Math 11
Calculus-Based Introductory Probability and Statistics

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

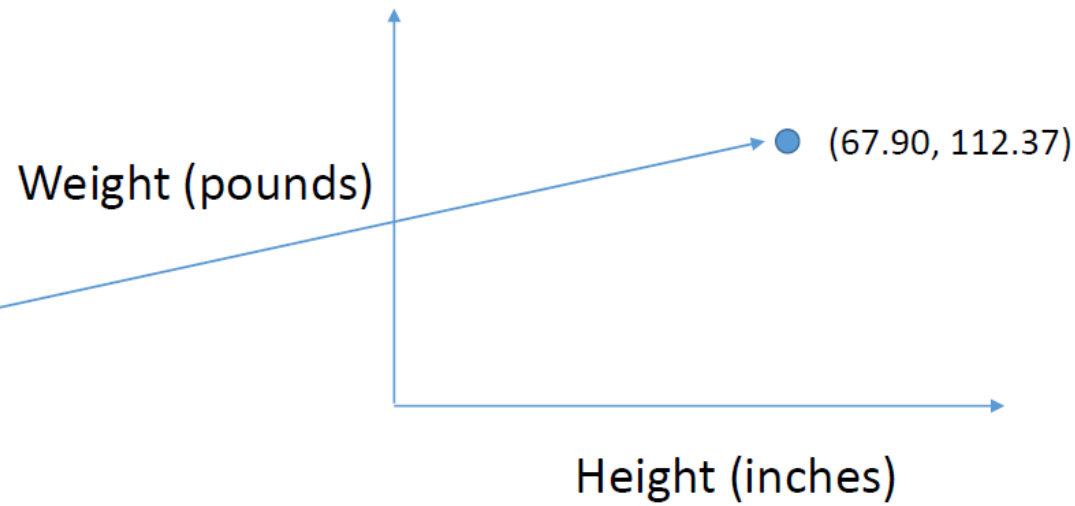
Today:

- Scatterplots
- Correlation coefficient
- Correlation VS Causation
- Linear Regression

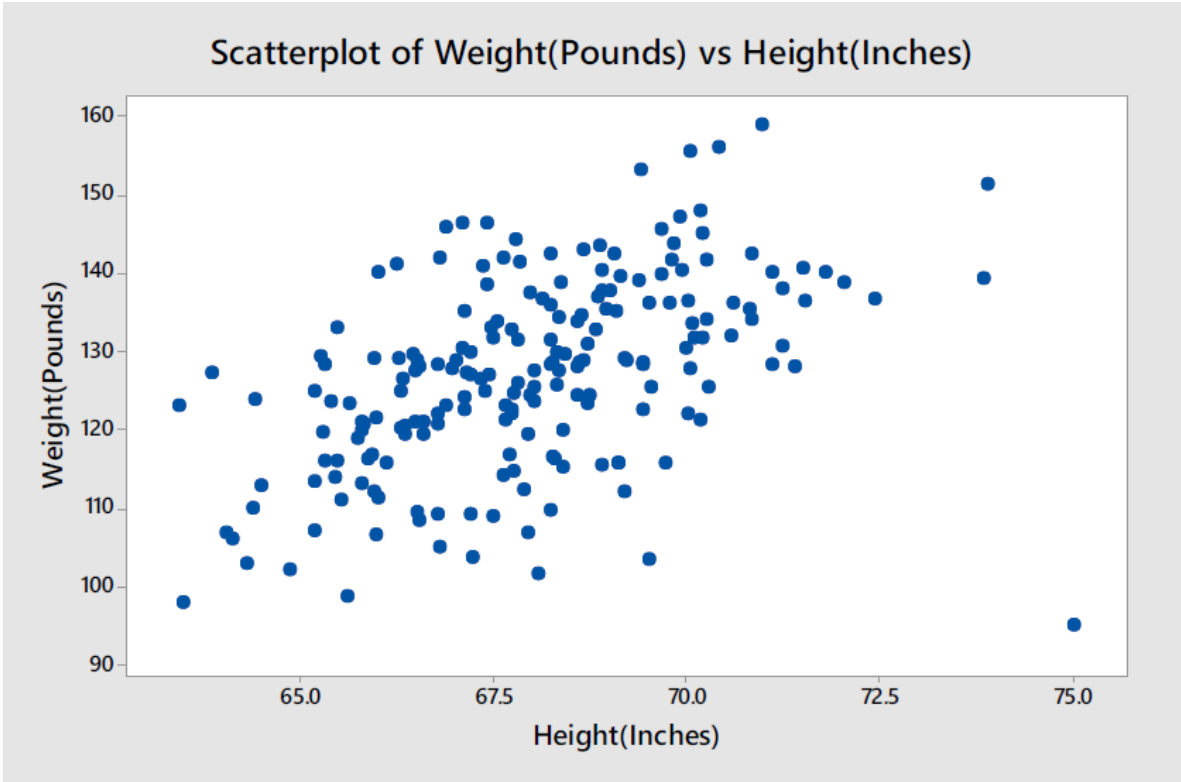
Scatterplots

Height(Inches)	Weight(Pounds)
65.78	112.99
71.52	136.49
69.40	153.03
68.22	142.34
67.79	144.30
68.70	123.30
69.80	141.49
70.01	136.46
67.90	112.37
66.78	120.67
66.49	127.45
67.62	114.14

Each row holds two pieces of data about the **same** person



Scatterplots



A scatterplot shows the relationship between two quantitative variables.

The x-axis variable is known as the **predictor** or explanatory variable.

The y-axis variable is the **response** variable.

How to Discus Scatterplots: FDSO

Form:

How to Discus Scatterplots: FDSO

Form:

Linear Curve Repeating/Oscillating

How to Discus Scatterplots: FDSO

Form:

Linear Curve Repeating/Oscillating

Direction:

Positive Negative No direction

How to Discuss Scatterplots: FDSO

Strength:

Very strong Strong Weak

How to Discuss Scatterplots: FDSO

Strength:

Very strong Strong Weak

Outliers:

As with a single quantitative variable, the notion of outlier is vague. Just think: "A point that stands far apart from the overall trend in the scatterplot."

How to Discuss Scatterplots: FDSO

Strength:

Very strong Strong Weak

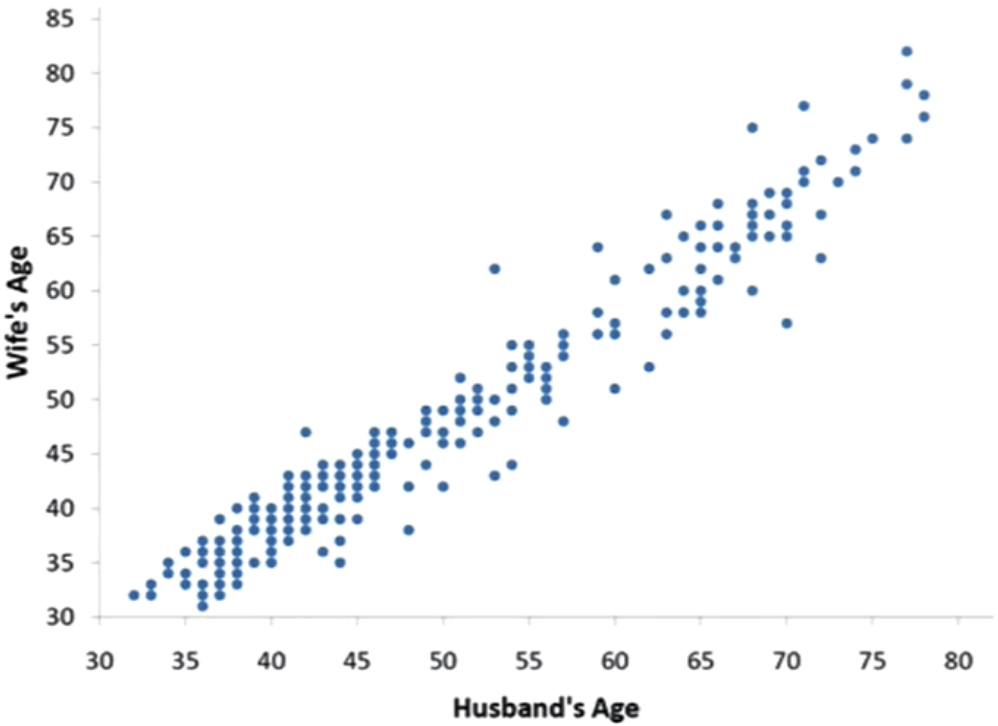
Outliers:

This scatterplot has a weak to moderate, positive, linear association. We have an outlier.

As with a single quantitative variable, the notion of outlier is vague. Just think: "A point that stands far apart from the overall trend in the scatterplot."

Your Turn! Describing Real Data Using FDSO

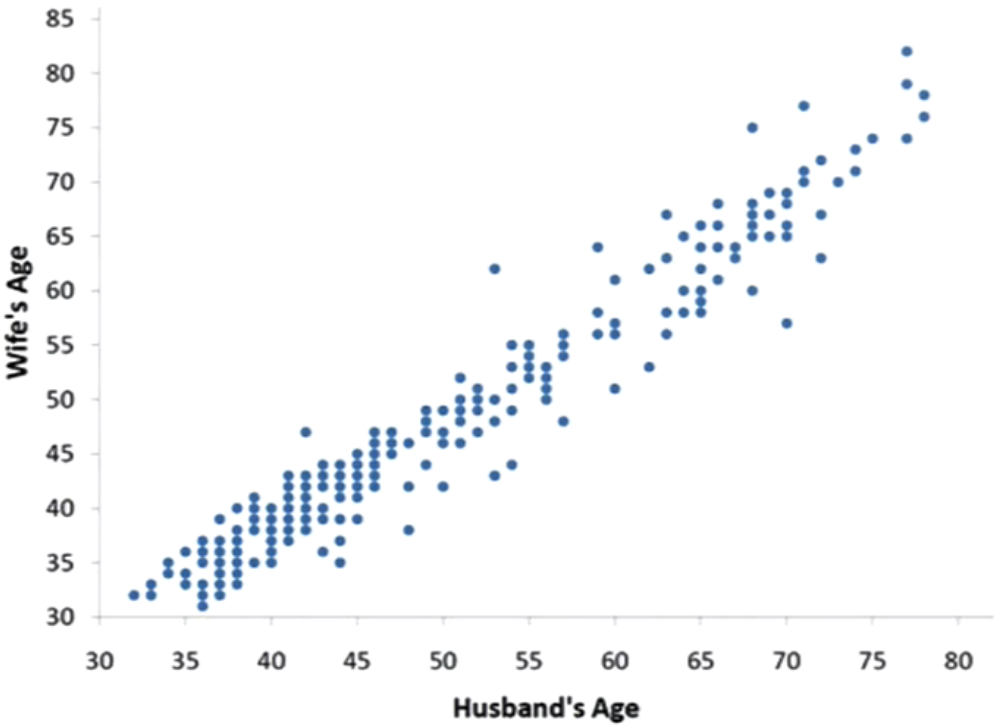
Age of husband and Age of wife



- Form:
- Direction:
- Strenght:
- Outliers:

Your Turn! Describing Real Data Using FDSO

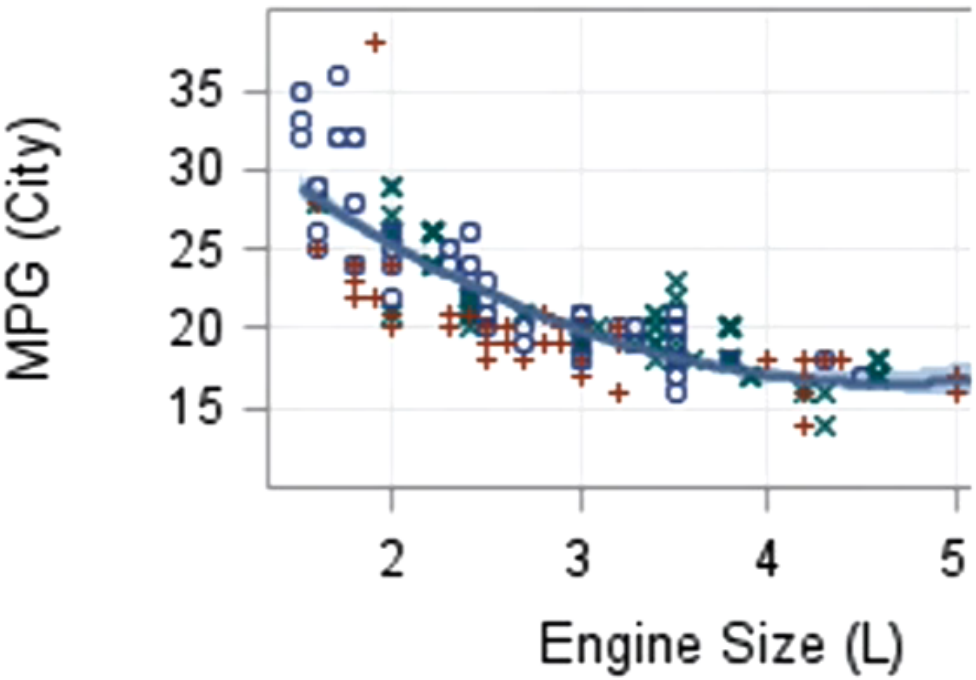
Age of husband and Age of wife



- Form: Linear
- Direction: Positive
- Strenght: Strong
- Outliers: None

Your Turn! Describing Real Data Using FDSO

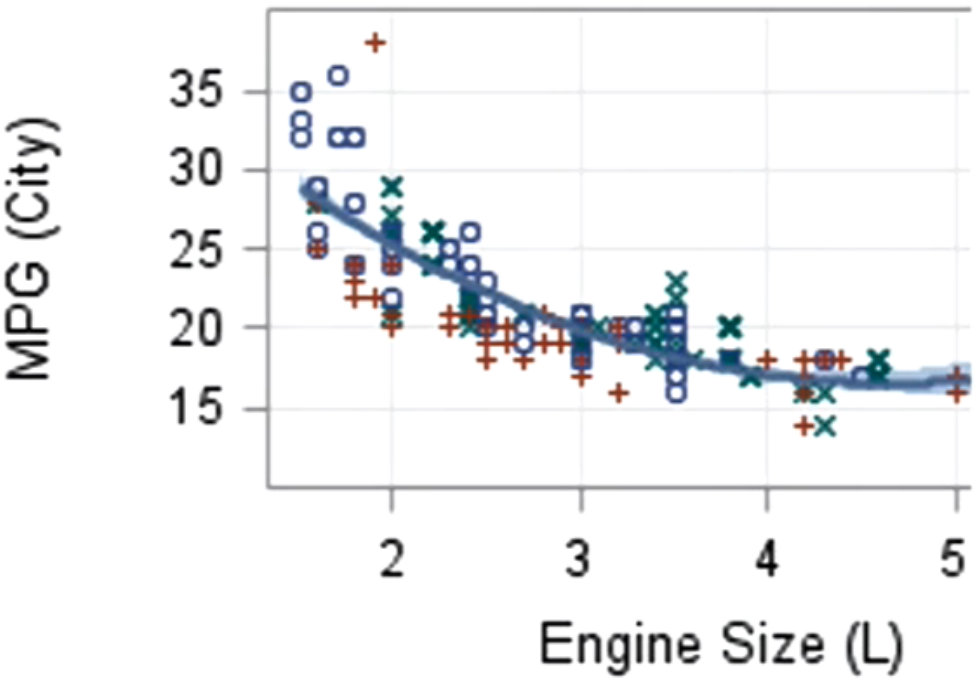
Engine size of a car (liters) and Fuel economy (MPG)



- Form:
- Direction:
- Strenght:
- Outliers:

Your Turn! Describing Real Data Using FDSO

Engine size of a car (liters) and Fuel economy (MPG)



- Form: Curved
- Direction: Negative
- Strenght: Moderate
- Outliers: Maybe the top red plus...?

Moving Beyond Vague Language

Correlation: A statistic that measures the Srength and Direction of a linear association (Form) between two quantitative variables where no Outliers are present.

Reported using either an uppercase R or a lowercase r:

$$R = r = -0.3.$$

Moving Beyond Vague Language

Correlation: A statistic that measures the Srength and Direction of a linear association (Form) between two quantitative variables where no Outliers are present.

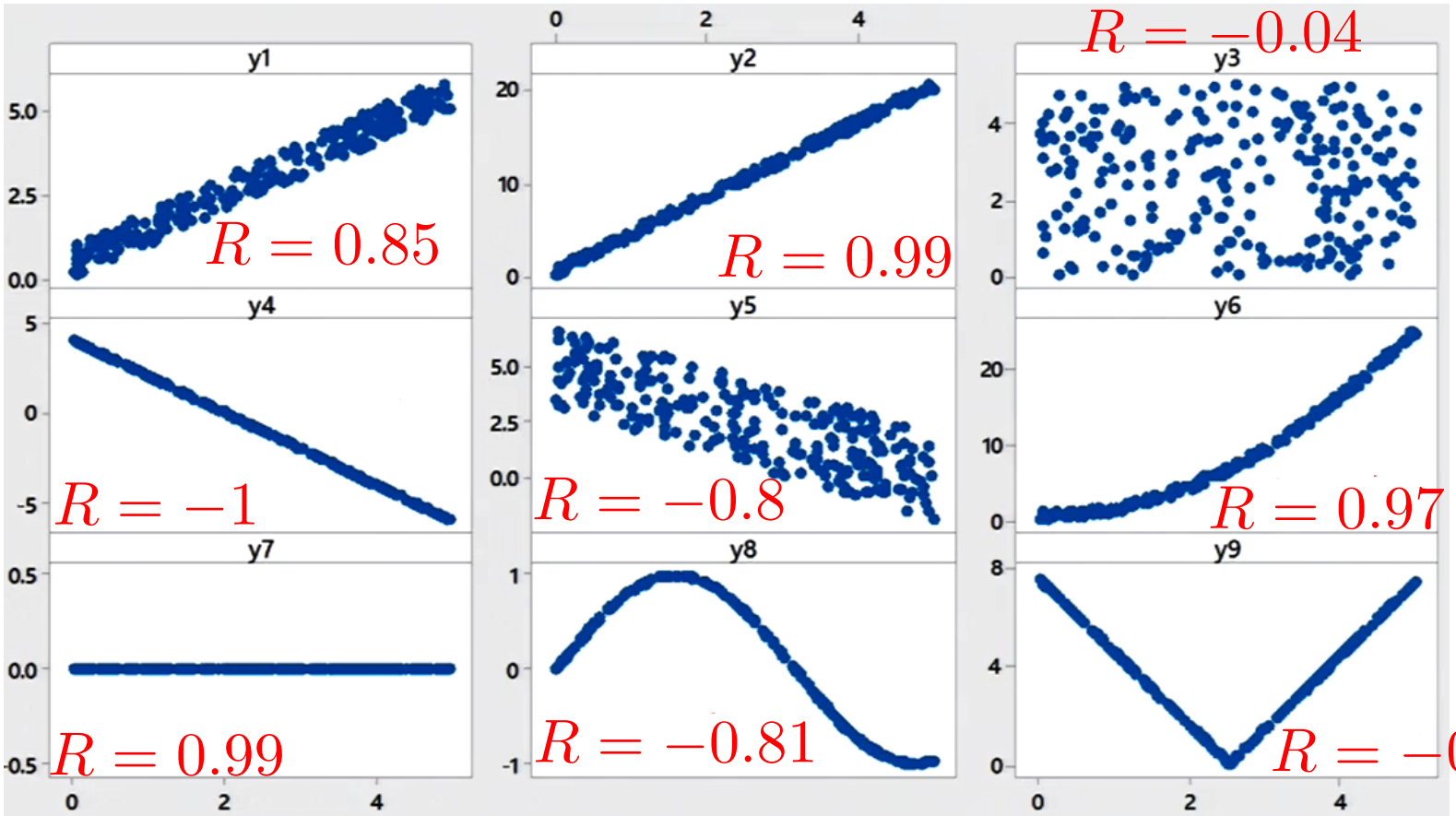
Reported using either an uppercase R or a lowercase r:

$$R = r = -0.3.$$

Note: The programming language R is NOT named after the R statistic, but for the first names of its inventers (Ross Ihaka and Robert Gentleman) and as a reference to a related language S.

Deriving Knowledge about R Using Examples

Looking at these examples, make observations about the R statistic.



Facts About R, The Correlation Statistic

- We always have $-1 \leq R \leq 1$.
- The Direction is encoded in the sign:

$$\begin{aligned} + &: \quad \text{positive association} \\ - &: \quad \text{negative association} \end{aligned}$$

- The Srength is encoded in the value:

$$\begin{aligned} \text{Stronger association} &: \quad \text{occurs when } R \text{ is closer to 1 or -1} \\ \text{Weaker association} &: \quad \text{occurs when } R \text{ is closer to 0} \end{aligned}$$

- 1 and -1 are possible: the data must perfectly lie on a line
- Choice of predictor/response doesn't matter:

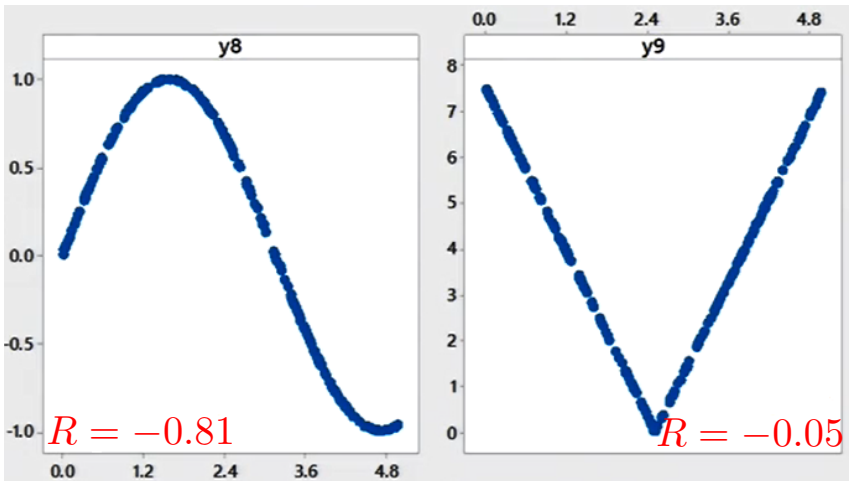
$$cor(X, Y) = cor(Y, X)$$

- Correlation is a unitless idea
- Correlation is unaffected by linear scale changes

$$cor(X, Y) = cor(X, 3.14Y) = cor(X, Y + 100) = cor(2X - 6.8, 95Y + 7)$$

Cautions About R

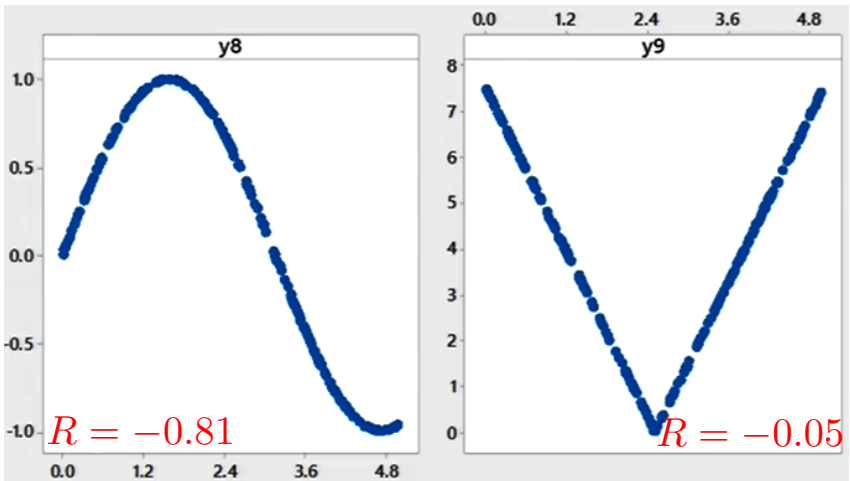
Never use R to measure correlation for associations that are non-linear.



Correlation doesn't measure strength in non-linear associations

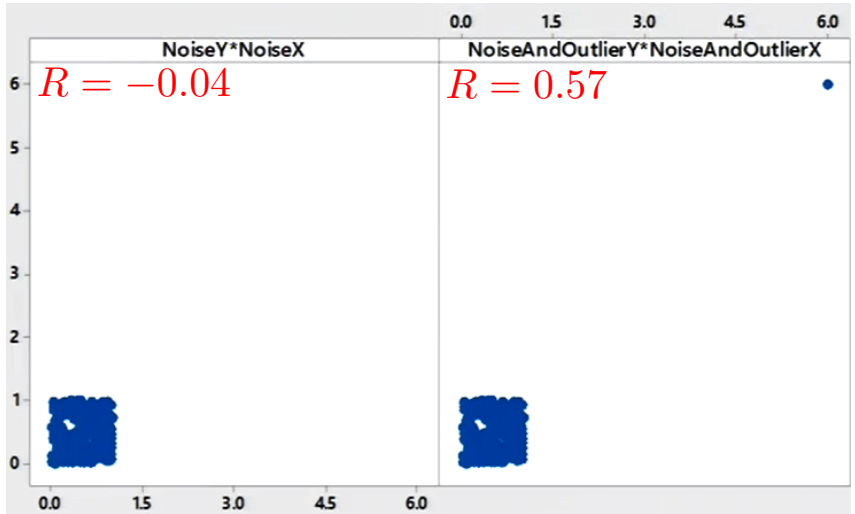
Cautions About R

Never use R to measure correlation for associations that are non-linear.



Correlation doesn't measure strength in non-linear associations

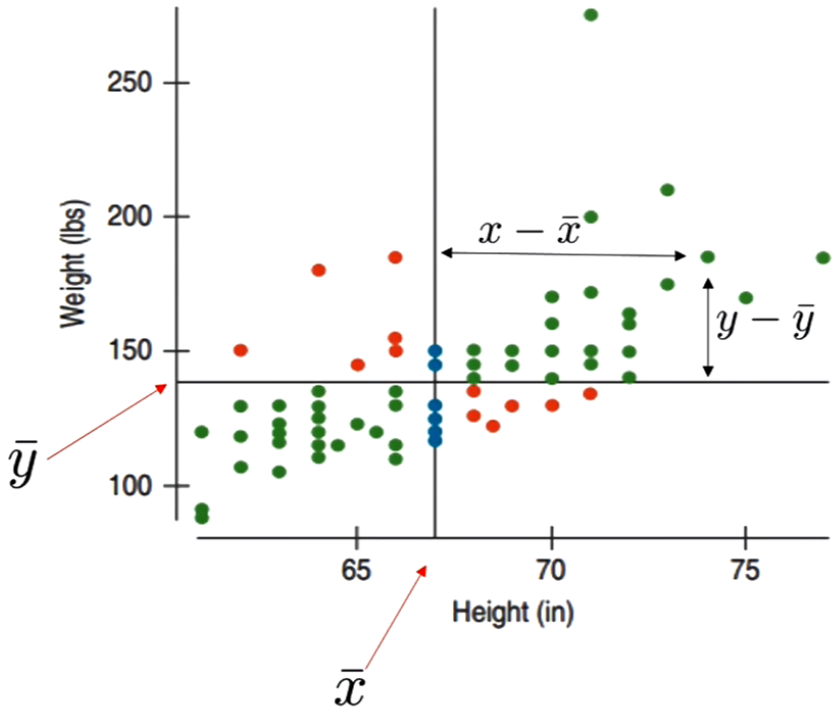
Never use R to measure correlation if outliers are present (even one!).



The effect of a single outlier can be huge.

How is R Calculated?

$$R = \frac{1}{n-1} \sum_{(x,y) \text{ pairs}} \frac{(x-\bar{x})}{s_x} \frac{(y-\bar{y})}{s_y}$$



Numerator: Factors that accumulate positive and negative directions based on the individual points

Denominator: Factors that eliminate scaling for each axis, and ensure that $-1 \leq R \leq 1$.

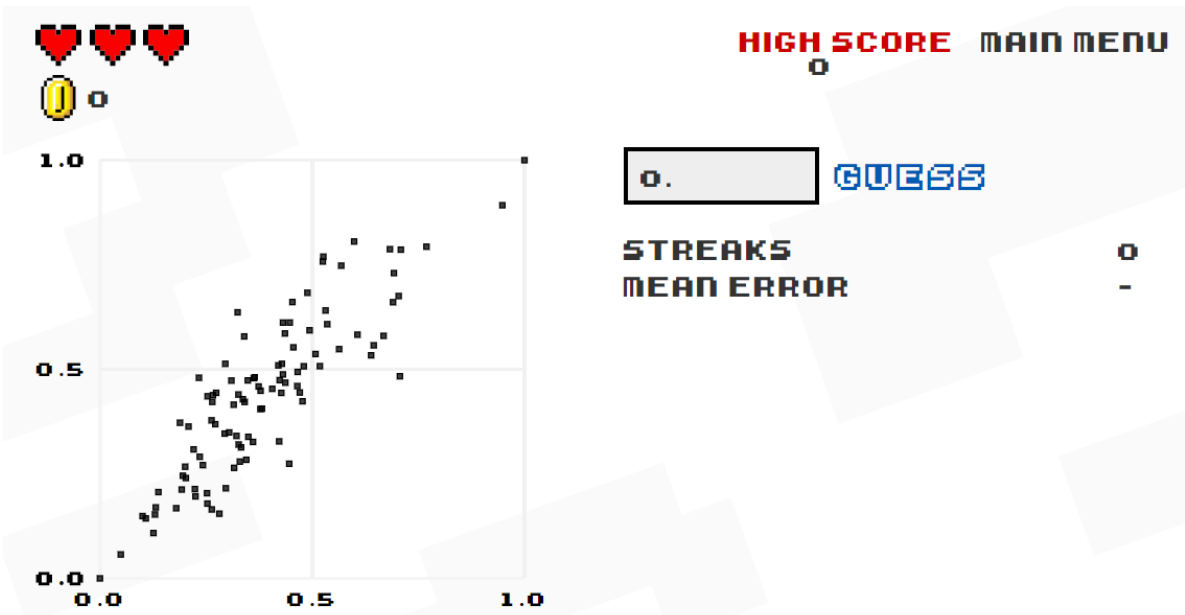
(R is almost always calculated in a statistical program, not by hand)

Your Turn!

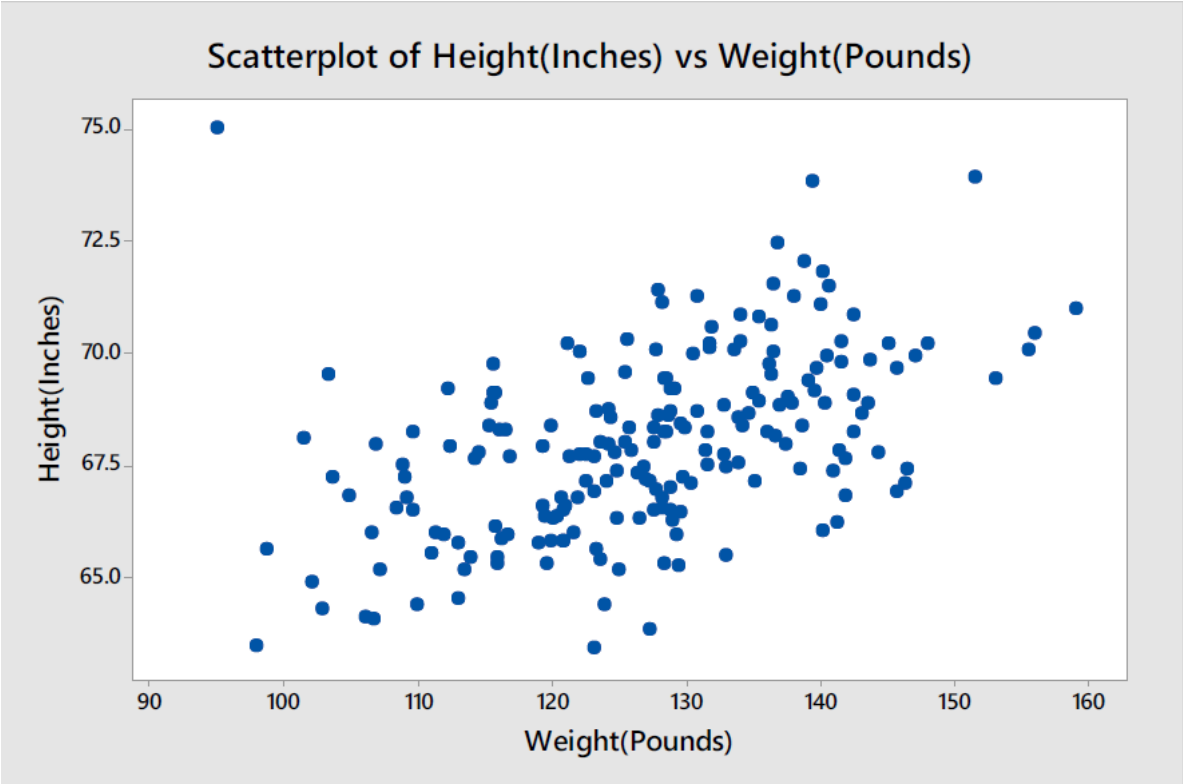
On your cell phone, go to

http://guessthecorrelation.com/

and play for one minute.



Correlation is NOT Causation



- **Correlation thinking**: Weight and Height are correlated ($R = 0.483$). There is a moderate, positive, linear association, with one potential outlier.
- **Causation thinking**: Weighing more causes you to be taller

Better Thinking

Sometimes a change in the predictor variable does DIRECTLY cause a change in the response variable. A correlation analysis CAN'T reveal this causation

$P \leftrightarrow R$

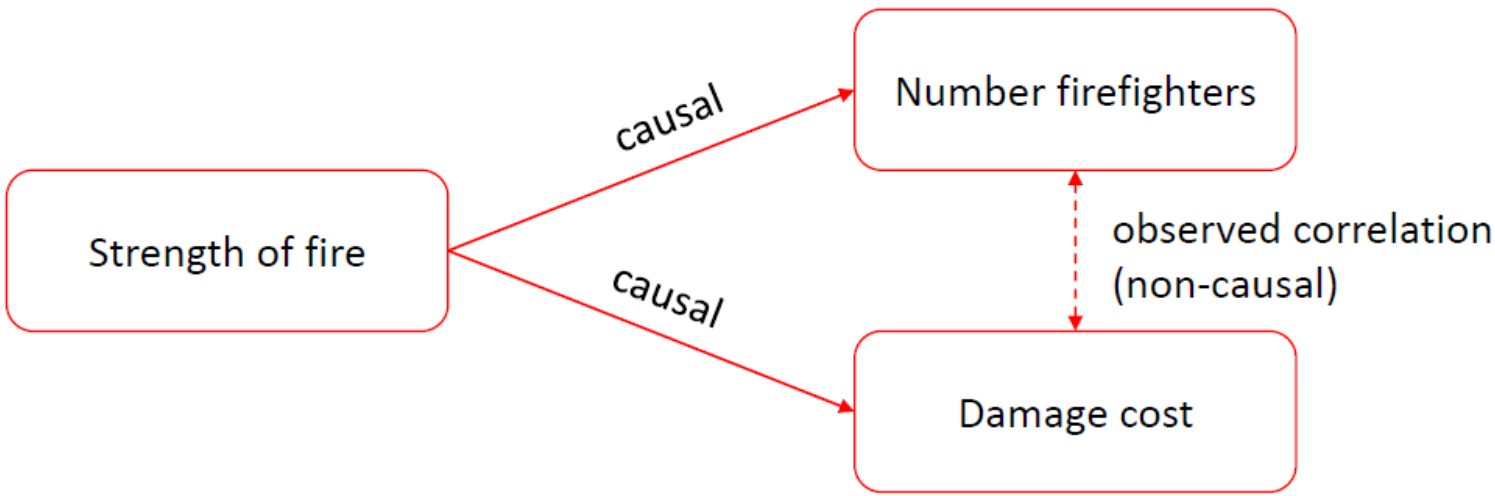
Better Thinking

Sometimes a change in the predictor variable does DIRECTLY cause a change in the response variable. A correlation analysis CAN'T reveal this causation
P $\leftrightarrow$ R

Sometimes a change in the predictor variable causes a change in some other variable (lurking variable) which then causes a change in the response variable. A correlation analysis CAN'T reveal this causation.
P $\leftrightarrow$ L $\leftrightarrow$ R

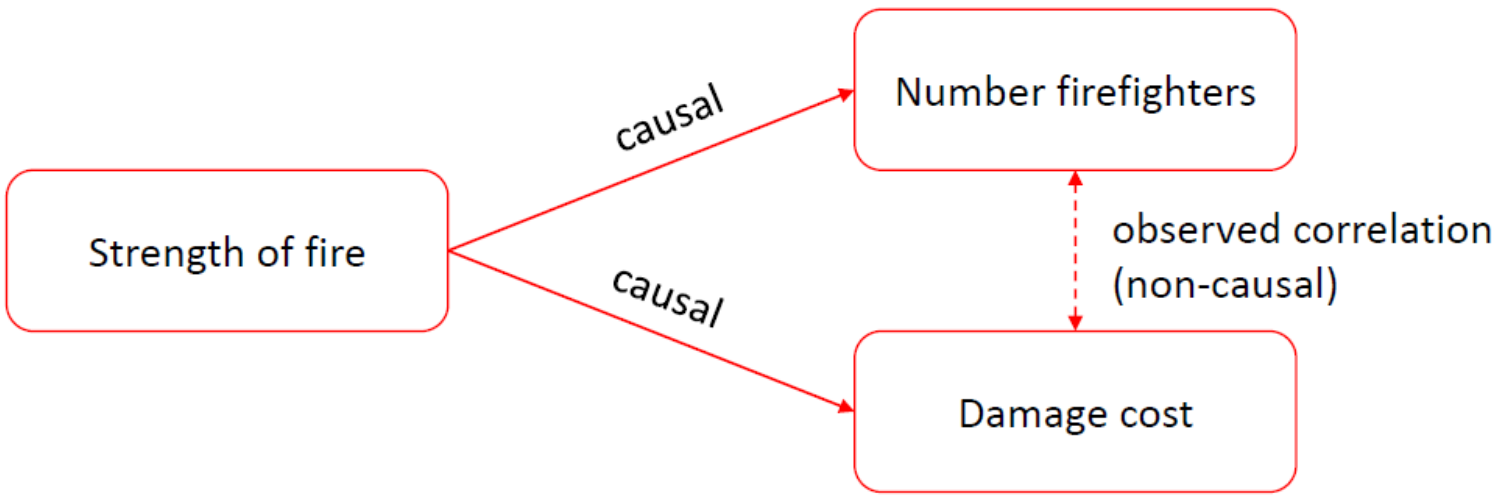
Guess The Lurker

1. Number of sunburns reported and ice-cream sales
2. Number of firefighters at a blaze and cost of damage from the blaze



Guess The Lurker

1. Number of sunburns reported and ice-cream sales
2. Number of firefighters at a blaze and cost of damage from the blaze



More extreme examples:

Spurious correlations

Correlation Matrix

Shows the correlation of every variable with every other variable.

The matrix often omits

- Self correlations since $cor(X, X) = 1$
- Symmetric correlations since $cor(X, Y) = cor(Y, X)$, these are the blank spots in the table

Time	Score	Change	Happiness	Effect
3.5236	4.37356	2.37998	3.32430	0.9397
12.7872	3.95285	-2.23578	1.14472	10.0369
5.6247	0.49852	1.51192	4.35026	2.1962
8.2373	3.27092	0.06258	2.48132	4.2383
0.1774	3.51523	3.94999	3.28453	0.0780
5.0768	4.53241	1.79803	4.59037	1.8850
2.0210	3.64329	3.24838	4.69524	0.6590
1.5945	4.24732	3.29474	3.38313	0.3083
4.4450	1.87556	2.11454	4.75345	1.5628
4.1318	2.78293	2.02919	2.77525	1.1612
16.8057	0.02387	-4.24947	0.10218	17.3202

	Time	Score	Change	Happiness
Score	-0.491			
Change	-0.999	0.481		
Happiness	-0.800	0.262	0.825	
Effect	0.963	-0.513	-0.968	-0.850

Correlation: Time, Score, Change, Happiness, Effect

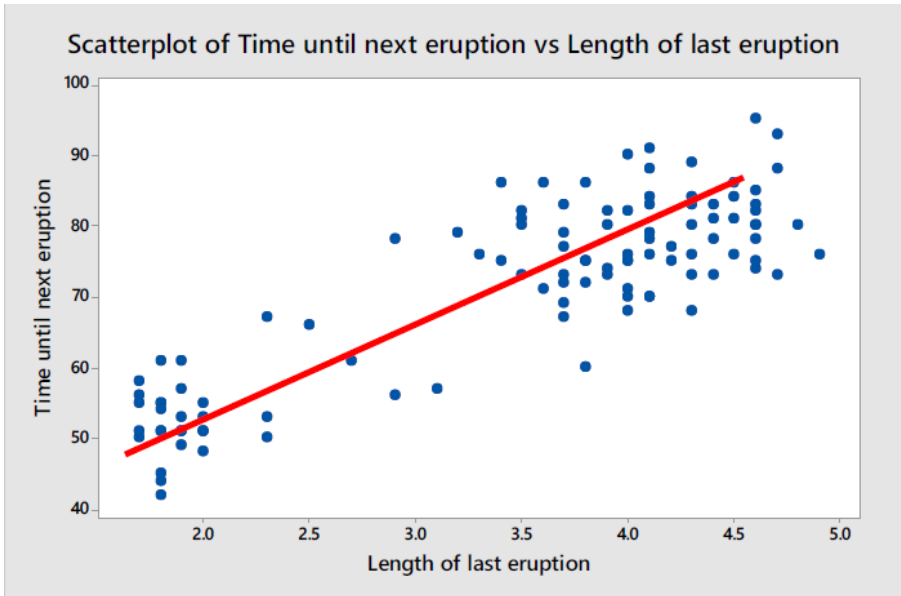
From Correlation to Prediction

Can we predict how long it will be (in minutes) until the next eruption of Old Faithful based on how long the current eruption lasted (in minutes)?

From Correlation to Prediction

Can we predict how long it will be (in minutes) until the next eruption of Old Faithful based on how long the current eruption lasted (in minutes)?

	TimeUntilNext	LengthOfLast
1	78	4.4
2	74	3.9
3	68	4.0
4	76	4.0
5	80	3.5
6	84	4.1
7	50	2.3
8	93	4.7
9	55	1.7



We'd like to be able to create a "linear model" through our data that "best fits" the data

Observed Value, Predicted Value, Residual

Given any point (x, y) in a scatterplot, we now have two key ideas: for x fixed,

- y: the **value observed** from actual data point
- $\hat{y}$: the **value predicted** from model

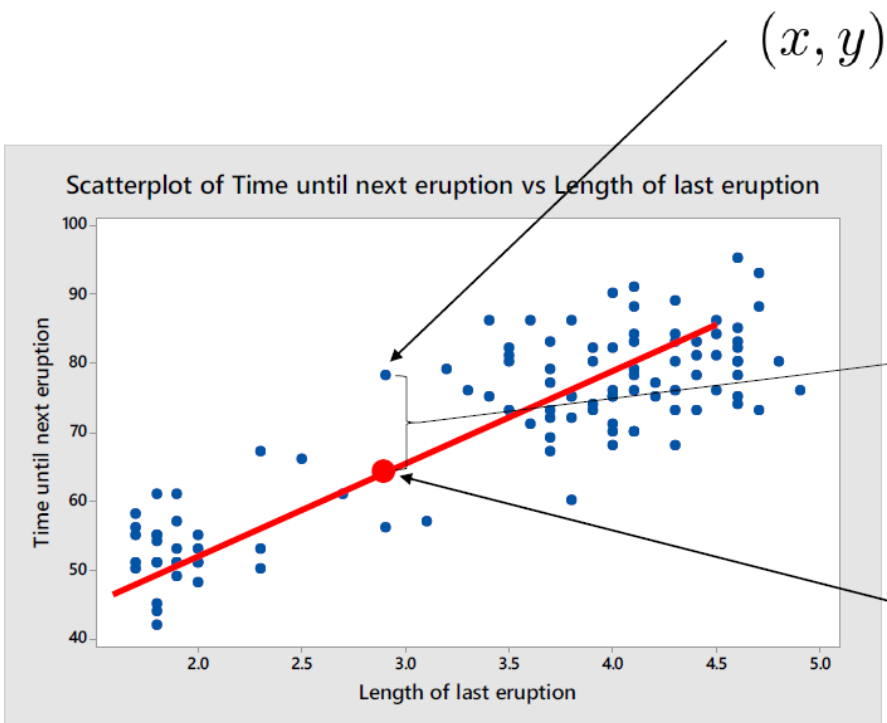
From these values, we can calculate the **residual** $e = y - \hat{y}$, which measure how off the model is at the value x.

Observed Value, Predicted Value, Residual

Given any point (x, y) in a scatterplot, we now have two key ideas: for x fixed,

- y: the **value observed** from actual data point
- $\hat{y}$: the **value predicted** from model

From these values, we can calculate the **residual** $e = y - \hat{y}$, which measure how off the model is at the value x.



What is the “Best Fit”?

Ideally, all the residuals would be 0 (a perfect model!).

What is the "Best Fit"?

Ideally, all the residuals would be 0 (a perfect model!).

Decause data are rarely perfect, we could define

$$\text{model error} = \sum (\text{residuals for each data point})?$$

Then, the line of best fit makes this expression minimal.

What is the "Best Fit"?

Ideally, all the residuals would be 0 (a perfect model!).

Decause data are rarely perfect, we could define

$$\text{model error} = \sum (\text{residuals for each data point})?$$

Then, the line of best fit makes this expression minimal.

Two choices to fix this issue are:

$$\text{model error} = \sum |\text{residuals}| \quad \text{or} \quad \text{model error} = \sum (\text{residuals})^2 .$$

What is the "Best Fit"?

Ideally, all the residuals would be 0 (a perfect model!).

Decause data are rarely perfect, we could define

$$\text{model error} = \sum (\text{residuals for each data point})?$$

Then, the line of best fit makes this expression minimal.

Two choices to fix this issue are:

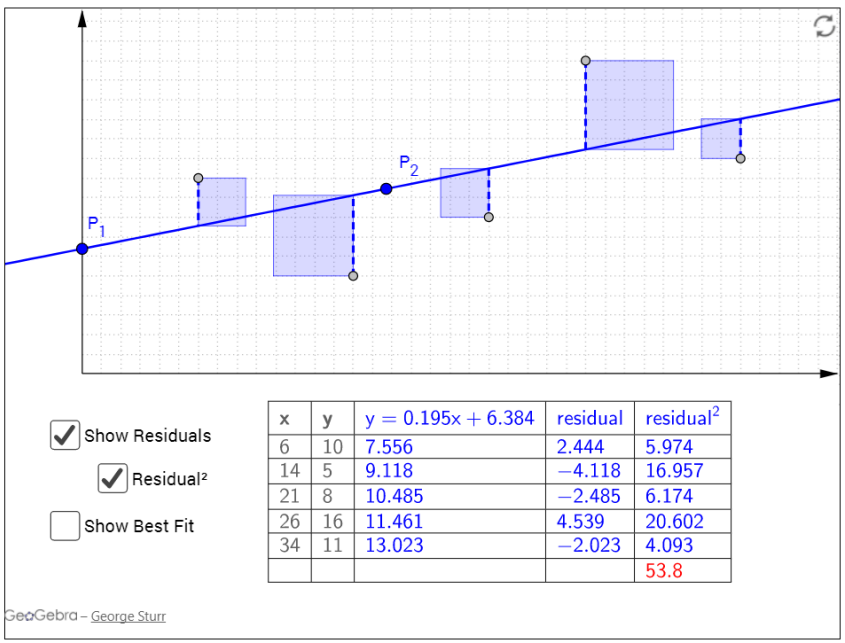
$$\text{model error} = \sum |\text{residuals}| \quad \text{or} \quad \text{model error} = \sum (\text{residuals})^2 .$$

The second choice is standard, penalizes large residuals, and is easily differentiable.

Exploring Linear Regression Visually

Try out this interactive GeoGebra tool:

https://www.geogebra.org/m/dlsxY1uX



x	y	y = 0.195x + 6.384	residual	residual²
6	10	7.556	2.444	5.974
14	5	9.118	−4.118	16.957
21	8	10.485	−2.485	6.174
26	16	11.461	4.539	20.602
34	11	13.023	−2.023	4.093
				53.8

- Move p_1 and p_2 until you *feel* you have the line of best fit. Check your guess by showing the line of best fit.
- Refresh the page. Try checking "*Show Residuals*" and "*Residuals*^{2}", and then minimize the sum of squared residuals (red number). Then check your guess.

Line of Best Fit
(AKA Linear Model AKA Regression Line)

Notation: for the regression Line: $\hat{y} = b_0 + b_1 \cdot x$

Line of Best Fit
(AKA Linear Model AKA Regression Line)

Notation: for the regression Line: $\hat{y} = b_0 + b_1 \cdot x$

Interpretation:

- Intercept b_0: This is the predicted value for y when $x = 0$.
- Slope b_1: Measures the steepness of the regression line.
  It says how much y changes for each 1 unit change of x.

Calculating The Regression Equation

Regression Line: $\hat{y} = b_0 + b_1 \cdot x$

$$b_1 = R \cdot \frac{s_y}{s_x}$$

Calculating The Regression Equation

Regression Line: $\hat{y} = b_0 + b_1 \cdot x$

$$b_1 = R \cdot \frac{s_y}{s_x}$$

We see that:

- R gets the correct sign on the slope
- s_y/s_x gets the correct units on the slope

Calculating The Regression Equation

Regression Line: $\hat{y} = b_0 + b_1 \cdot x$

$$b_1 = R \cdot \frac{s_y}{s_x}$$

We see that:

- R gets the correct sign on the slope
- s_y/s_x gets the correct units on the slope

After calculating b_1 you get

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

This formula holds because the regression line always passes through $(\bar{x}, \bar{y})$.

Back to Old Faithful

Using technology, we get:

$$\widehat{\text{Time until next}} = 33.83 + 10.74 \cdot (\text{Length of last})$$

Interpret what the intercept and slop mean in this context.

Back to Old Faithful

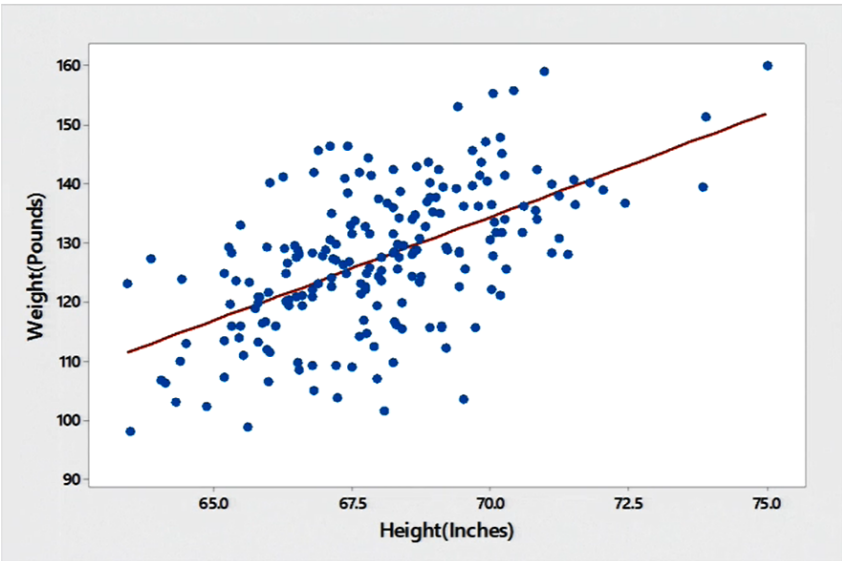
Using technology, we get:

$$\widehat{\text{Time until next}} = 33.83 + 10.74 \cdot (\text{Length of last})$$

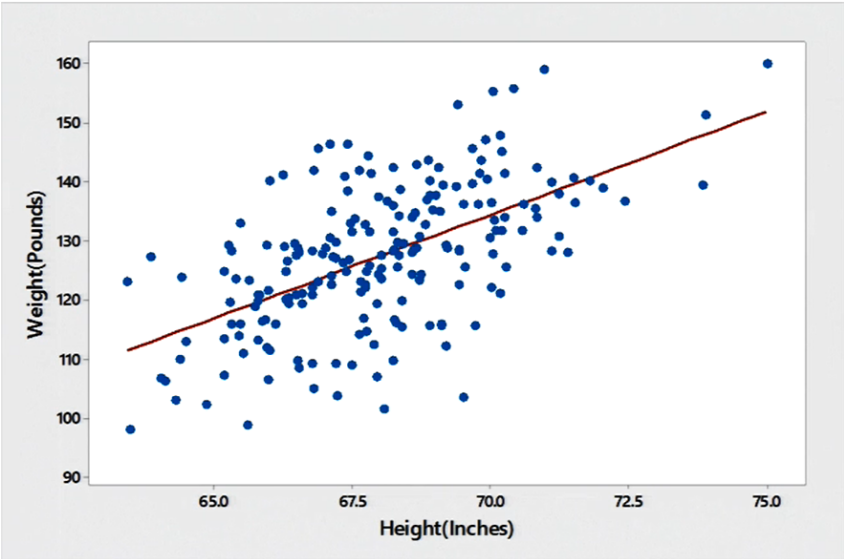
Interpret what the intercept and slop mean in this context.

- **The intercept** is 33.83 minutes, which means that if the last eruption lasted 0 minutes (!), then we will wait about 34 minutes until the next eruption begins.
  (Sometimes intercepts don't make real-world sense)
- **The slope** is 10.74 (minutes until next/minutes of last). This means that each additional minute of eruption time leads to about 11 more minutes of waiting for the next eruption.

A Classic Example

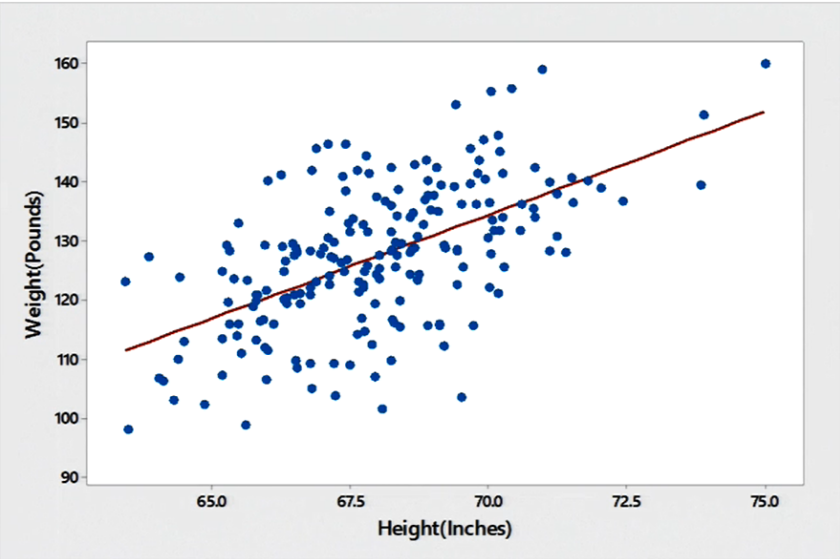


A Classic Example



$$\widehat{\text{Weight}} = -111 + 3.51 \times \text{Height}$$

A Classic Example

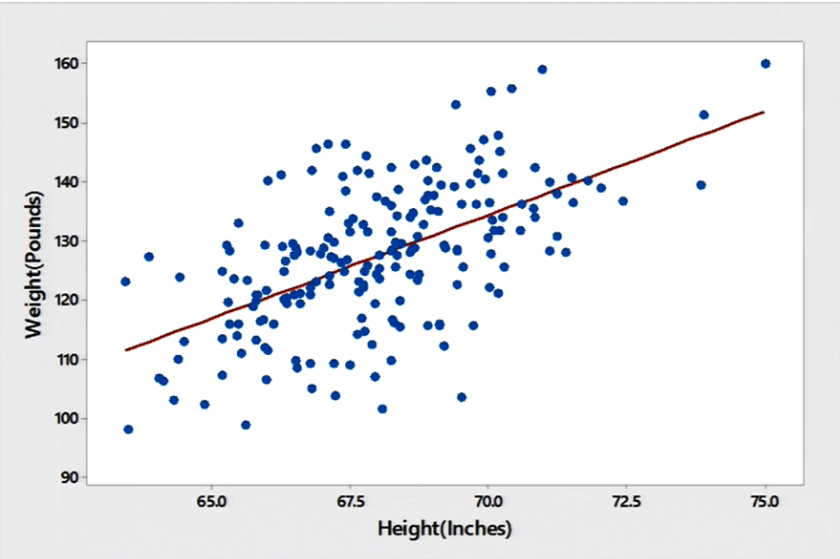


The intercept suggests that a 0 inch tall person should weigh -111 lbs.
This makes no real-world sense, but is a theoretical starting point for the model.

The slope suggests that for every inch increase in height, we expect a person to be about 3.5 lbs heavier.
Similarly, for every inch decrease in height we expect a decrease in 3.5 lbs.

$$\widehat{\text{Weight}} = -111 + 3.51 \times \text{Height}$$

A Classic Example



$$\widehat{\text{Weight}} = -111 + 3.51 \times \text{Height}$$

The intercept suggests that a 0 inch tall person should weigh -111 lbs.
This makes no real-world sense, but is a theoretical starting point for the model.

The slope suggests that for every inch increase in height, we expect a person to be about 3.5 lbs heavier.
Similarly, for every inch decrease in height we expect a decrease in 3.5 lbs.

$$\text{Slope} = \frac{\Delta y}{\Delta x}$$

$$3.5 \text{ lbs/inch} = \frac{3.5 \text{ lbs}}{1 \text{ inch}}$$

Why Build A Model?

– Perhaps y is really hard or expensive to measure, but well associated with x which is easy to measure.

– Perhaps y can only be measured after the fact (e.g. damage done by a tornado), but you need a sense for this before the fact.

– A model allows you to move from your data set to the larger universe of possibilities

– Parts of a model might answer questions you have about an issue (e.g. slope of height-weight graph gives the "weight of an inch of a person")

U.S. Navy

TABLE 1
MAXIMUM WEIGHT FOR HEIGHT SCREENING TABLE

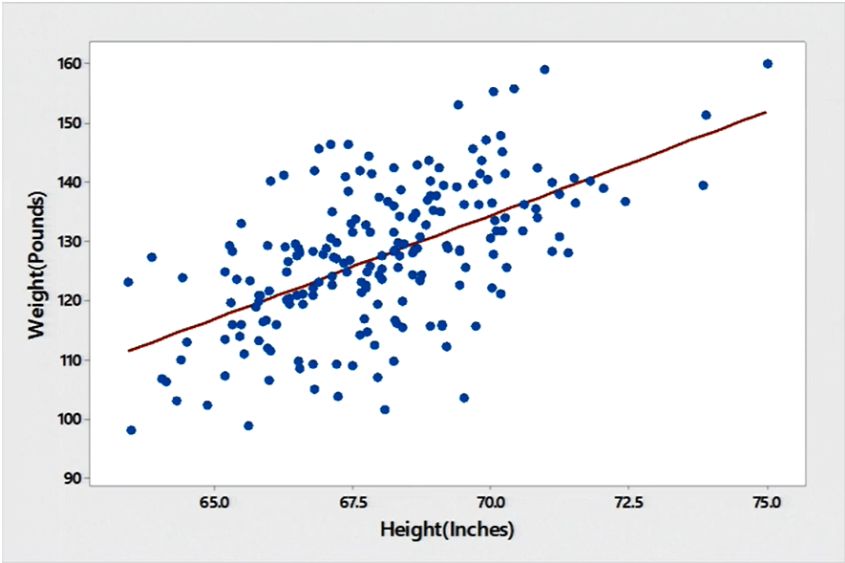
Men Maximum Weight (pounds)	Member's Height (inches) (fractions rounded up to nearest whole inch)	Women Maximum Weight (pounds)
97	51	102
102	52	106
107	53	110
112	54	114
117	55	118
122	56	123
127	57	127
131	58	131
136	59	136
141	60	141
145	61	145
150	62	149
155	63	152
160	64	156
165	65	160
170	66	163
175	67	167
181	68	170
186	69	174

We see from this chart that every inch of height for a male equals about 5 or 6 pounds, and every inch for a female weighs about 3 or 4 pounds. This is exactly the slope of the regression line!

Using the Regression Model

$$\widehat{\text{Weight}} = -111 + 3.51 \times \text{Height}$$

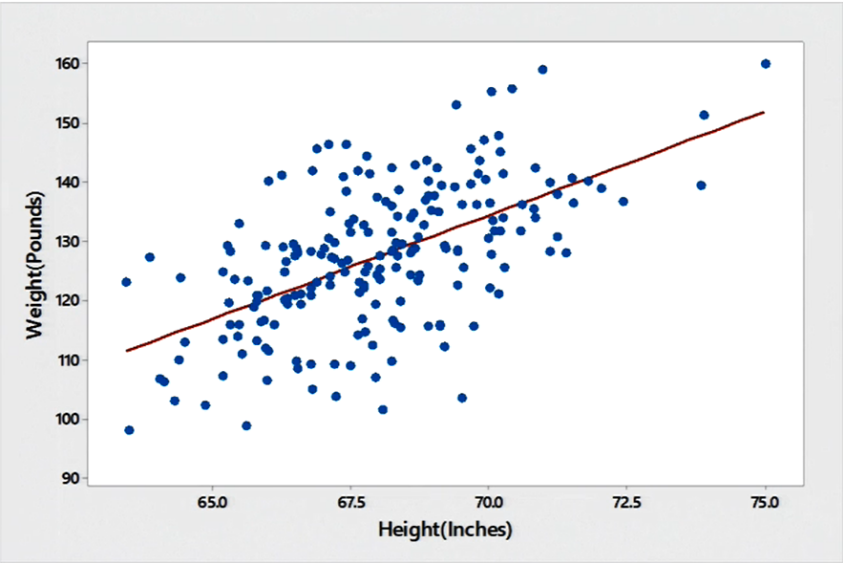
Try an example:
Convert your height to inches, see what the model predicts.
What is the residual based on your actual weight?



Using the Regression Model

$$\widehat{\text{Weight}} = -111 + 3.51 \times \text{Height}$$

Try an example:
Convert your height to inches, see what the model predicts.
What is the residual based on your actual weight?

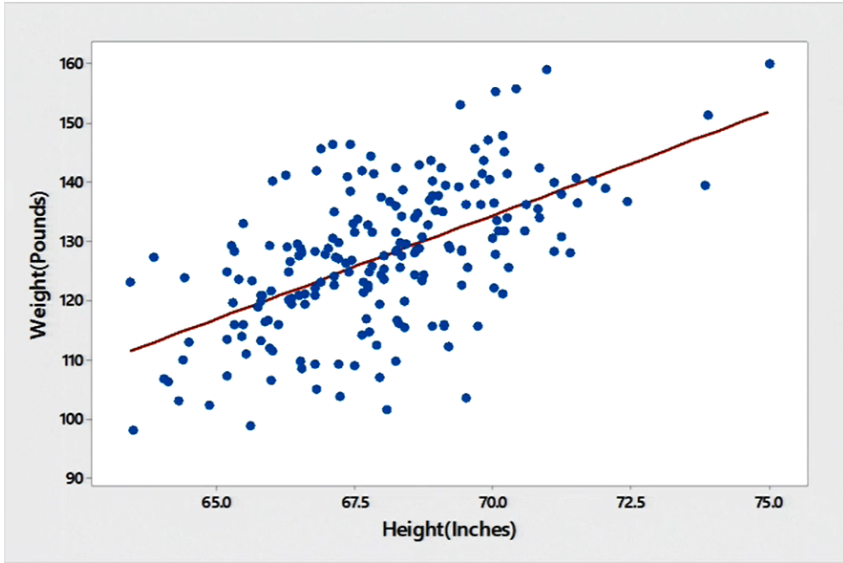


My data: 190cm converts to 75 inches.
The associated predicted weight is $-111 + 3.5 \cdot 75 = 151.5$.

Using the Regression Model

$$\widehat{\text{Weight}} = -111 + 3.51 \times \text{Height}$$

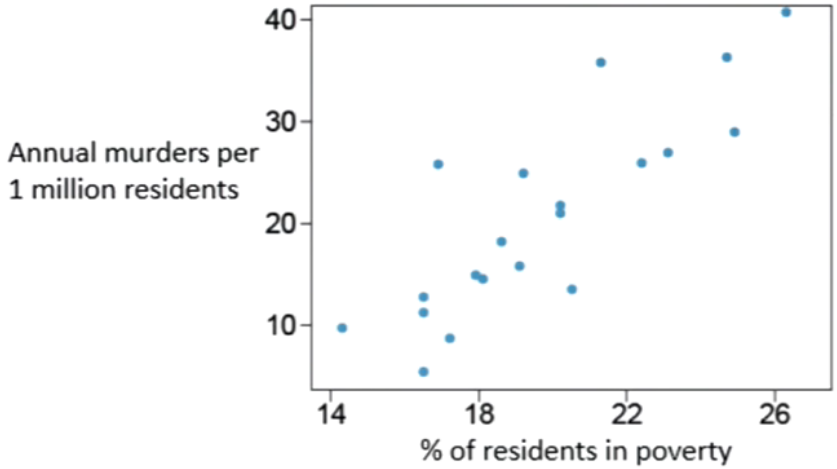
Try an example:
Convert your height to inches, see what the model predicts.
What is the residual based on your actual weight?



My data: 190cm converts to 75 inches.
The associated predicted weight is $-111 + 3.5 \cdot 75 = 151.5$.

My actual weight is 202 lbs, so the residual is $202 - 151.5 = 50.5$ lbs.
So my data point lies above the regression line (since residual >0).
The model (strongly) under-predicted.

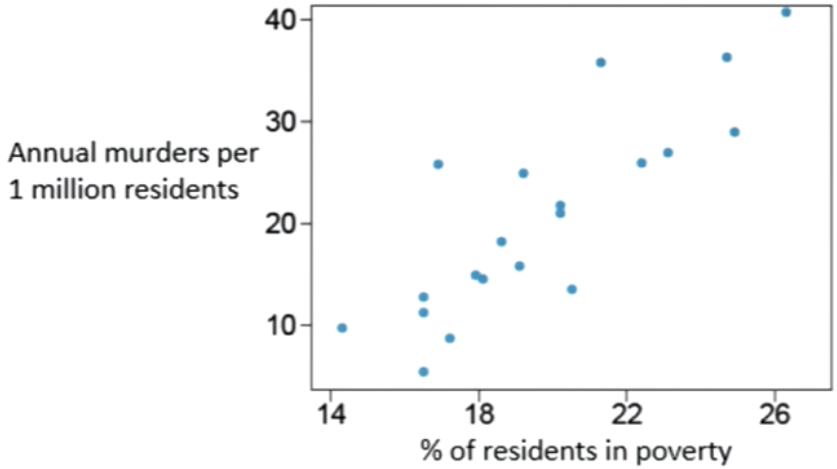
Your Turn!



What might each dot represent?

1. A person in the U.S.
2. A small town in the U.S.
3. A metropolitan area in the U.S.
4. One of America's 20 richest cities

Your Turn!



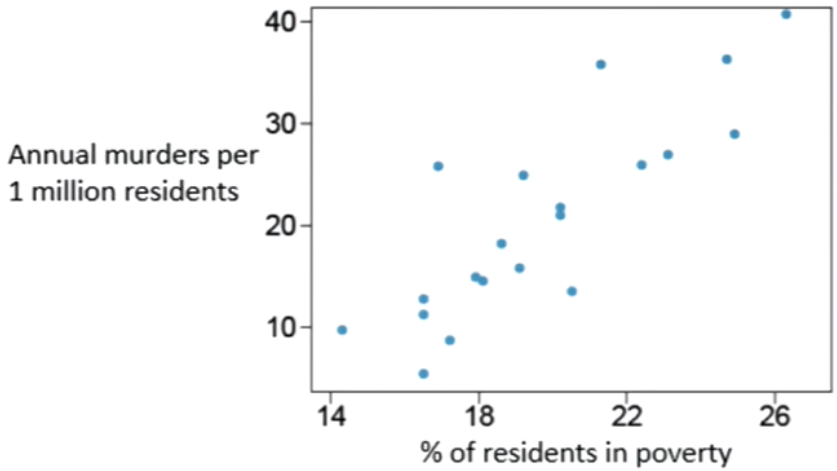
What might each dot represent?

1. A person in the U.S.
2. A small town in the U.S.
3. A metropolitan area in the U.S.
4. One of America's 20 richest cities

Answer: 3.

Individuals don't have poverty rates, so 1. is wrong.
Small towns don't have a million people, so the y axis wouldn't make sense in 2.
Rich cities have low poverty rates, so the x axis wouldn't make sense in 4.

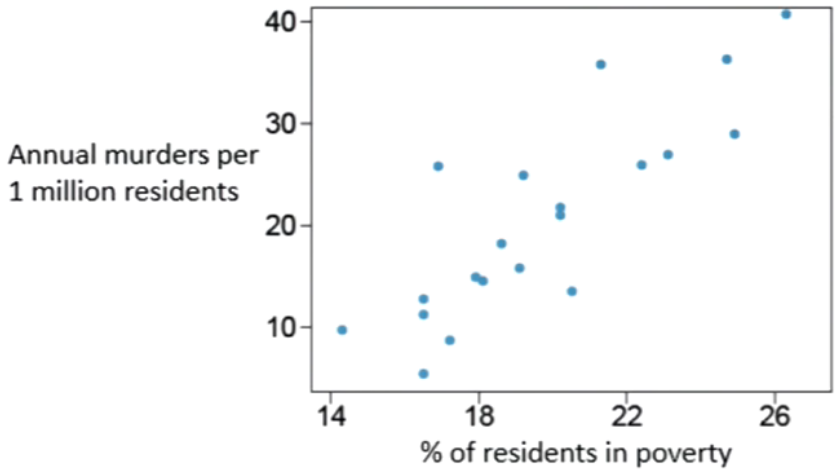
Your Turn!



Guess the correlation coefficient for this scatterplot.

1. $R \simeq 0$
2. $R \simeq 0.25$
3. $R \simeq 0.55$
4. $R \simeq 0.85$
5. $R \simeq 1$

Your Turn!

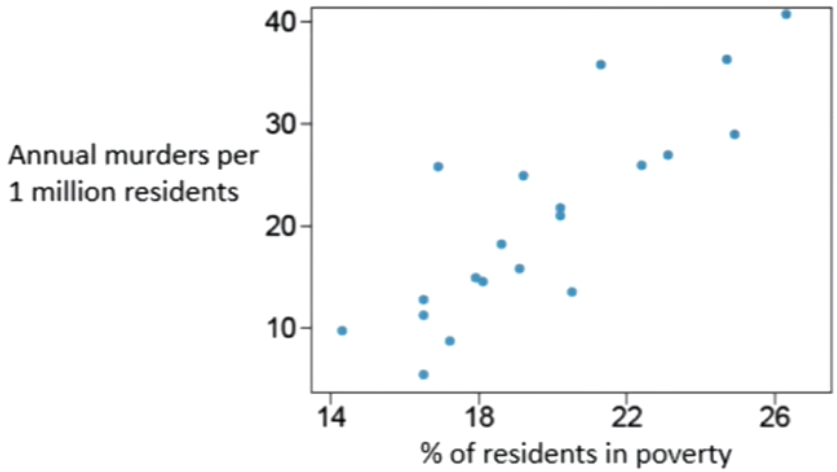


Guess the correlation coefficient for this scatterplot.

1. $R \simeq 0$
2. $R \simeq 0.25$
3. $R \simeq 0.55$
4. $R \simeq 0.85$
5. $R \simeq 1$

Answer: 4. The actual value is $R = 0.84$.

Your Turn!



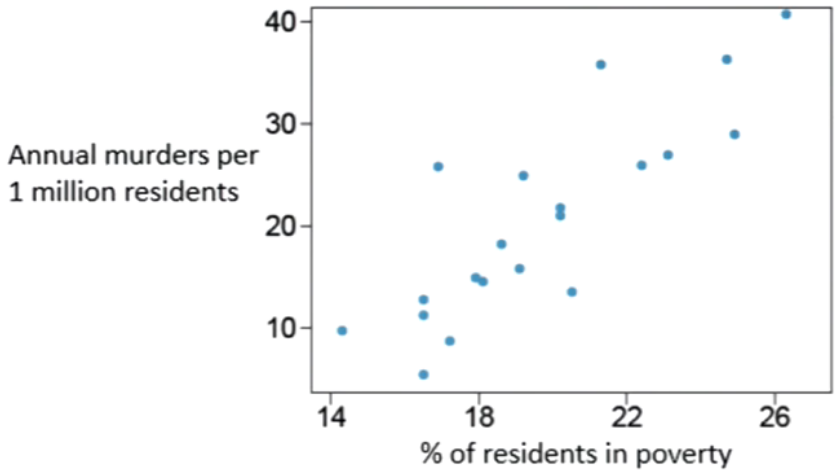
You are told the regression line is

$$\text{Annual murder rate/million people} = -30 + 2.6 \cdot \text{Poverty Rate}.$$

What annual murder rate (per million people) do we expect in a city with a 20% poverty rate?

1. 4
2. 12
3. 22
4. 31

Your Turn!



You are told the regression line is

$$\text{Annual murder rate/million people} = -30 + 2.6 \cdot \text{Poverty Rate}.$$

What annual murder rate (per million people) do we expect in a city with a 20% poverty rate?

1. 4
2. 12
3. 22
4. 31

Answer: 3., since $-30 + 2.6 \cdot 20 = 22$.

Your Turn!

Which statements are true? Recall that the prediction is

$$\text{Annual murder rate/million people} = -30 + 2.6 \cdot \text{Poverty Rate}.$$

1. A city with no poverty would have a murder rate of -30 people/million.
2. For every 1 unit increase in poverty, 2.6 more people will be murdered per year (for each million people in the city).
3. If you want to know the murder rate (per million people) of any city in the U.S., plug in the poverty rate into this equation.
4. The best values to plug in for the poverty rate are vetween 14 and 26.
5. The only values we may plug in for the poverty rate are between 14 and 26.

Your Turn!

Which statements are true? Recall that the prediction is

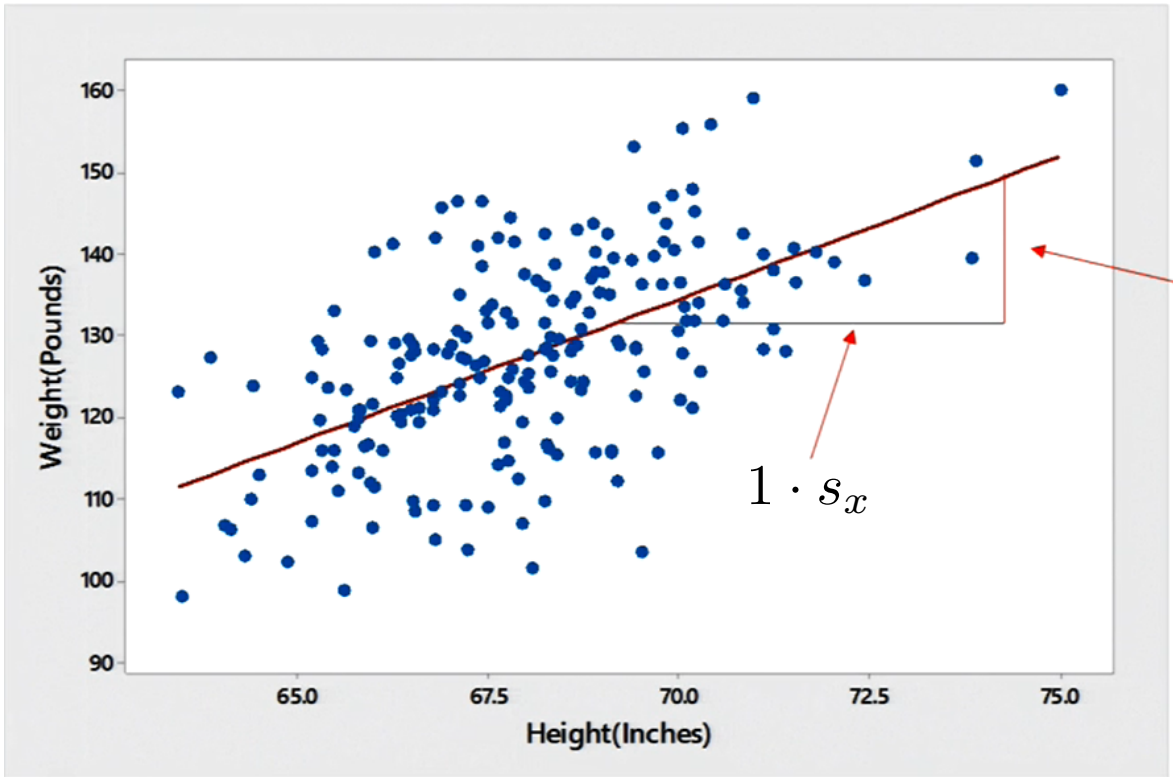
$$\text{Annual murder rate/million people} = -30 + 2.6 \cdot \text{Poverty Rate}.$$

1. A city with no poverty would have a murder rate of -30 people/million.
2. For every 1 unit increase in poverty, 2.6 more people will be murdered per year (for each million people in the city).
3. If you want to know the murder rate (per million people) of any city in the U.S., plug in the poverty rate into this equation.
4. The best values to plug in for the poverty rate are vetween 14 and 26.
5. The only values we may plug in for the poverty rate are between 14 and 26.

1. True. That's the interpretation of the intercept.
2. True. That's the interpreation of the slope.
3. False. Our prediction may only be valid for big cities.
4. True. Since most of the data used to build the model are between 14 and 26, we get the best results in this range.
5. False. Too strong language to be true.

More on the Slope

$$b_1 = R \cdot \frac{s_y}{s_x} = \frac{R \cdot s_y}{1 \cdot s_x} = \frac{\Delta y}{\Delta x}$$



In other words:

- If you're 1·SD above the mean height, you'll be R·SD's above the mean weight.
- If you're 2·SD above the mean height, you'll be $2R$·SD's above the mean weight.

Regression to the Mean

Recall that

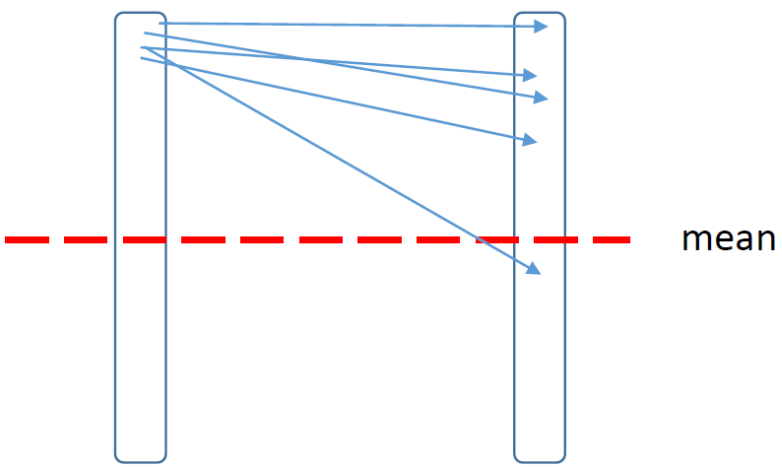
$$-1 \leq R \leq 1.$$

Hence, moving 1SD from the mean of the x-variable takes us less than 1·SD (precisely R·SD) from the mean in the y-variable.

So, the world of x-values gets compressed (SD-wise) as the linear model converts them over to y-predictions.

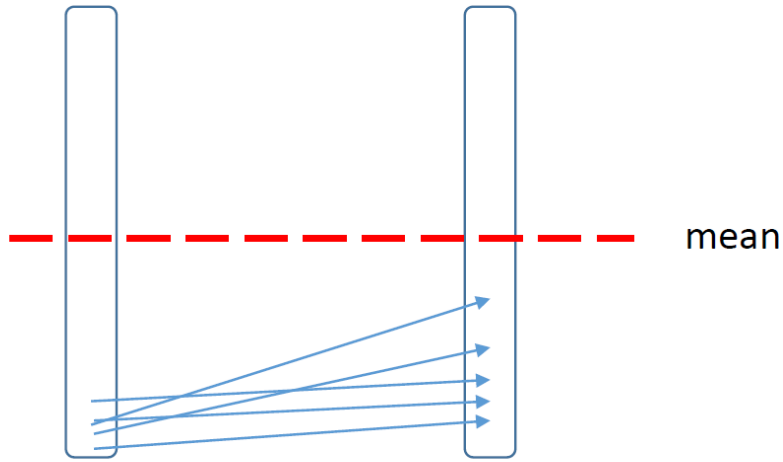
The phrase "regression to the mean" is used to describe this phenomenon, and is where the term "linear regression" comes from.

Regression to the Mean Examples



Scores on test 1

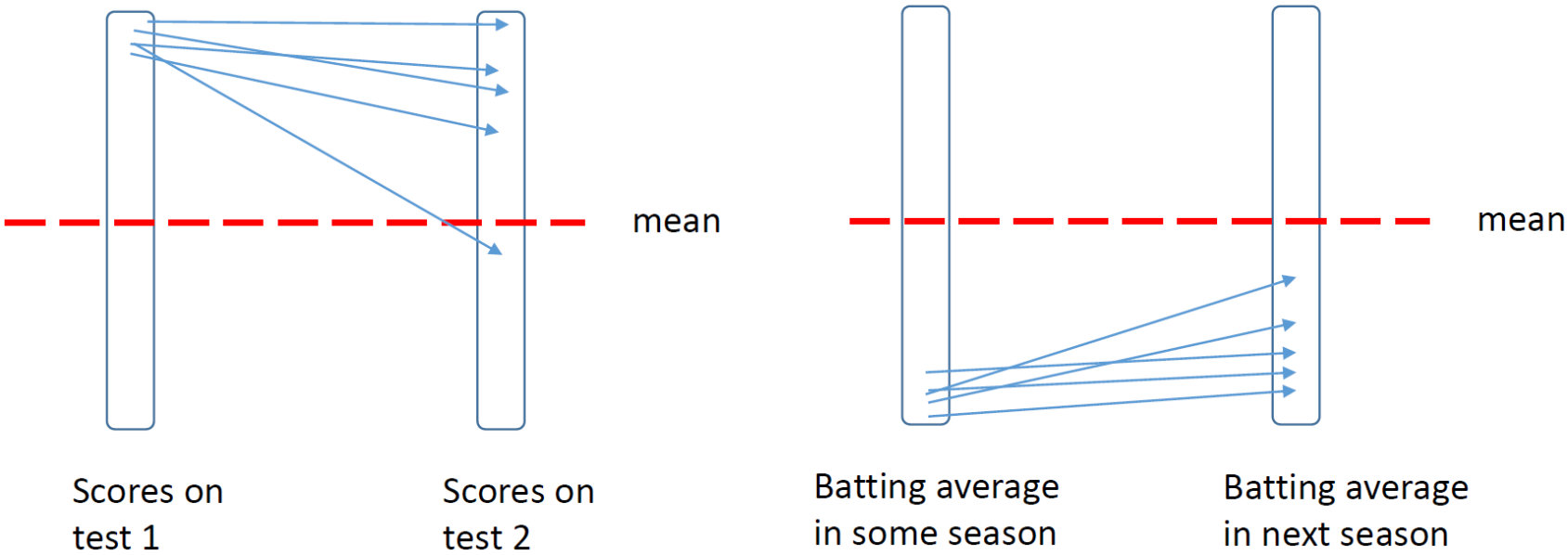
Scores on test 2



Batting average in some season

Batting average in next season

Regression to the Mean Examples



Why this occurs: Being exceptional on one measure (the idea on the left) requires exceptionalism AND luck. If you focus on these people, you are focusing on those who had both exceptionalism and luck.

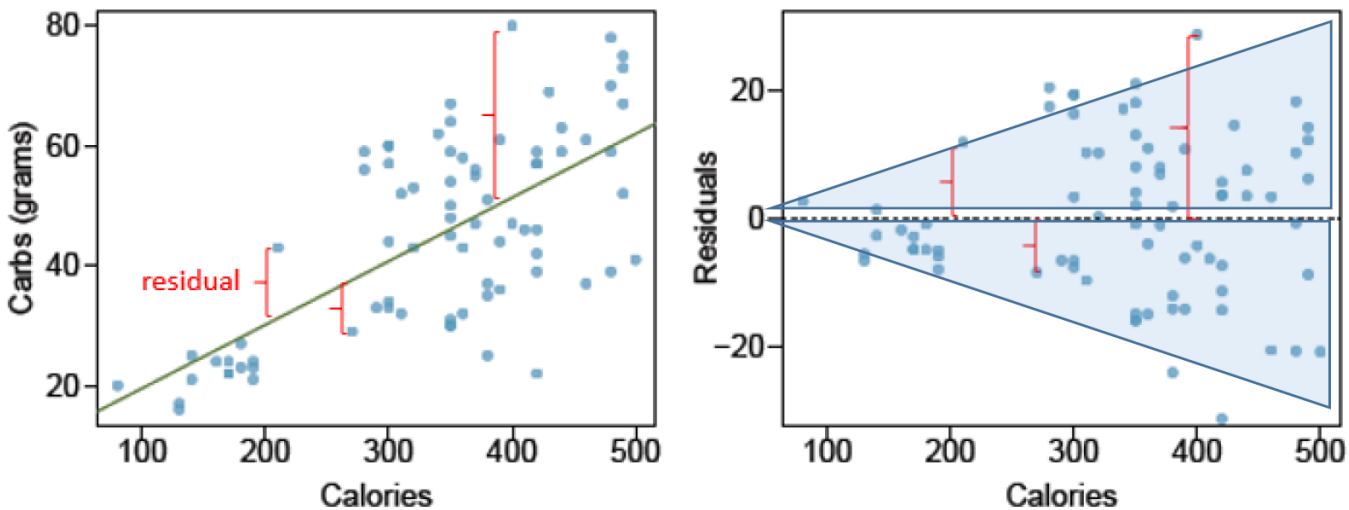
When you look at them on the other measure (the idea on the right), they are still exceptional, but probably won't have the luck this time around.

Conditions for Creating a Regression Model

Since correlations are involved, we need our three conditions from before:
1) Quantitative Variables
2) Straight Enough
3) No Outliers
Be we also have one new condition: 4) Residual Noise. We want the residual plot to look like "noise". It should have no pattern.

Conditions for Creating a Regression Model

Since correlations are involved, we need our three conditions from before:
1) Quantitative Variables
2) Straight Enough
3) No Outliers
Be we also have one new condition: 4) Residual Noise. We want the residual plot to look like "noise". It should have no pattern.



Here, we do see a pattern, as indicated by the fanning out effect.

A New Statistic: R^2 (Percent Variance Explained)

If you calculate the correlation from the Old Faithful example, you get $R = 0.854$.

A New Statistic: R^2 (Percent Variance Explained)

If you calculate the correlation from the Old Faithful example, you get $R = 0.854$.

For a given linear model, $R^2 = r^2$ is the proportion of the variation in the y-variable that is accounted for (or explained) by the variation in the x-variable.

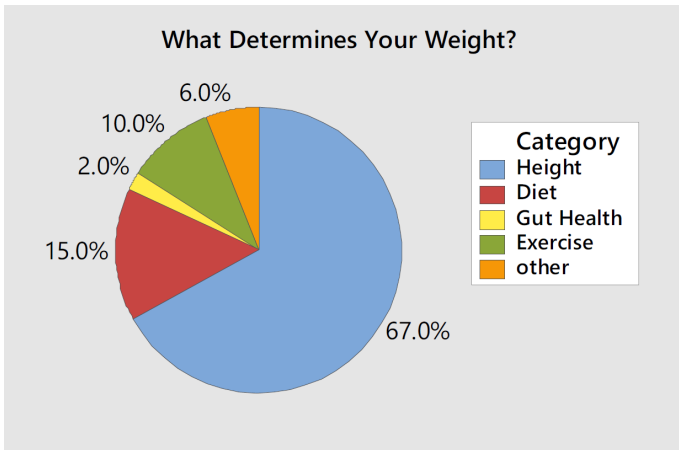
So, $R^2 = 0.854^2 = 0.73 = 73\%$ of how long we must wait is completely determined by how long the last eruption lasted.

A New Statistic: R^2 (Percent Variance Explained)

If you calculate the correlation from the Old Faithful example, you get $R = 0.854$.

For a given linear model, $R^2 = r^2$ is the proportion of the variation in the y-variable that is accounted for (or explained) by the variation in the x-variable.

So, $R^2 = 0.854^2 = 0.73 = 73\%$ of how long we must wait is completely determined by how long the last eruption lasted.



As another example, the R^2 in the height-weight regression is 0.67. So 67% of the variability in weight is because of height differences.

Warning! Danger!

If you want to switch the roles of the predictor and response variables, you cannot just rearrange your existing linear model.

$$\hat{y} = b_0 + b_1 x,$$

which gives

$$x = -\frac{b_0}{b_1} + \frac{1}{b_1}\hat{y}.$$

But what we really want is $\hat{x} = c_0 + c_1 \cdots y$.

Warning! Danger!

If you want to switch the roles of the predictor and response variables, you cannot just rearrange your existing linear model.

$$\hat{y} = b_0 + b_1 x,$$

which gives

$$x = -\frac{b_0}{b_1} + \frac{1}{b_1}\hat{y}.$$

But what we really want is $\hat{x} = c_0 + c_1 \cdots y$.

In the Height-Weight example, we actually have

$$\widehat{Weight} = -111 + 3.51 \cdot (Height)$$
$$\widehat{Height} = 55.9 + 0.0949 \cdot (Weight)$$

You can check that these equations are NOT simply rearrangements of each other.