

UNIT 2: TWO VARIABLE DATA

WHAT IS OUR GOAL FOR UNIT 2?

- Representing Relationships Between Bivariate Categorical Data**
- Representing Relationships Between Bivariate Quantitative Data**

BIVARIATE CATEGORICAL DATA

- Is there a relationship between two Categorical Variables?
- We will represent relationships using **tables** (same as treat example before), **Graphs**, and **Statistics** (numbers)

BIVARIATE CATEGORICAL DATA

- Our Example: X: Shirt Colour ● ○ ■ , Y: Status ☠️ 😊

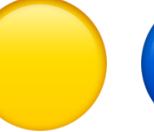


If George R.R. Martin wrote for Star Trek

Matthew Barsalou published an article in *Significance* that studies this from a statistical perspective



BIVARIATE CATEGORICAL DATA

- Our Example: X: Shirt Colour    , Y: Status  

Crew Member	Area	Shirt Color	Status
Brendan	Operations, Engineering and Security	Red 	DEAD 
Leif	Command And Helm	Gold 	DEAD 
Shailah	Science and Medical	Blue 	Alive 

Dataset is composed of 430 cremates

Enterprise NCC 1701 casualties from episodes aired between September 8, 1966 and June 03, 1969 based on casualty figures from Memory Alpha.

BIVARIATE CATEGORICAL DATA

- First we tabulate data into a **contingency table** (also known as a two way table)

		😊	💀	
🔵	😊	129	7	136
	💀	46	9	55
🟡	215	24	239	
🔴	390	40	430	

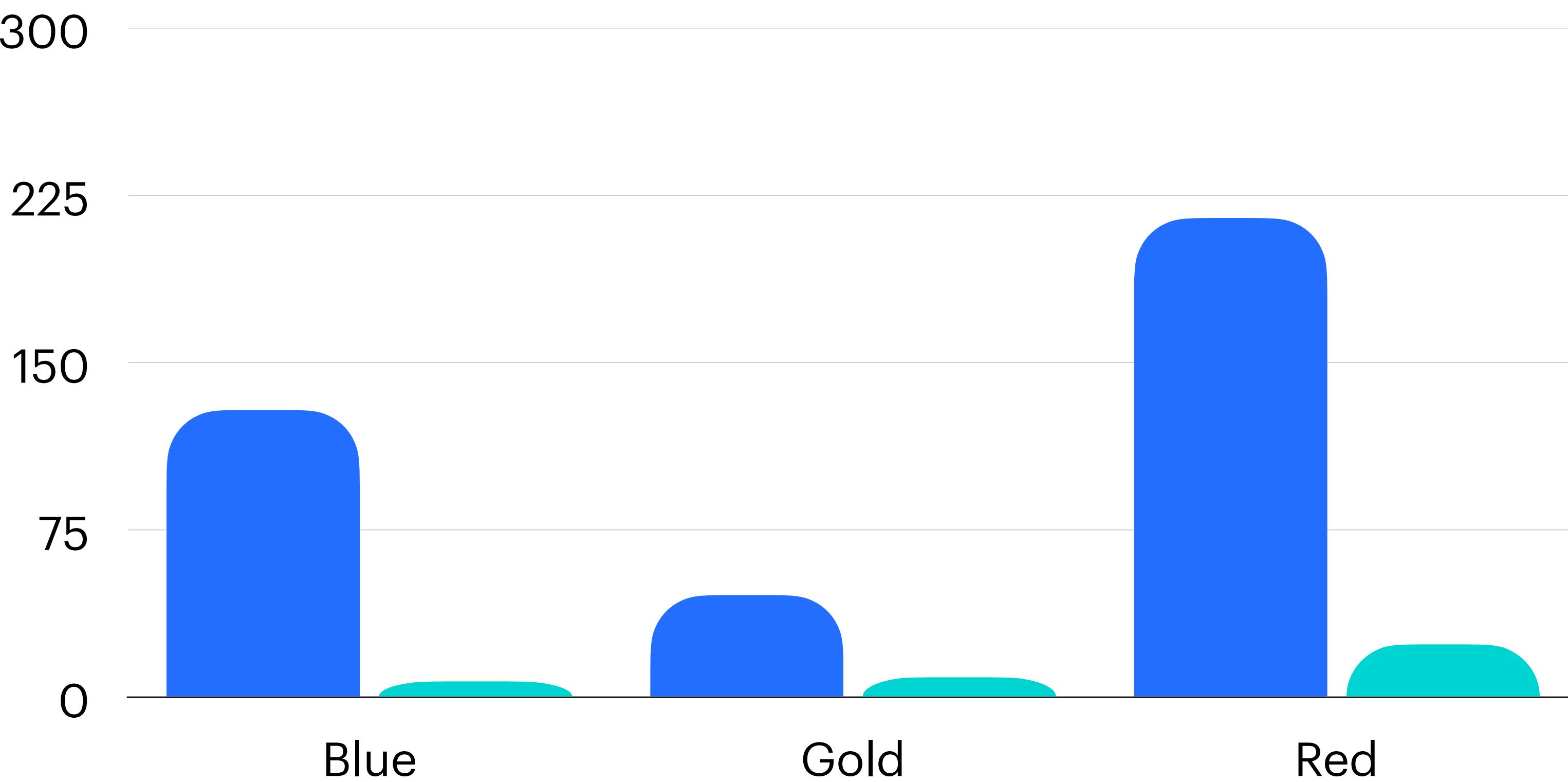
- **Marginal Distribution**
- **Joint Distribution**

BIVARIATE CATEGORICAL DATA

- First we tabulate data into a **contingency table** (also known as a two way table)

		😊	💀
🌐	129	7	136
🟡	46	9	55
🔴	215	24	239
	390	40	430

It's hard to notice association when using frequencies



BIVARIATE CATEGORICAL DATA

Conditional Probability

	129	7	136
	46	9	55
	215	24	239
	390	40	430

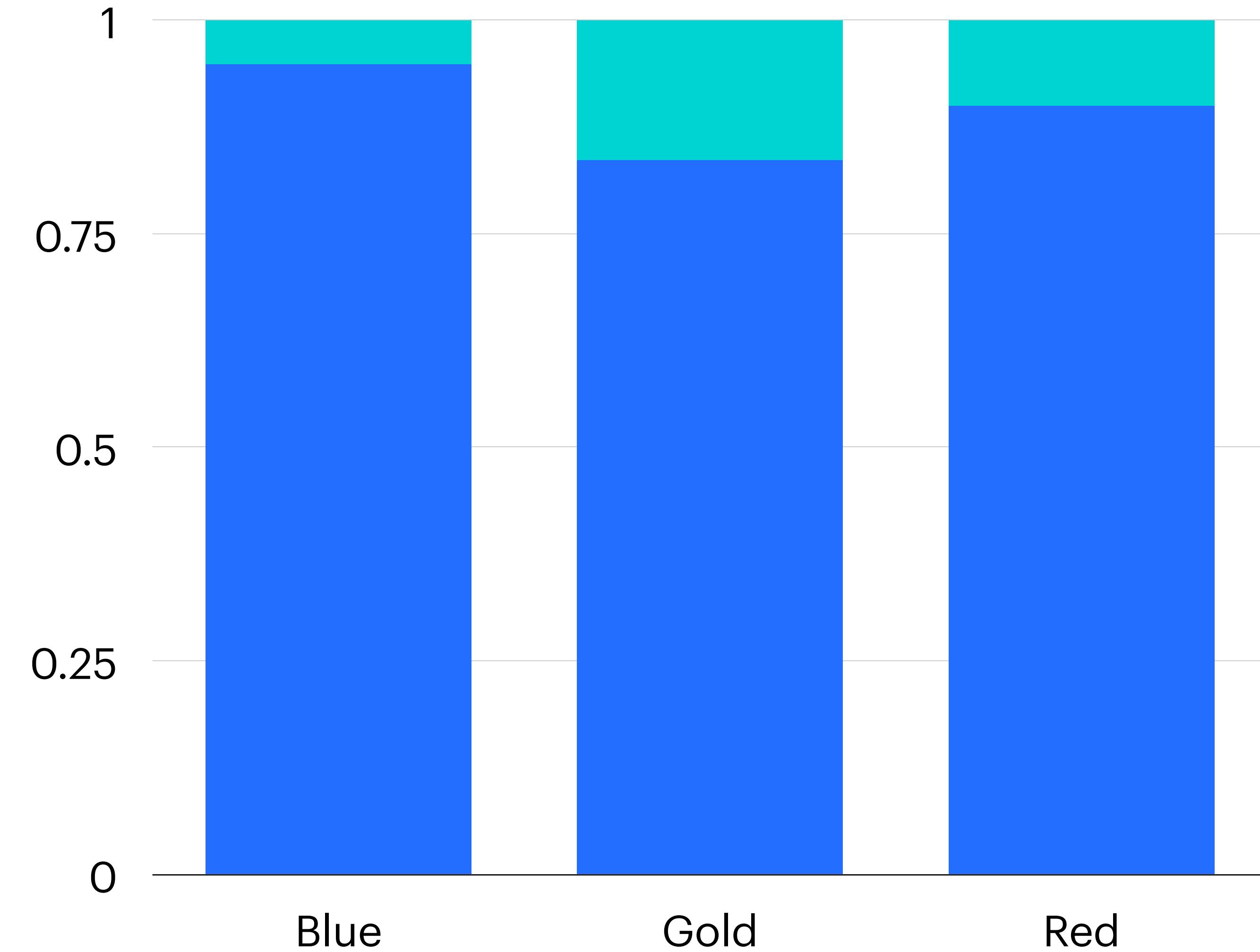
Questions

1. What is the probability of dying, given you are a Red Shirt?
2. What is the percentage of crew members that have red shirts and died?
3. What is the percentage of blue shirts who survived?
4. What is the probability of dying Given you are a Gold Shirt?

BIVARIATE CATEGORICAL DATA

- Next we may find **conditional relative frequencies**

	0.9485294	0.0514706	1
	0.8363636	0.1636364	1
	0.8995816	0.1004184	1



BIVARIATE CATEGORICAL DATA

Distribution of Conditional Relative Frequencies

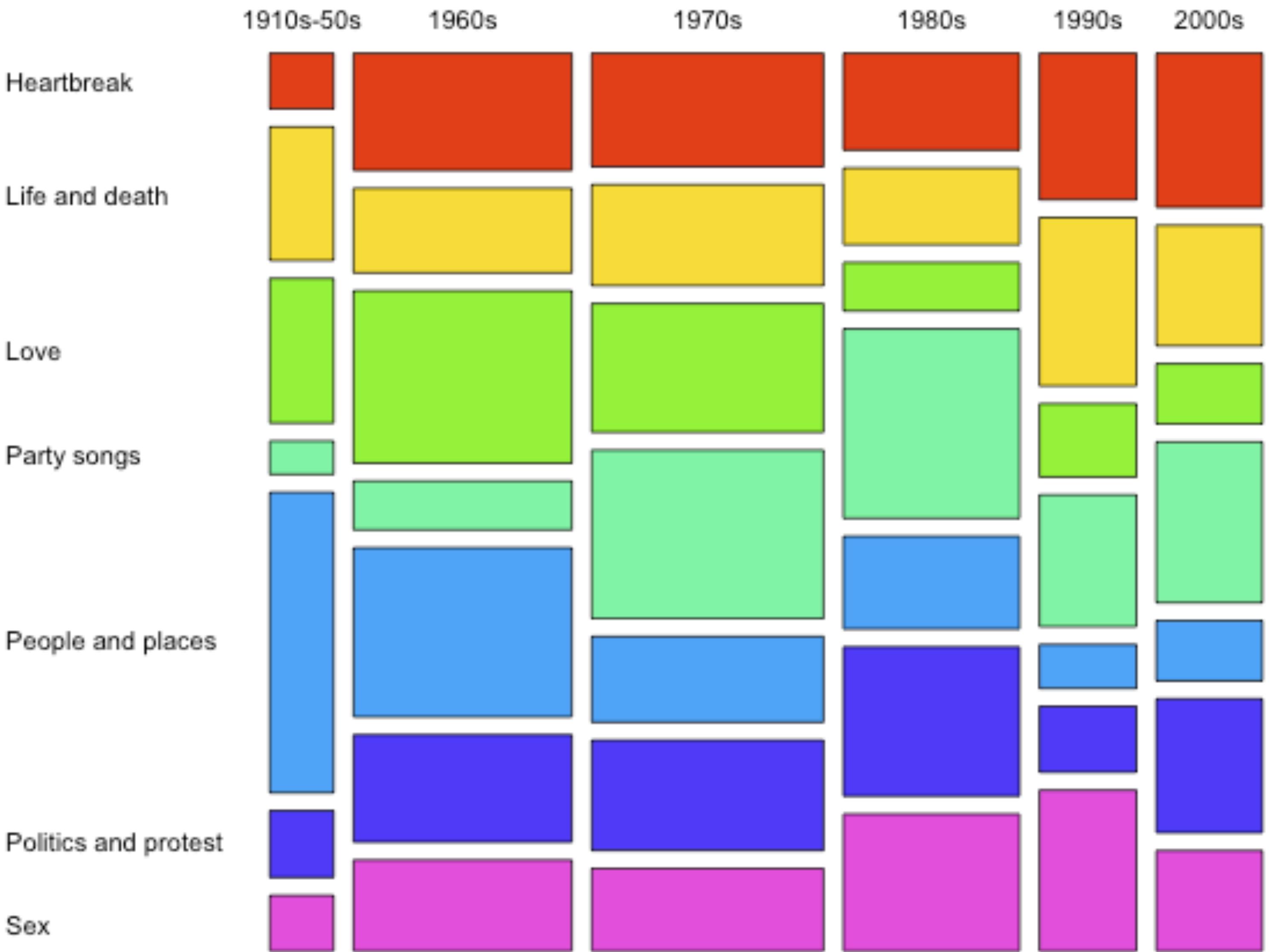
	0.9485294	0.0514706	1
	0.8363636	0.1636364	1
	0.8995816	0.1004184	1

If shirt colour is **Independent** of Status, then the probability of dying should be the same regardless of shirt colour.

Chi-Square Tests (For Later)

BIVARIATE CATEGORICAL DATA

- Another type of graph used is **Mosaic Plots**. Widths describe how many observations fall in each category.
- Mosaic plot showing cross-sectional distribution through time of different musical themes in the Guardian's list of "1000 songs to hear before you die"



BIVARIATE CATEGORICAL DATA

Examples:

- Question 1, Page 107
- Question 2, Page 108
- Question 3, Page 112

Homework: Read Pages 97-104 Barron's,
Quiz 6, Quiz 7

BIVARIATE QUANTITATIVE DATA

- We represent relationships between two numeric variables (sample data) using a **scatter plot**.
- When describing a relationship there are several things we must consider
 - **Form**
 - **Direction**
 - **Strength**

EXAMPLE

Is there a relationship between the amount of sugar (in grams) and the number of calories in movie-theatre candy? Here are the data from a sample of 12 types of candy.

Name	Sugar (g)	Calories
Butterfinger Minis	45	450
Junior Mints	107	570
M&M'S	62	480
Milk Duds	44	370
Peanut M&M'S	79	790
Raisinets	60	420
Reese's Pieces	61	580
Skittles	87	450
Sour Patch Kids	92	490
SweeTarts	136	680
Twizzlers	59	460
Whoppers	48	350

Using your TI-84, plot the data.

How Would You Describe the Relationship?

BIVARIATE QUANTITATIVE DATA

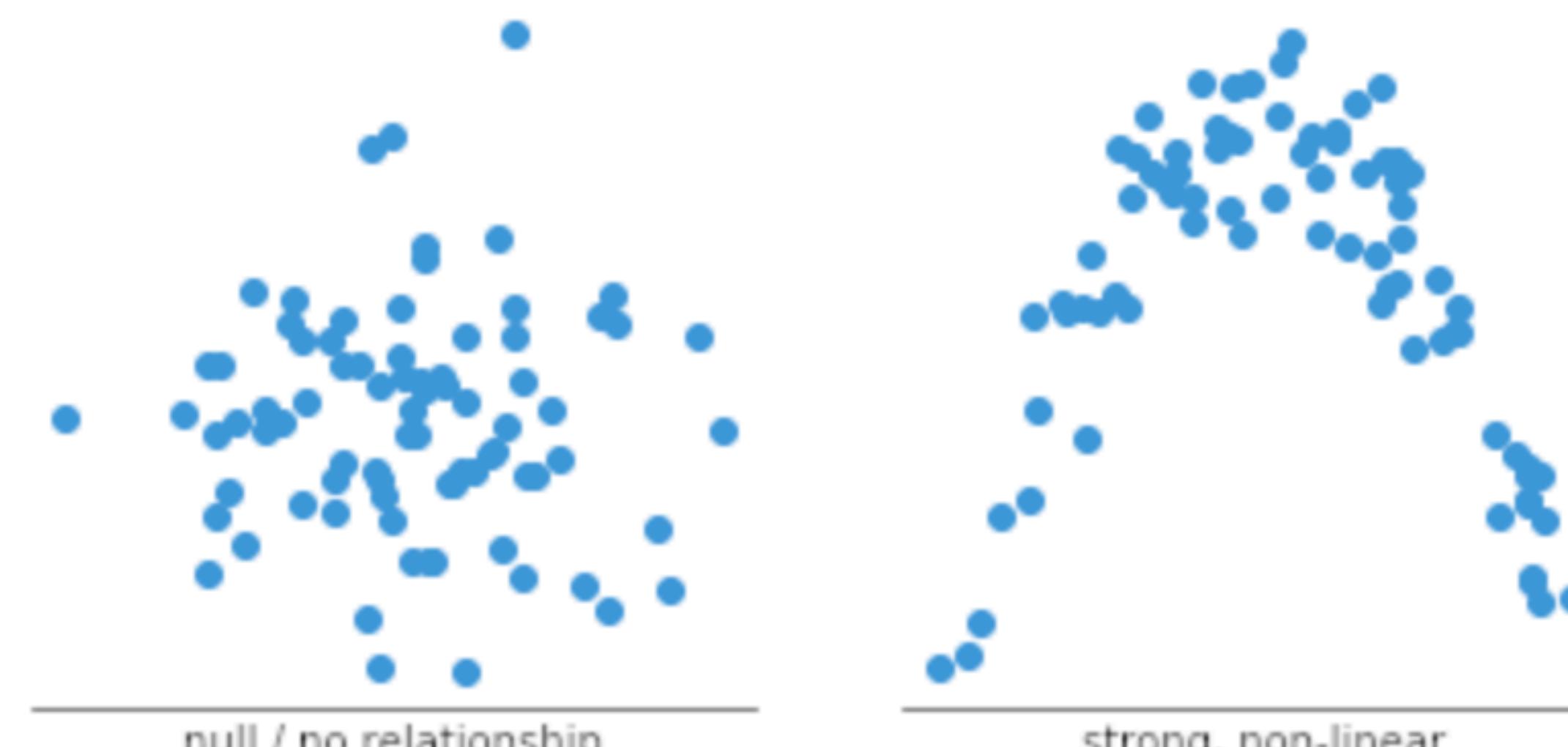
Form of relationship



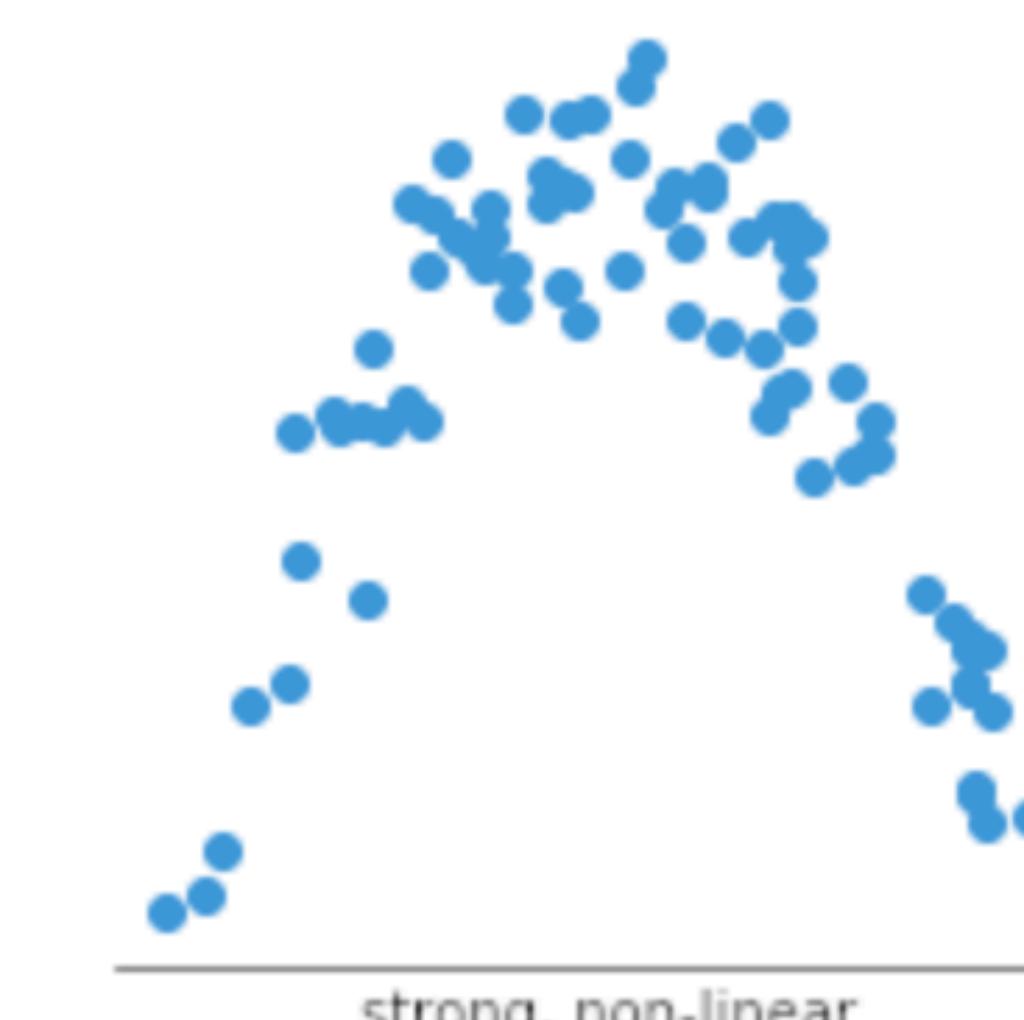
strong, positive, linear



moderate, negative, linear



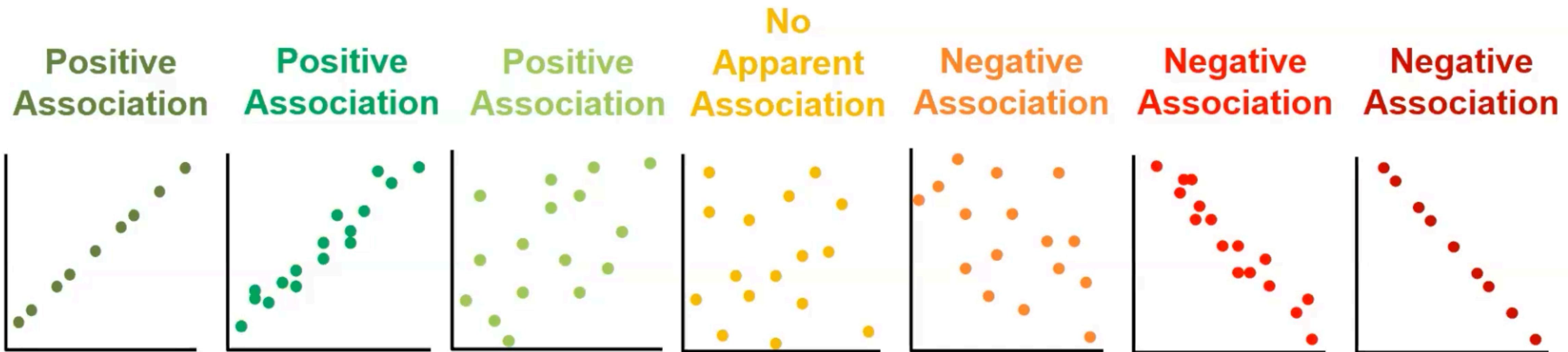
null / no relationship



strong, non-linear

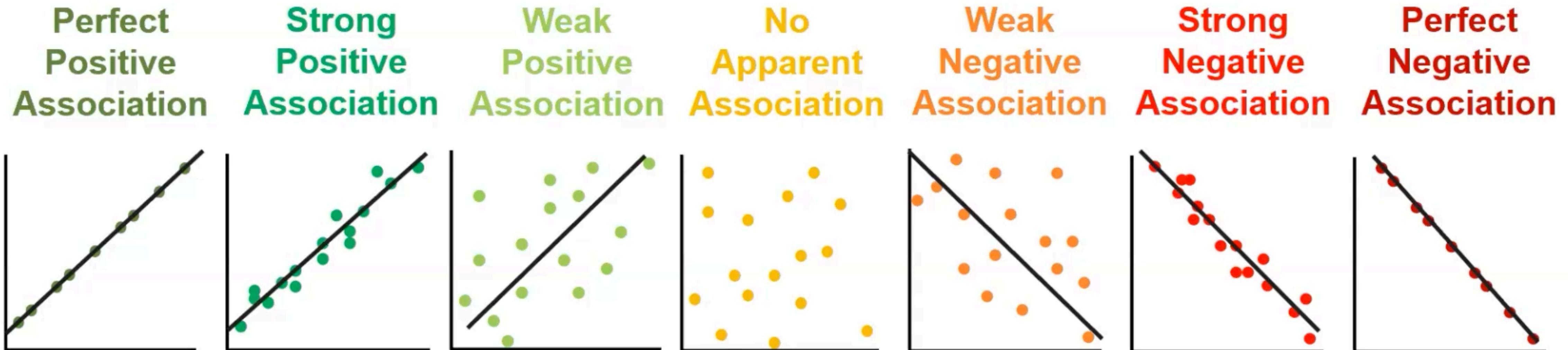
BIVARIATE QUANTITATIVE DATA

Direction of relationship



BIVARIATE QUANTITATIVE DATA

Strength of relationship



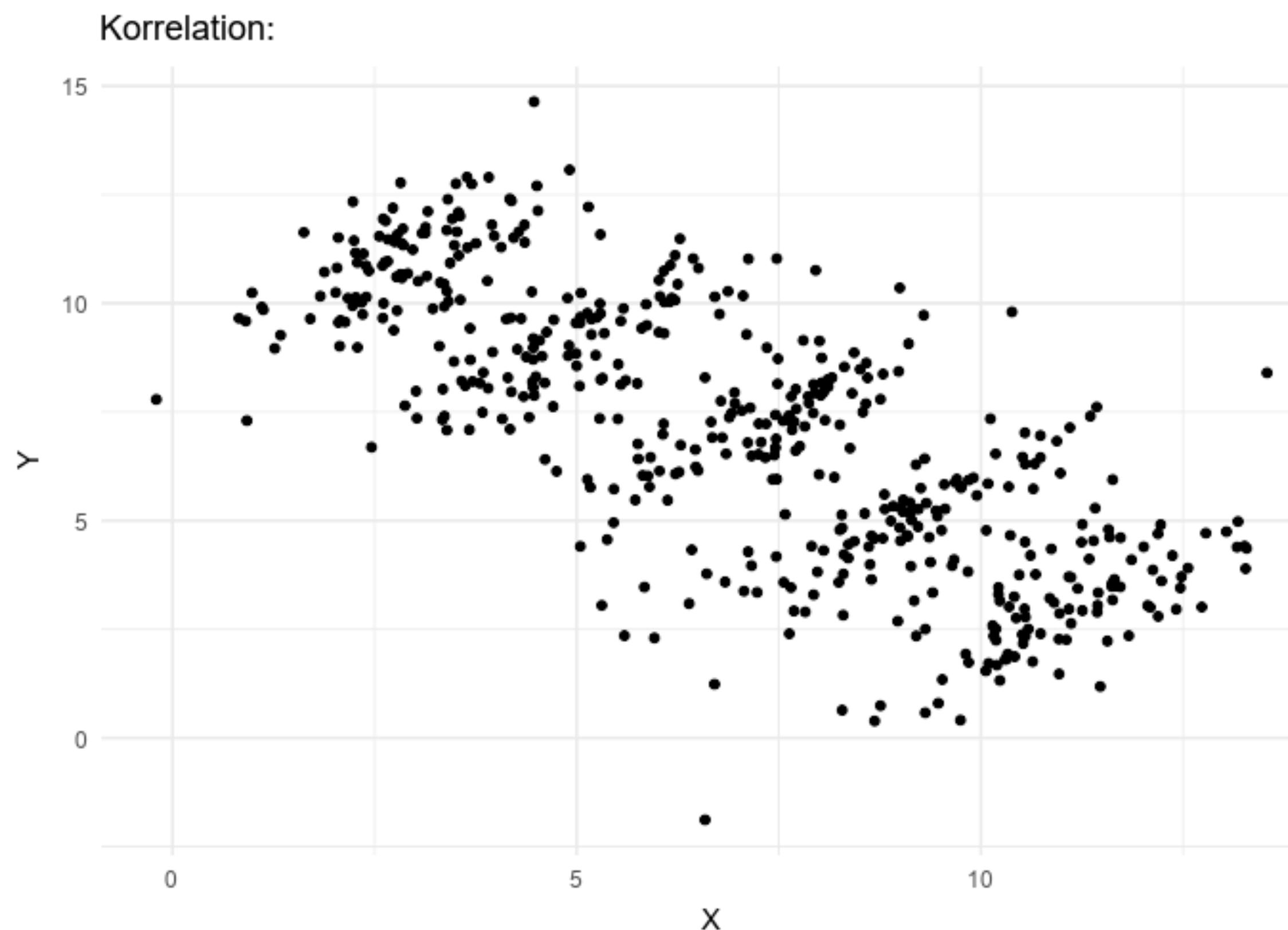
BIVARIATE QUANTITATIVE DATA

Influential Points of relationship



BIVARIATE QUANTITATIVE DATA

- **Simpson's Paradox:** There is an association within groups of data but the trend disappears, or reverses when groups are combined.
- **Example:** Q7 P.111



BIVARIATE QUANTITATIVE DATA

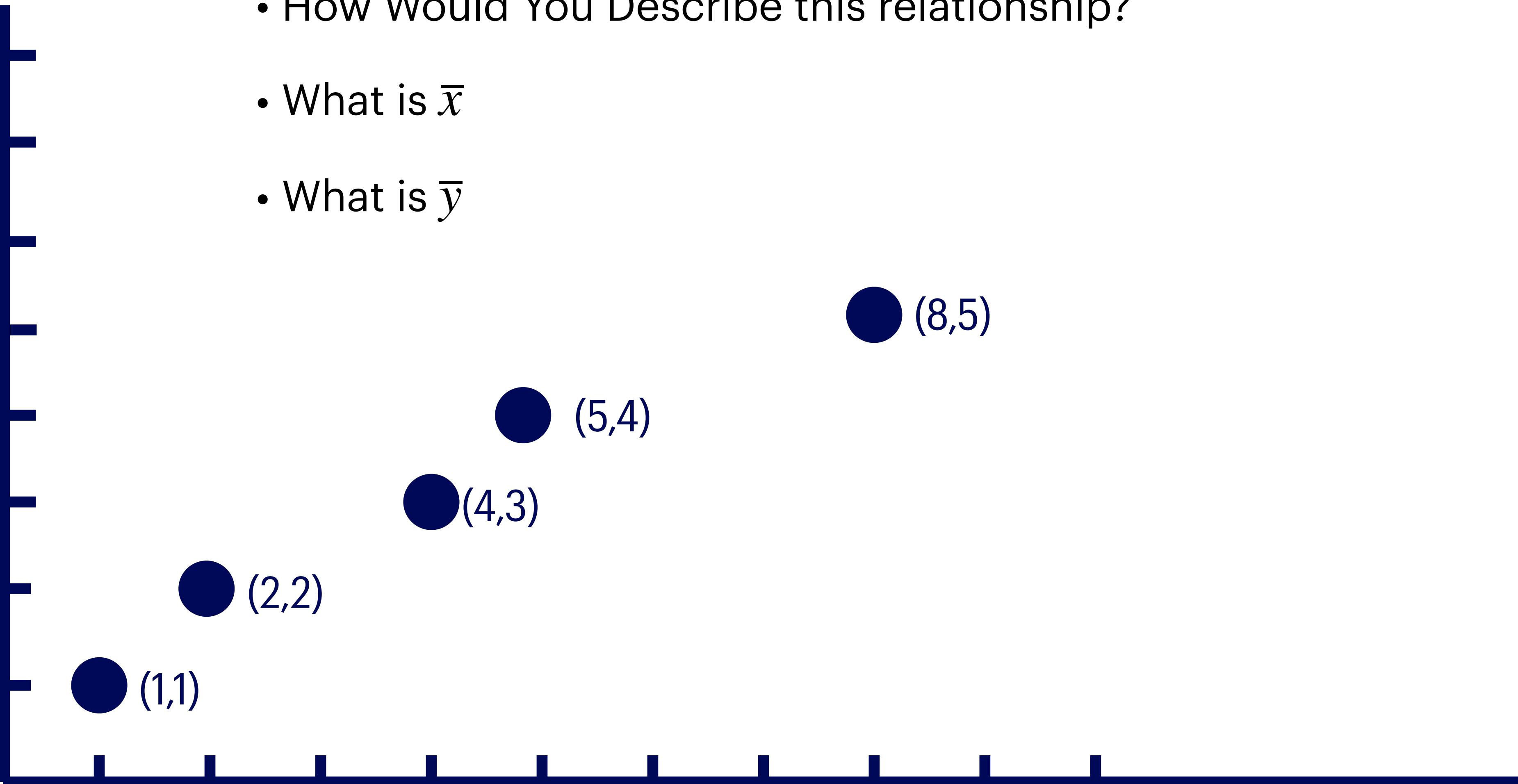
What is the Sample Covariance?

$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

BIVARIATE QUANTITATIVE DATA

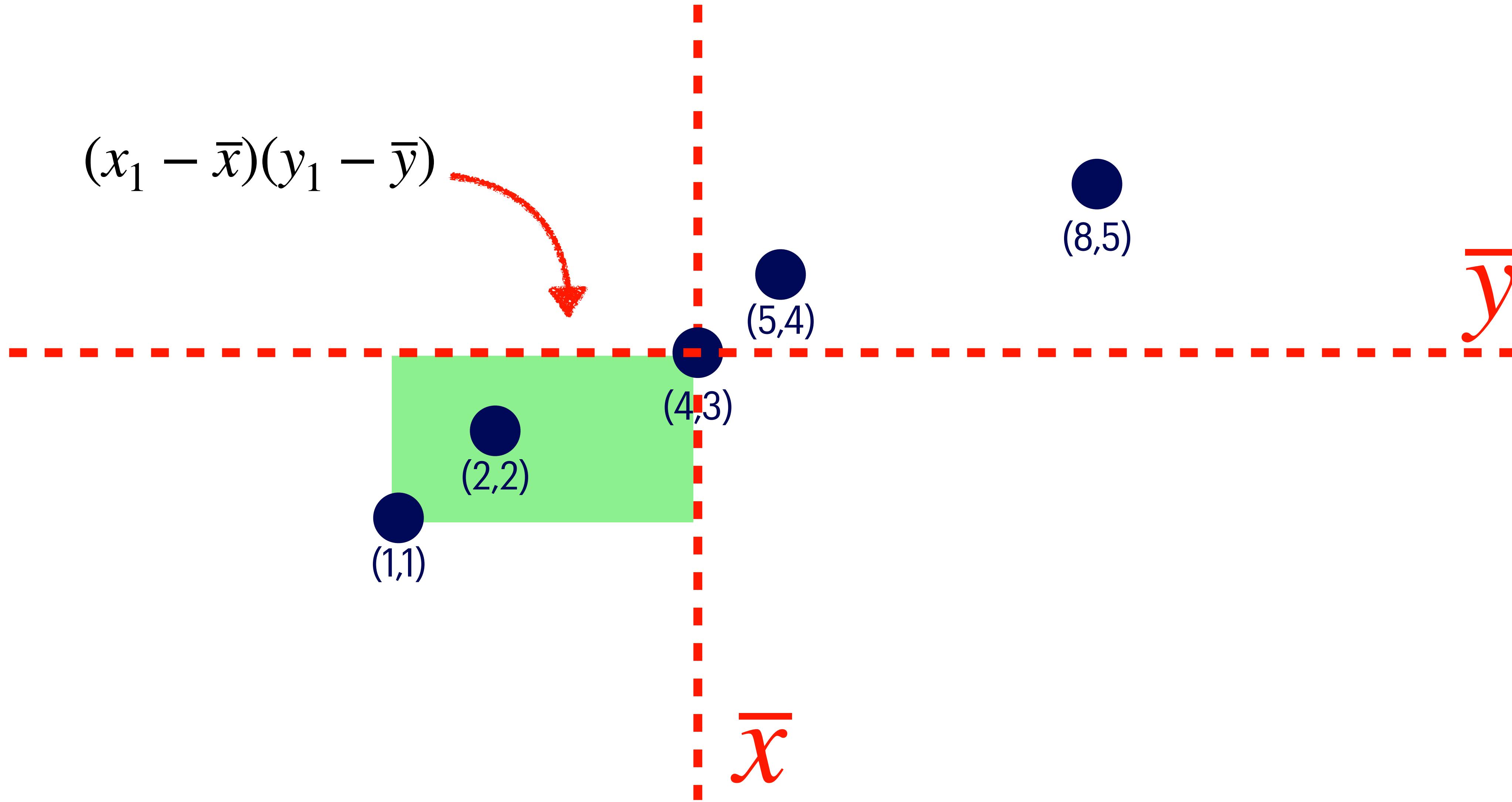
What is the Sample Covariance?

- How Would You Describe this relationship?
- What is \bar{x}
- What is \bar{y}



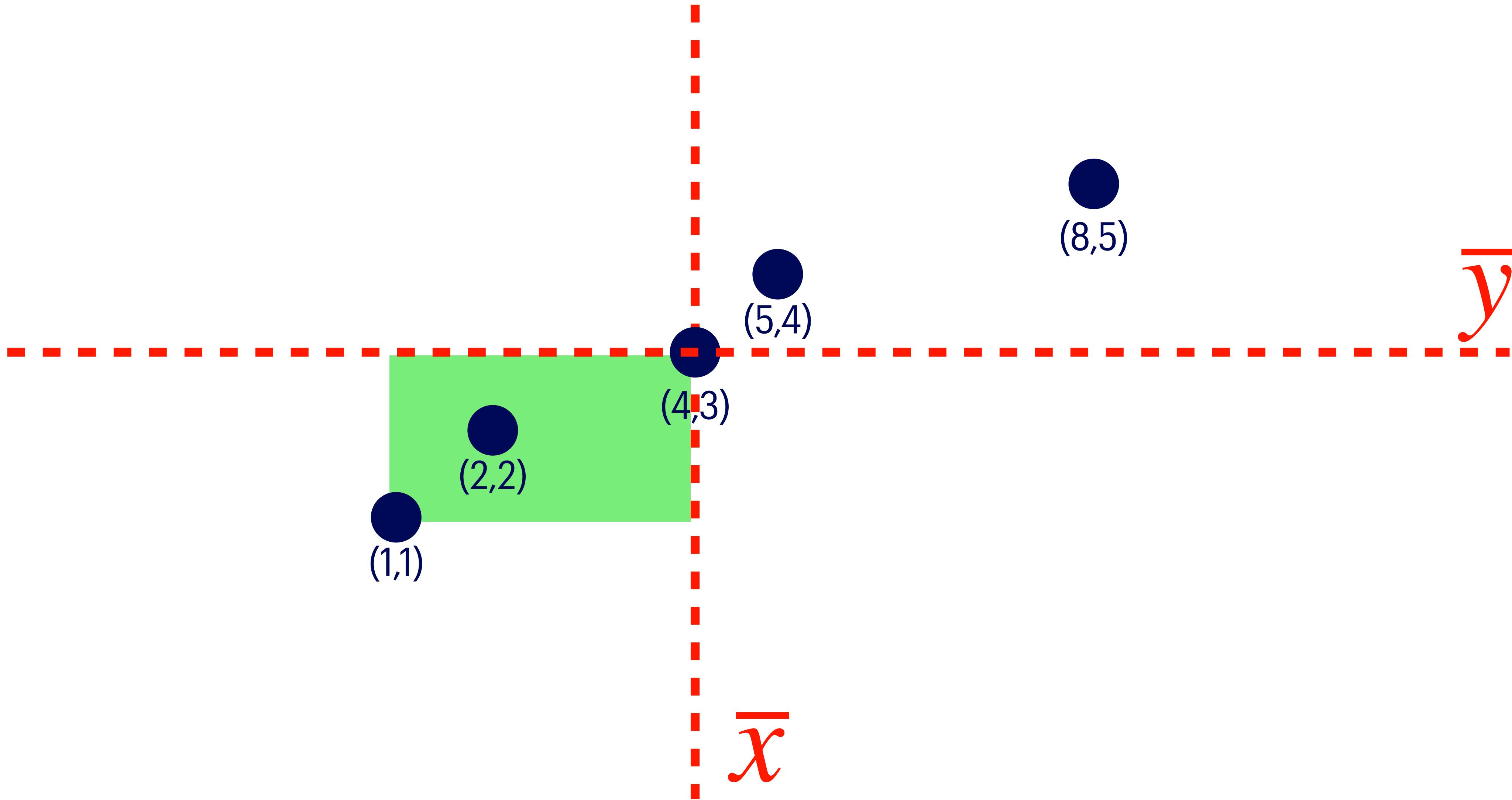
BIVARIATE QUANTITATIVE DATA

What is the Sample Covariance?



BIVARIATE QUANTITATIVE DATA

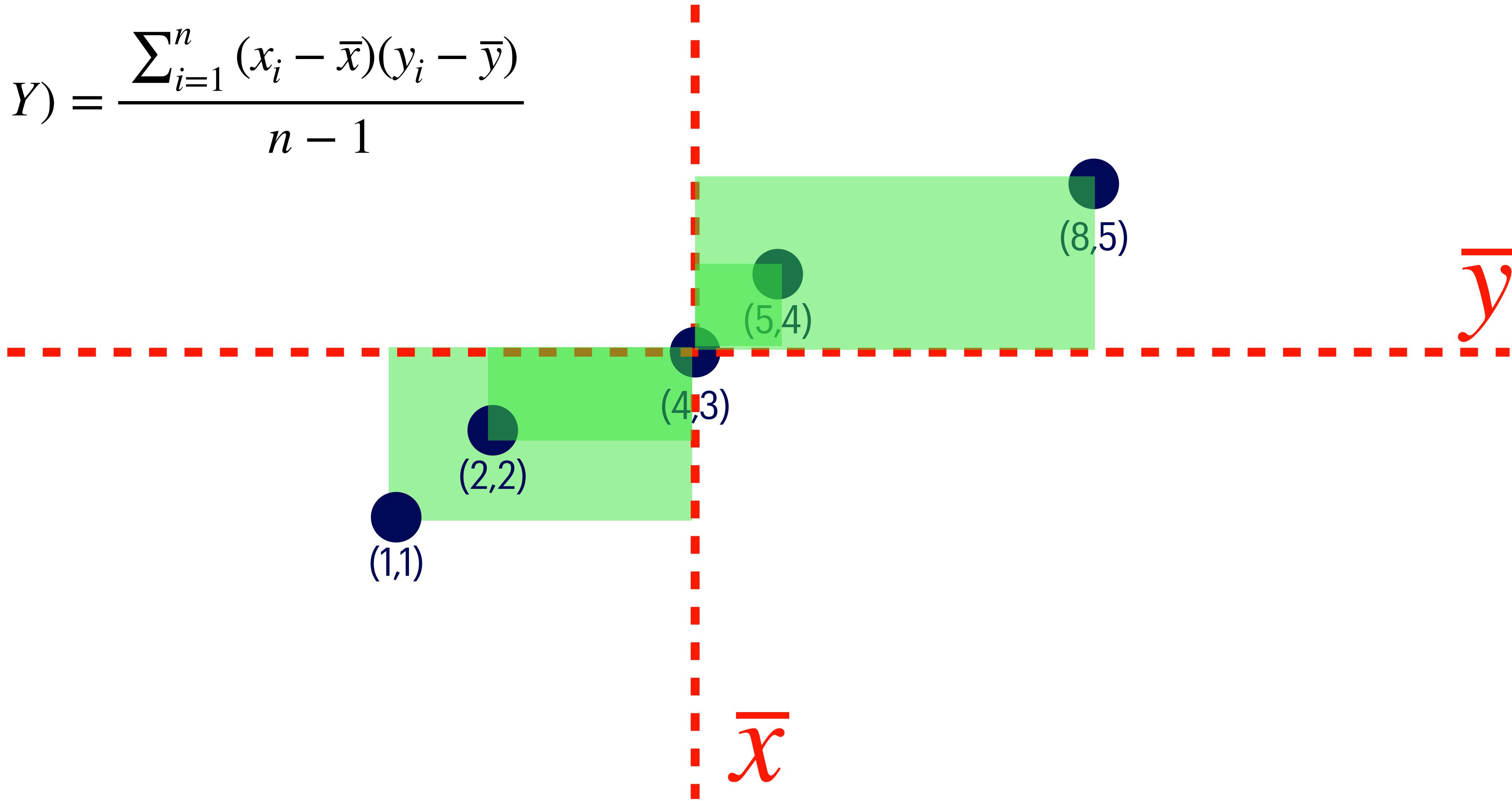
What is the Sample Covariance?



BIVARIATE QUANTITATIVE DATA

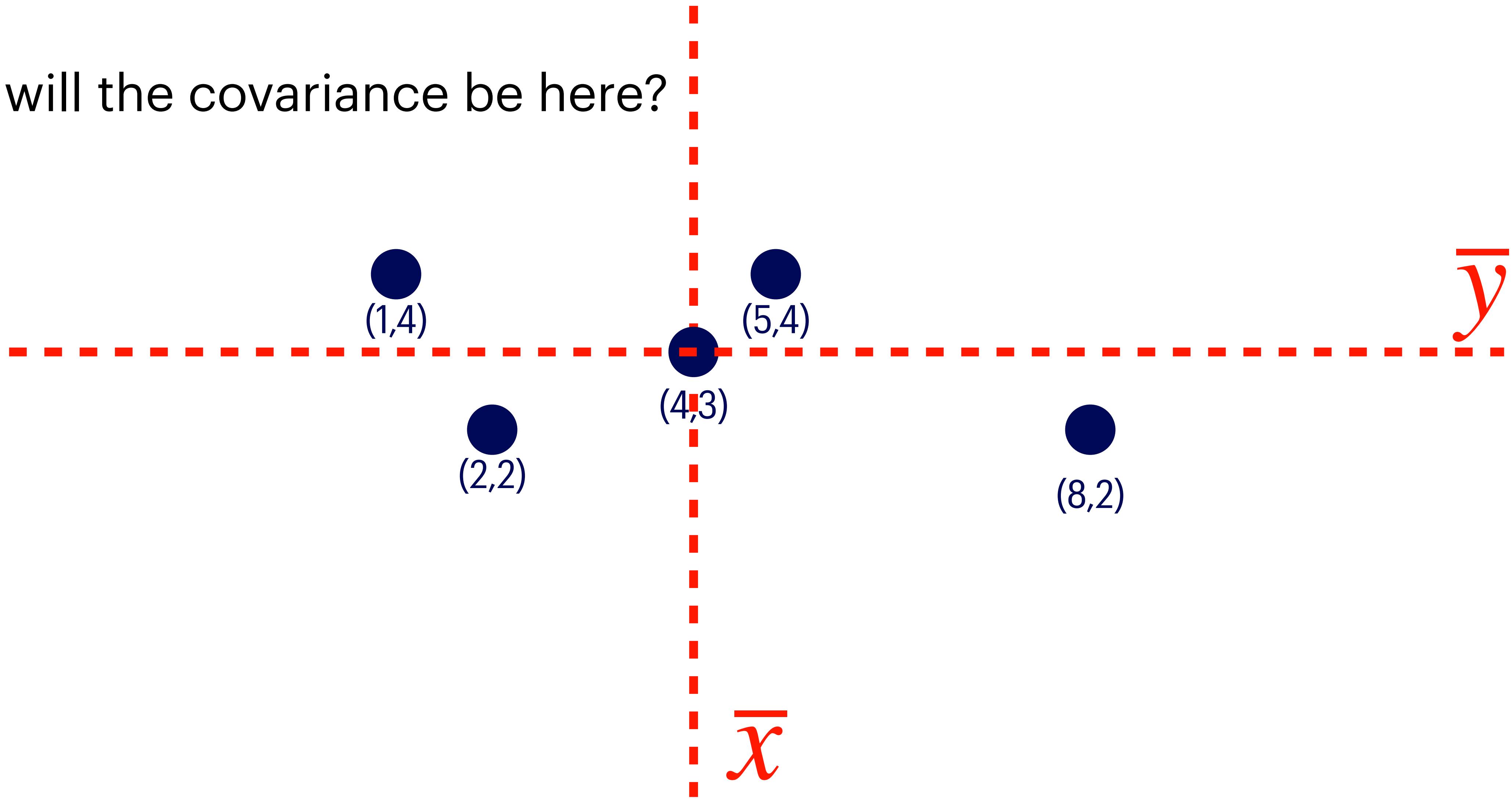
What is the Sample Covariance?

$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$



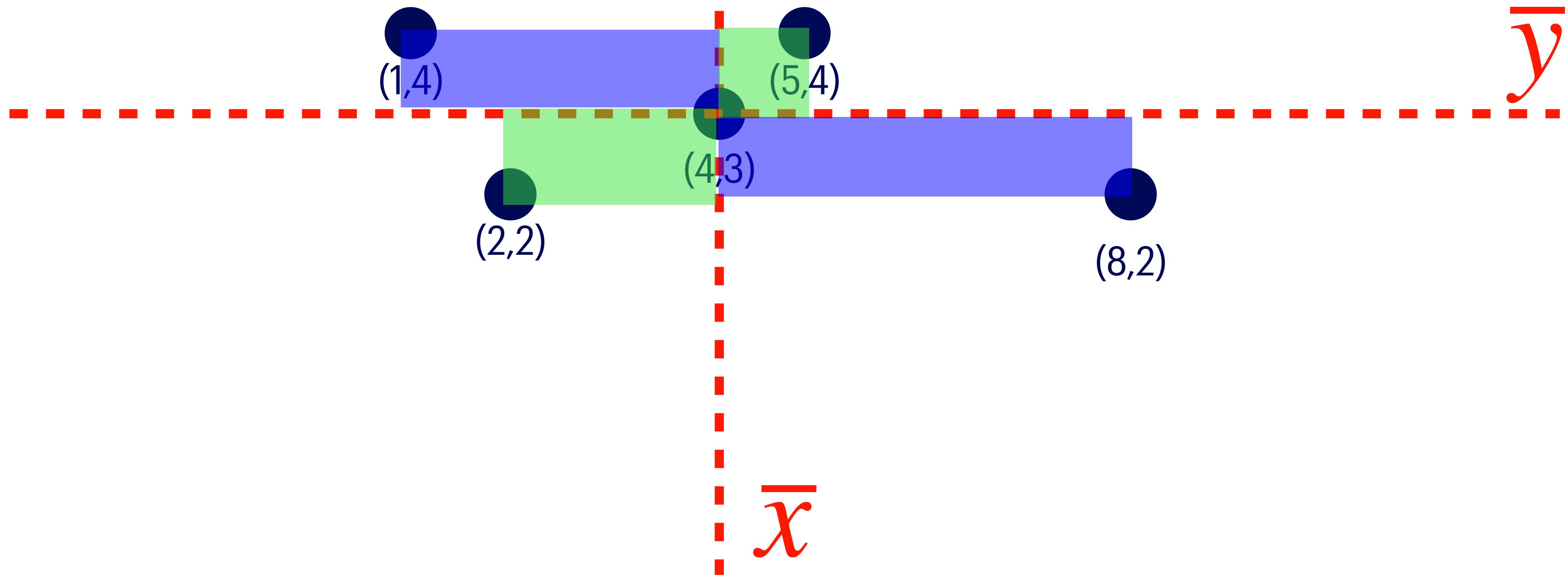
What is the Sample Covariance?

- What will the covariance be here?

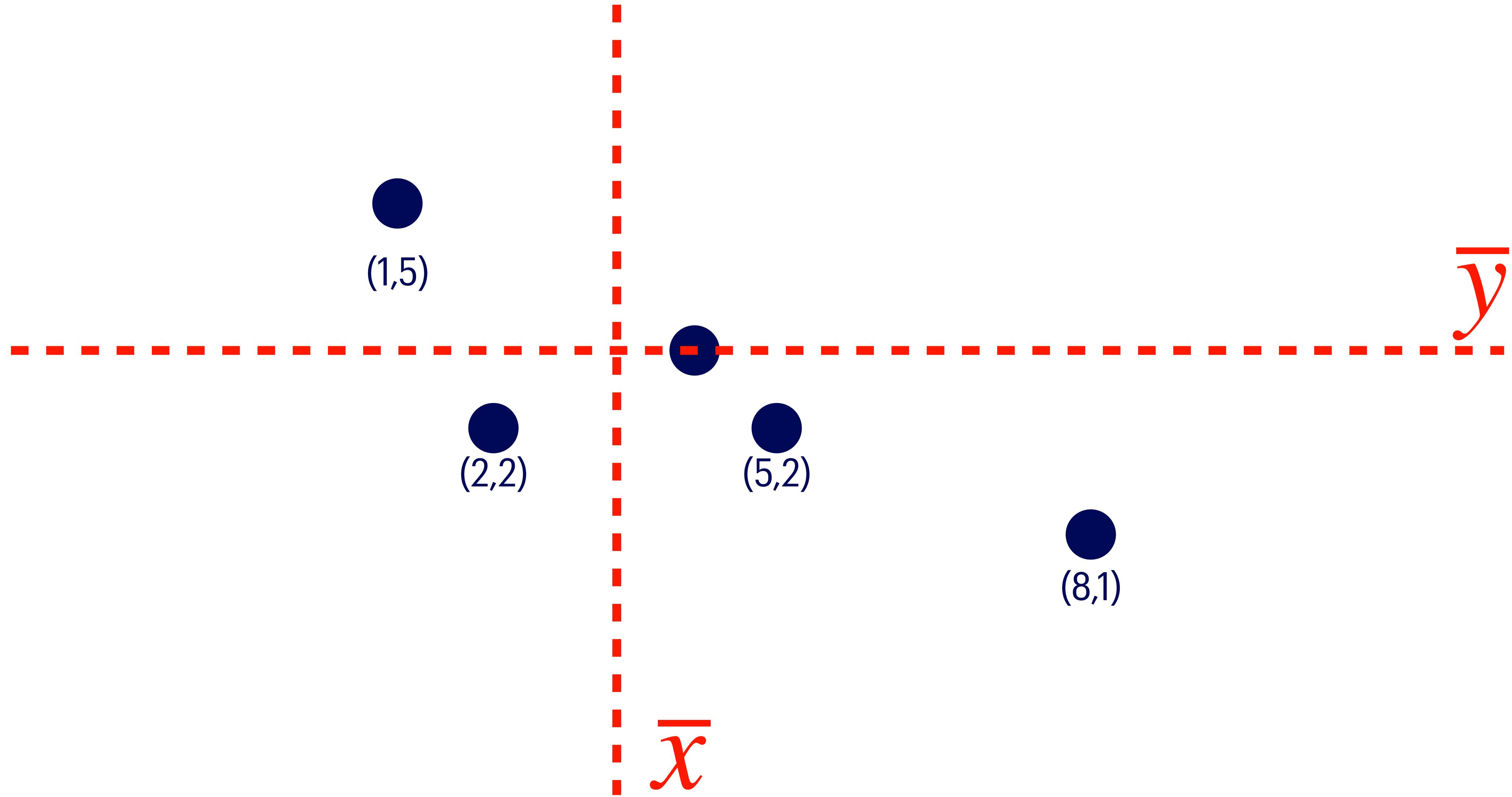


What is the Sample Covariance?

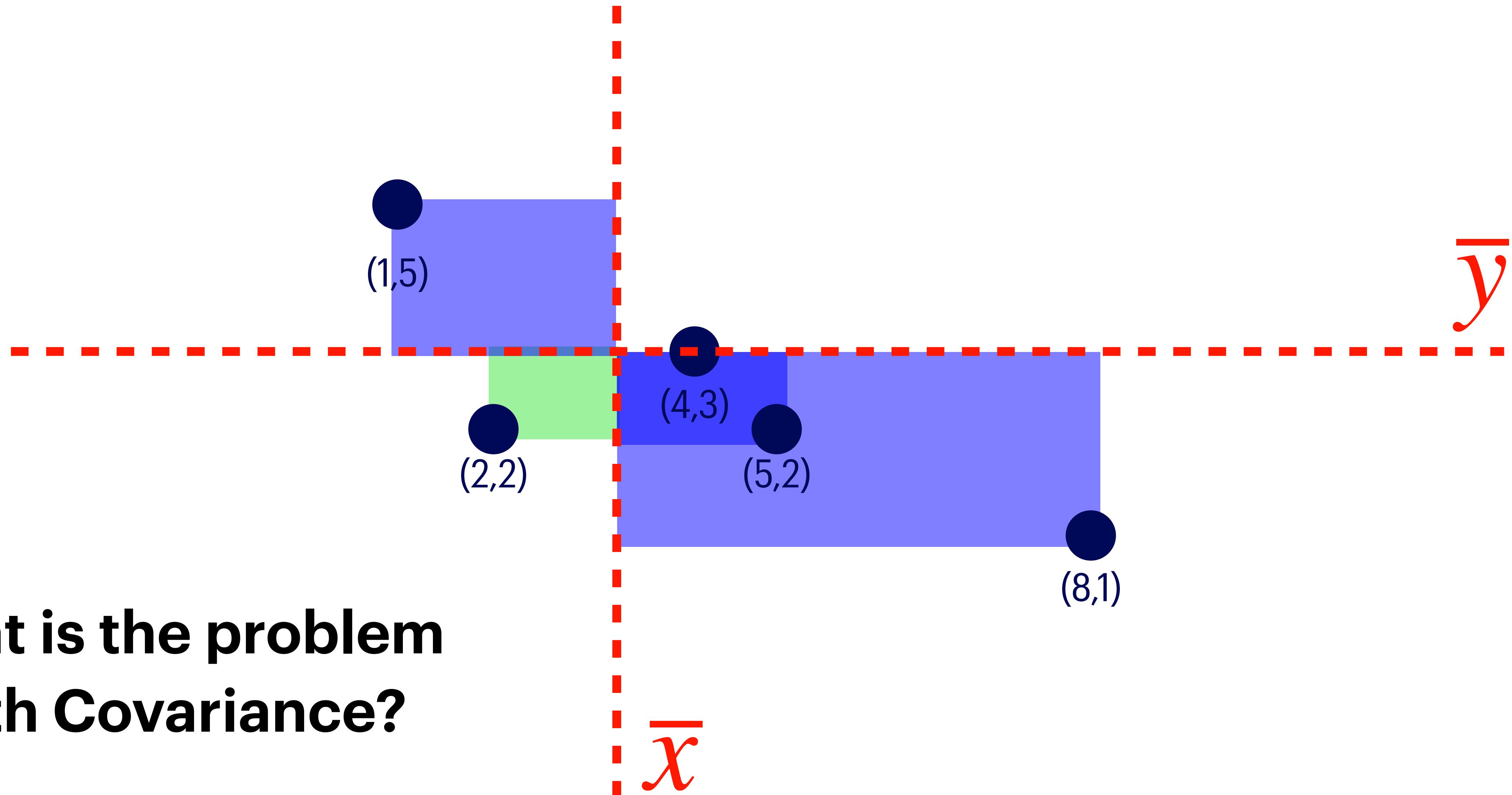
Negative and positive differences cancel out



What is the Sample Covariance?



What is the Sample Covariance?



What is the problem
with Covariance?

More Examples: [http://digitalfirst.bfwpub.com/
stats_applet/stats_applet_5_correg.html](http://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html)

BIVARIATE QUANTITATIVE DATA

What is the Sample Correlation?

$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$r = \frac{Cov(X, Y)}{s_x s_y} = \frac{1}{n - 1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

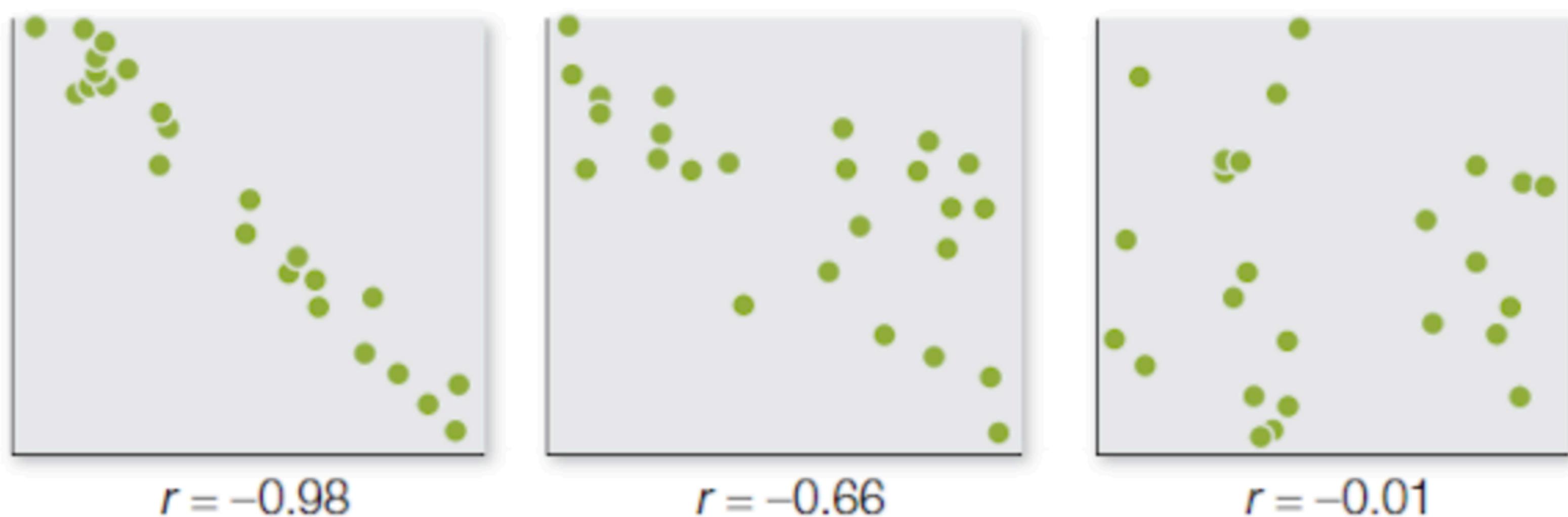
Note that r is
a statistic

BIVARIATE QUANTITATIVE DATA

What Does Correlation Measure?

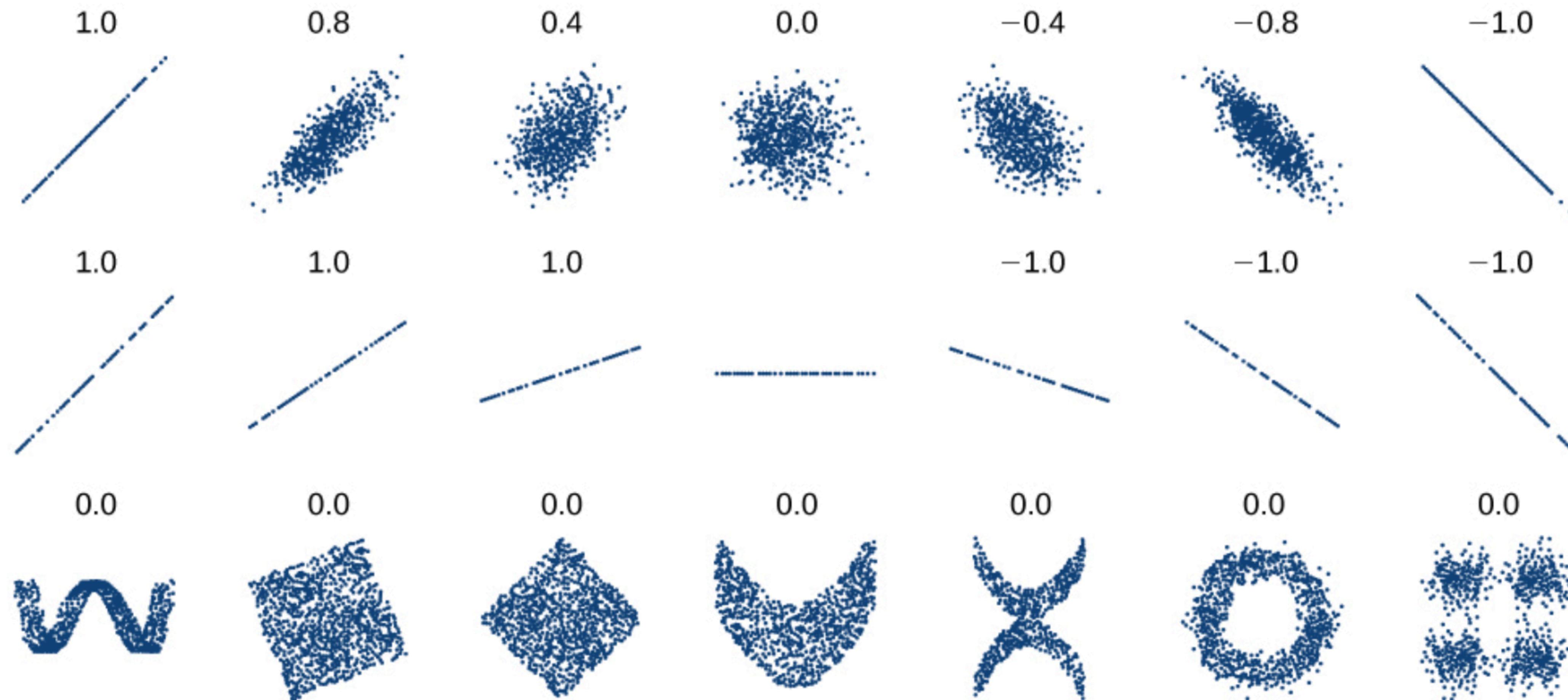
- Direction
- Strength

Guess the Correlation
[https://
www.rosmachance.
com/applets/2021/
guesscorrelation/
GuessCorrelation.html](https://www.rosmachance.com/applets/2021/guesscorrelation/GuessCorrelation.html)



BIVARIATE QUANTITATIVE DATA

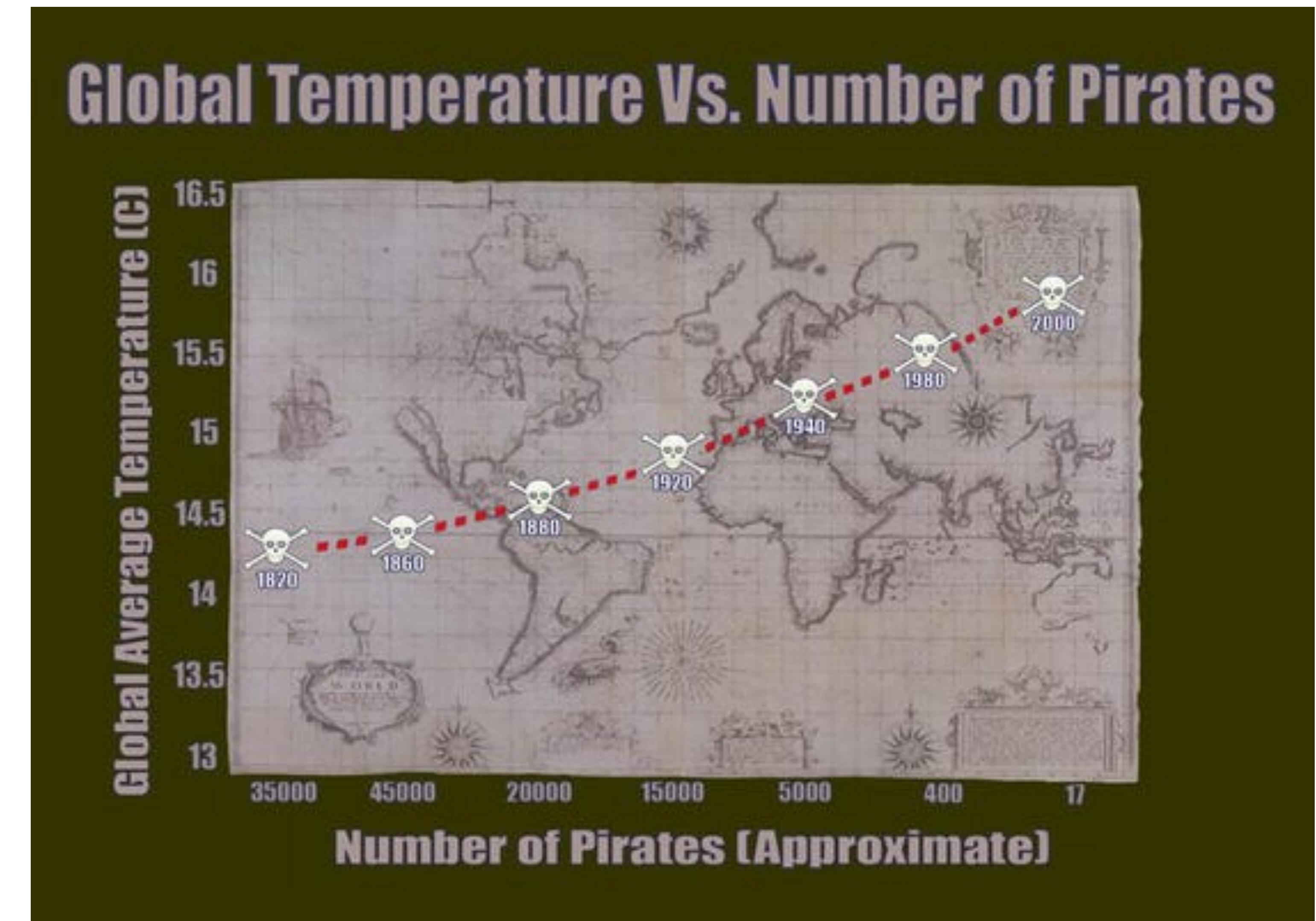
ALWAYS PLOT THE DATA



BIVARIATE QUANTITATIVE DATA

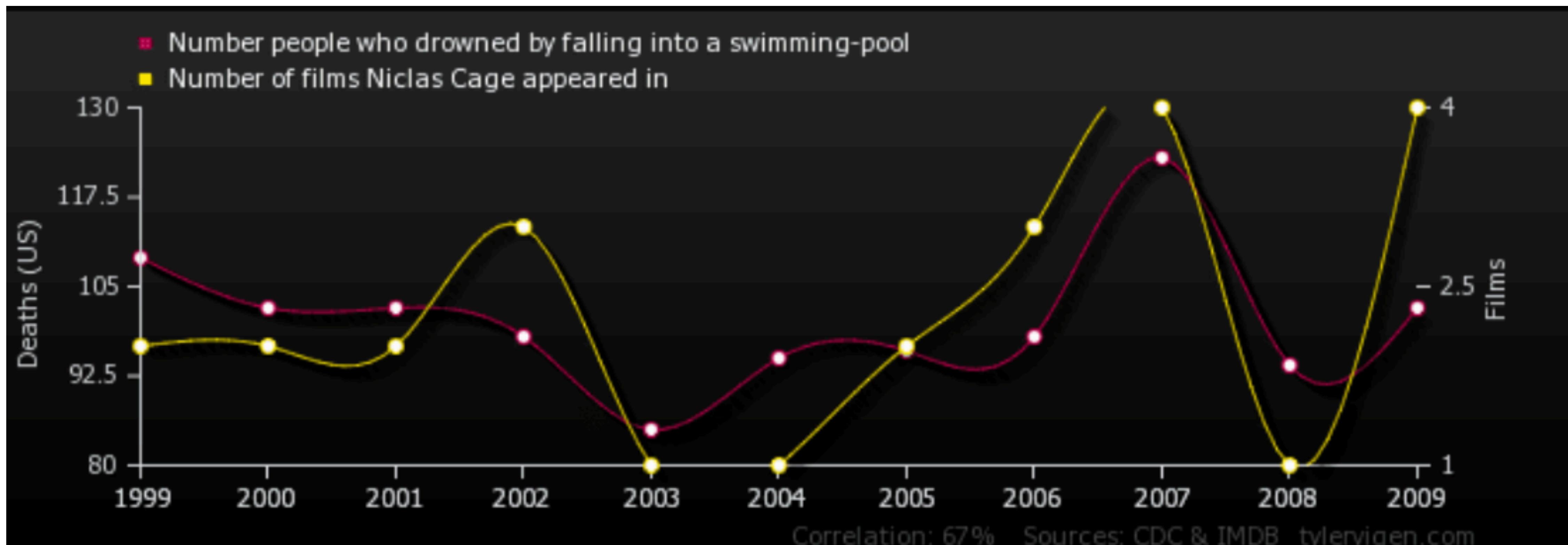
What does correlation tell us about causation?

- Is a lack of pirates causing global warming?
- Are Ice Cream Salesman responsible for increased drowning fatalities?



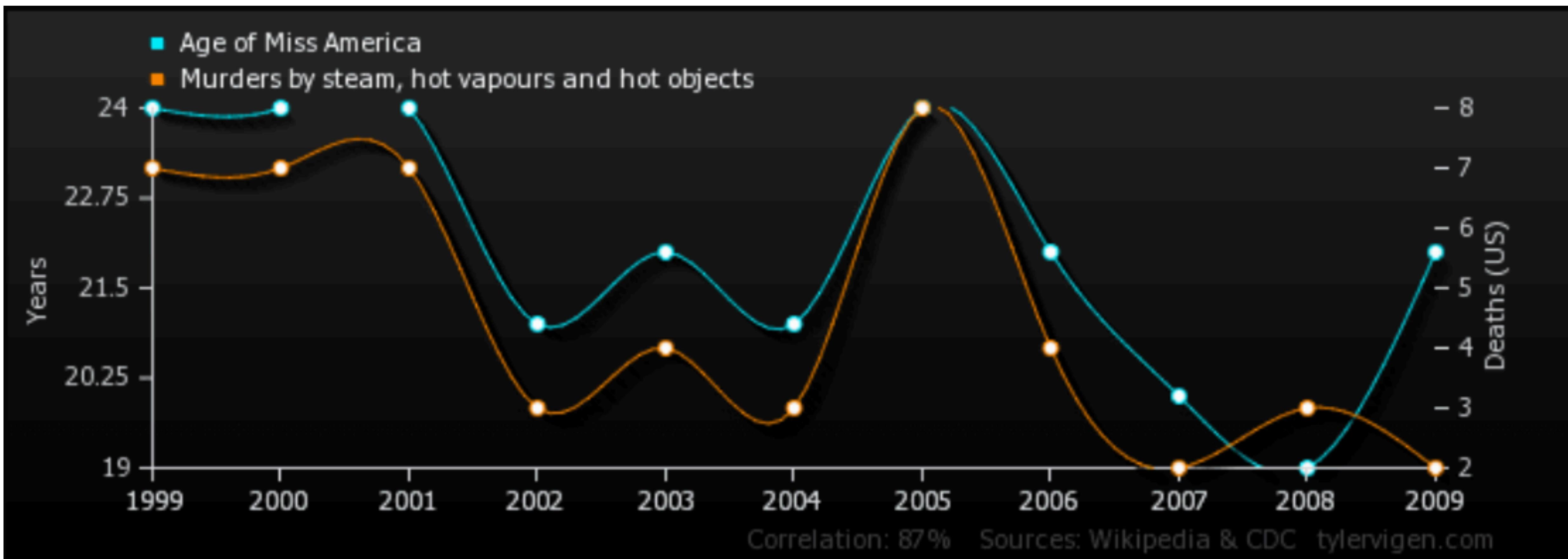
BIVARIATE QUANTITATIVE DATA

Correlation ≠ Causation Examples



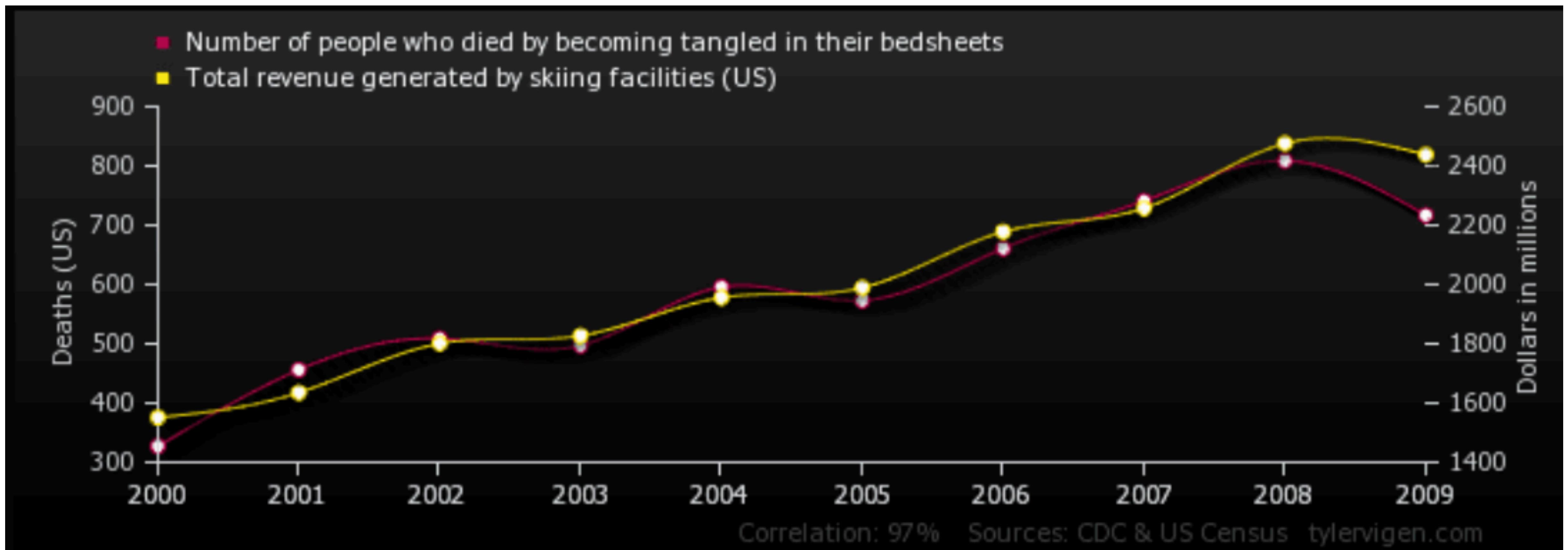
BIVARIATE QUANTITATIVE DATA

Correlation ≠ Causation Examples



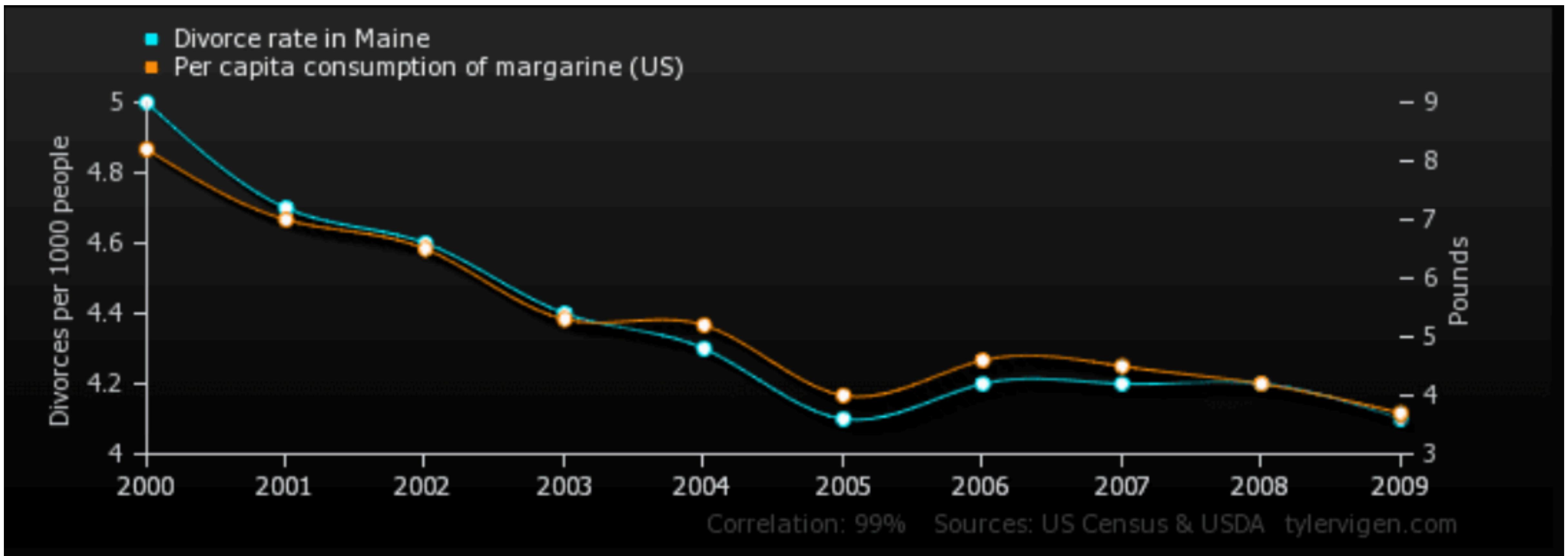
BIVARIATE QUANTITATIVE DATA

Correlation ≠ Causation Examples



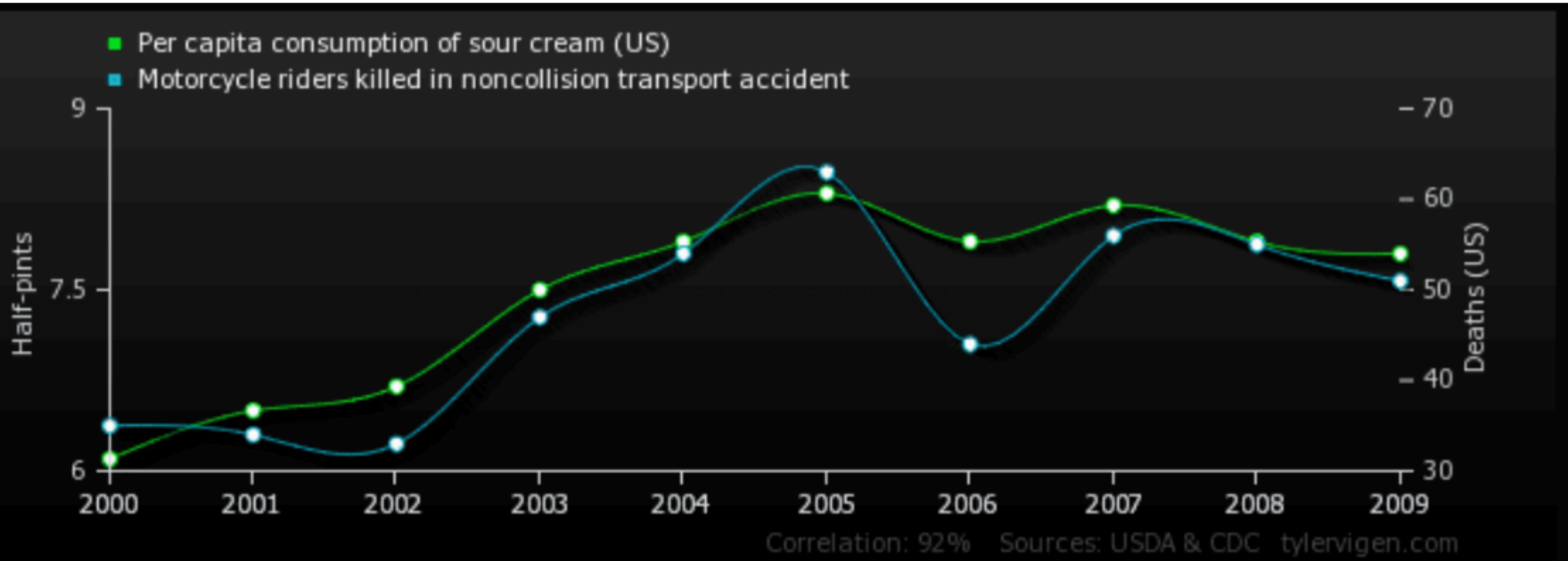
BIVARIATE QUANTITATIVE DATA

Correlation ≠ Causation Examples



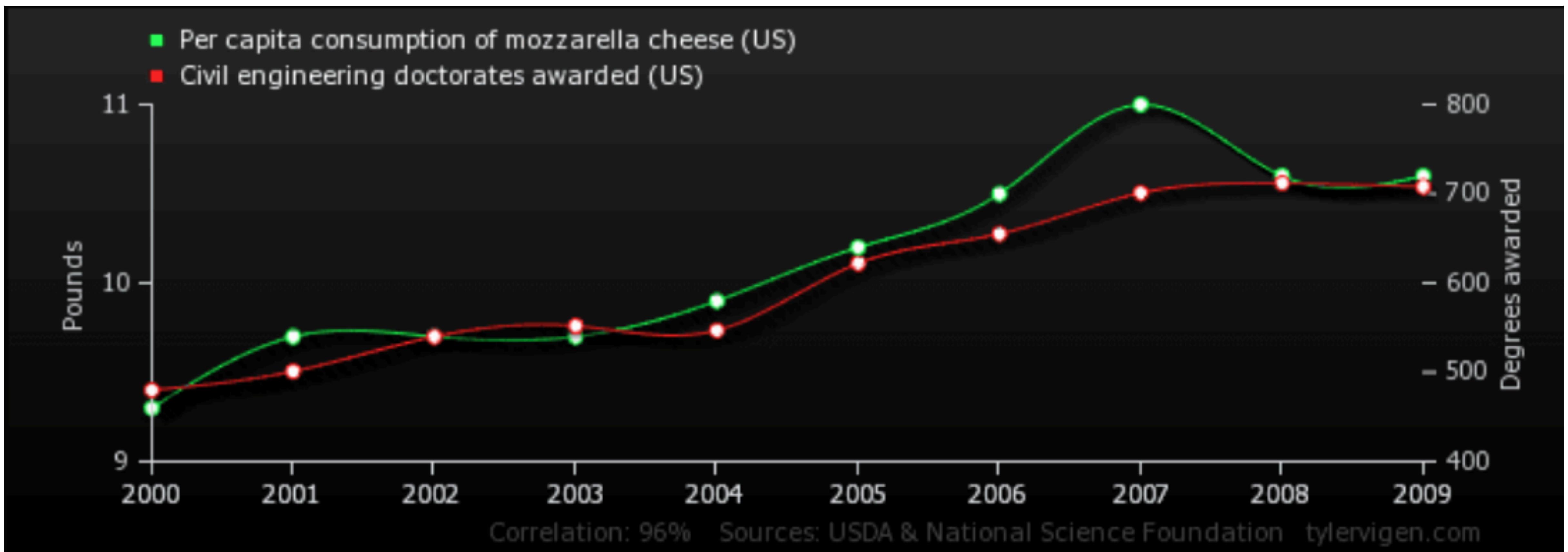
BIVARIATE QUANTITATIVE DATA

Correlation ≠ Causation Examples



BIVARIATE QUANTITATIVE DATA

Correlation ≠ Causation Examples



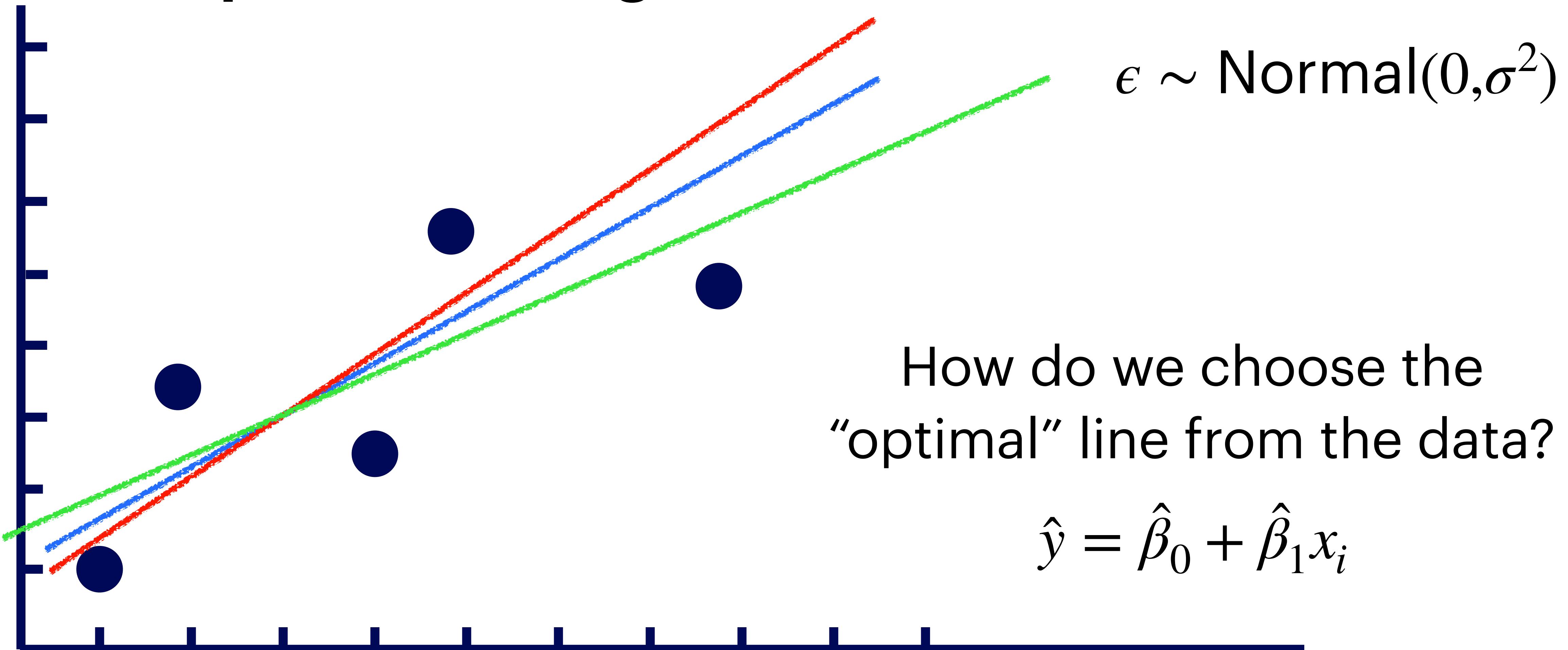
BIVARIATE QUANTITATIVE DATA

How do we “prove” something is a causal relationship?

Experiments will be discussed in more detail later
(unit 3)

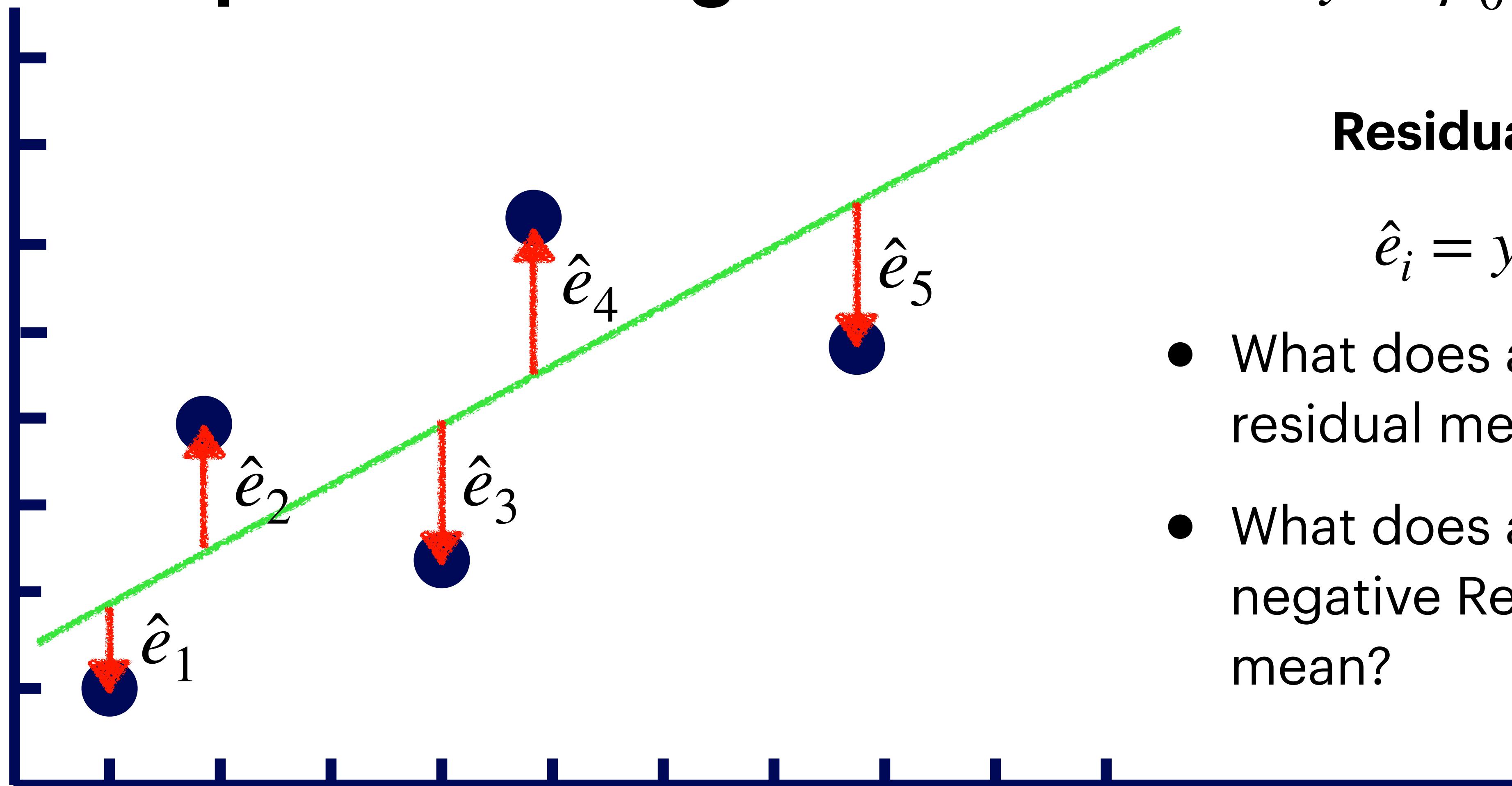
BIVARIATE QUANTITATIVE DATA

Simple Linear Regression Model $y = \beta + \beta_1 x + \epsilon$



BIVARIATE QUANTITATIVE DATA

Simple Linear Regression Model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$



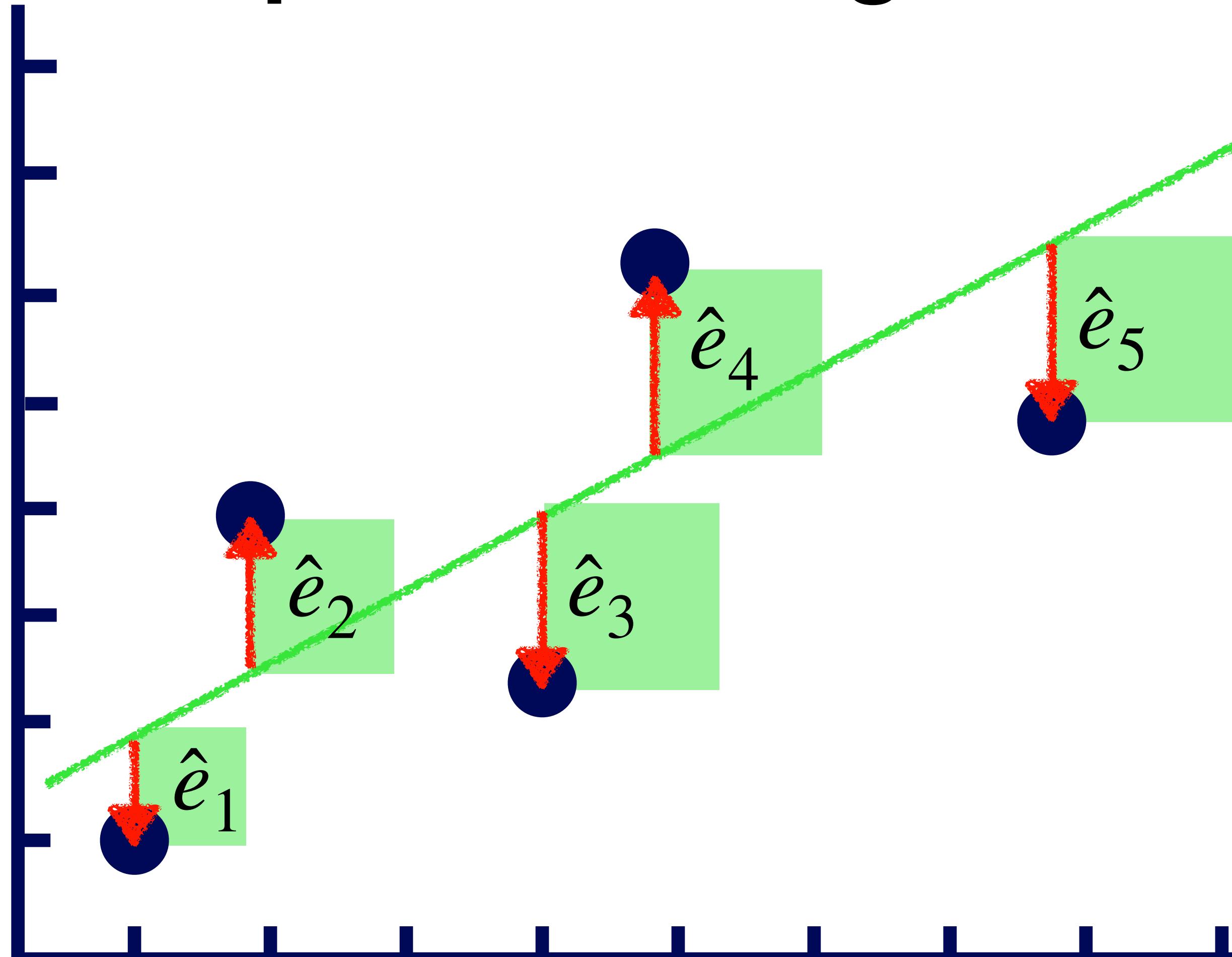
Residuals

$$\hat{e}_i = y_i - \hat{y}_i$$

- What does a positive residual mean?
- What does a negative Residual mean?

BIVARIATE QUANTITATIVE DATA

Simple Linear Regression Model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$



**Sum of Square
Residuals**

$$\sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Want to minimize sum of square residuals w.r.t $\hat{\beta}_0$, and $\hat{\beta}_1$ to get linear model

BIVARIATE QUANTITATIVE DATA

The Derivation

$$0 = \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$0 = \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

<https://www.desmos.com/calculator/lywhybetzt>

BIVARIATE QUANTITATIVE DATA

Formulas

$$\hat{\beta}_1 = r \frac{s_y}{s_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b = r \frac{s_y}{s_x}$$

$$b_1 = r \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

$$b_0 = \bar{y} - b_1 \bar{x}$$

BIVARIATE QUANTITATIVE DATA

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Don't forget that our line of best fit will always pass through (\bar{x}, \bar{y})

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

BIVARIATE QUANTITATIVE DATA

$\hat{\beta}_0$ Represents the average value of "y" when "x" is zero. This is often meaningless

$\hat{\beta}_1$ Represents the average increase in "y" for a **one unit** change in "x". Think Rise/One

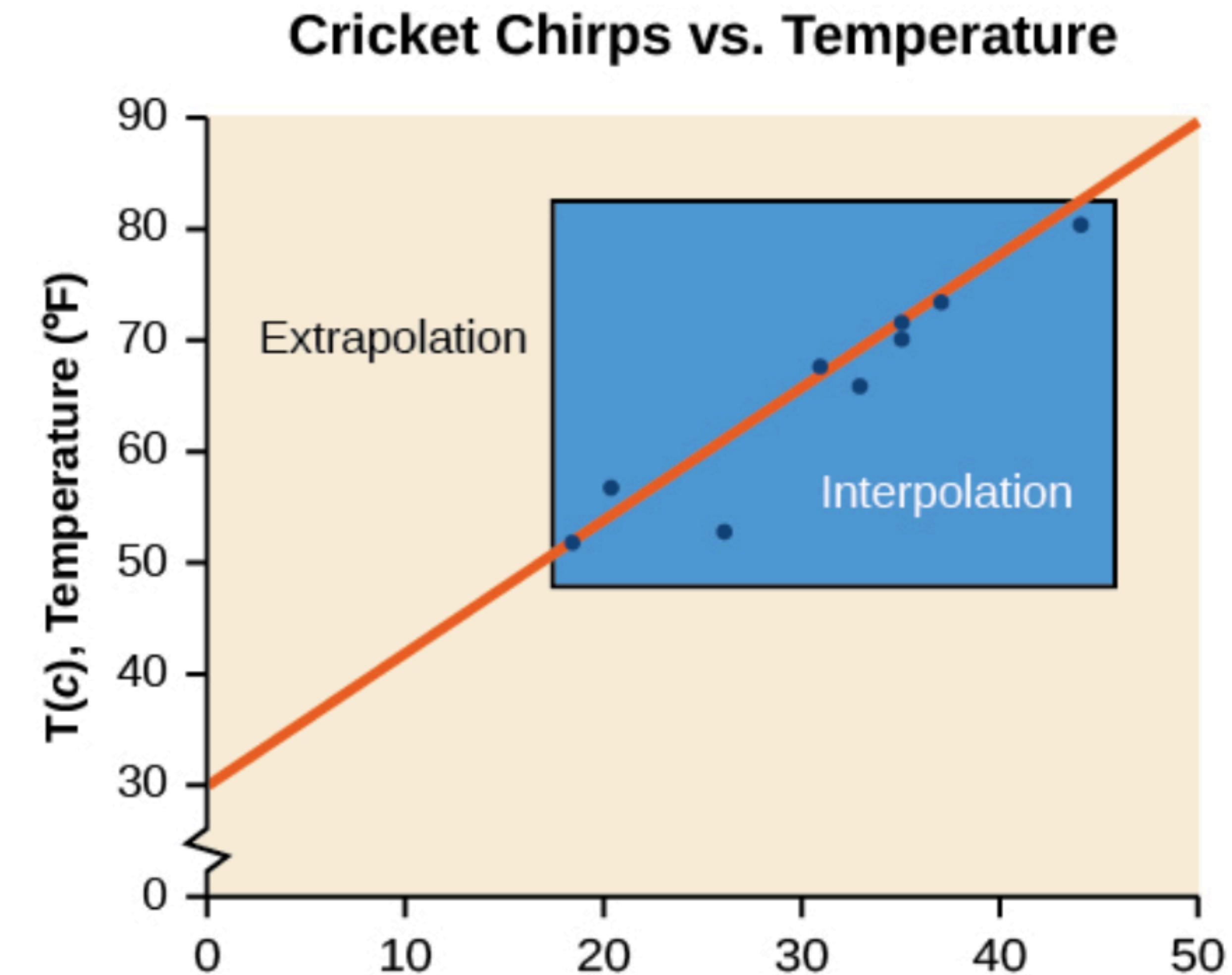
BIVARIATE QUANTITATIVE DATA

Making Predictions: What does a prediction mean?

Average value of y given value of x . “Using our model we would predict an average temperature of Y for x Cricket Chirps in 15 seconds.

What is extrapolation?

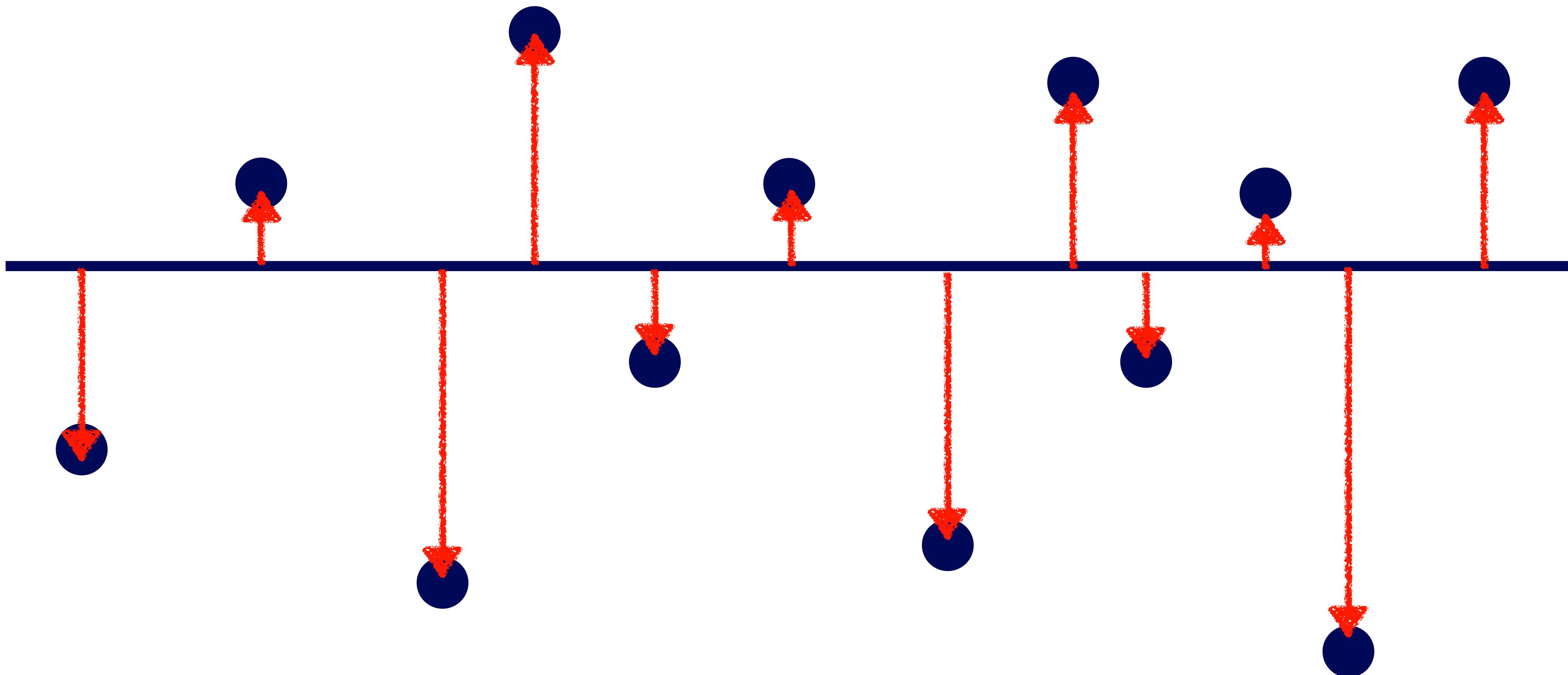
What does 0 cricket chirps in 15 seconds tell us?



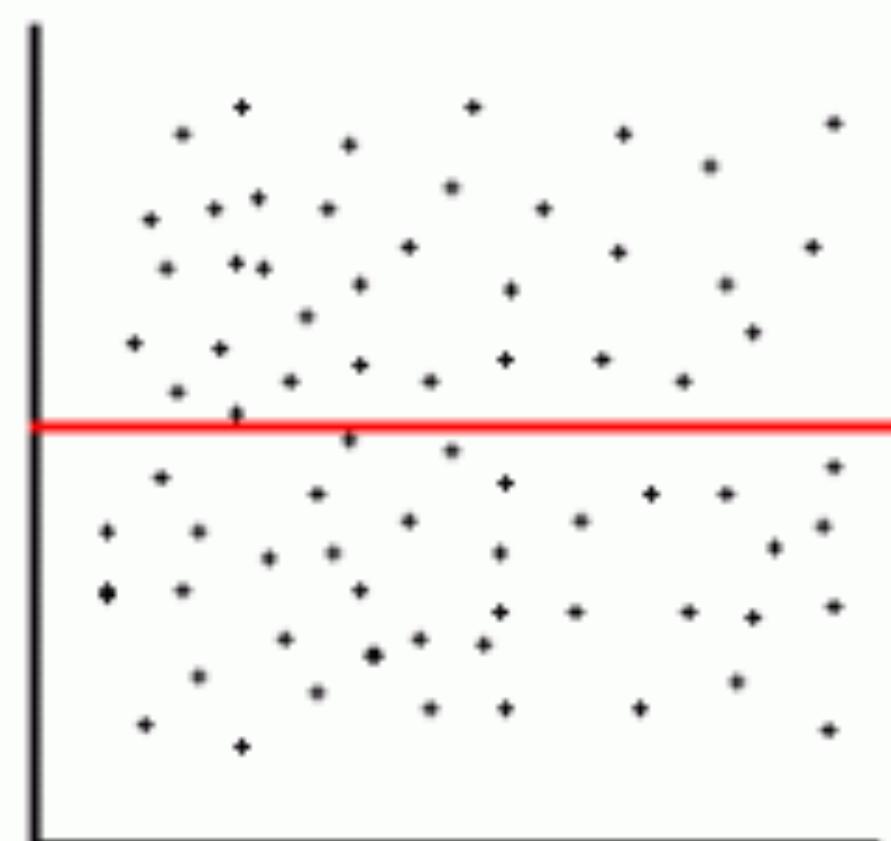
BIVARIATE QUANTITATIVE DATA

Model Validation: What is a residual plot?

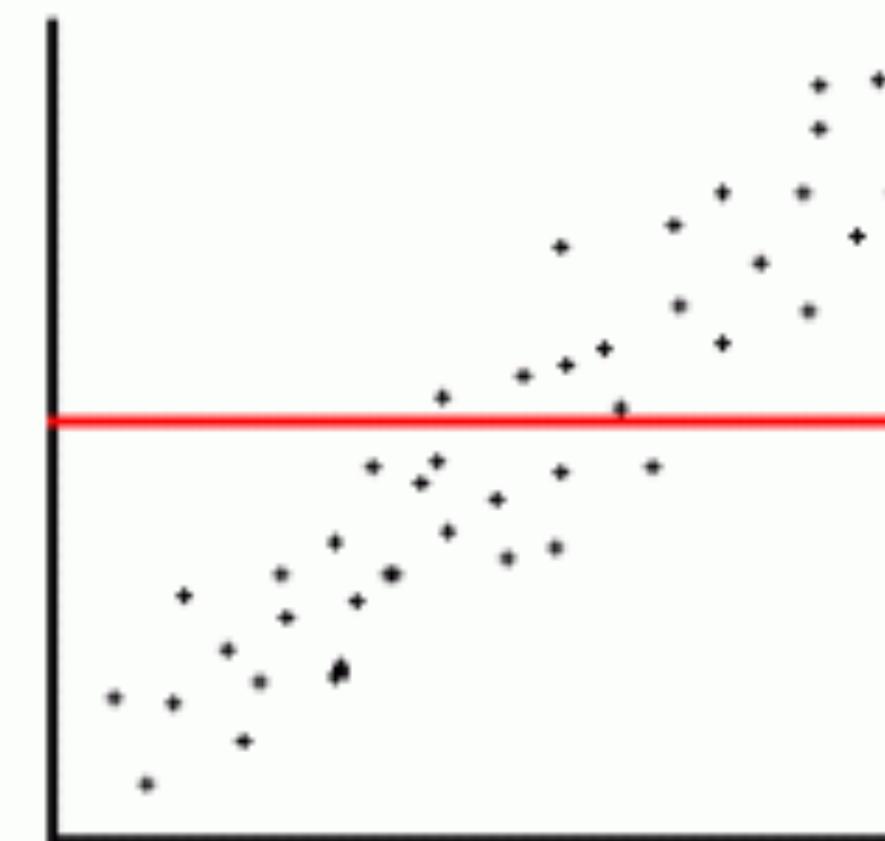
Recall Our Model: $y = \beta_0 + \beta_1 x + \epsilon$ $\epsilon \sim \text{Normal}(0, \sigma^2)$



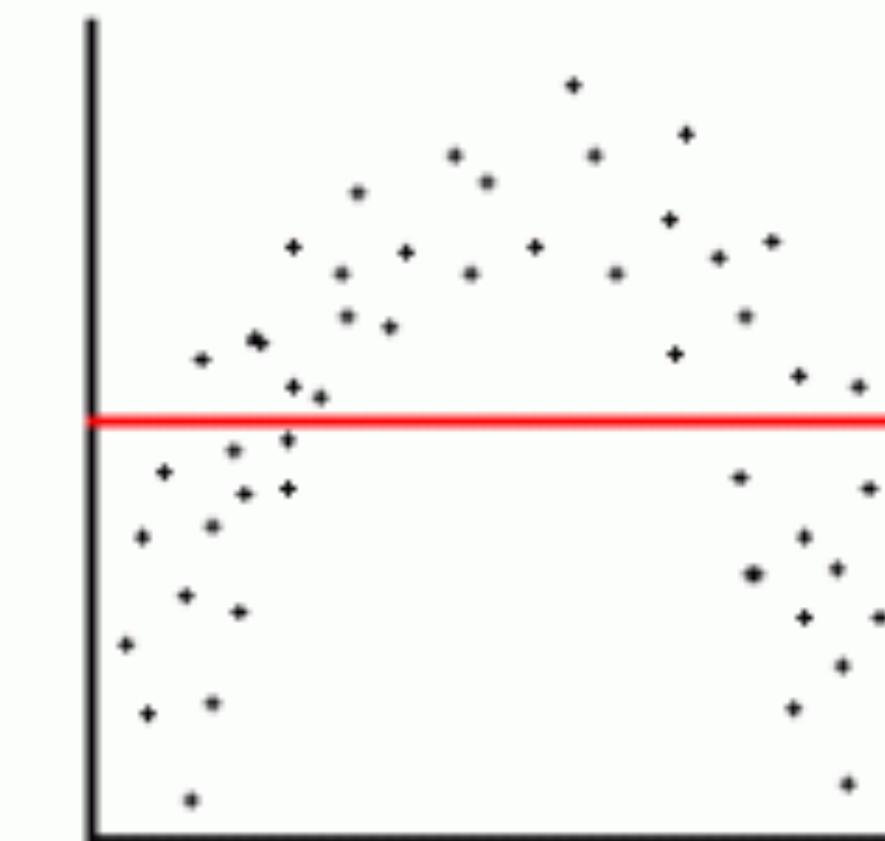
RESIDUAL PLOTS



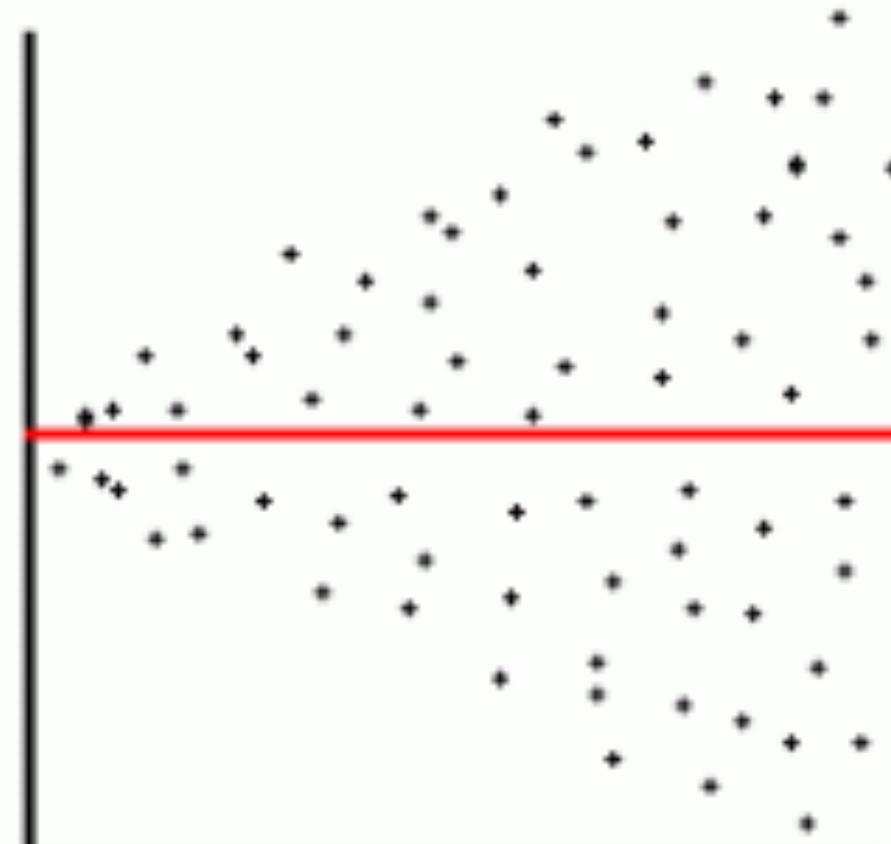
(a) Unbiased and Homoscedastic



(b) Biased and Homoscedastic



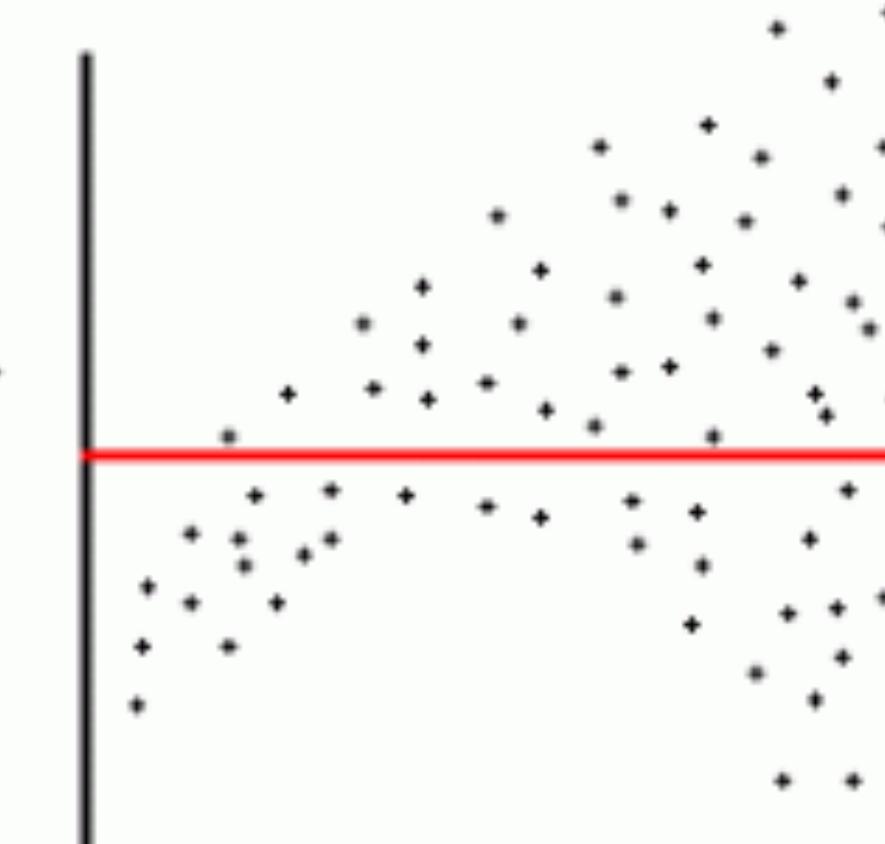
(c) Biased and Homoscedastic



(d) Unbiased and Heteroscedastic

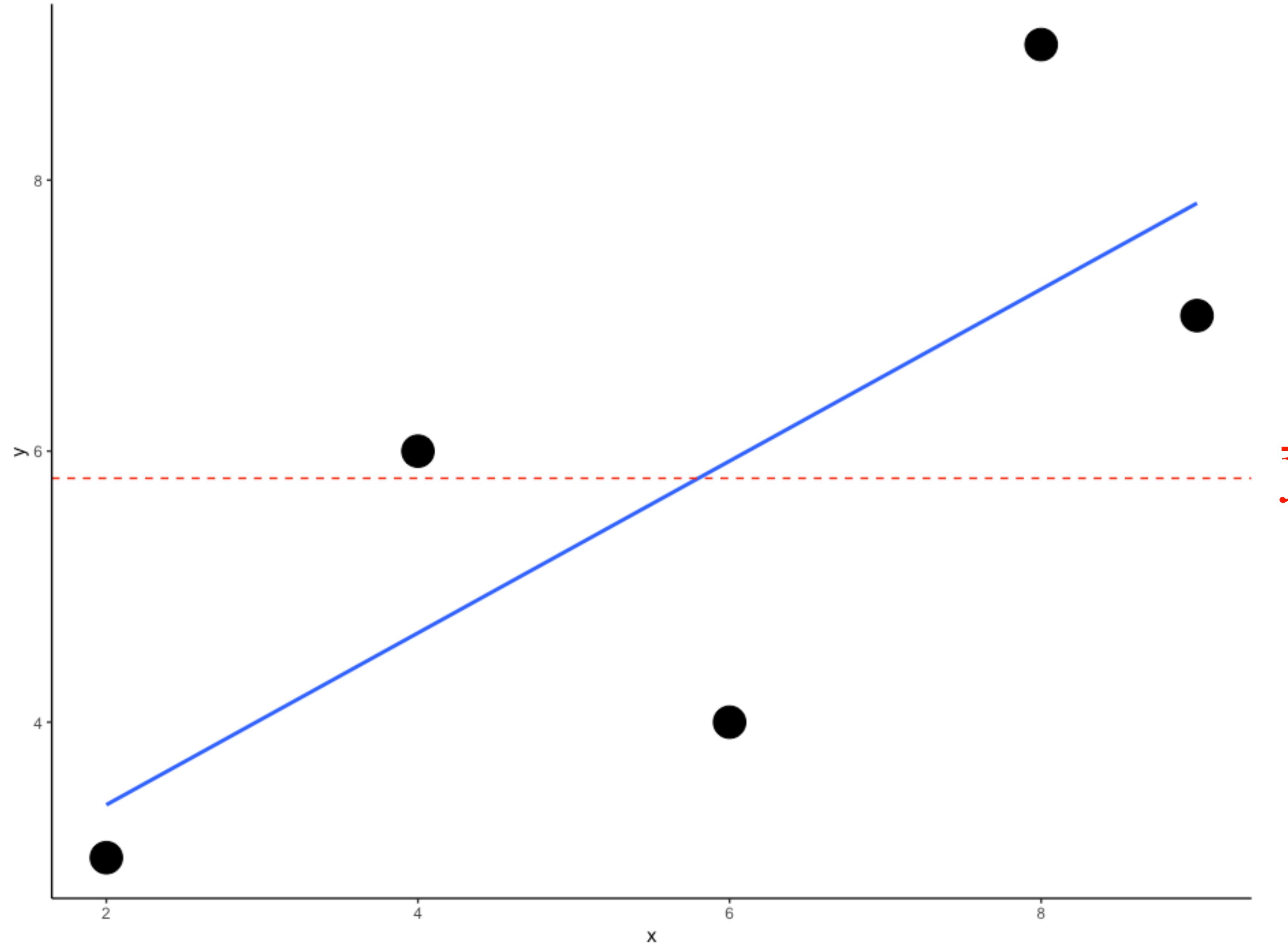


(e) Biased and Heteroscedastic



(f) Biased and Heteroscedastic

BIVARIATE QUANTITATIVE DATA

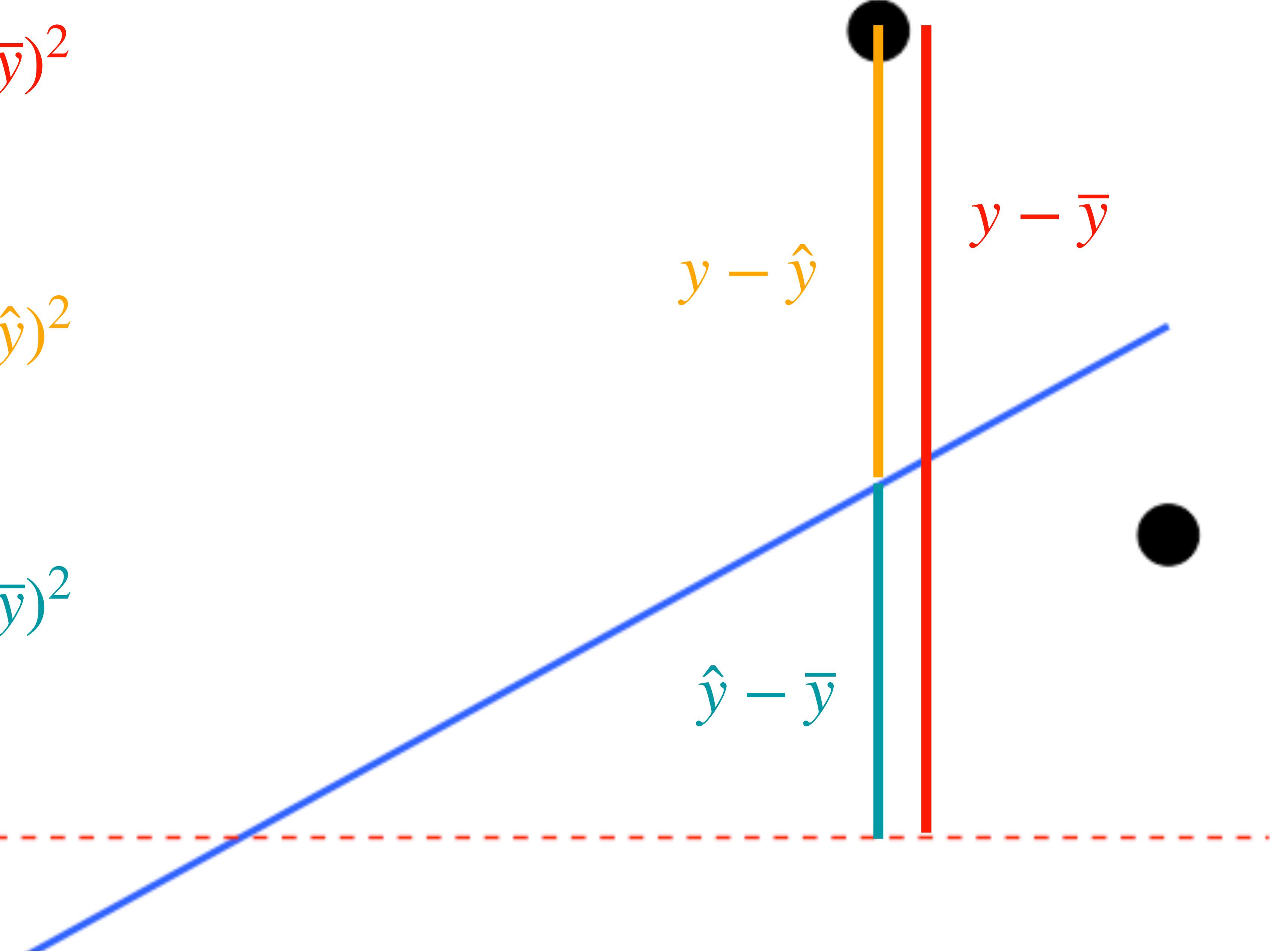


Model Validation:
What is r^2 ?

$$SST = \sum_{i=1}^n (y - \bar{y})^2$$

$$SSE = \sum_{i=1}^n (y - \hat{y})^2$$

$$SSR = \sum_{i=1}^n (\hat{y} - \bar{y})^2$$

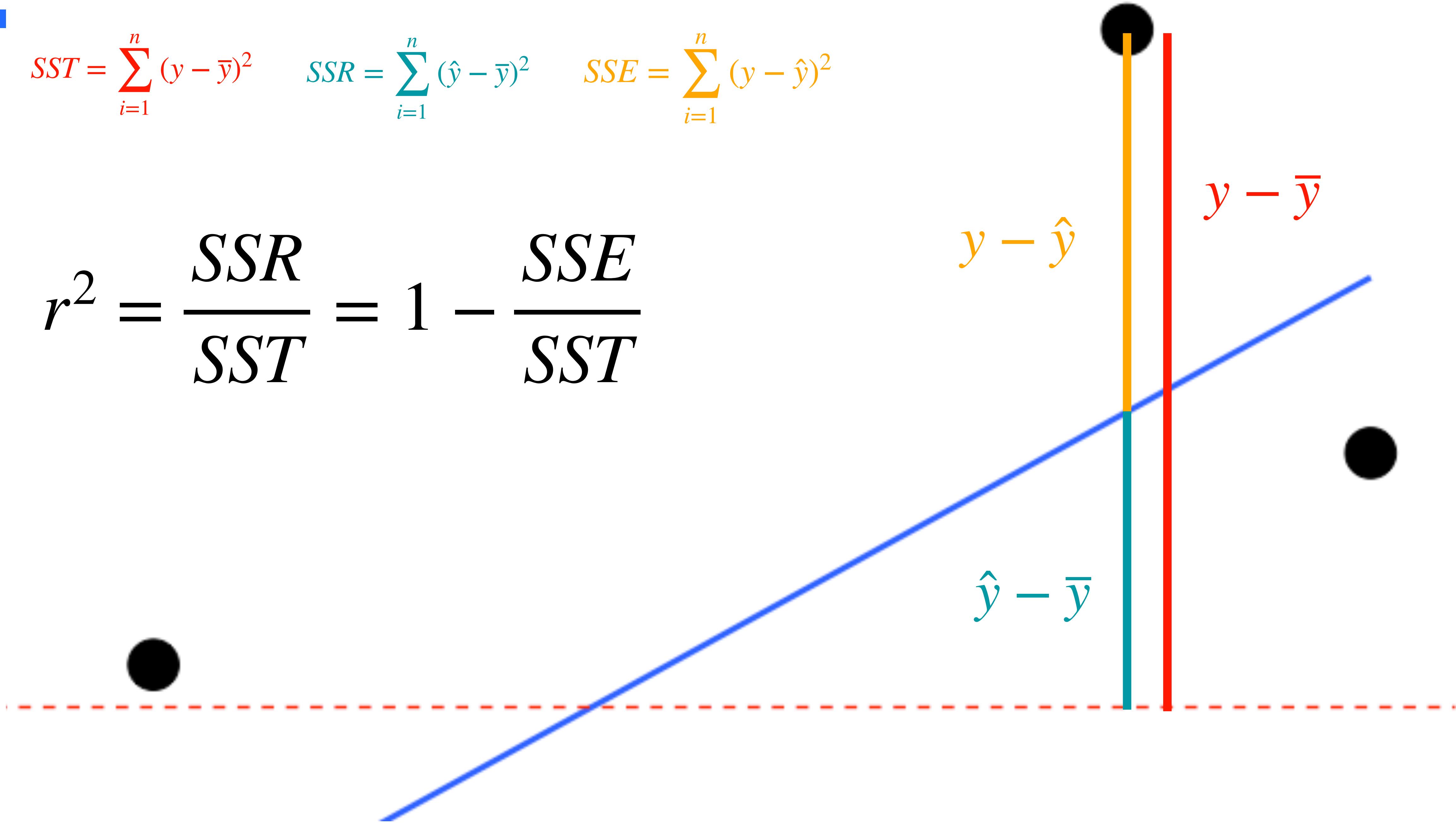


$$SST = \sum_{i=1}^n (y - \bar{y})^2$$

$$SSR = \sum_{i=1}^n (\hat{y} - \bar{y})^2$$

$$SSE = \sum_{i=1}^n (y - \hat{y})^2$$

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

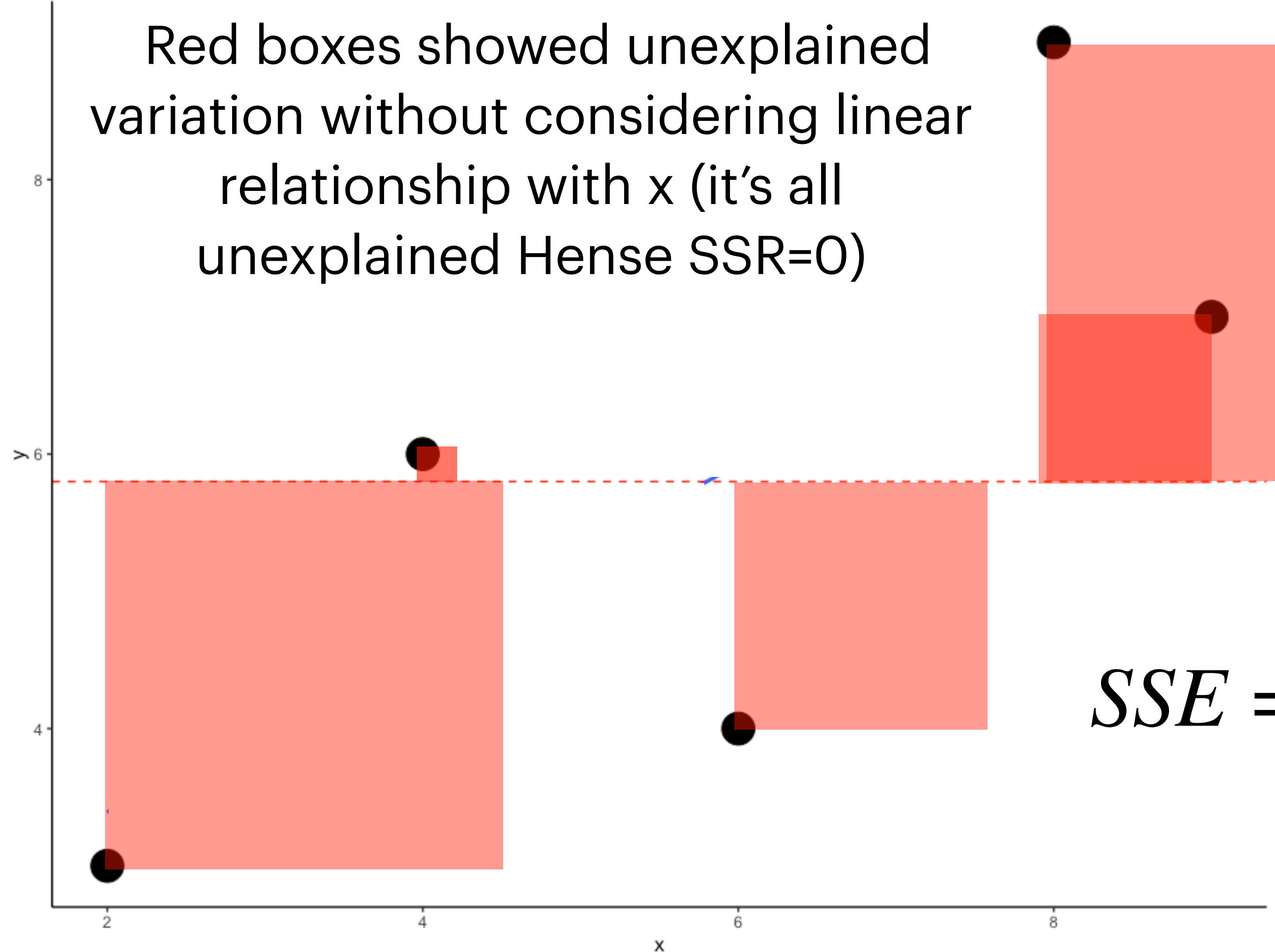


$$SST = \sum_{i=1}^n (y - \bar{y})^2 \quad SSR = \sum_{i=1}^n (\hat{y} - \bar{y})^2 \quad SSE = \sum_{i=1}^n (y - \hat{y})^2$$

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

r^2 describes the percentage of variation in “y” that can be explained by “y’s” linear relationship with “x”

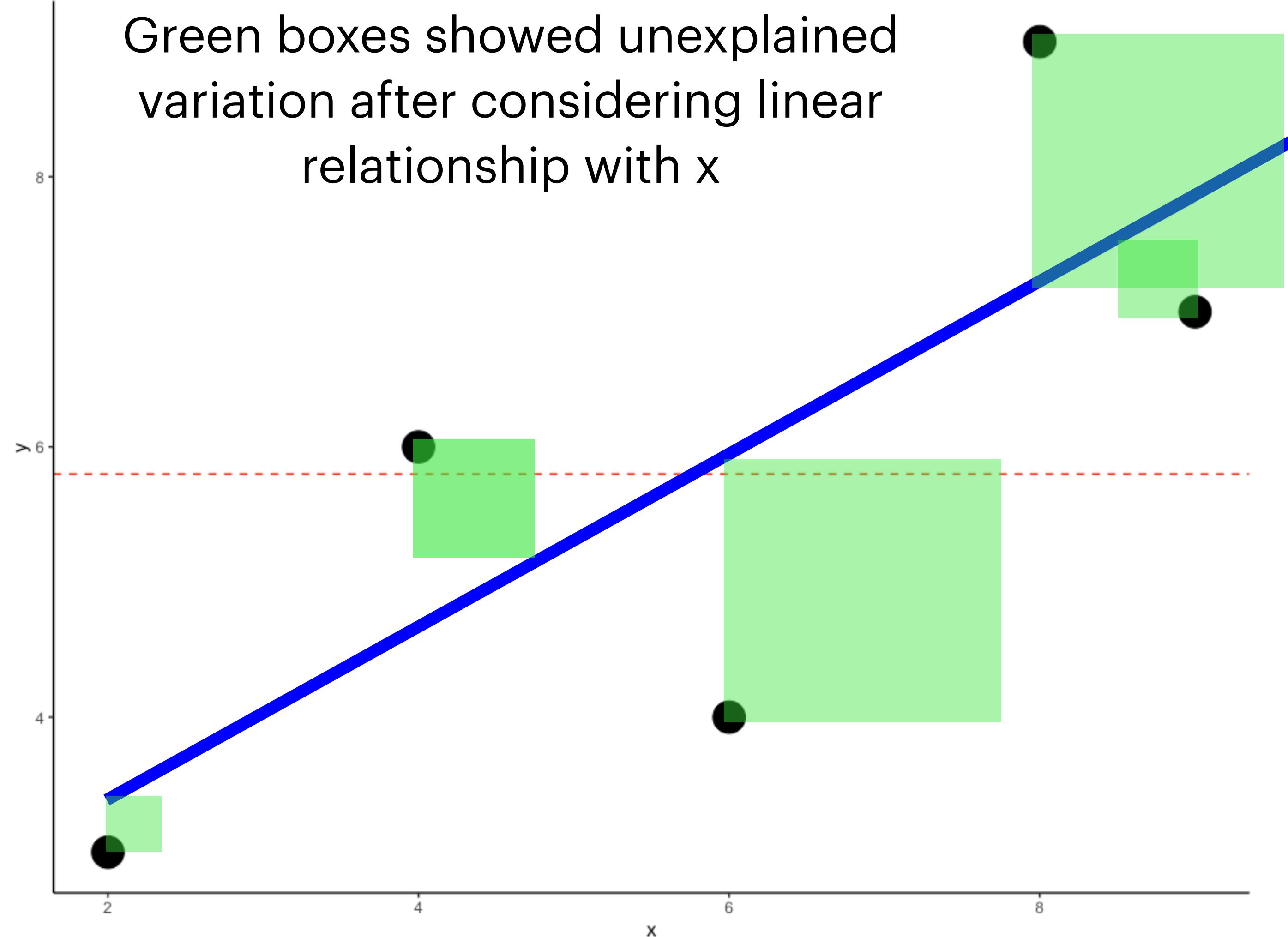
Red boxes showed unexplained variation without considering linear relationship with x (it's all unexplained Hence SSR=0)



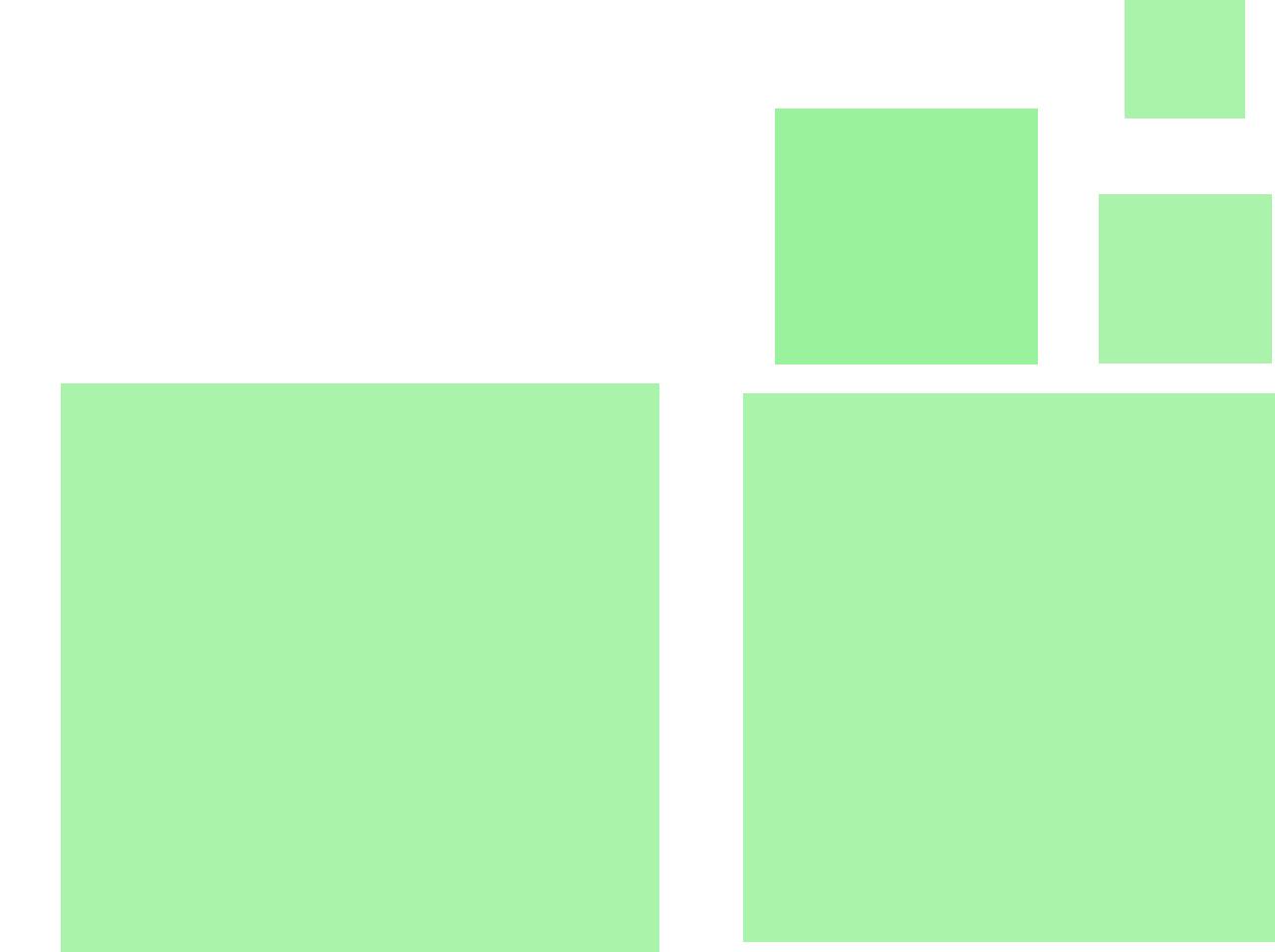
$$SSE = \sum_{i=1}^n (y - \hat{y})^2$$



Green boxes showed unexplained variation after considering linear relationship with x



$$SSE = \sum_{i=1}^n (y - \hat{y})^2$$

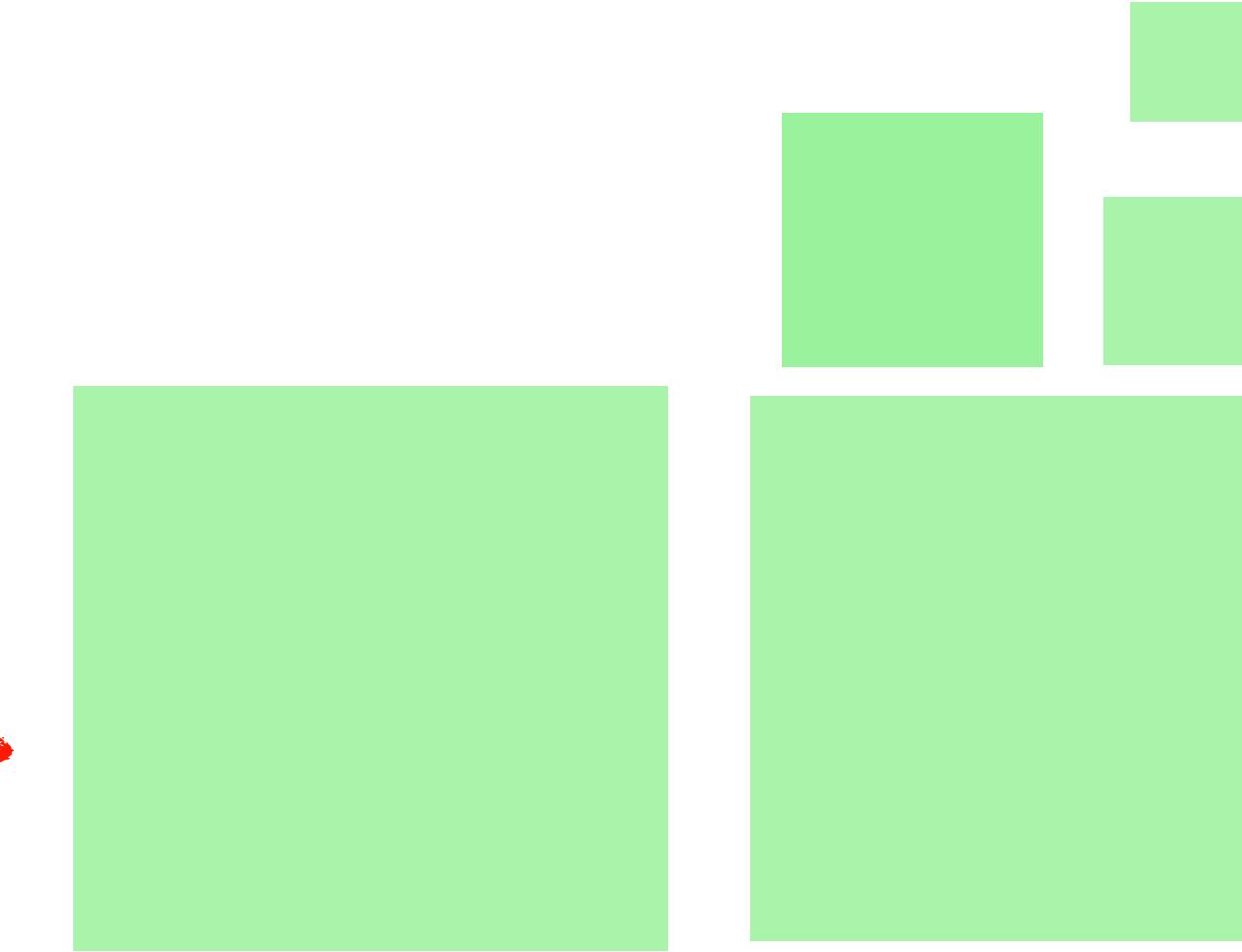




$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

When we consider the linear relationship the unexplained variance is reduced by $\frac{SSE}{SST}$ percent. The percentage “explained” by the model is $\frac{SSR}{SST}$

$$SSE = \sum_{i=1}^n (y - \hat{y})^2$$



Standard Error of the regression Line tells us the average residual length, in other words the average amount our model over/under predicts.

$$s = \sqrt{\frac{SSE}{n - 1}}$$

Not expected to calculate by hand

BIVARIATE CATEGORICAL DATA

Examples:

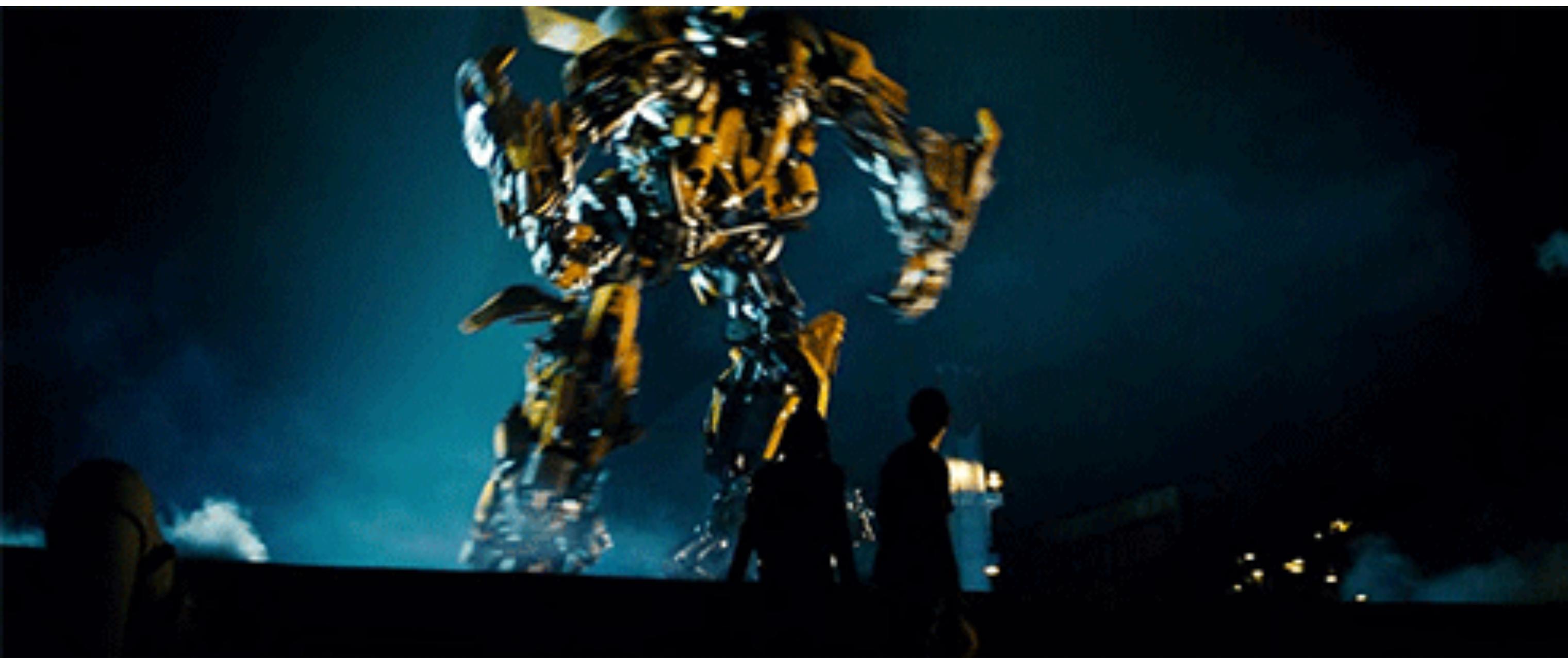
- Deep Thoughts Unit 2 Q1-Q4
- Question 1 Page 143

Homework: Read Pages 113-130 Barron's,
Quiz 8, Quiz 9

BIVARIATE QUANTITATIVE DATA

Transformations

Scatter Plot is Non-Linear



Linear Model Appropriate

There are many different transformations we might use

Method	Transform	Regression equation	Predicted value (\hat{y})
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	$DV = \log(y)$	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	$DV = \sqrt{y}$	$\sqrt{y} = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	$DV = 1/y$	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	$IV = \log(x)$	$y = b_0 + b_1 \log(x)$	$\hat{y} = b_0 + b_1 \log(x)$
Power model	$DV = \log(y)$ $IV = \log(x)$	$\log(y) = b_0 + b_1 \log(x)$	$\hat{y} = 10^{b_0 + b_1 \log(x)}$

Transformations

Example: the length of a year for a planet, based on its distance from the sun. Here are the data:

Distance (millions of miles)	Year (# of Earth-years)
36	0.24
67	0.61
93	1
142	1.88
484	11.86
887	29.46
1784	84.07
2796	164.82
3666	247.68

Transformations

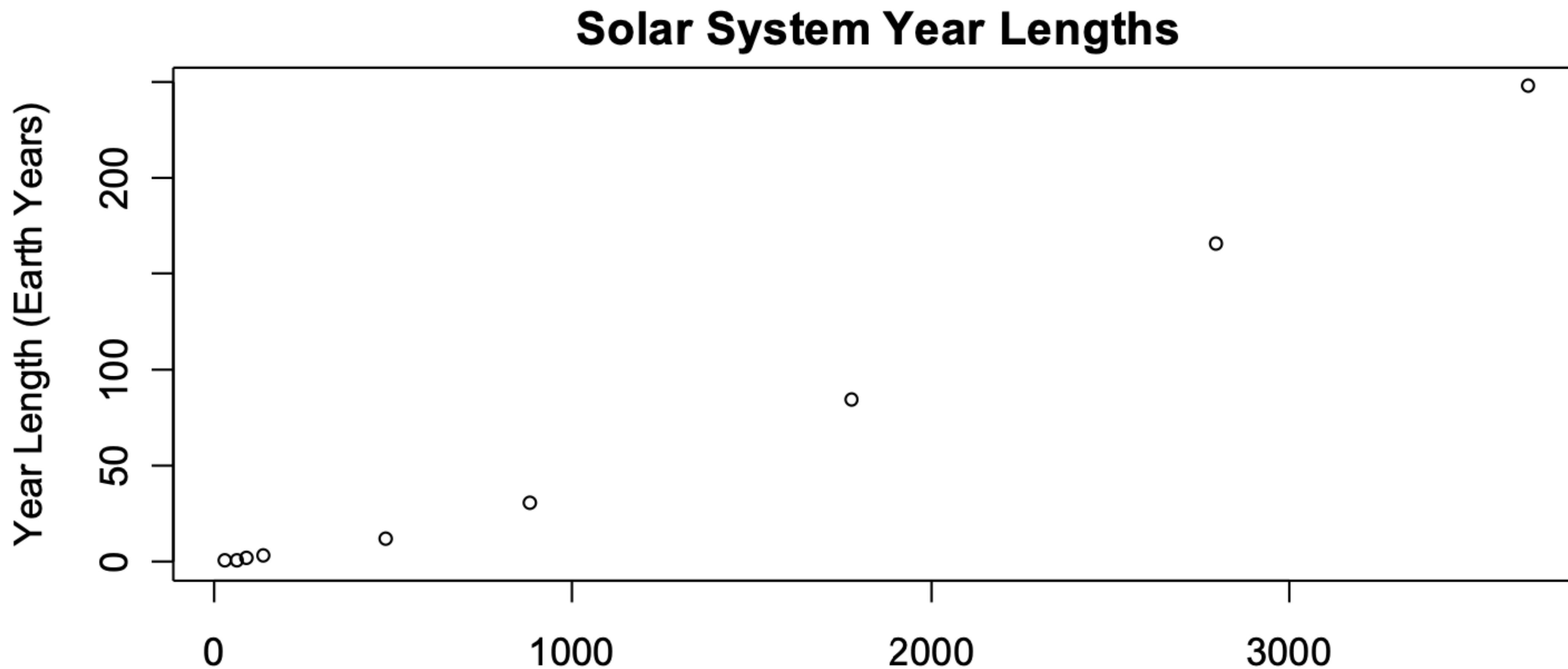
1. Let's run a simple linear regression.

What is r^2 ?

Is the Model Appropriate?

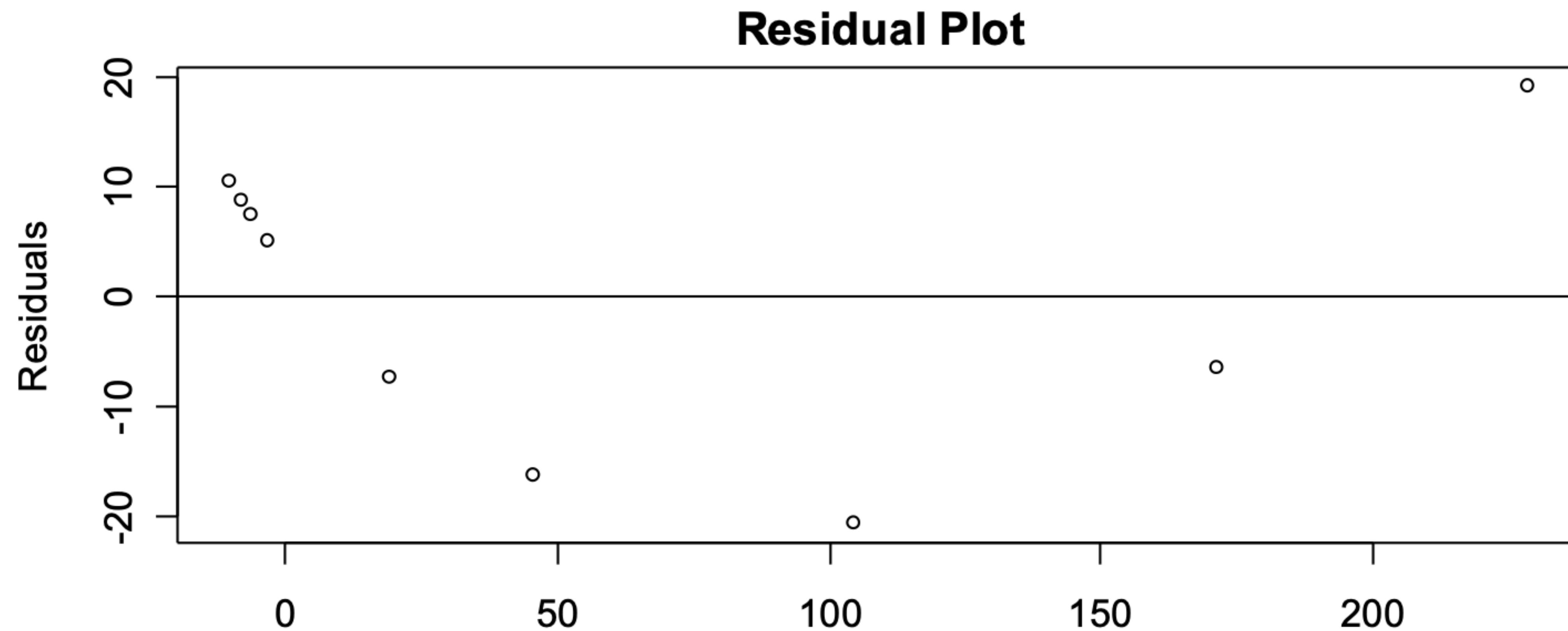
Transformations

Scatter Plot Looks **non-linear**



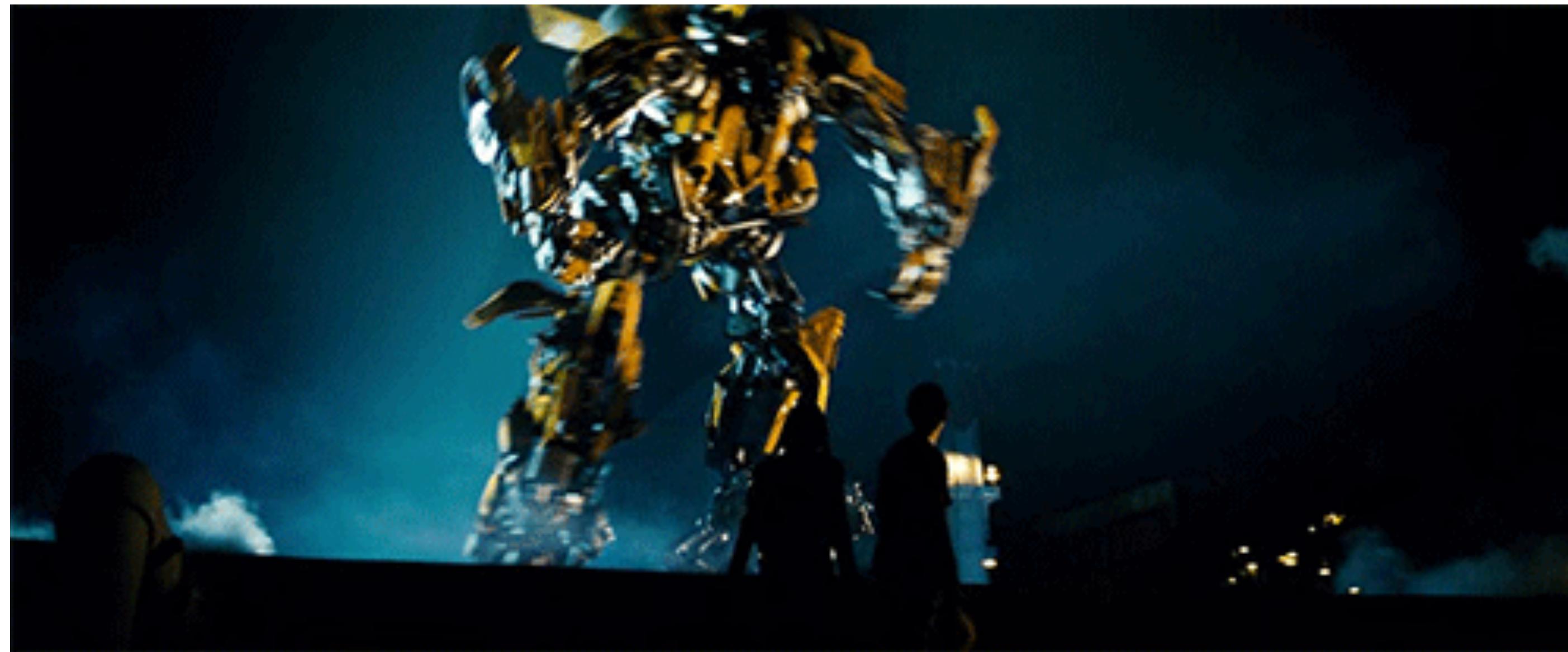
Transformations

Residual plot makes non-linear pattern even more clear



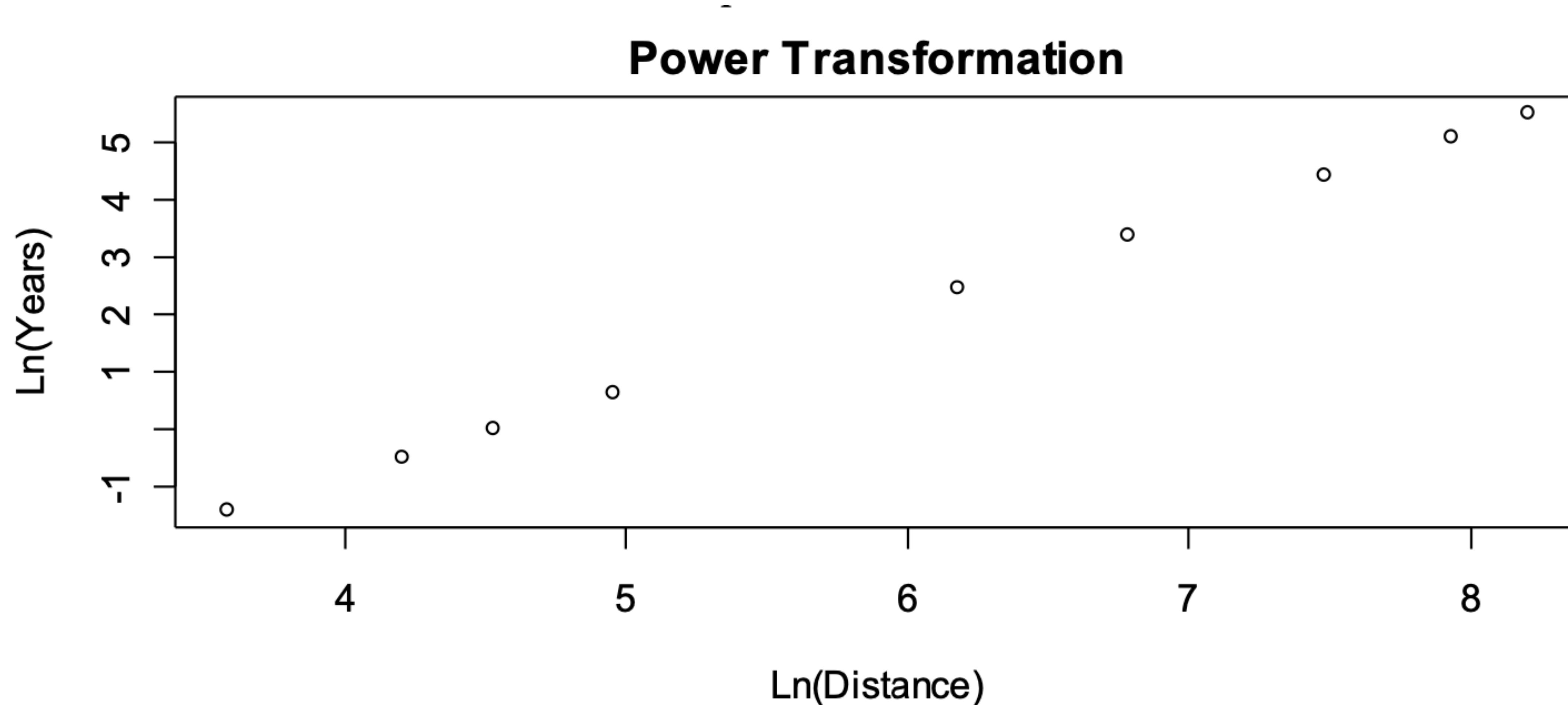
Transformations

1. Let's run a simple linear regression.
2. **Problem:** EW, that's not linear. Lets apply a power transformation



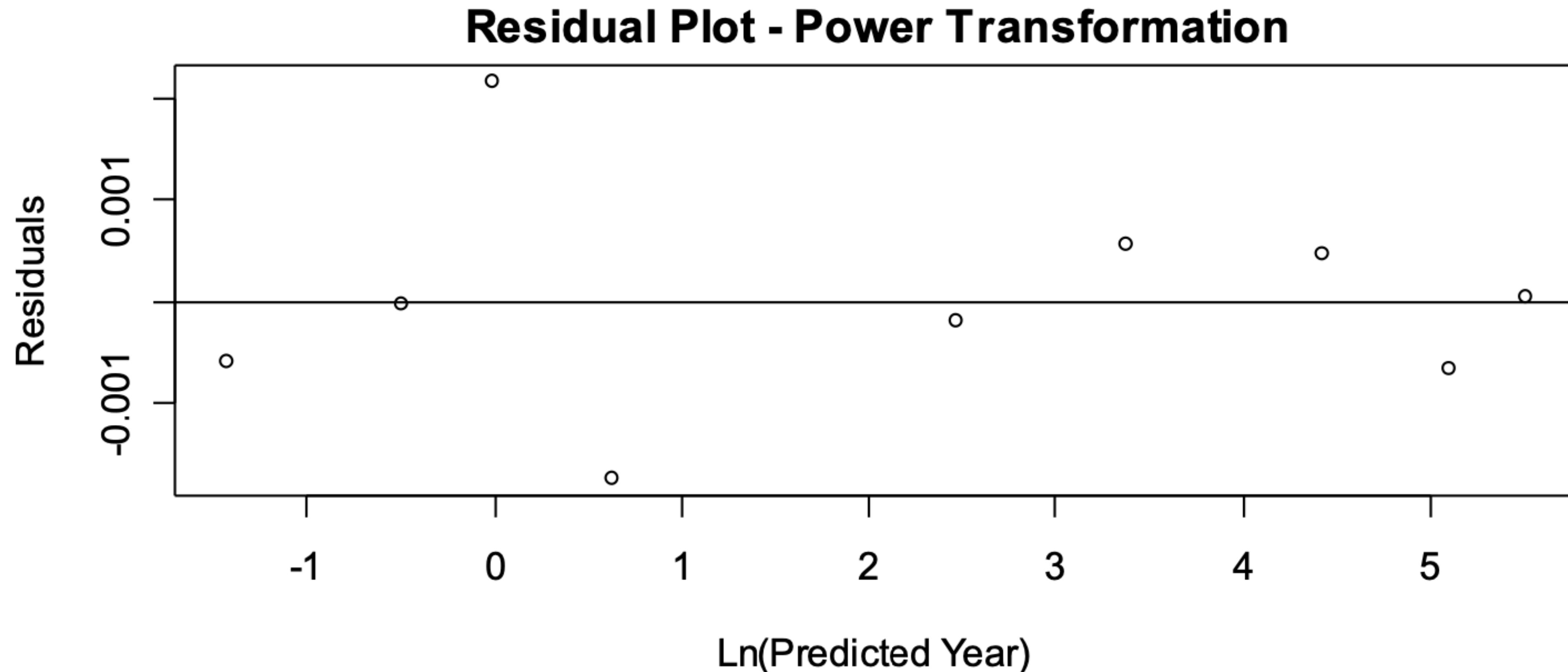
Transformations

New Scatter Plot Looks much more linear



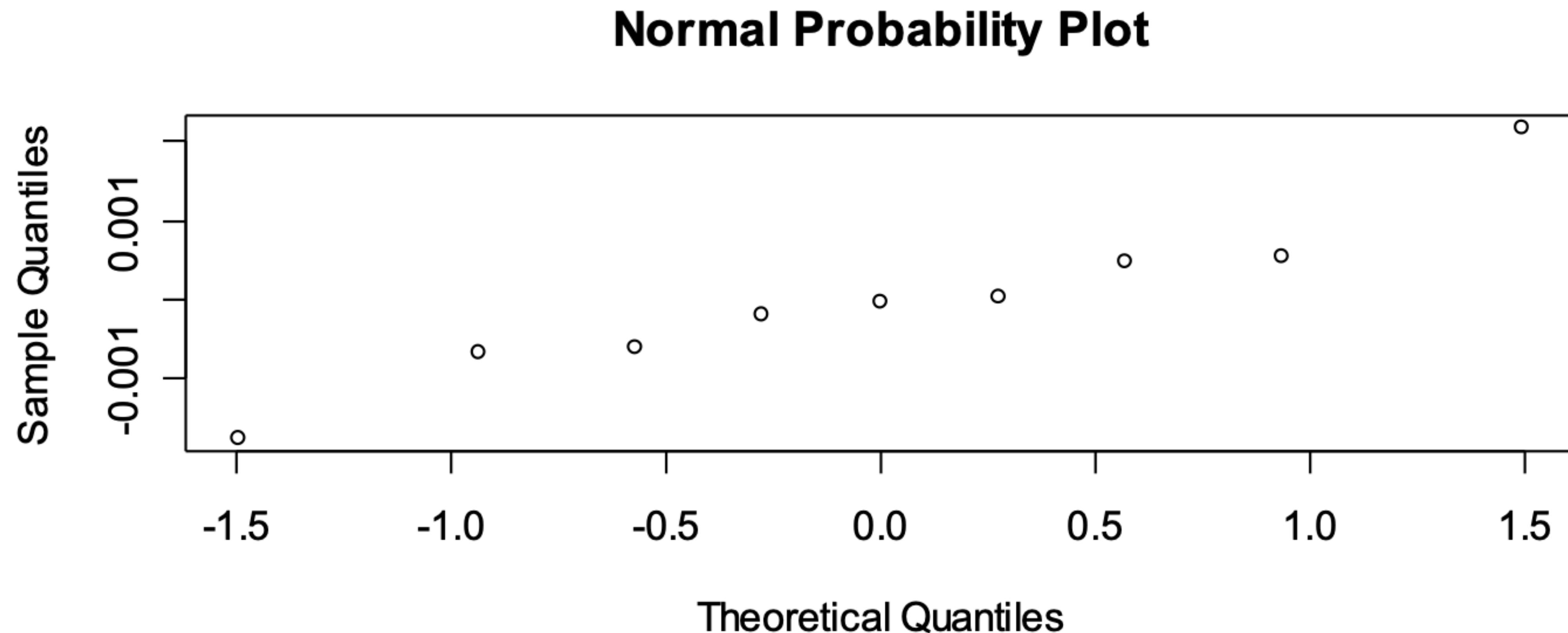
Transformations

Residual plot improves significantly



Transformations

Normality in residuals isn't bad either!



Transformations

1. Lets run a simple linear regression.
2. **Problem:** EW, that's not linear. Lets apply a power transformation
3. Run simple linear regression with transformed data.
4. Model: $\ln(\hat{y}) = -6.8046 + 1.5008 \cdot \ln(x)$

BIVARIATE QUANTITATIVE DATA

Model: $\ln(\hat{y}) = -6.8046 + 1.5008 \cdot \ln(x)$

Let's use this model to predict the year length of a planet that doesn't exist. The halfway point between Mars and Jupiter is around 313 million miles from Sol. What will this model predict for a year length if a planet occupied this position?

BIVARIATE QUANTITATIVE DATA

Model: $\ln(\hat{y}) = -6.8046 + 1.5008 \cdot \ln(x)$

$$\ln(\hat{y}) = -6.8046 + 1.5008 \cdot \ln(313)$$

$$\ln(\hat{y}) = 1.8192$$

$$\hat{y} = e^{1.8192} = 6.167$$

BIVARIATE QUANTITATIVE DATA

What you need to know

- Recognize the need for a transformation
- Justify a transformations appropriateness

Examples:

- Barron's pg. 130 Example 2.26
- Deep Thoughts Q5-Q6