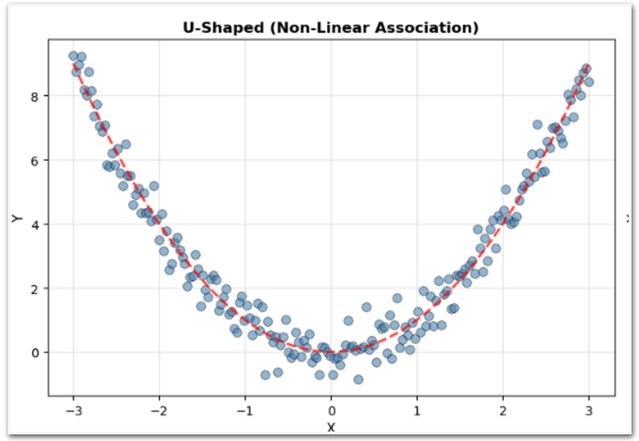
Any discernible “shape” in the scatter

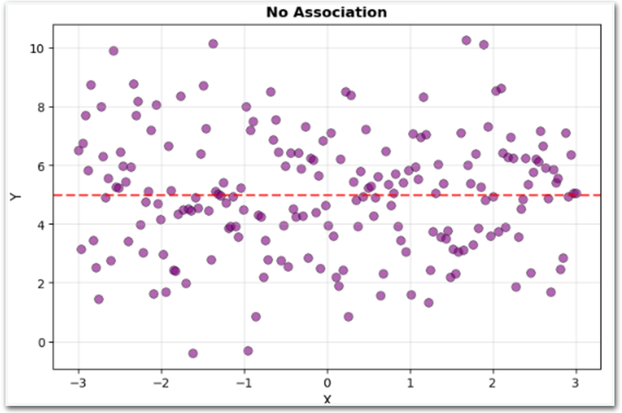
- Linear
- Non-linear

Visualize, then quantify









Prediction

Guess the future

- Based on incomplete information

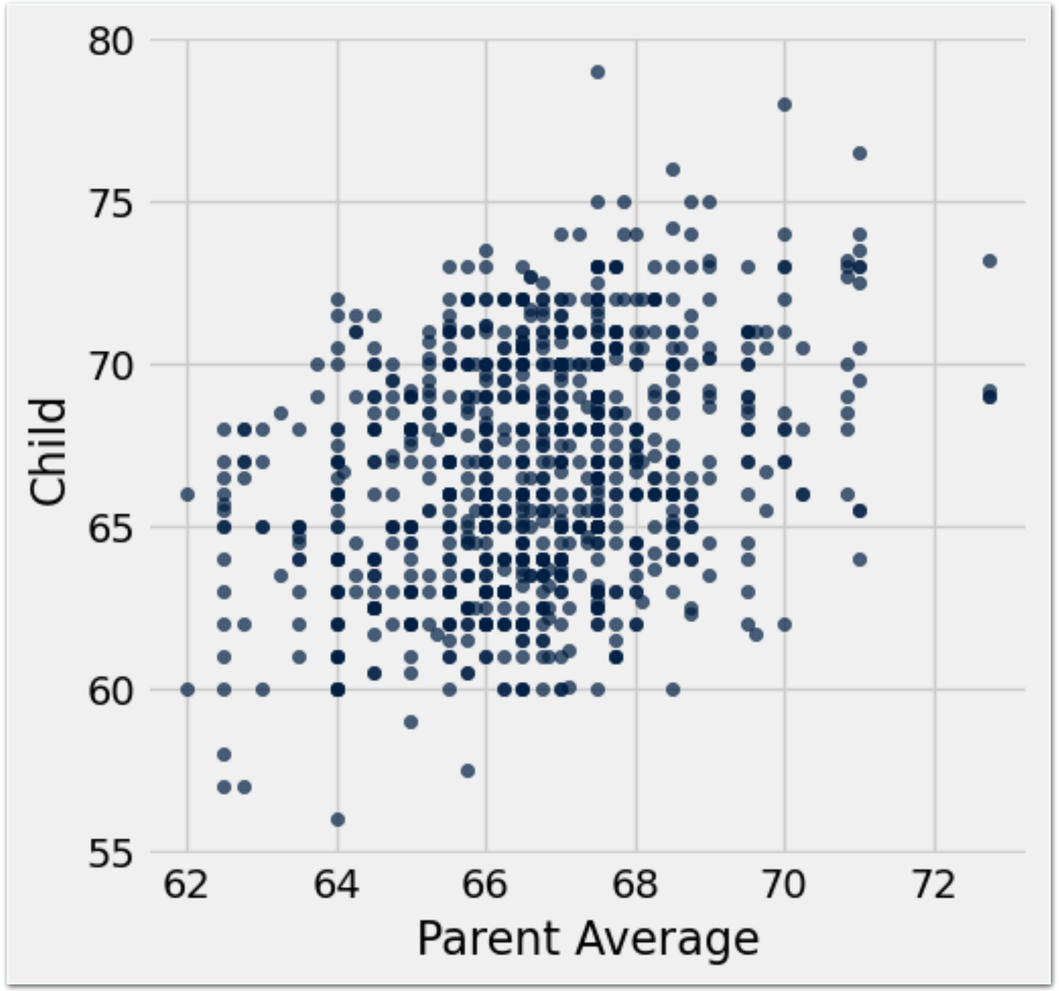
- One way of making predictions:

  - 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

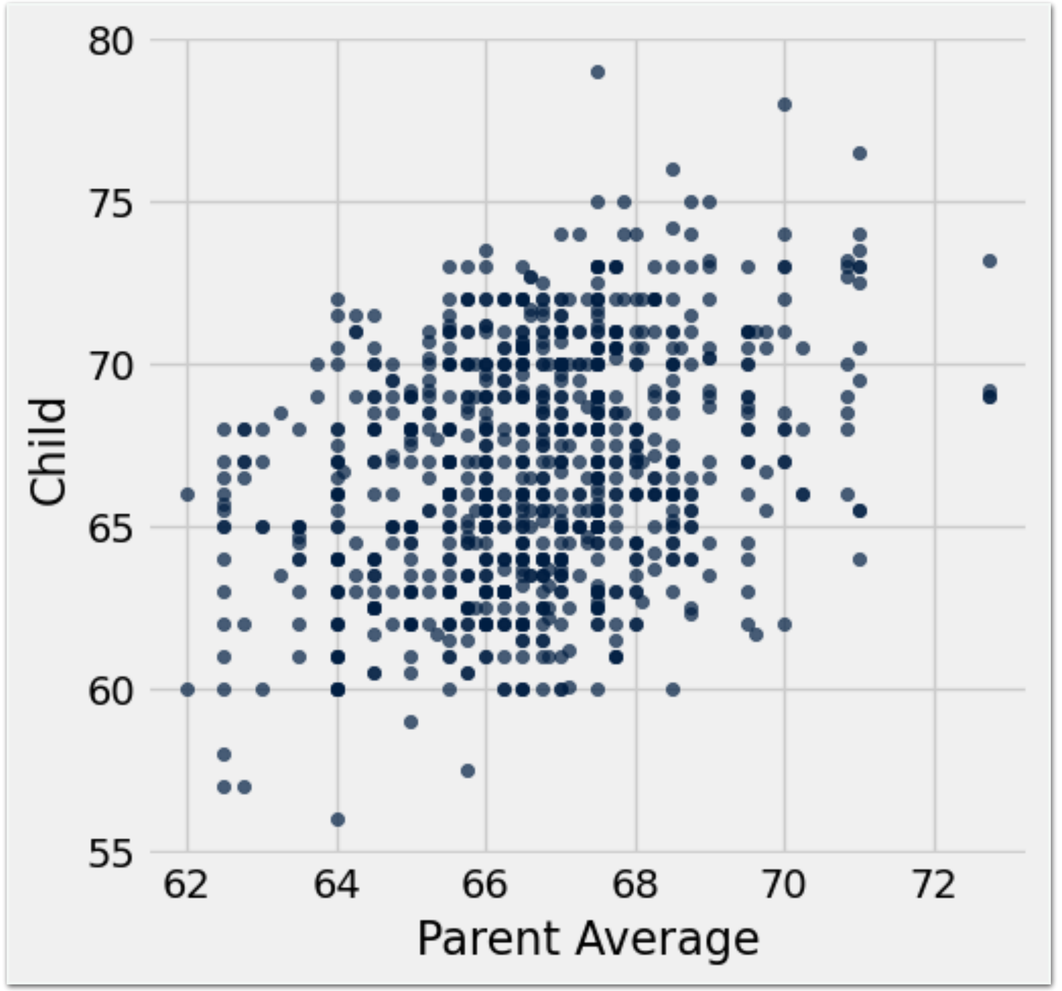
Goal: Predict the height of a new child based on that child's midparent height (average parent height)



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?

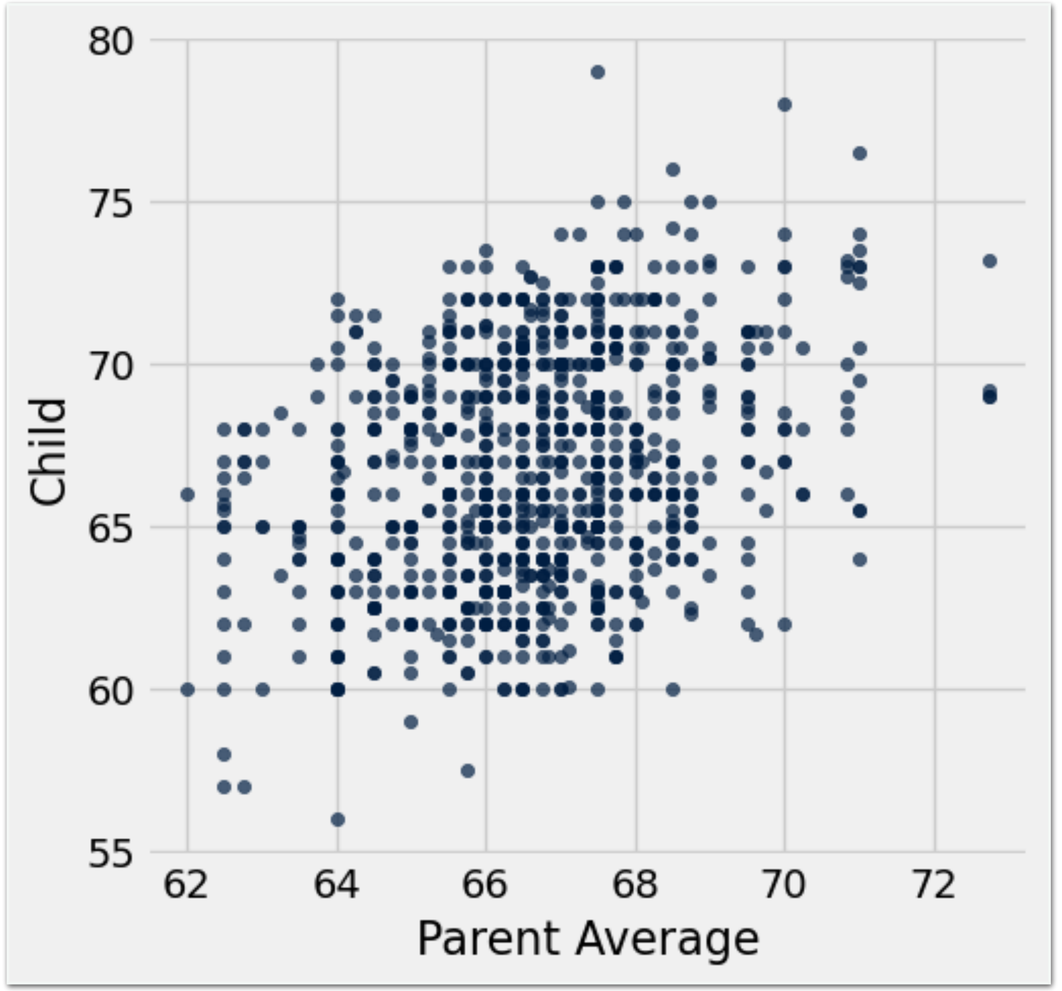


Example: Galton's Heights

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

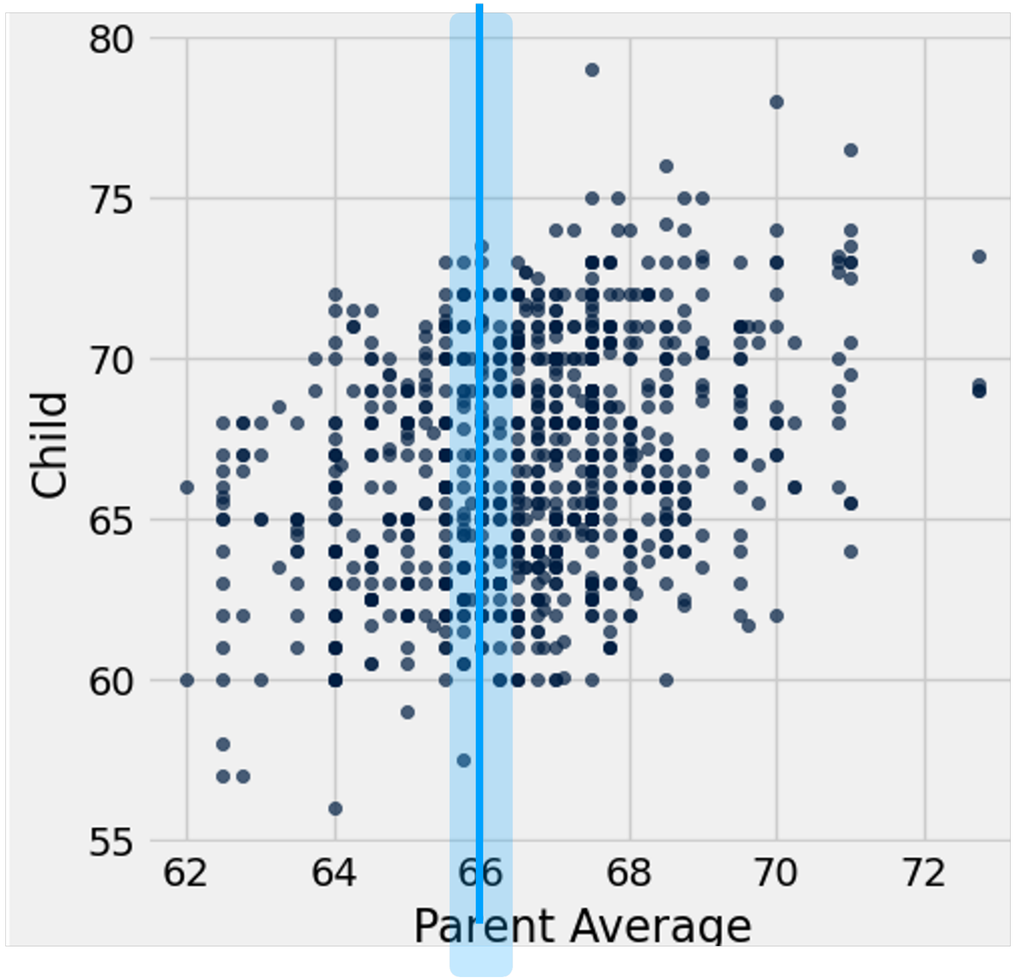


Notebook Demo: Child’s Height

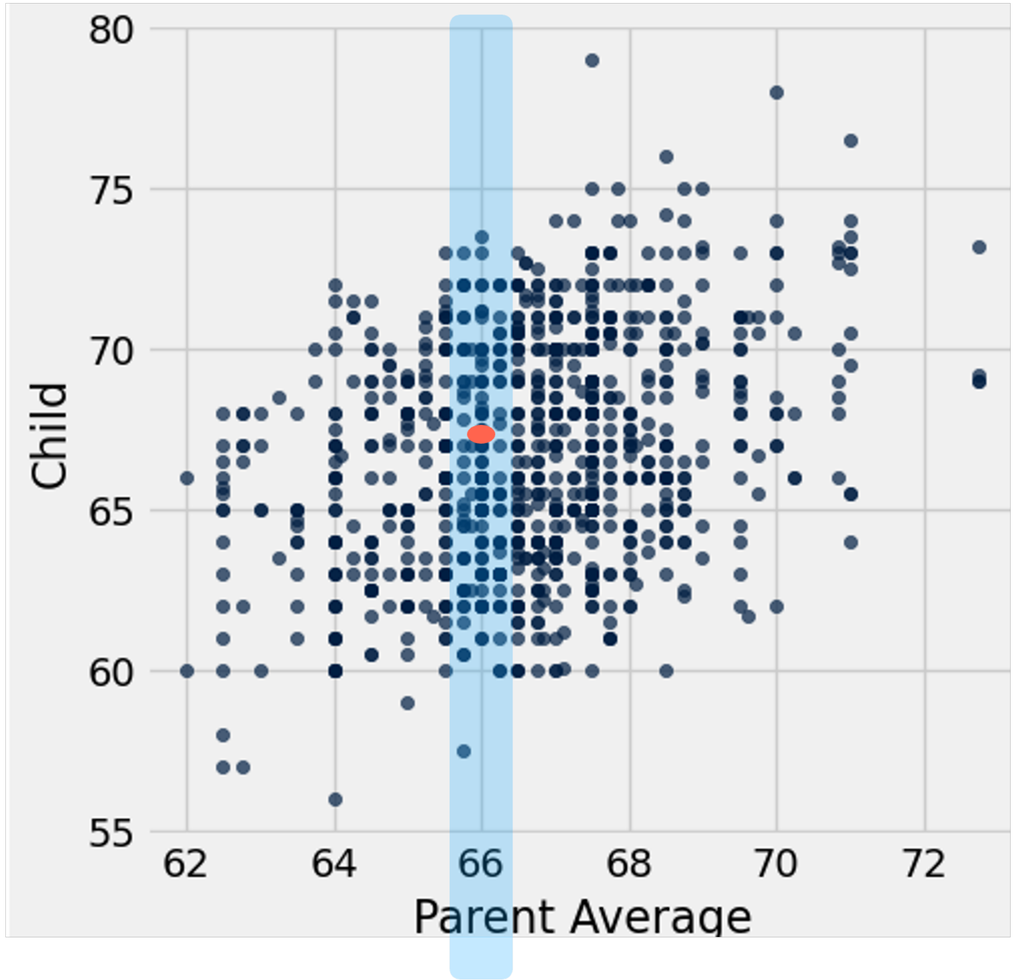
Graph of Average

- 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

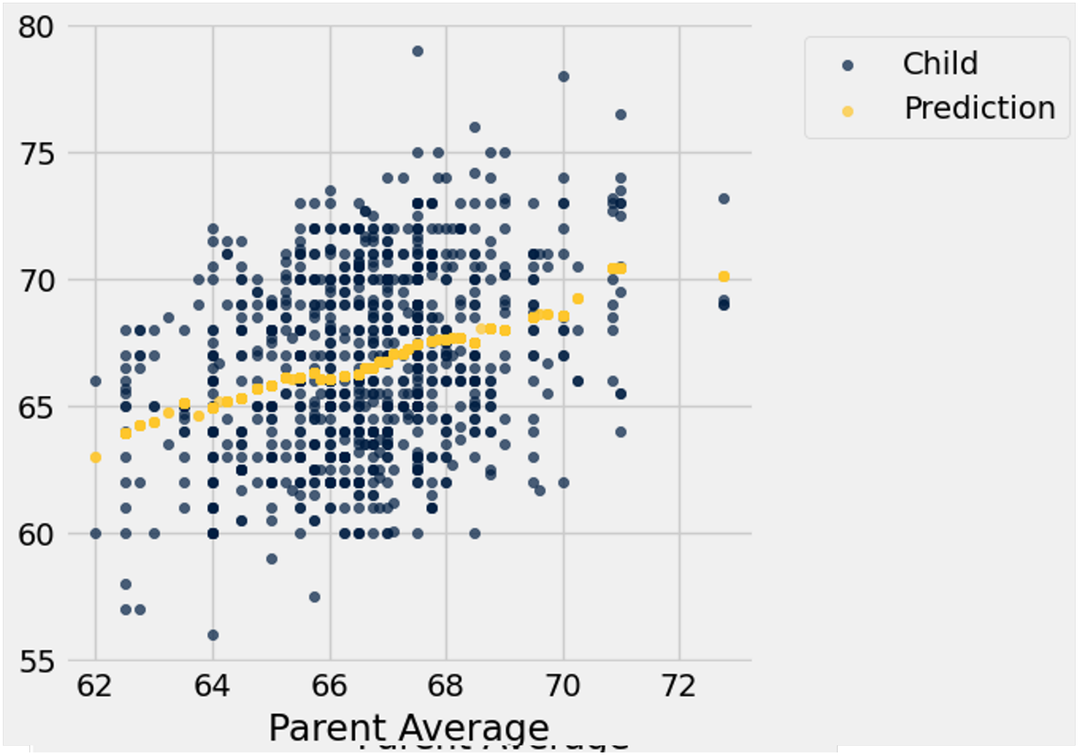


Graph of Average



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



Correlation Coefficient r

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

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

Notebook demo: Correlation Plots

Formula 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 the equation for the **linear regression line**

Recall: Standard Units

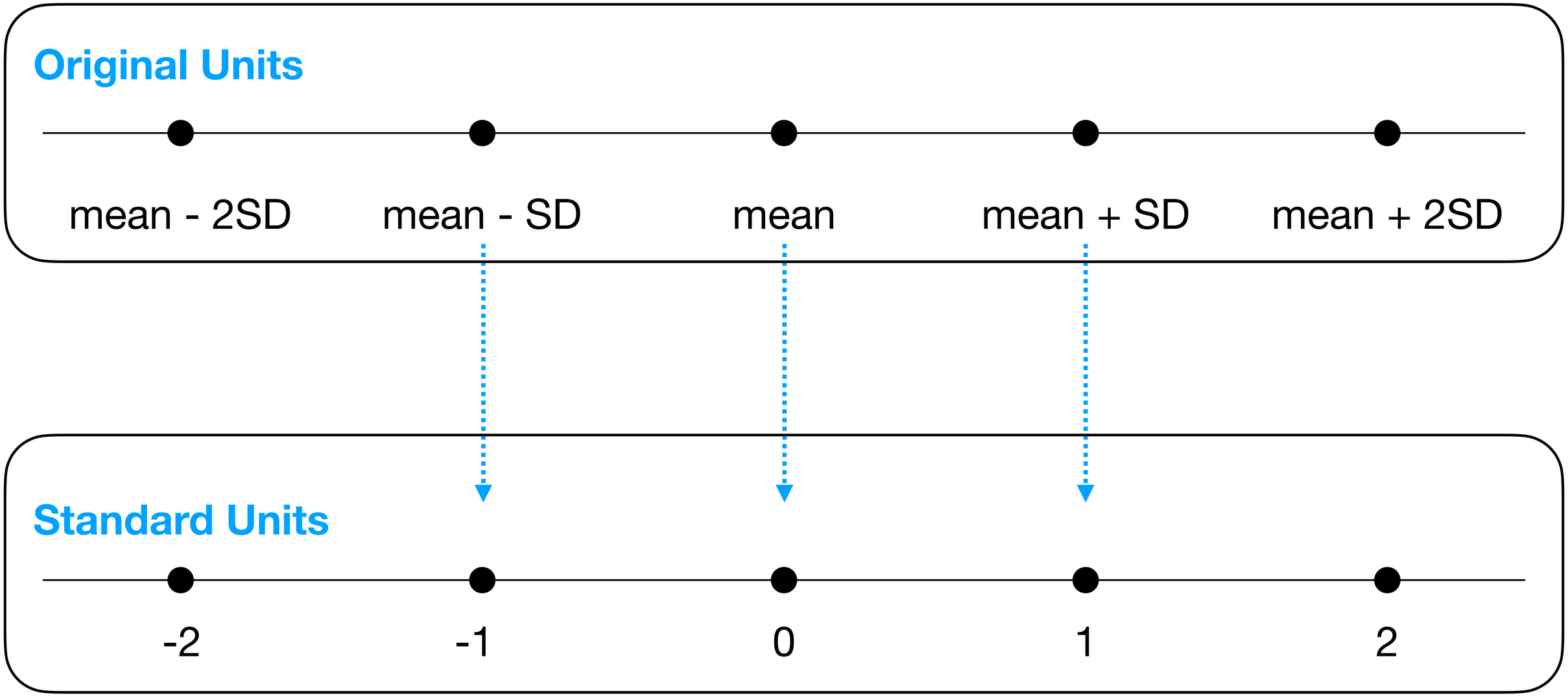
- The quantity z (from "$\text{average} \pm z\ \text{SDs}$" in Chebychev's inequality) measures **standard units**

  - **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):

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

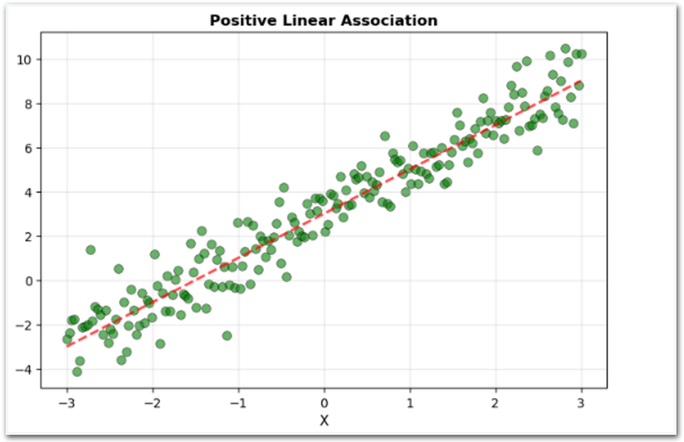
Recall: Converting to Standard Units

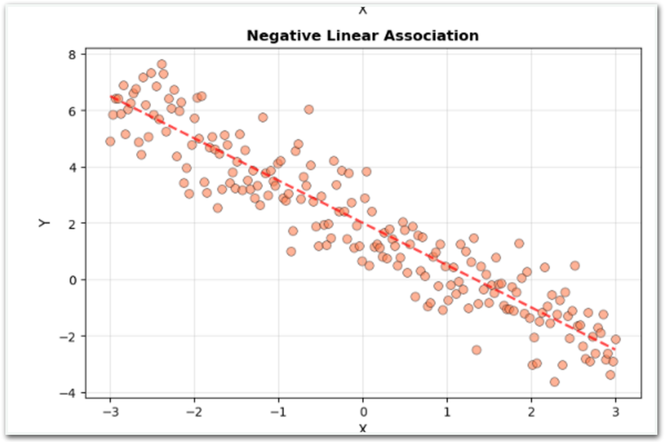


Notebook demo: Calculating r

Notes about using r

- r measures **linear** association
  - We'll see formally later on, but r is the slope of a straight line drawn through our scatter plot
  - Check that your scatter plot looks roughly linear before computing r





Notes about using r

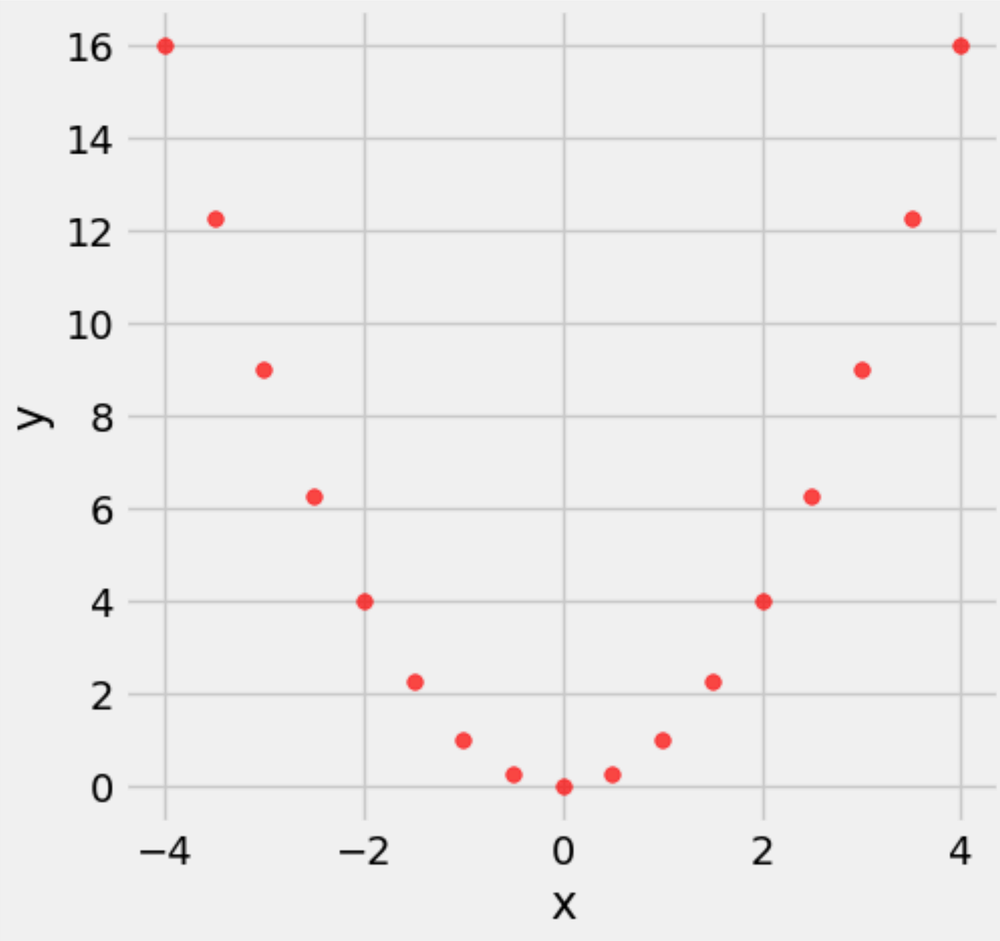
- Association does *not* imply causation
  - Two variables may be correlated, but one may not cause the other
  - Non-linearity and outliers can affect correlation
- Correlations based on aggregated data (**ecological correlations**) can be misleading
  - Correlations between individuals may be lower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

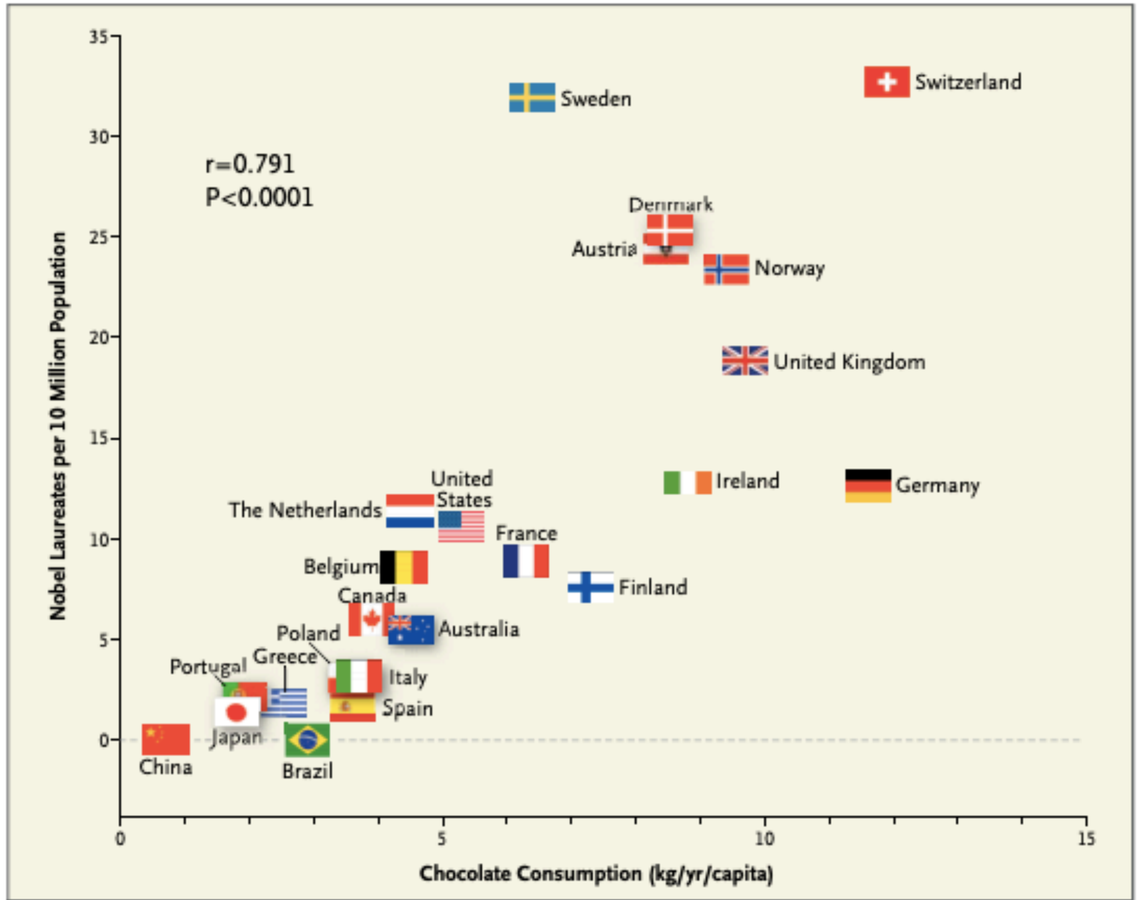


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

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

Regression Line: Slope & Intercept

Predicting Values with the Regression Line

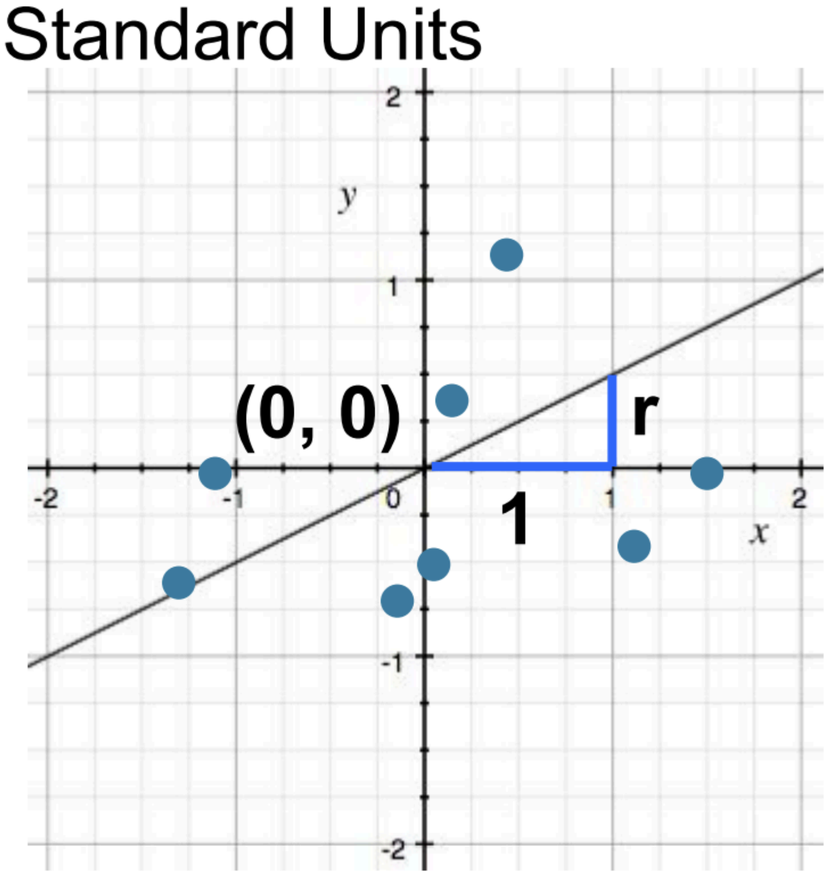
We know how to compute the regression line in standard units $y_{\text{su}} = r \times x_{\text{su}}$

Recall: Lines can generally be expressed by

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

How can we use this to predict values?

- We need to cover this back to original units!



Predicting Values with the Regression Line

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

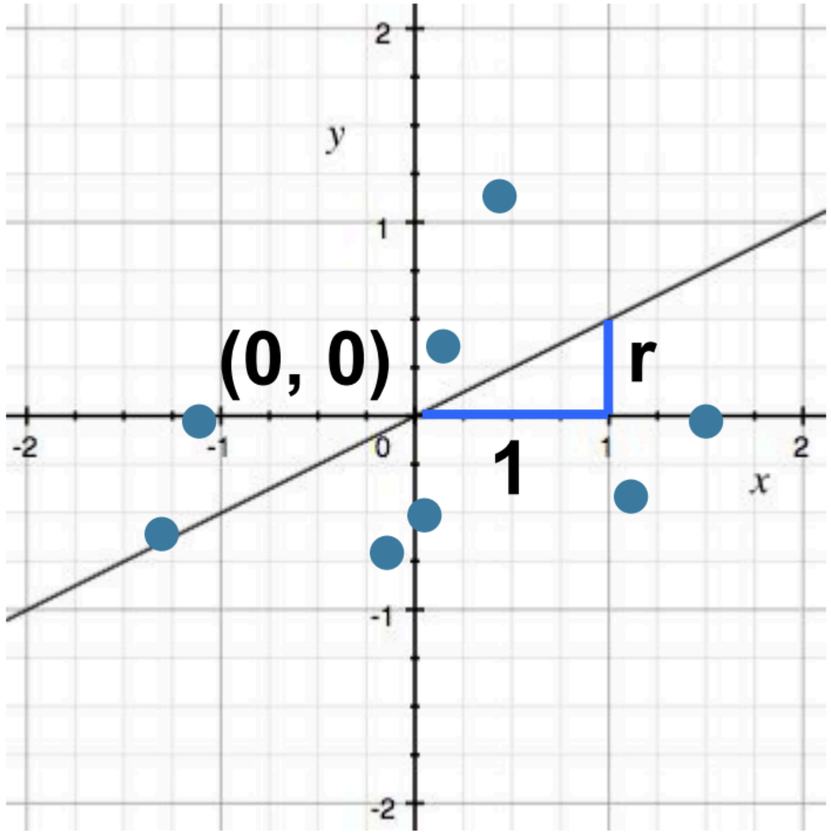
$$z = \frac{v - \mu}{\text{SD}}$$

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

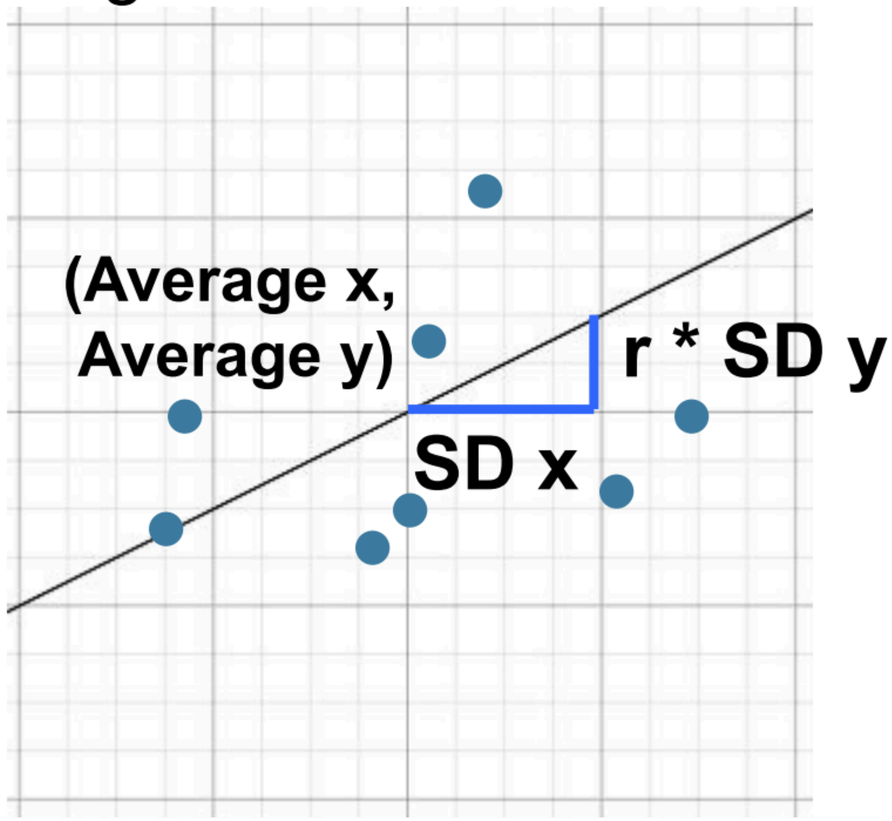
$$\frac{\text{estimate of } y - \text{avg}(y)}{\text{SD of } y} = r \times \frac{x - \text{avg}(x)}{\text{SD of } x}$$

Predicting Values with the Regression Line

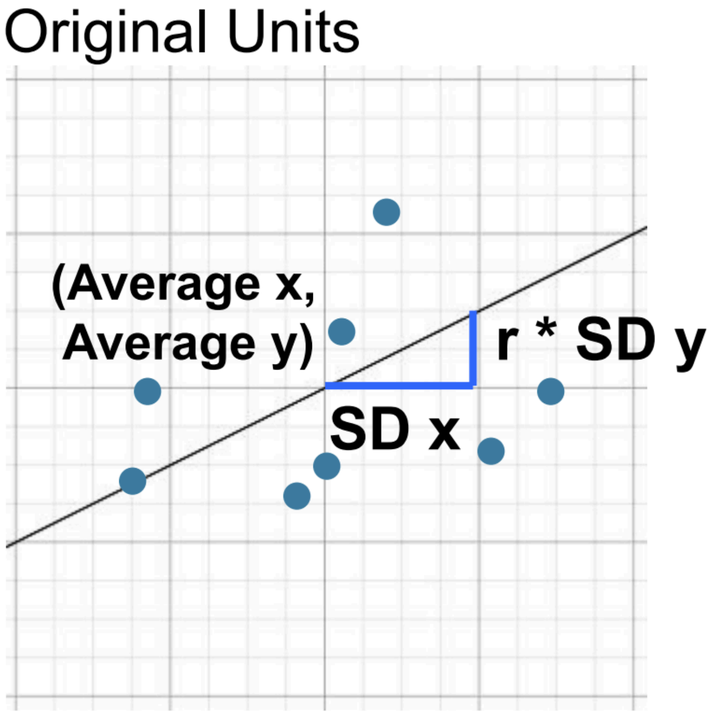
Standard Units



Original Units



Predicting Values with the Regression Line



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

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

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

Notebook Demo: Regression Line for Height

Next time

- Least Squares and Residuals