Math 11
Calculus-Based Introductory Probability and Statistics

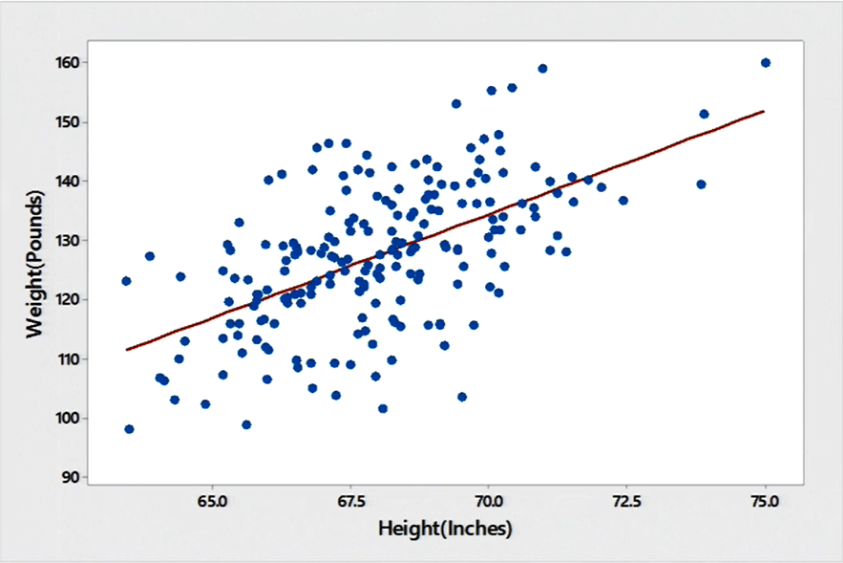
Eddie Aamari
S.E.W. Assistant Professor

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

Today:

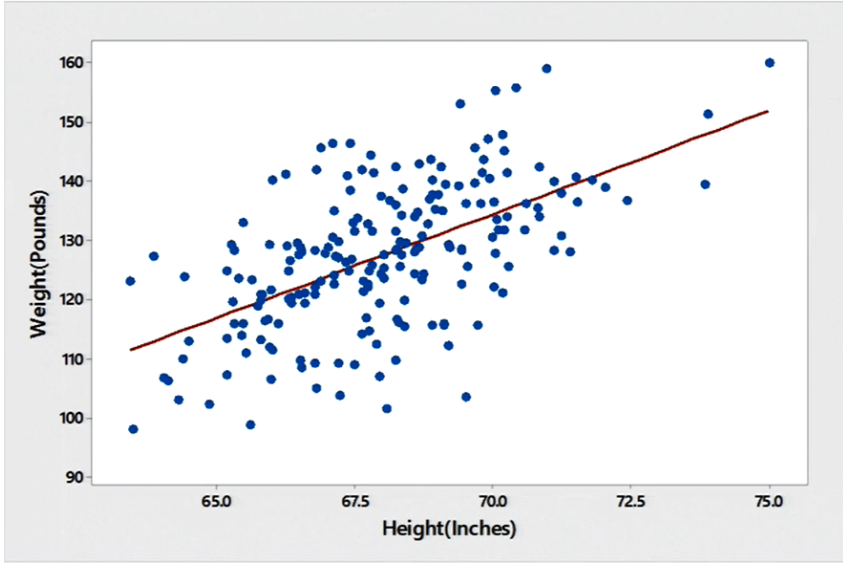
- Linear regression

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

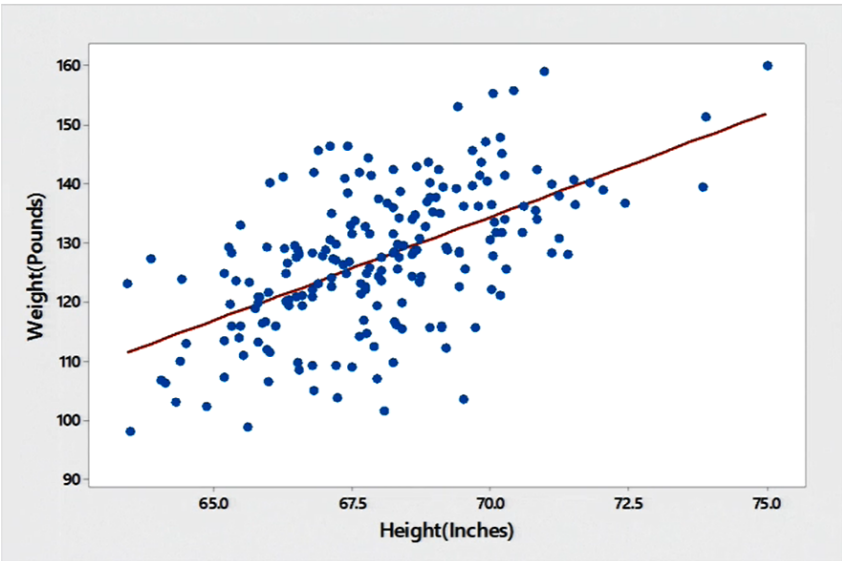
Using technology, we get:

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

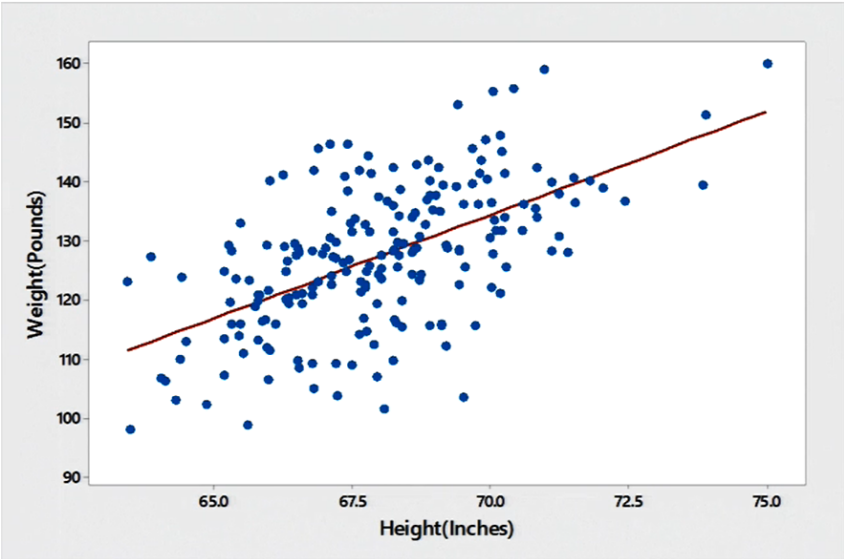
Interpret what the intercept and slop mean in this context.

- **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

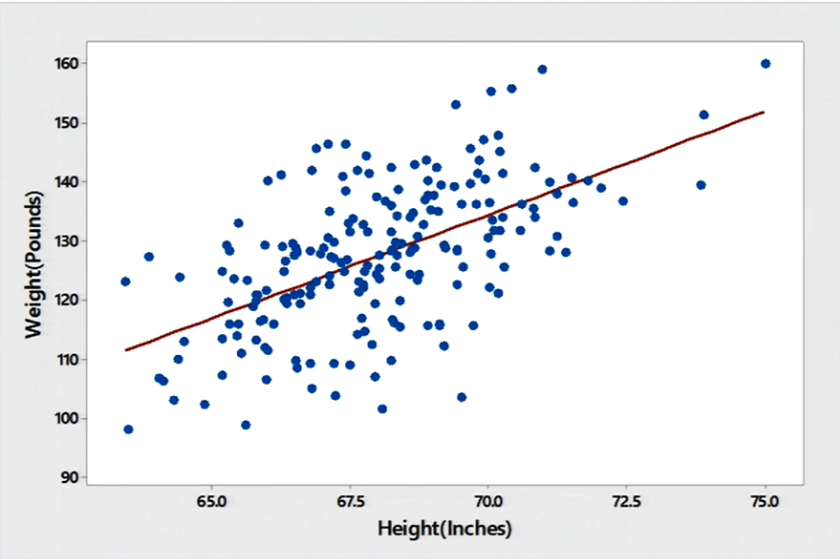


A Classic Example



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

A Classic Example

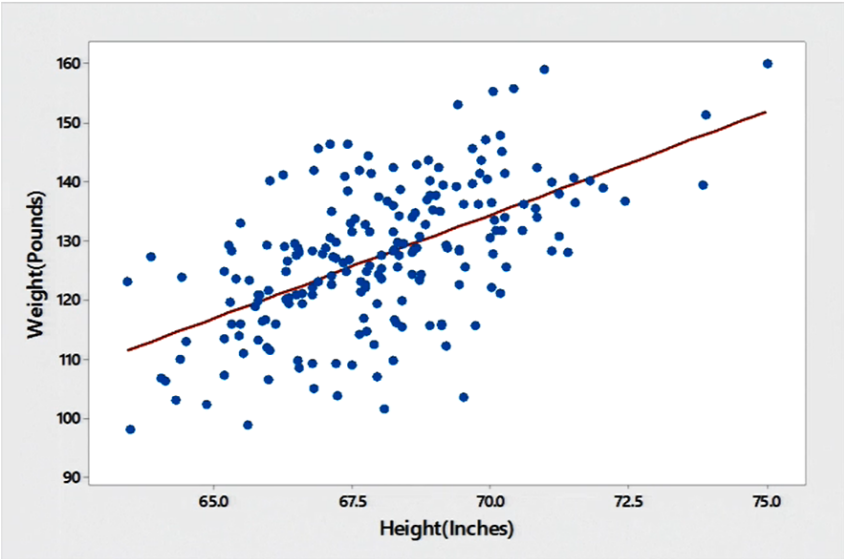


The intercept suggests that a 0 inch tall person should weigh -111 lbs.
This makes no real-world sense, but is a theoretical starting point for the model.

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

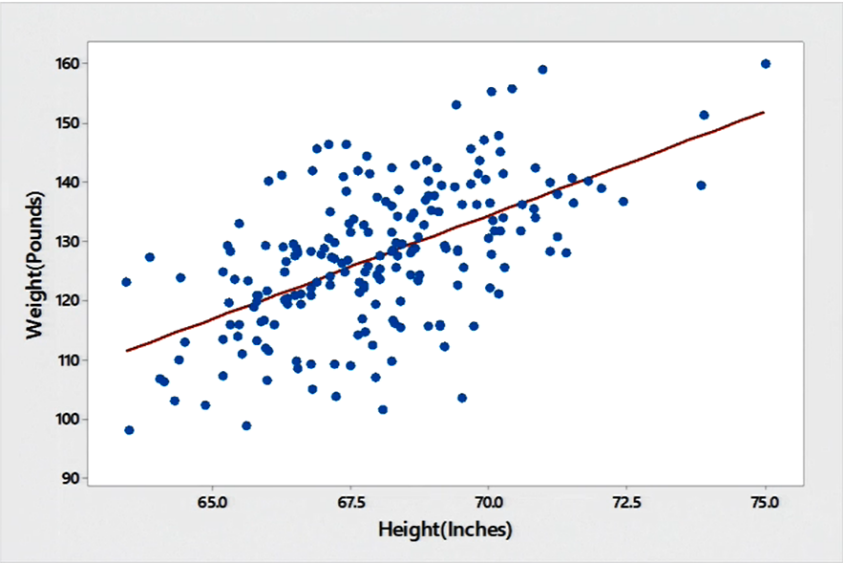
Men Maximum Weight (pounds)	Member's Height (inches) (fractions rounded up to nearest whole inch)	Women Maximum Weight (pounds)
97	51	102
102	52	106
107	53	110
112	54	114
117	55	118
122	56	123
127	57	127
131	58	131
136	59	136
141	60	141
145	61	145
150	62	149
155	63	152
160	64	156
165	65	160
170	66	163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

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

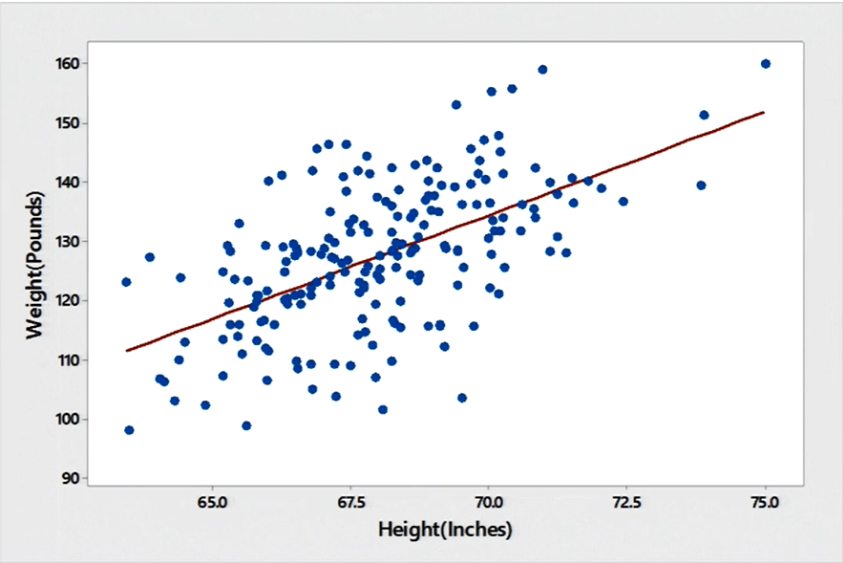
Try an example:
Convert your height to inches, see what the model predicts.
What is the residual based on your actual weight?



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?

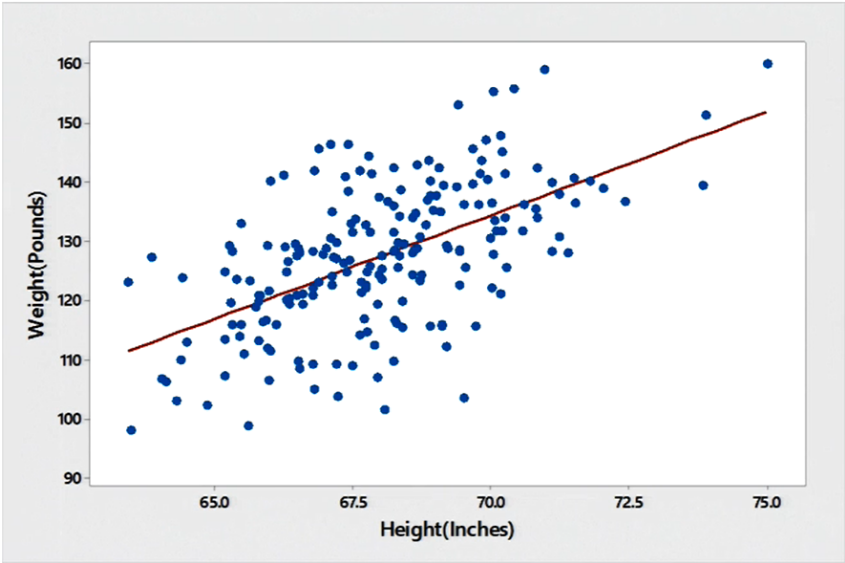


My data: 190cm converts to 75 inches.
The associated predicted weight is $-111 + 3.5 \cdot 75 = 151.5$.

Using the Regression Model

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

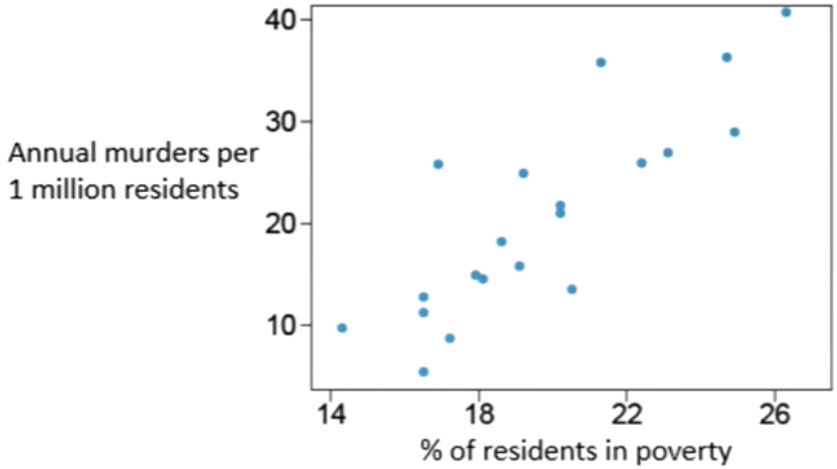
Try an example:
Convert your height to inches, see what the model predicts.
What is the residual based on your actual weight?



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.

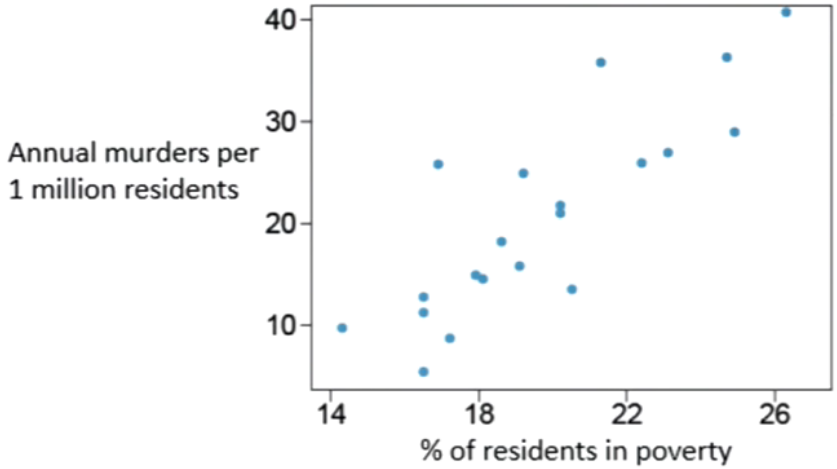
Your Turn!



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!



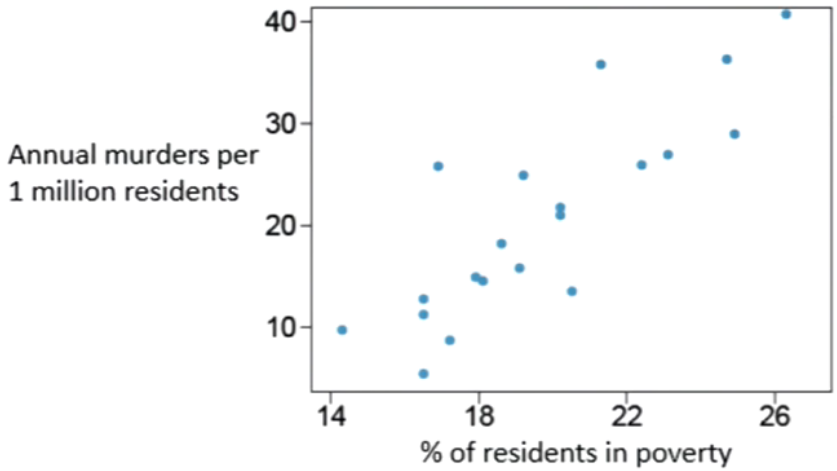
What might each dot represent?

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.

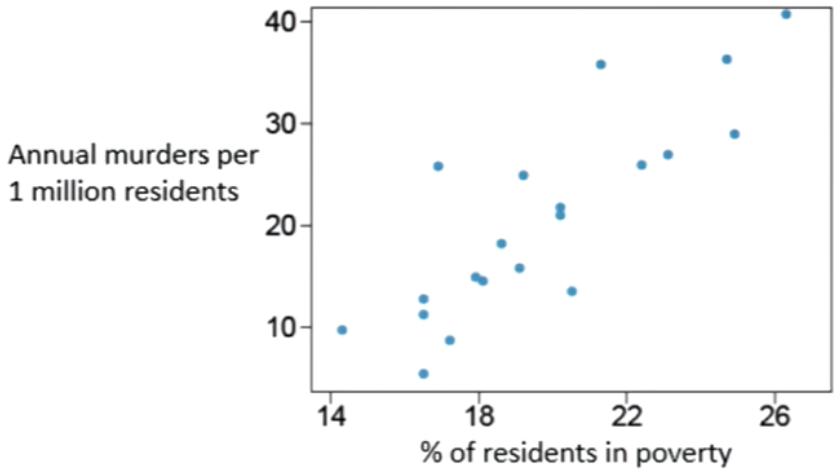
Your Turn!



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!

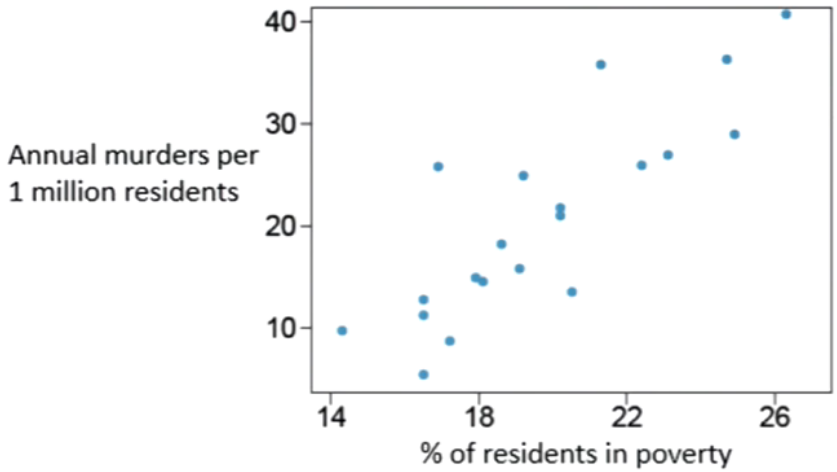


Guess the correlation coefficient for this scatterplot.

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!



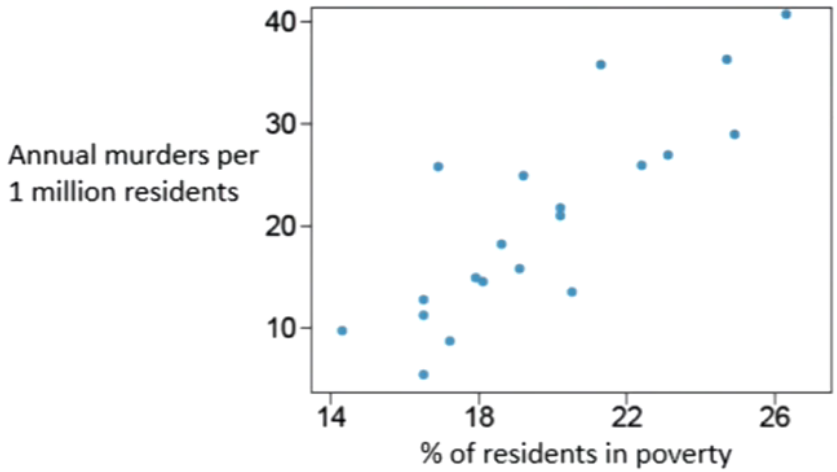
You are told the regression line is

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

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

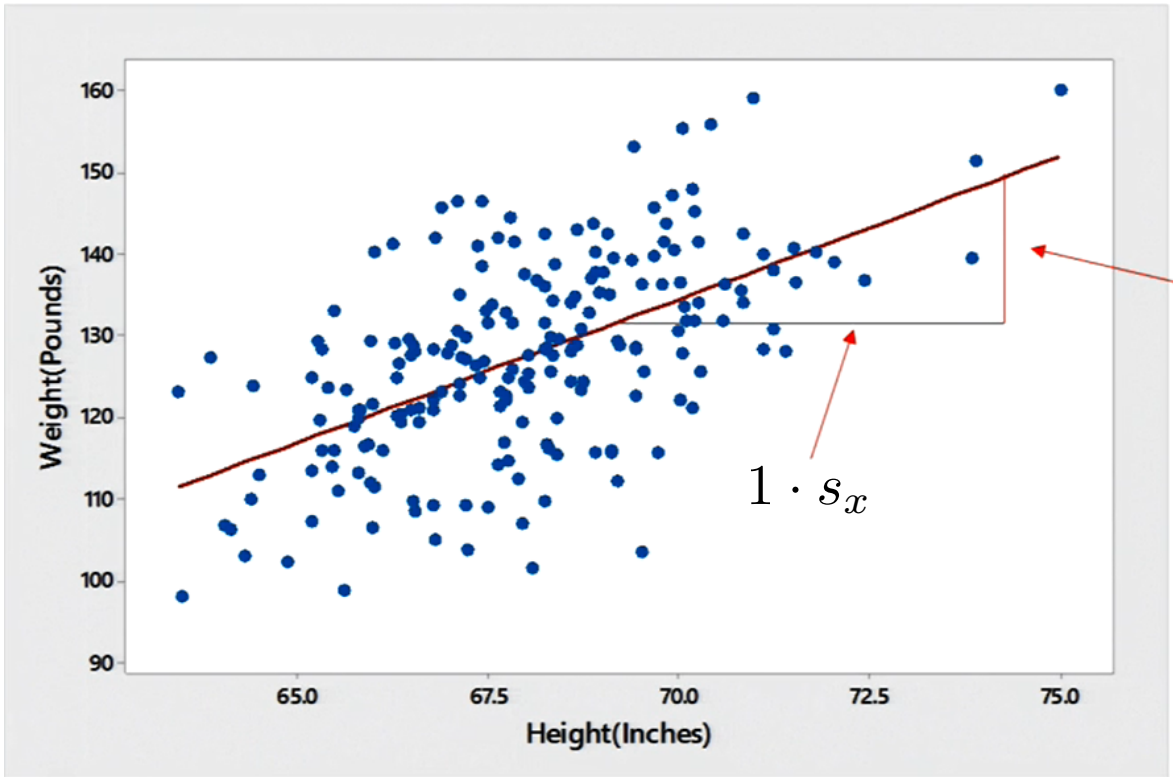
$$\text{Annual murder rate/million people} = -30 + 2.6 \cdot \text{Poverty Rate}.$$

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

$$
\begin{aligned}
b_1 &= R \cdot \frac{s_y}{s_x} \\
&= \frac{R \cdot s_y}{1 \cdot s_x} \\
&= \frac{\Delta y}{\Delta x}
\end{aligned}
$$



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

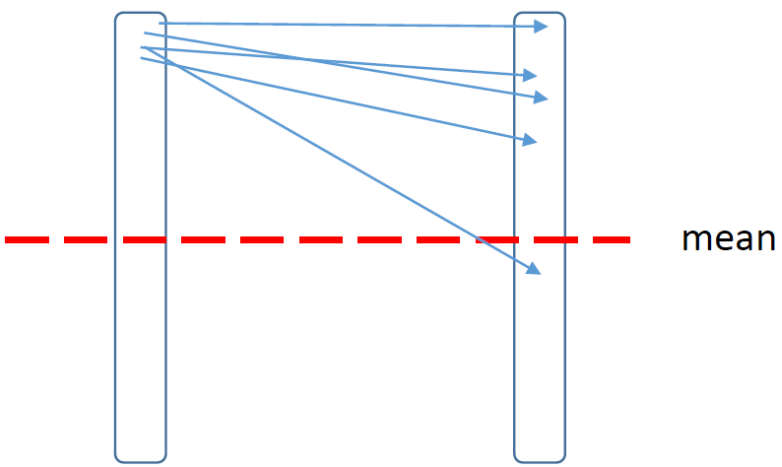
$$-1 \leq R \leq 1.$$

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.

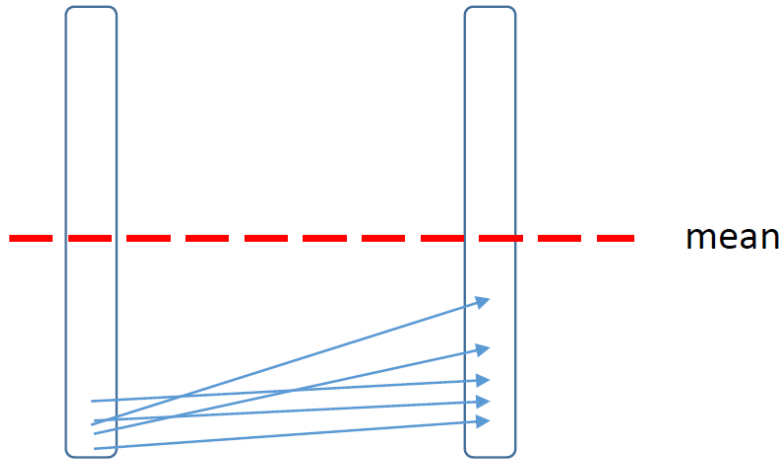
The phrase "regression to the mean" is used to describe this phenomenon, and is where the term "linear regression" comes from.

Regression to the Mean Examples



Scores on test 1

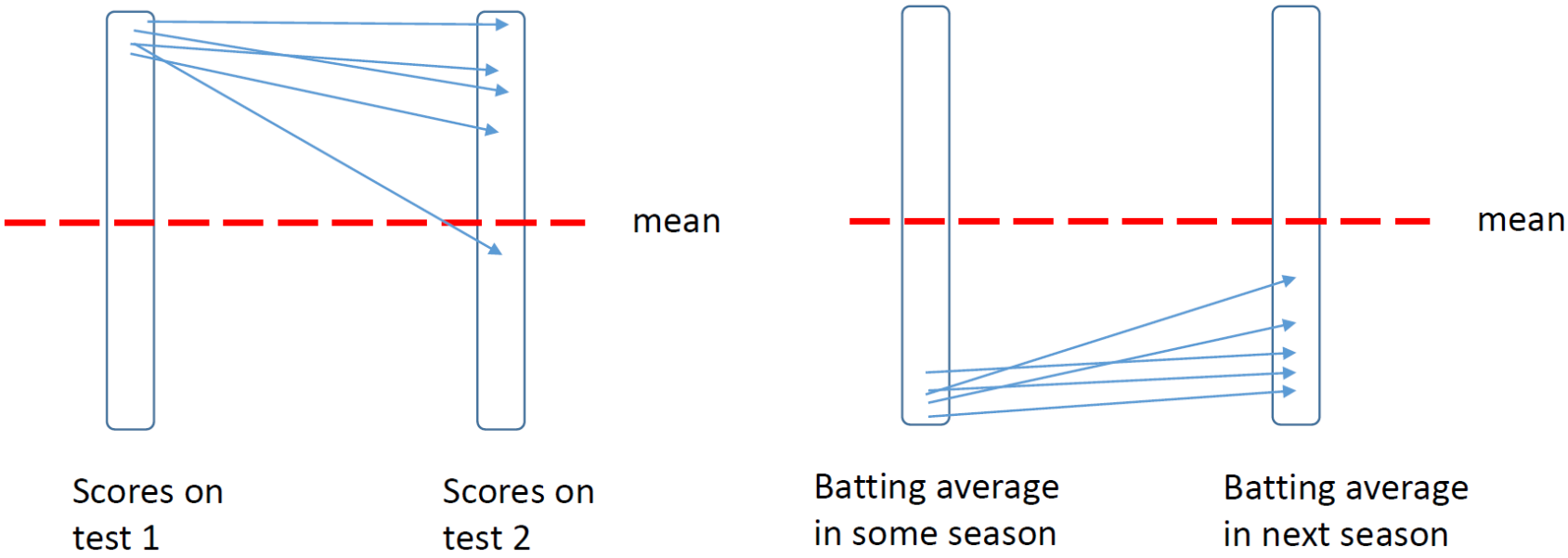
Scores on test 2



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.

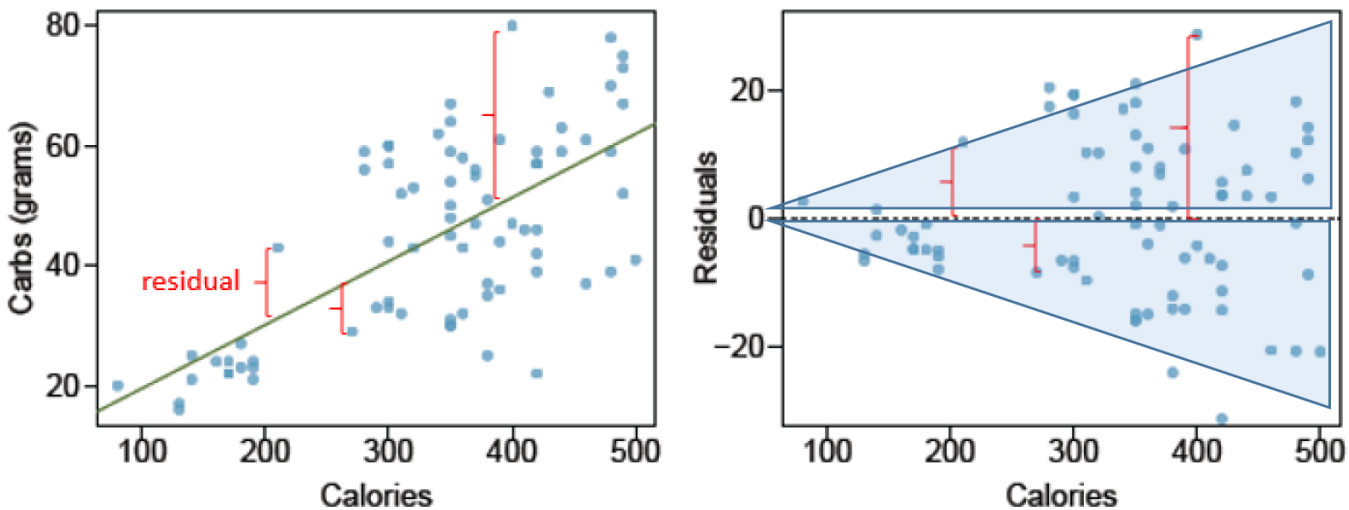
When you look at them on the other measure (the idea on the right), they are still exceptional, but probably won't have the luck this time around.

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.

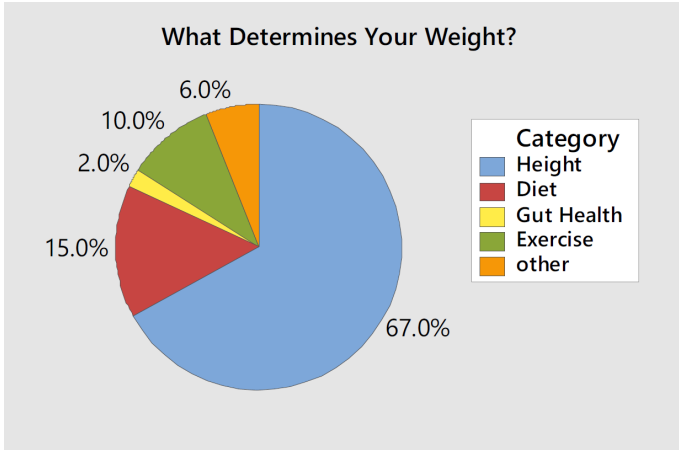
So, $R^2 = 0.854^2 = 0.73 = 73\%$ of how long we must wait is completely determined by how long the last eruption lasted.

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

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

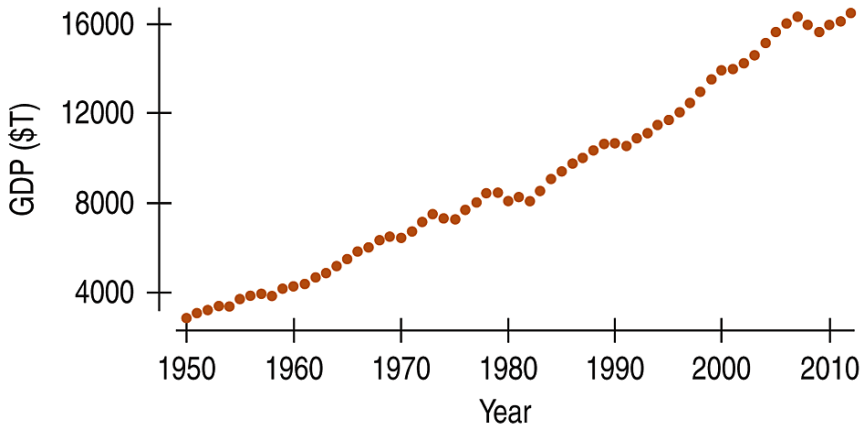
You can check that these equations are NOT simply rearrangements of each other.

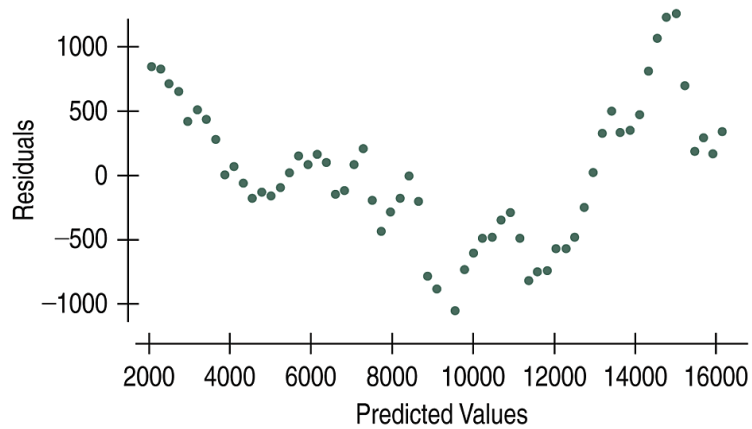
Getting Better at Regression

Sad Reality: As simple as linear regression is, most people use it incorrectly:

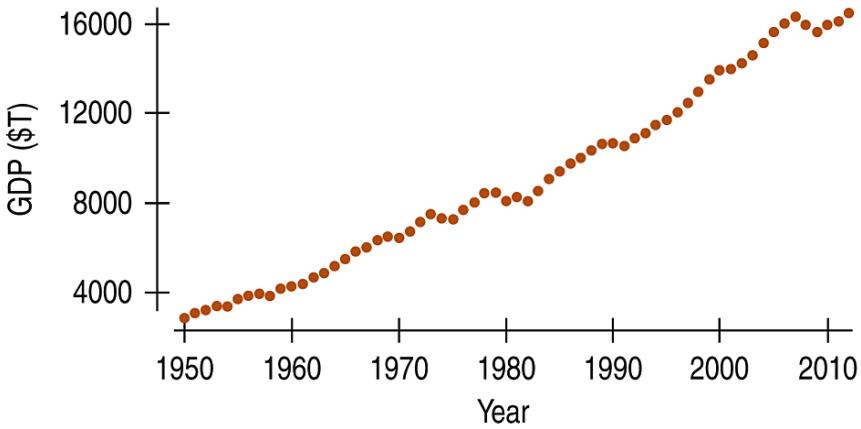
1. They fail to look at the residuals and make sure the model is reasonable.
2. They extrapolate without caution.
3. They don't consider outliers carefully enough.
4. They build a model on data that isn't "straight enough".

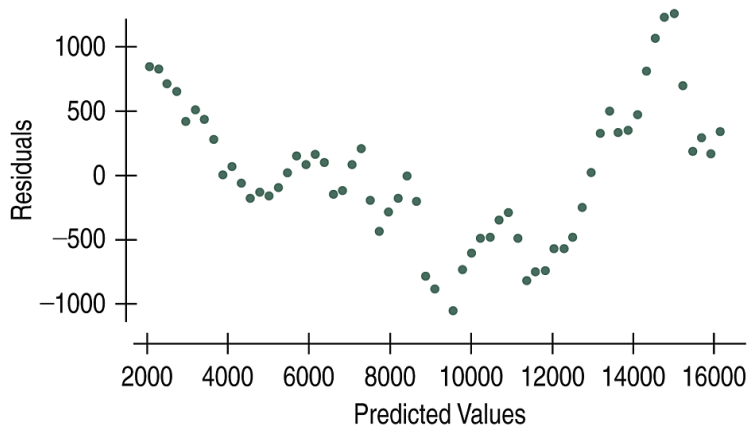
Look at the Residuals (Seriously!)





Look at the Residuals (Seriously!)

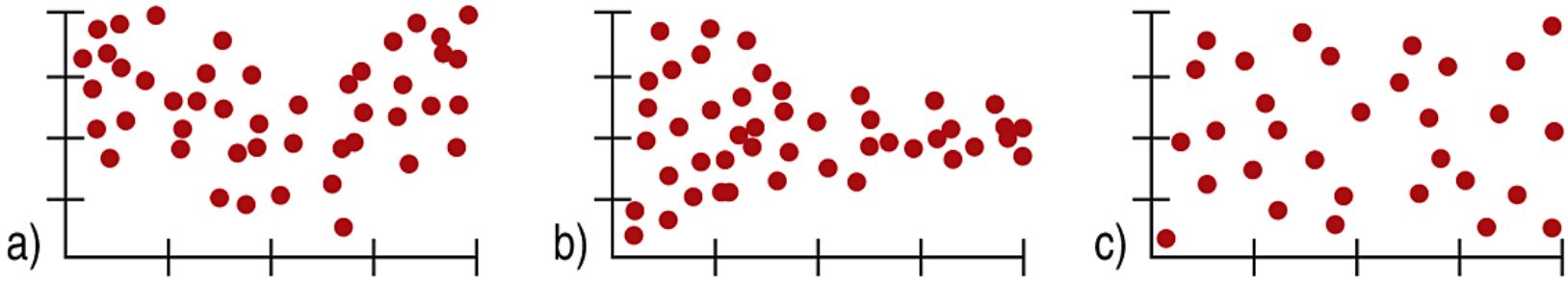




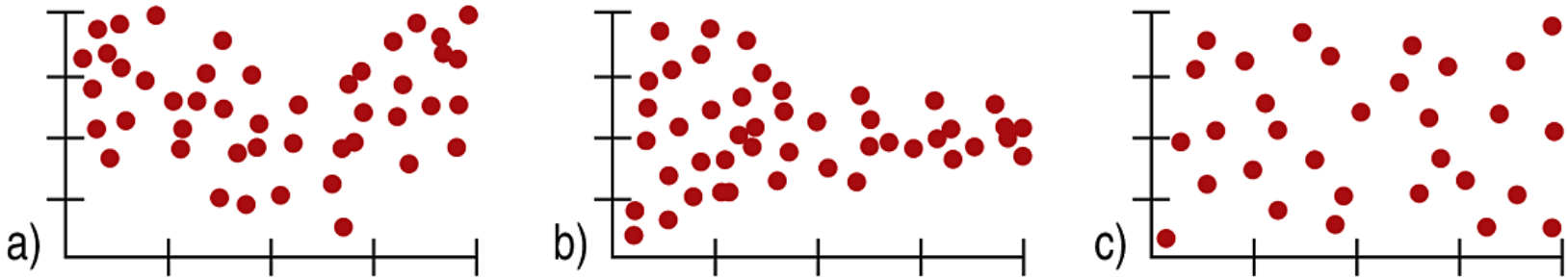
Lesson 1 If the residuals show any type of pattern, your current linear model is not appropriate.

Lesson 2 A high R^2 value is not an indication that a linear model is appropriate!
Here, $R^2 = 98.3\%$.

More on Residuals (Do They Look Like Noise?)



More on Residuals (Do They Look Like Noise?)



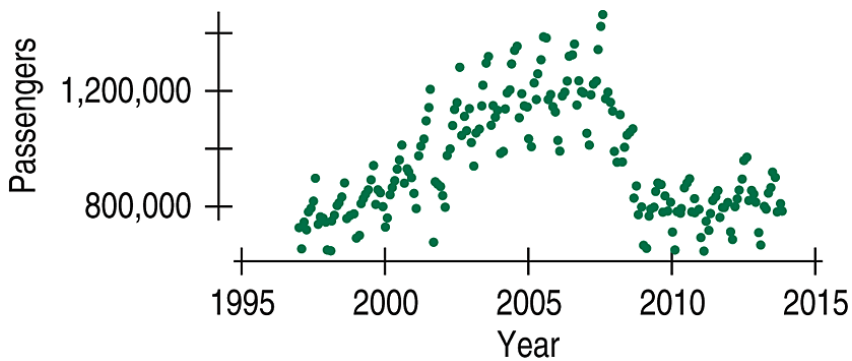
a) This is noisy, but there is a clear downward, then upward, pattern.

b) This plot tends to show more variation for smaller x values, and less for greater x values. A linear model is not appropriate.

c) This truly looks like noise! A linear model is appropriate assuming the other regression conditions are met.

Subgroups In Your Data

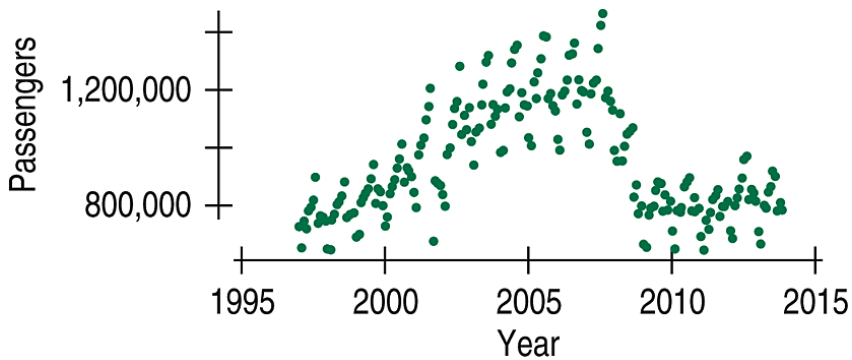
Often, you can identify subgroups in your original data or in the residuals. In this case, split your data into different parts and do several linear regression instead of one, clunky, regression.



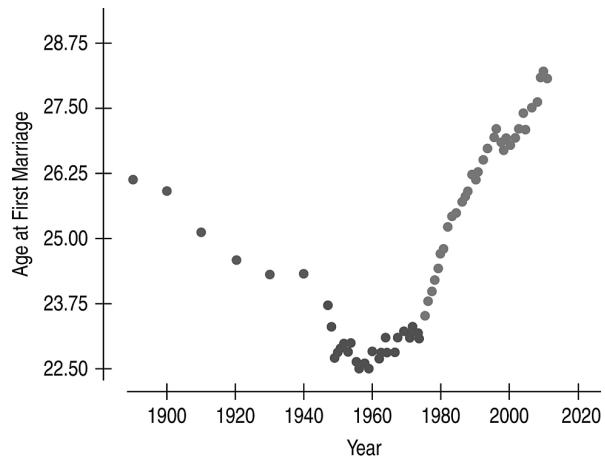
# of passengers thru Oakland's airport each month

Subgroups In Your Data

Often, you can identify subgroups in your original data or in the residuals. In this case, split your data into different parts and do several linear regression instead of one, clunky, regression.

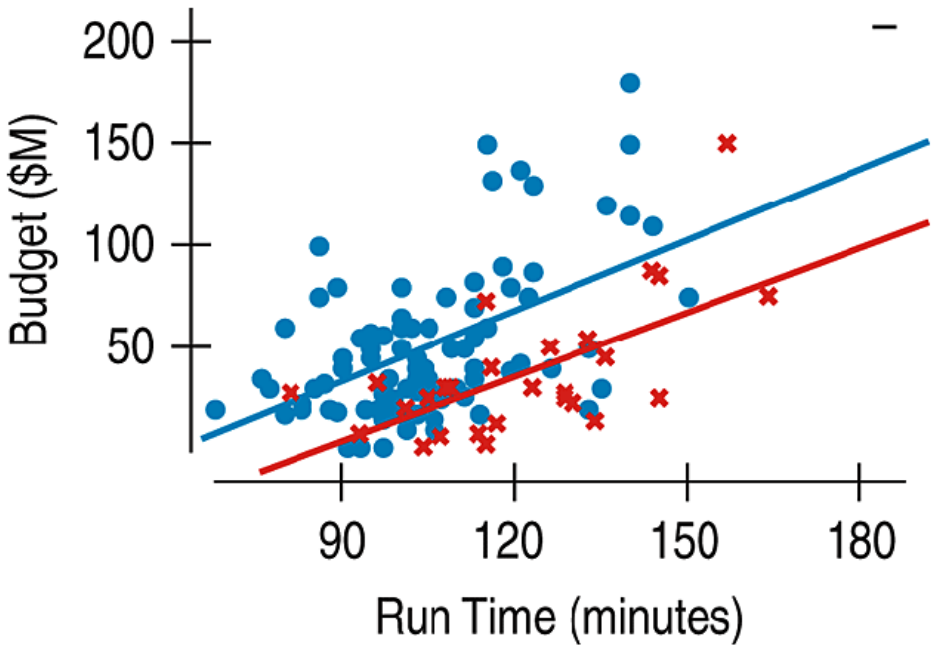


# of passengers thru Oakland's airport each month



Average age at which people got married

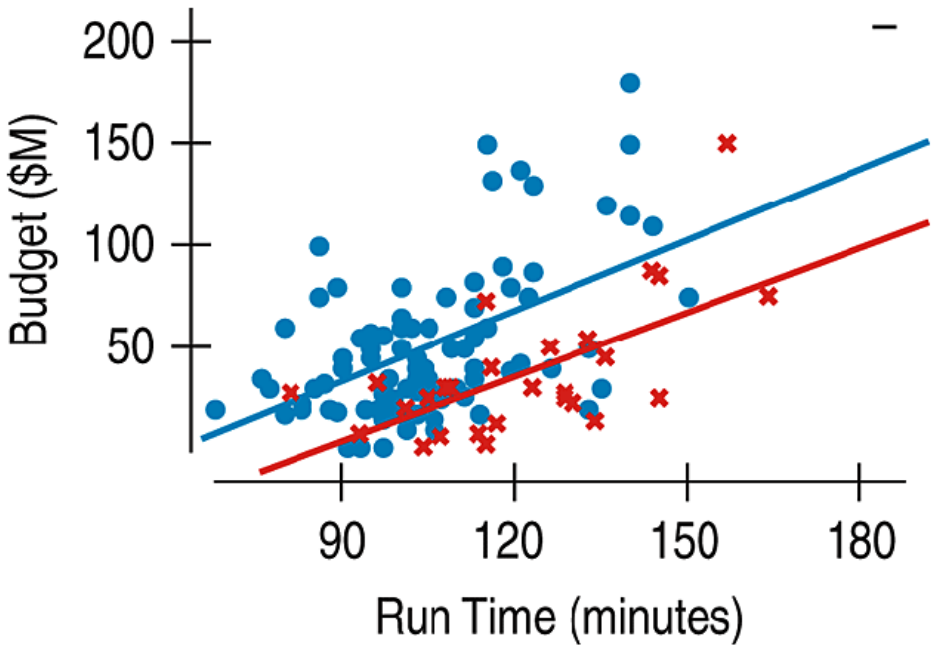
Subgroups May Not Be Visible Unless You Think of Them



Cost and Run Time of Major Movie Releases Since 2005

Red X: Drama
Blue dot: Non-drama
Dash: King Kong (2005)

Subgroups May Not Be Visible Unless You Think of Them

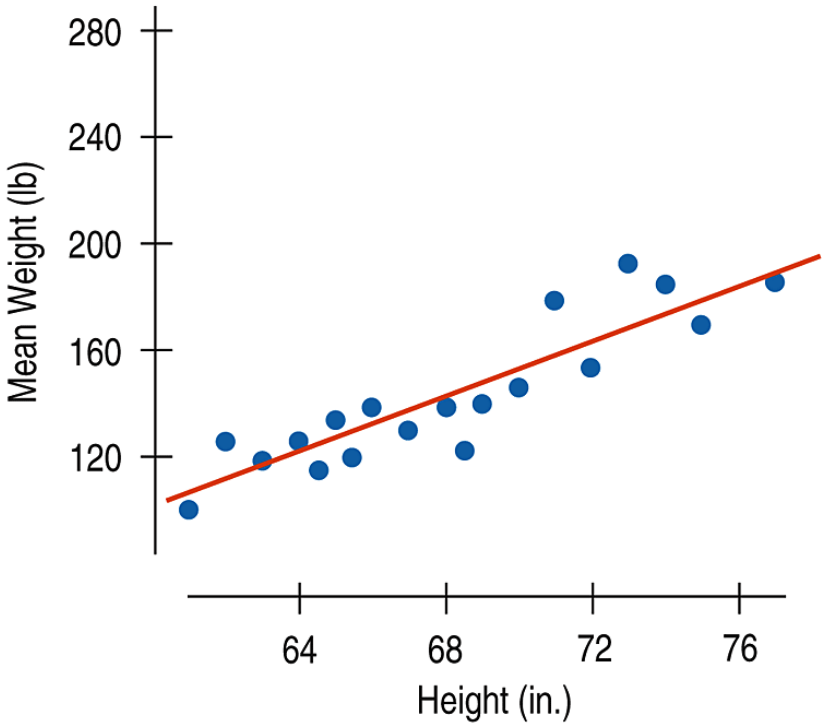


Cost and Run Time of Major Movie Releases Since 2005

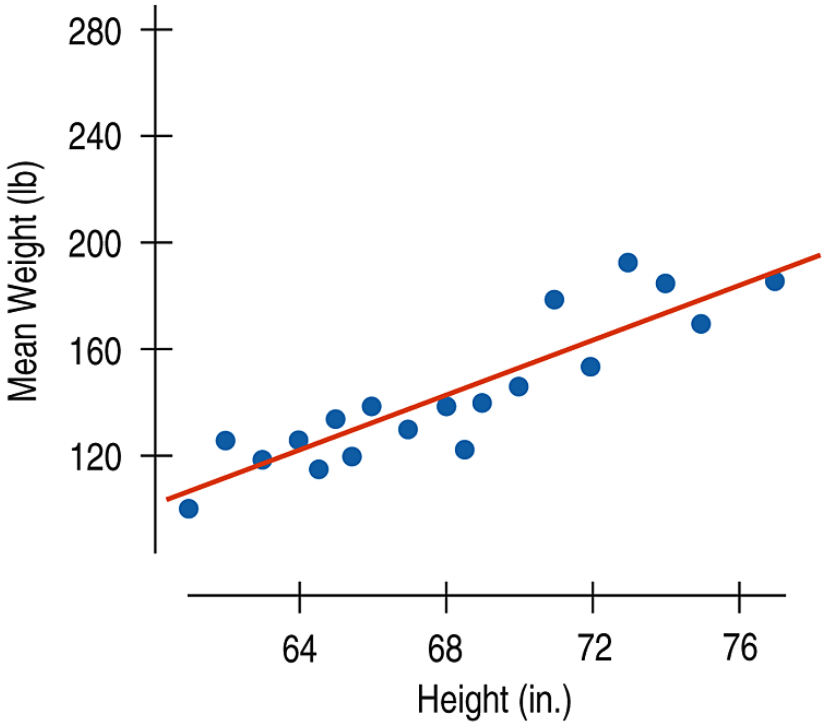
Red X: Drama
Blue dot: Non-drama
Dash: King Kong (2005)

Lesson 3 Don't assume your data are all part of one homogeneous population.
Think about possible subgroups to make the analysis better.

Interpolation VS Extrapolation



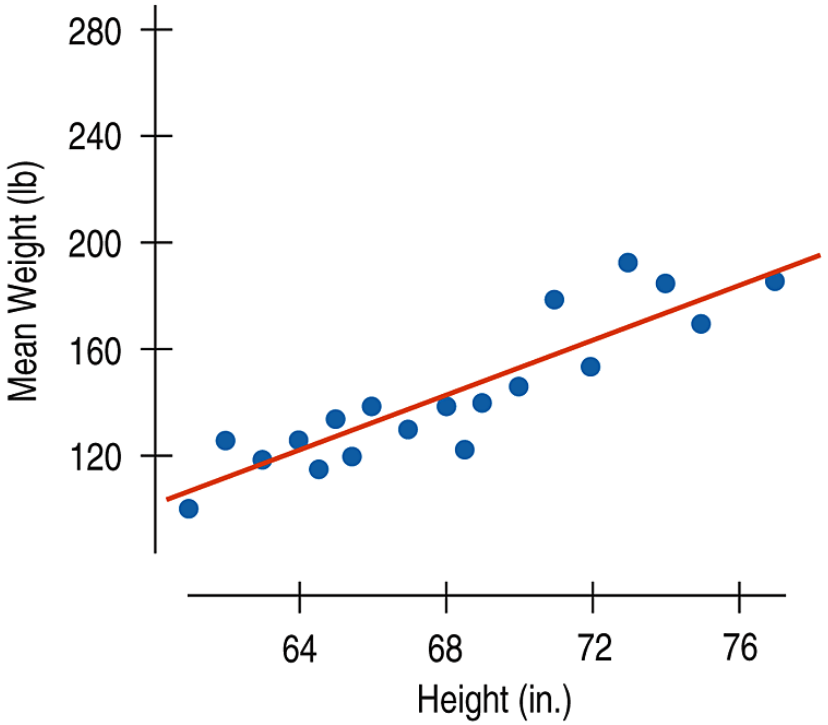
Interpolation VS Extrapolation



Interpolation: Using your model to predict a new y value for an x value that is within the span of x data you modeled. (Here, inside the 61-77 inch range)

Extrapolation: Using your model to predict a new y value for an x value that is outside the span of x data you modeled. (Here, outside the 61-77 inch range)

Interpolation VS Extrapolation



Interpolation: Using your model to predict a new y value for an x value that is within the span of x data you modeled. (Here, inside the 61-77 inch range)

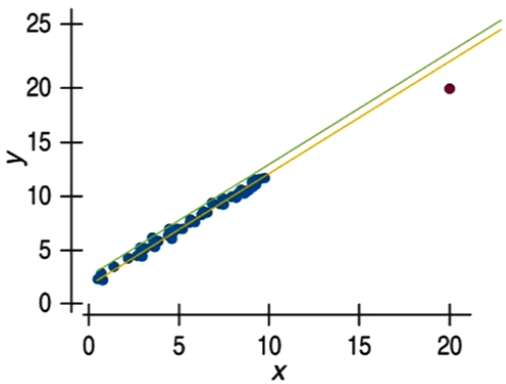
Extrapolation: Using your model to predict a new y value for an x value that is outside the span of x data you modeled. (Here, outside the 61-77 inch range)

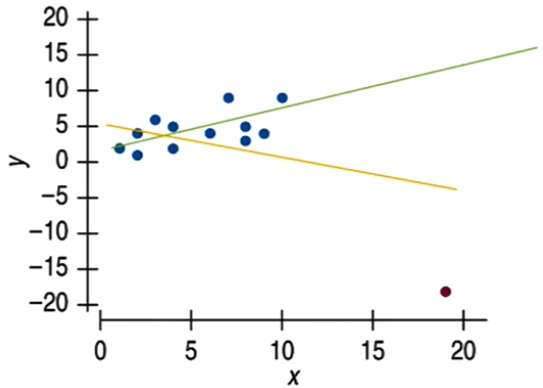
Lesson 4 Extrapolation is dangerous because it assumes the relationship holds beyond the data range you have seen and used for a model.

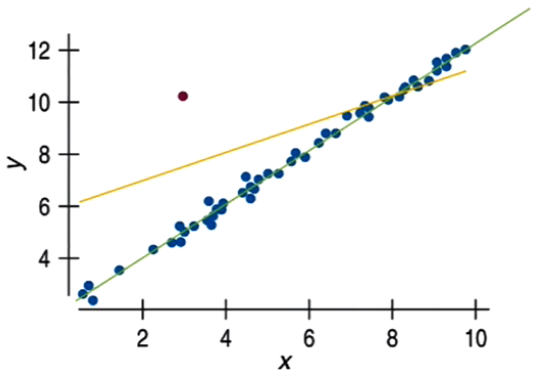
Should I Run a Regression With an Outlier?

For each graph,

- the orange line is the regression line using all the point (including the outlier)
- the green line is the regression line when the outlier is excluded.



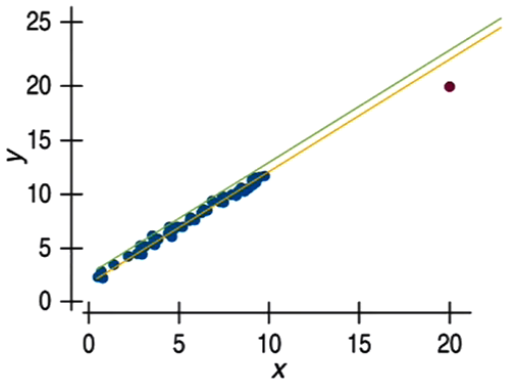


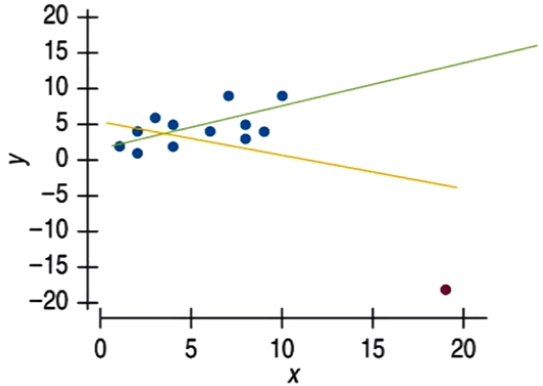


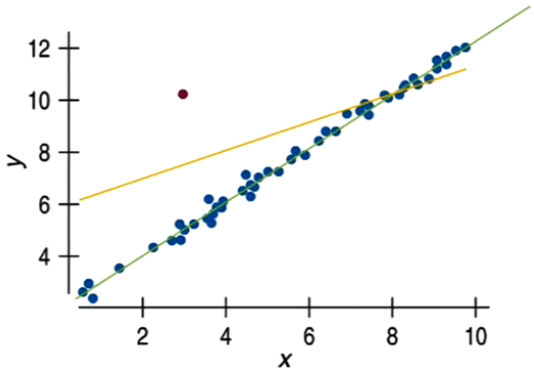
Should I Run a Regression With an Outlier?

For each graph,

- the orange line is the regression line using all the point (including the outlier)
- the green line is the regression line when the outlier is excluded.







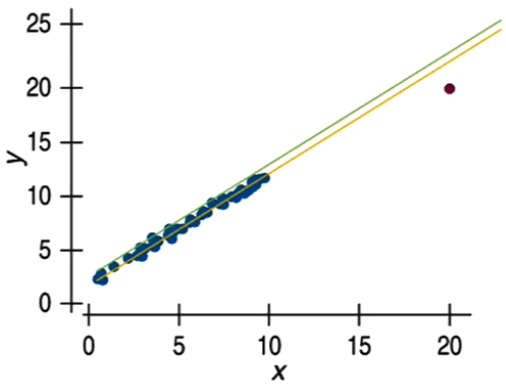
Two important ideas:

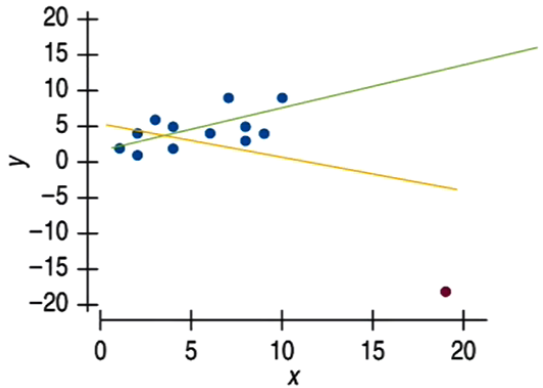
- A high leverage point is one where x is far from the mean of the x values.
- A high influence point is one that gives a significantly different slope for the regression line when it is included, versus excluded.

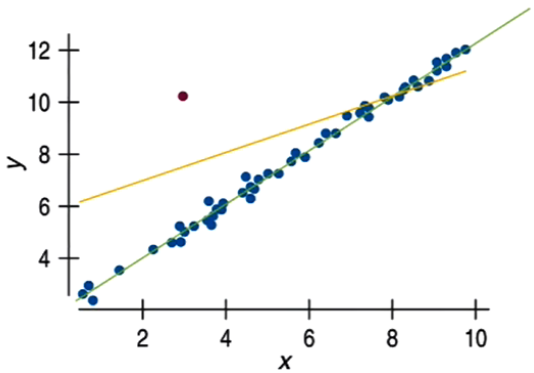
Should I Run a Regression With an Outlier?

For each graph,

- the orange line is the regression line using all the point (including the outlier)
- the green line is the regression line when the outlier is excluded.







High leverage
Low influence

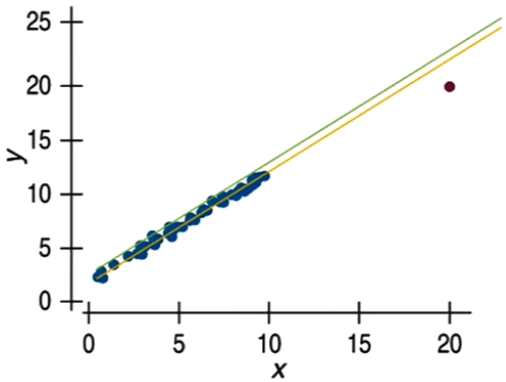
Two important ideas:

- A high leverage point is one where x is far from the mean of the x values.
- A high influence point is one that gives a significantly different slope for the regression line when it is included, versus excluded.

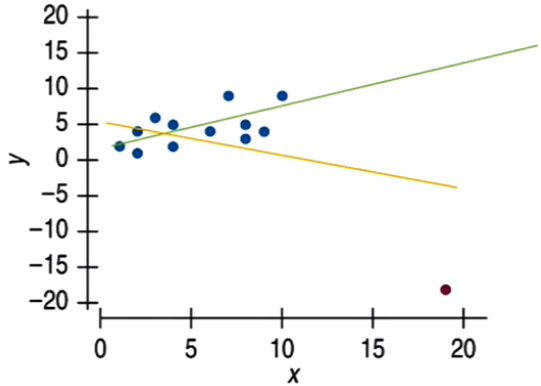
Should I Run a Regression With an Outlier?

For each graph,

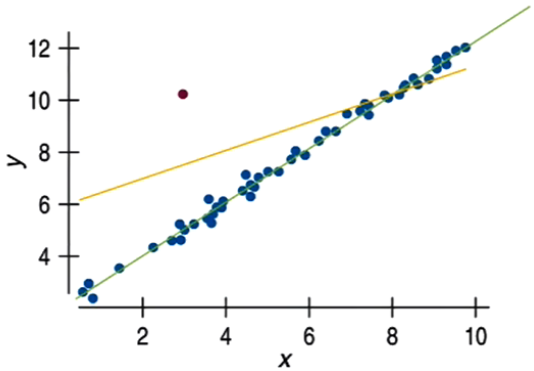
- the orange line is the regression line using all the point (including the outlier)
- the green line is the regression line when the outlier is excluded.



High leverage
Low influence



High leverage
High influence



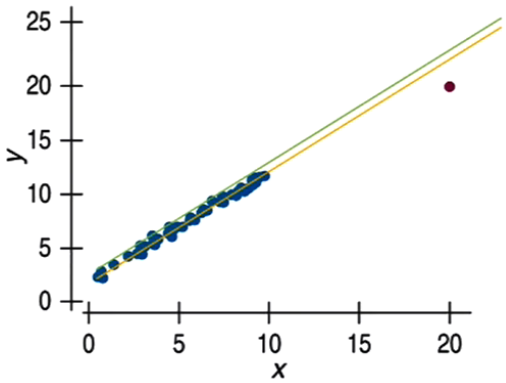
Two important ideas:

- A high leverage point is one where x is far from the mean of the x values.
- A high influence point is one that gives a significantly different slope for the regression line when it is included, versus excluded.

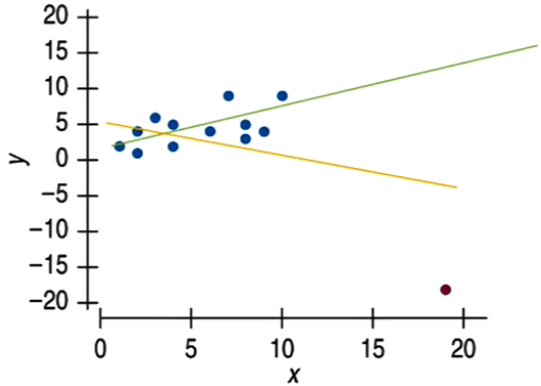
Should I Run a Regression With an Outlier?

For each graph,

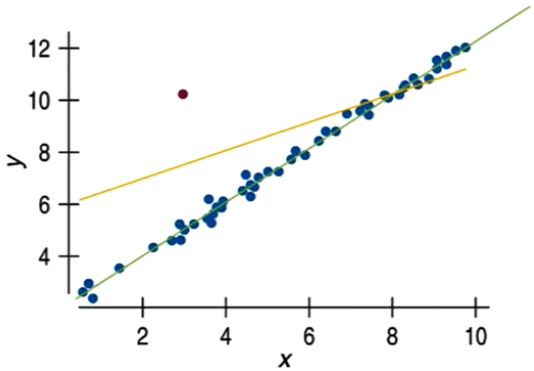
- the orange line is the regression line using all the point (including the outlier)
- the green line is the regression line when the outlier is excluded.



High leverage
Low influence



High leverage
High influence



Low leverage
Moderate influence

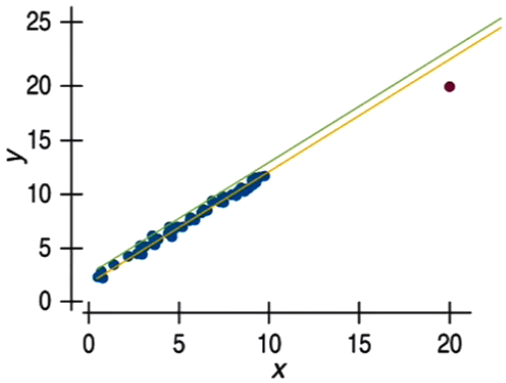
Two important ideas:

- A high leverage point is one where x is far from the mean of the x values.
- A high influence point is one that gives a significantly different slope for the regression line when it is included, versus excluded.

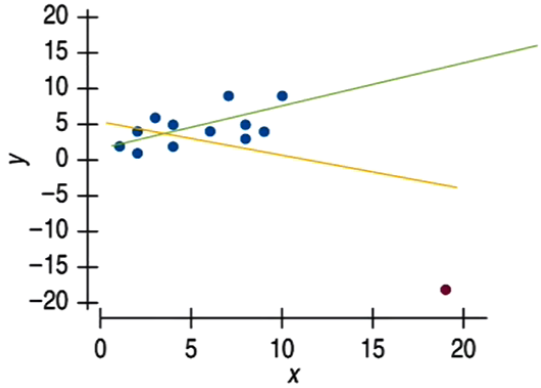
Should I Run a Regression With an Outlier?

For each graph,

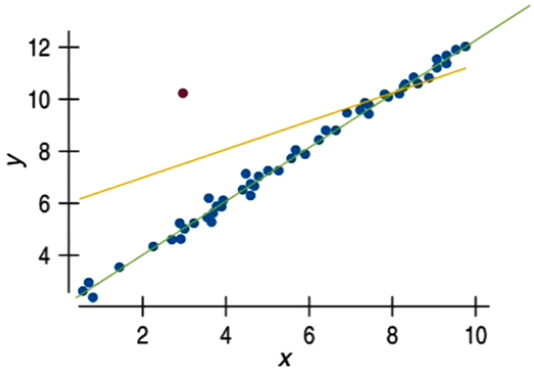
- the orange line is the regression line using all the point (including the outlier)
- the green line is the regression line when the outlier is excluded.



High leverage
Low influence



High leverage
High influence



Low leverage
Moderate influence

Two important ideas:

- A high leverage point is one where x is far from the mean of the x values.
- A high influence point is one that gives a significantly different slope for the regression line when it is included, versus excluded.

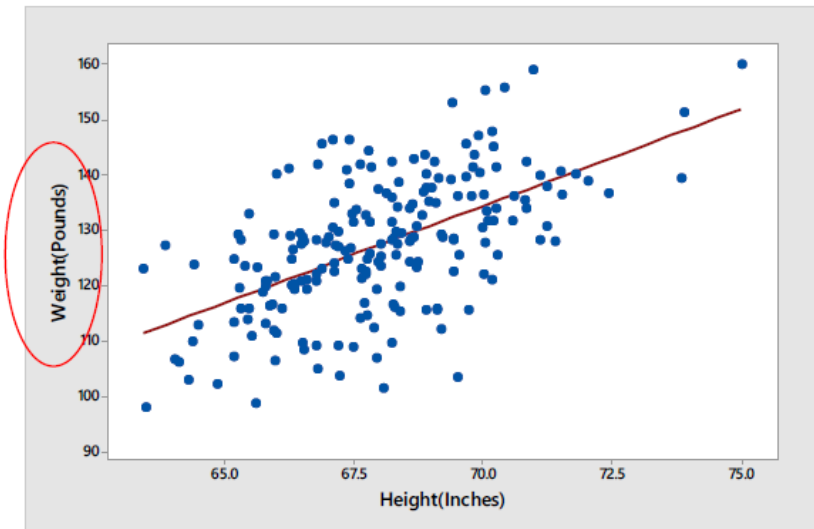
Lesson 5 Don't run regression when a high influence outlier is present

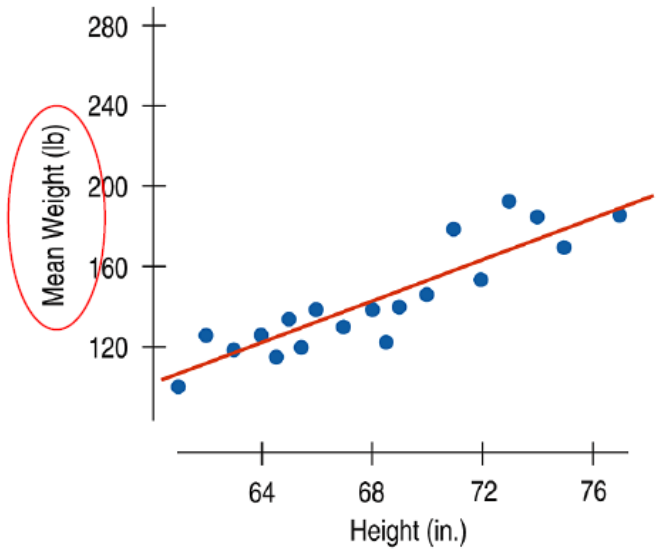
Inflating R^2: Three Common Practices

1. Dividing the data into subgroups that are more homogeneous. (Often GOOD)
2. Tossing outliers and doing the analysis without them. (GOOD or BAD, depending)
3. Using summarized data rather than un-summarized data. (Often BAD)

Inflating R^2: Three Common Practices

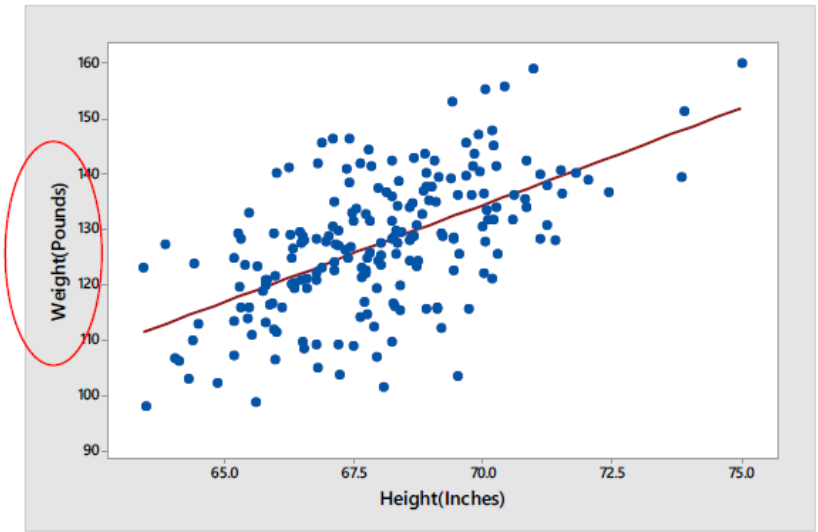
1. Dividing the data into subgroups that are more homogeneous. (Often GOOD)
2. Tossing outliers and doing the analysis without them. (GOOD or BAD, depending)
3. Using summarized data rather than un-summarized data. (Often BAD)

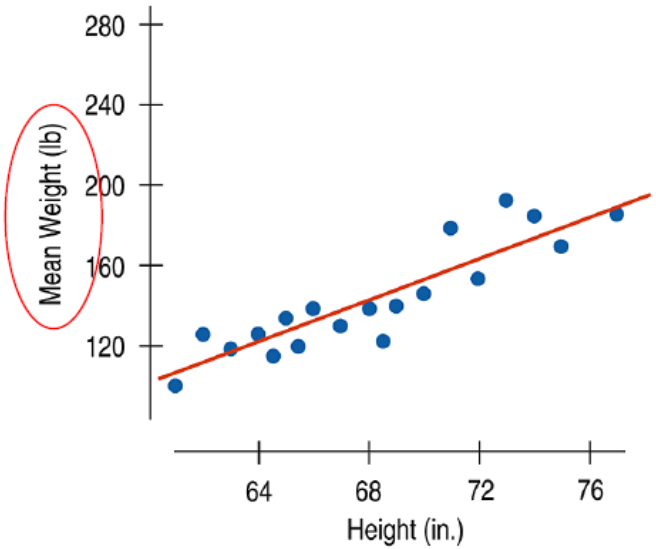




Inflating R^2: Three Common Practices

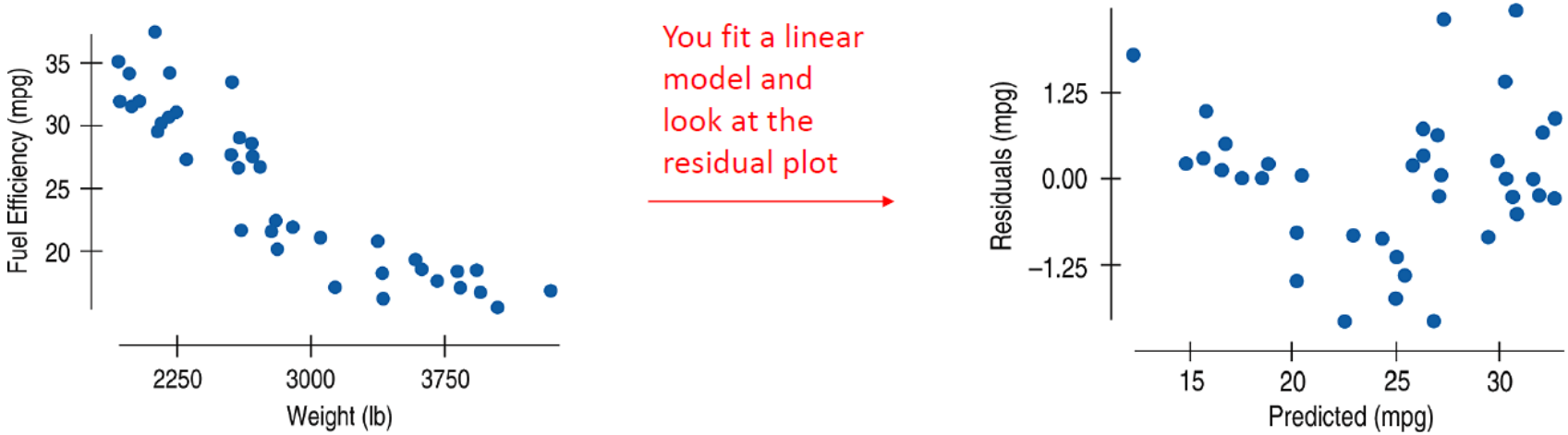
1. Dividing the data into subgroups that are more homogeneous. (Often GOOD)
2. Tossing outliers and doing the analysis without them. (GOOD or BAD, depending)
3. Using summarized data rather than un-summarized data. (Often BAD)





Lesson 2 (again) A high R^2 does not indicate a linear model is appropriate

Getting Your Data Straight



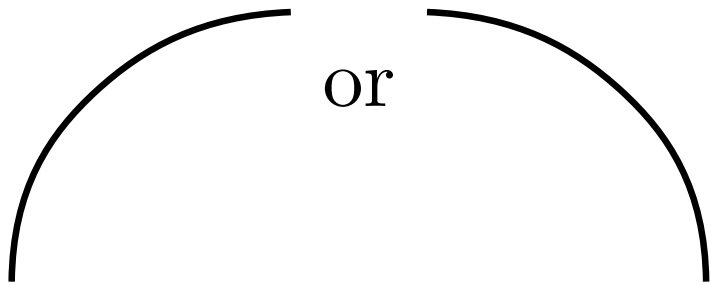
When the original data or the residuals convince you that the data are not straight enough, apply a mathematical function to the y-value.
(You might apply a function to the x-values, or both the x and y values)

What Function Should I Apply? The Tower of Power

Power	**Function You Apply**
2	y^2
1	y no function applied this is the raw data, and your home base for the Tower of Power
$1/2$	$\sqrt{y}$
"0"	$\log y$
$-1/2$	$\dfrac{-1}{\sqrt{y}}$
-1	$\dfrac{-1}{y}$
-2	$\dfrac{-1}{y^2}$

Tukey's Circle

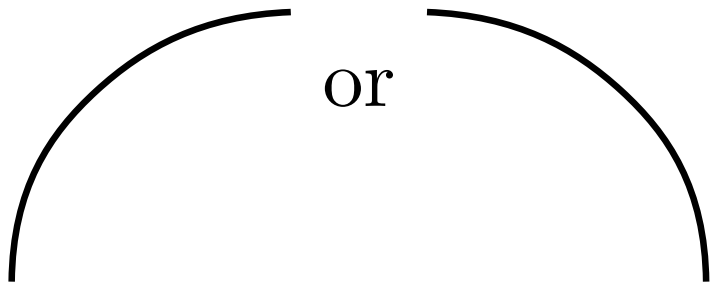
If the graph looks like



then apply a function higher on the Tower of Power than is currently being used.

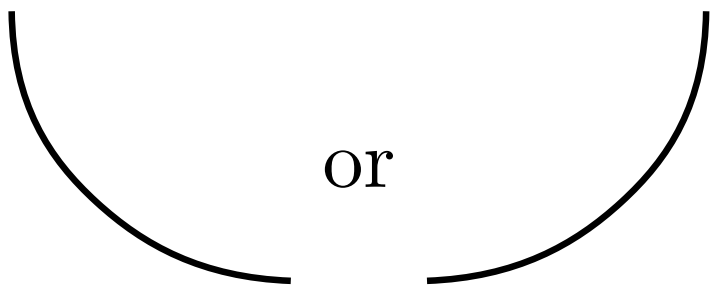
Tukey's Circle

If the graph looks like



then apply a function higher on the Tower of Power than is currently being used.

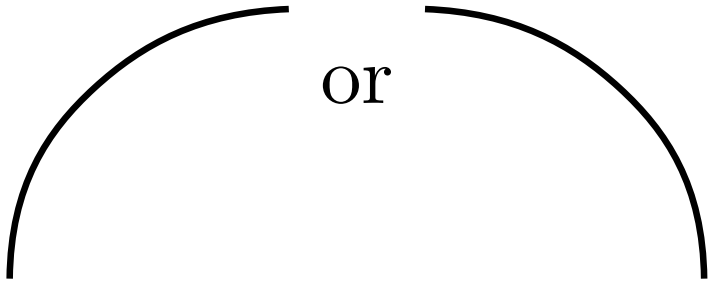
If the graph looks like



then apply a function lower on the Tower of Power than is currently being used.

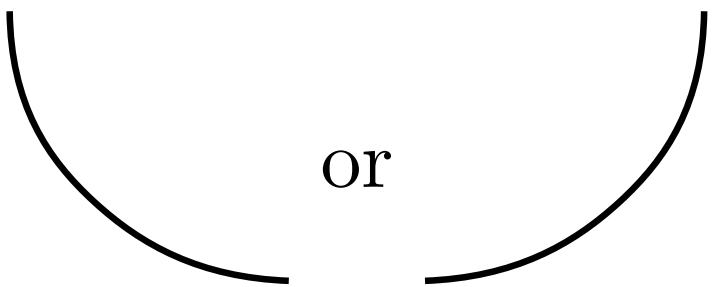
Tukey's Circle

If the graph looks like

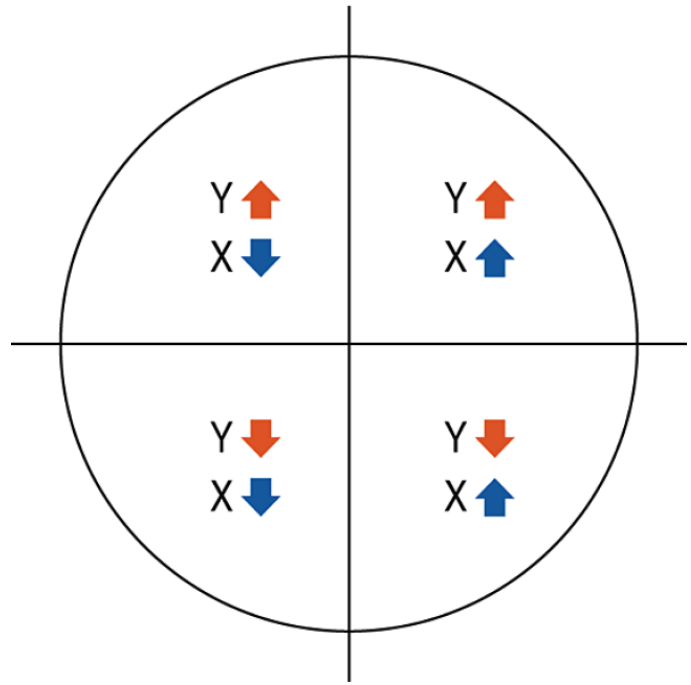


then apply a function higher on the Tower of Power than is currently being used.

If the graph looks like



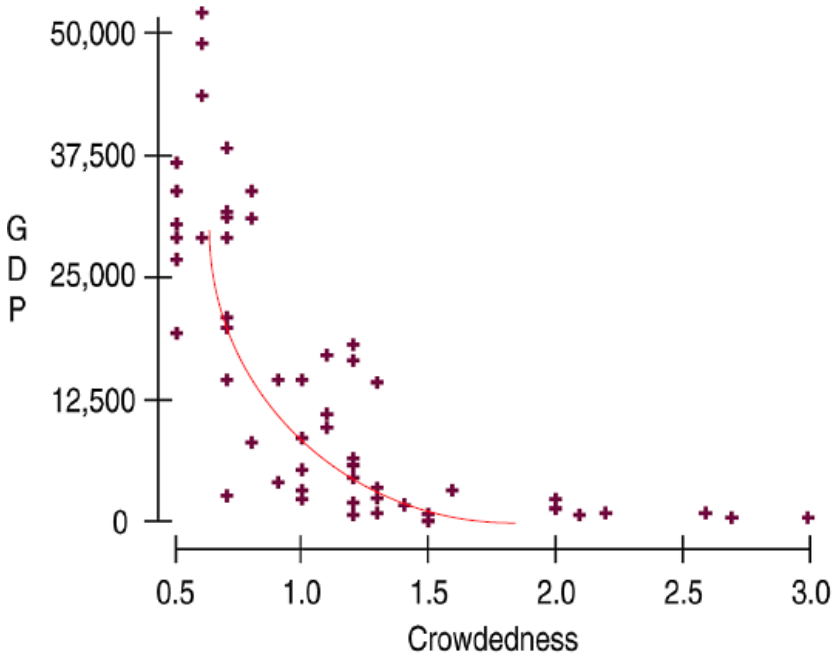
then apply a function lower on the Tower of Power than is currently being used.



Tukey's Rule of Thumbs for Re-Expression

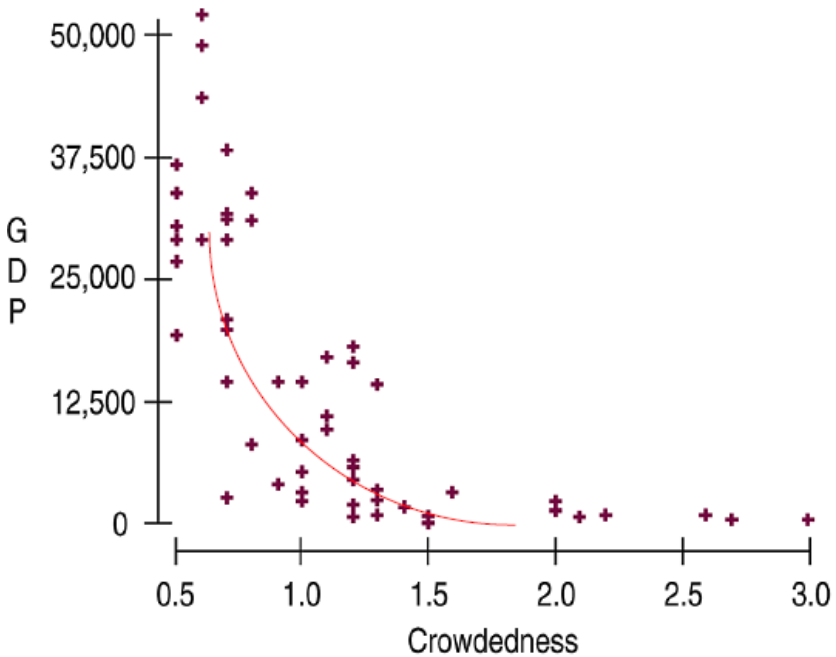
Transforming Data: An Example

Crowdedness (average number of people per room in a house) vs. GDP (per capita) in 56 countries



Transforming Data: An Example

Crowdedness (average number of people per room in a house) vs. GDP (per capita) in 56 countries



y vs. x (Current place in Tower of Power)

$\sqrt{y}$ vs. x

$\log y$ vs. x

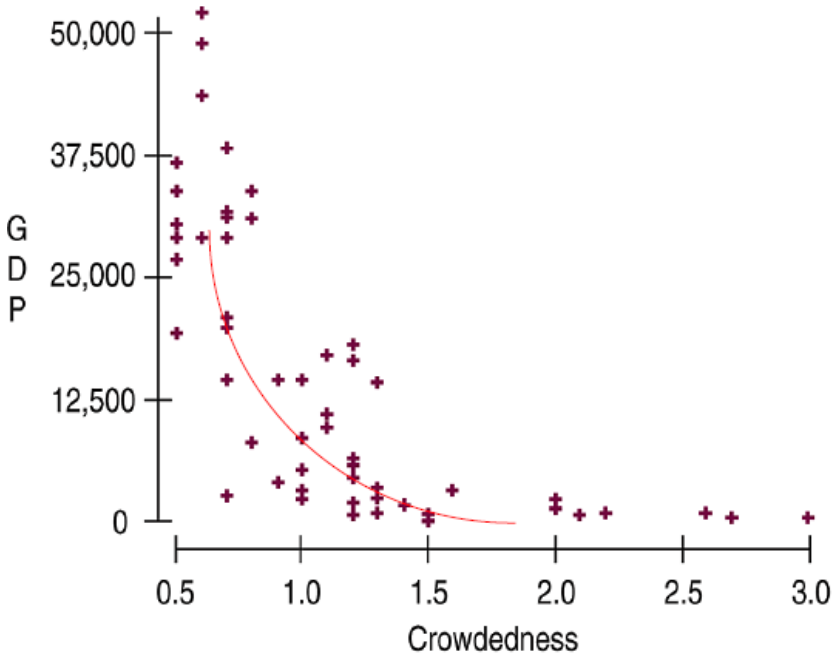
$\frac{-1}{\sqrt{y}}$ vs. x

$\frac{-1}{y}$ vs. x

Tukey says move down some amount

Transforming Data: An Example

Crowdedness (average number of people per room in a house) vs. GDP (per capita) in 56 countries



y vs. x (Current place in Tower of Power)

$\sqrt{y}$ vs. x

$\log y$ vs. x

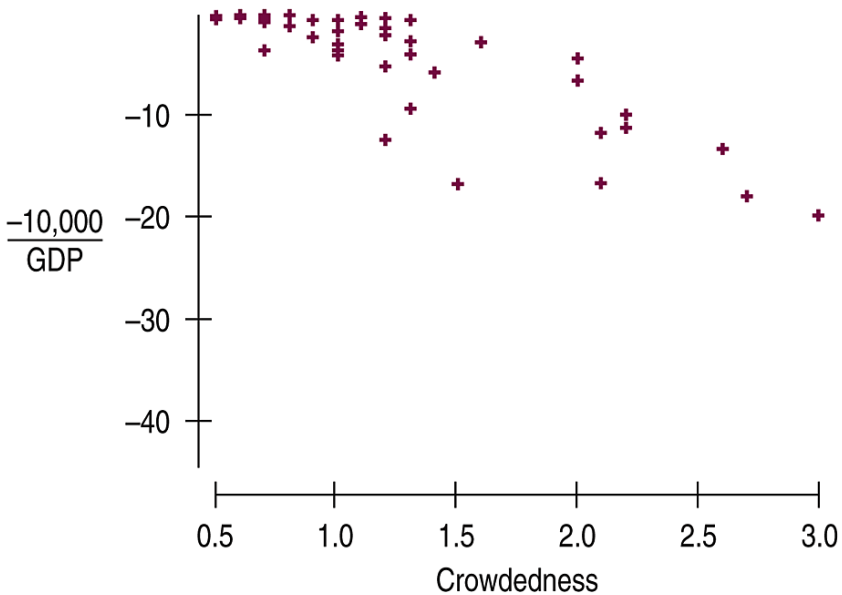
$\frac{-1}{\sqrt{y}}$ vs. x

$\frac{-1}{y}$ vs. x

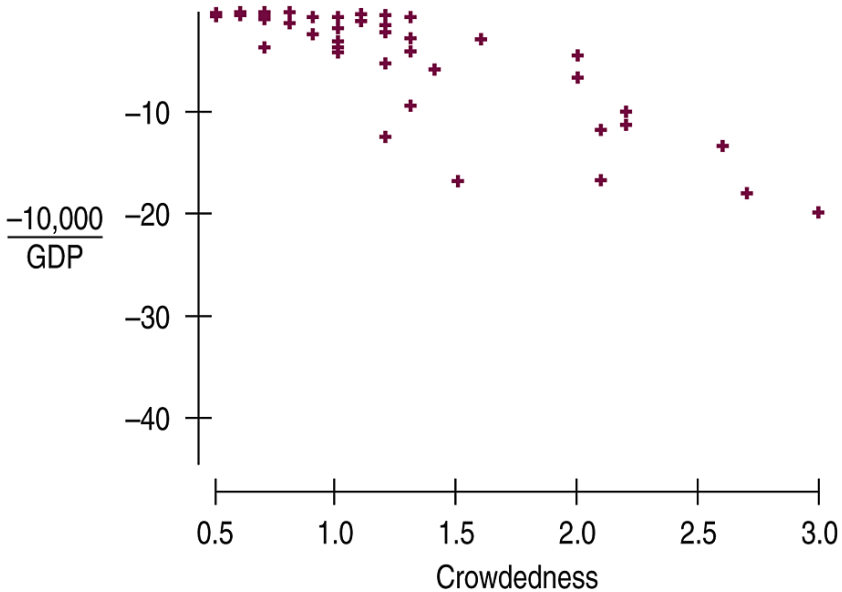
Tukey says move down some amount

Suppose we try this transformation $-1/y$. (see next slide)

Transforming Data: An Example



Transforming Data: An Example



y vs. x

$\sqrt{y}$ vs. x

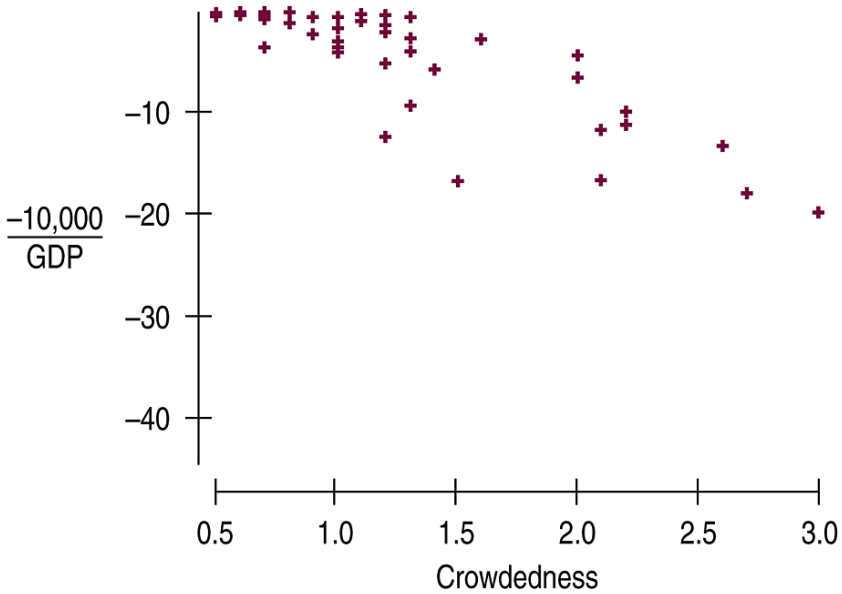
$\log y$ vs. x

$\frac{-1}{\sqrt{y}}$ vs. x

Tukey says move up some amount

$\frac{-1}{y}$ vs. x (HERE NOW!)

Transforming Data: An Example



y vs. x

$\sqrt{y}$ vs. x

$\log y$ vs. x

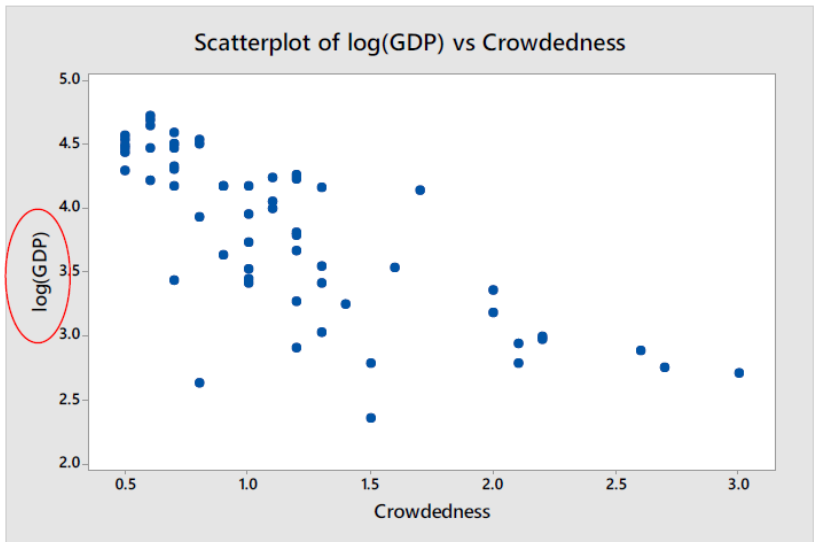
$\frac{-1}{\sqrt{y}}$ vs. x

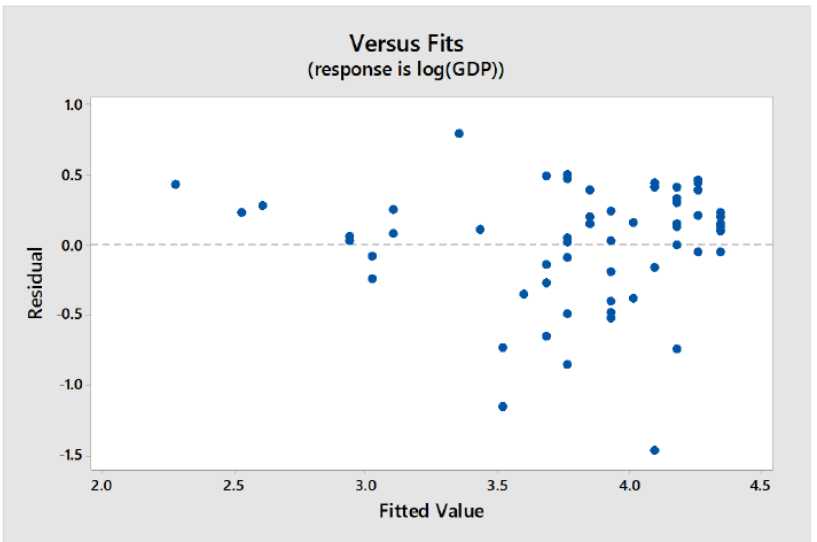
Tukey says move up some amount

$\frac{-1}{y}$ vs. x (HERE NOW!)

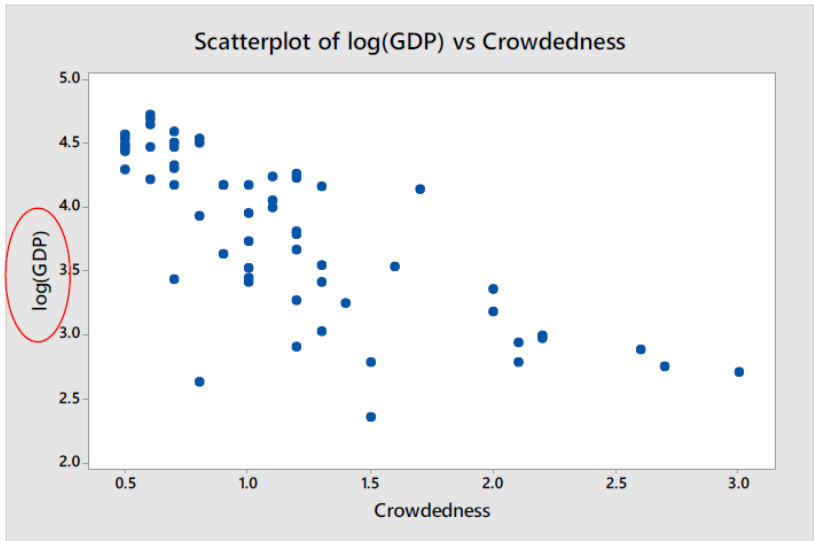
Let's try $\log y$ next...
(see next slide)

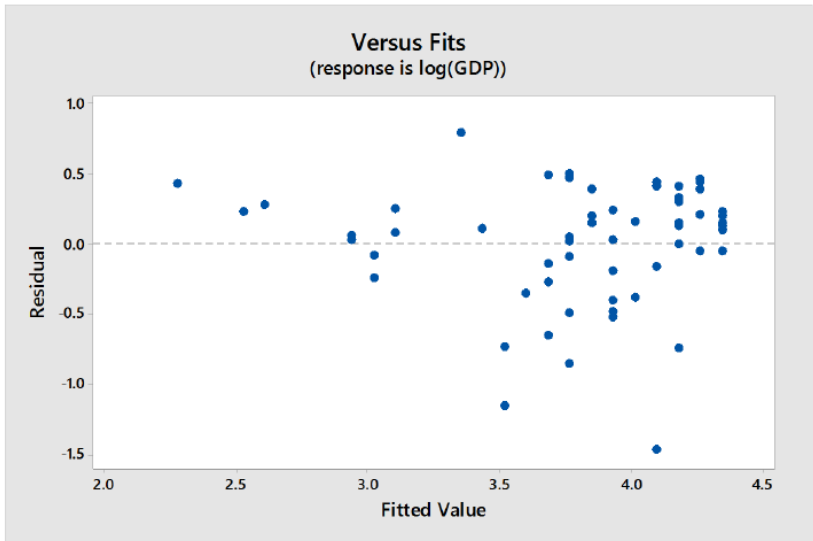
Transforming Data: An Example





Transforming Data: An Example





There is no magic to choosing the right transformation. If the changes to the y variable don't give a straight-enough graph, try some of the x transformations.
(or use both x and y transformations)

Working With a Transformed Model

Software gives us a linear model:

$$\widehat{\log(\text{GDP})} = 4.755 - 0.8264(\text{Crowdedness}).$$

Predict the per capita GDP for a country with a crowdedness of 5.

Working With a Transformed Model

Software gives us a linear model:

$$\widehat{\log(\text{GDP})} = 4.755 - 0.8264(\text{Crowdedness}).$$

Predict the per capita GDP for a country with a crowdedness of 5.

Plugging in 5 gives us $\widehat{\log(\text{GDP})} = 4.755 - 0.8264 \cdot 5$.

Thus $\widehat{\log(\text{GDP})} = 0.623$, and so GDP $= 10^{0.623} = 4.20$.

We expect a country with an average of 5 people living per room to have a per-capita GDP of \$4.20!

Working With a Transformed Model

Software gives us a linear model:

$$\widehat{\log(\text{GDP})} = 4.755 - 0.8264(\text{Crowdedness}).$$

Predict the per capita GDP for a country with a crowdedness of 5.

Plugging in 5 gives us $\widehat{\log(\text{GDP})} = 4.755 - 0.8264 \cdot 5$.

Thus $\widehat{\log(\text{GDP})} = 0.623$, and so GDP $= 10^{0.623} = 4.20$.

We expect a country with an average of 5 people living per room to have a per-capita GDP of \$4.20!

Note: Pakistan has the highest crowdedness score of 3. So, we just extrapolated, even if we didn't realize it.

Why Not Fit a Curve to the Data?

- This is possible but requires additional technical machinery
- You lose the intuitive meaning of slope
- You lose the ease of a linear model; transformations help convert things to a linear world
- Non-technical audiences struggle with anything beyond the world of the linear