Name_____

AP Statistics CED 2.1 Daily Video 1 (Skill 1.A)

Introducing Statistics – Are Variables Related?

What Will We Learn?

How will we decide if there is a relationship between two categorical variables?

How will we decide if there is a relationship between two quantitative variables?

A Quick Review

Categorical Variable:

A variable that takes values that are _____ or _____.

Quantitative Variable:

A variable that takes _____ values for a _____ or _____ quantity.

Examples: Are age group and educational attainment related?

	25 to 34 years old	35 to 54 years old	55+ years old
No high school diploma	3,264	7,982	10,729
High school diploma	19,169	33,082	44,696
Associate's or bachelor's degree	17,468	40,035	42,333
Master's degree or higher	5,311	12,407	12,163

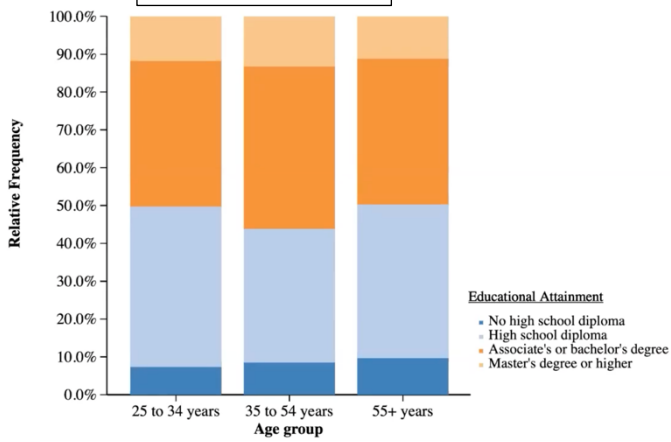
Source: 2019 Census data. Numbers are in thousands.

Categorical or Quantitative Variables? _____

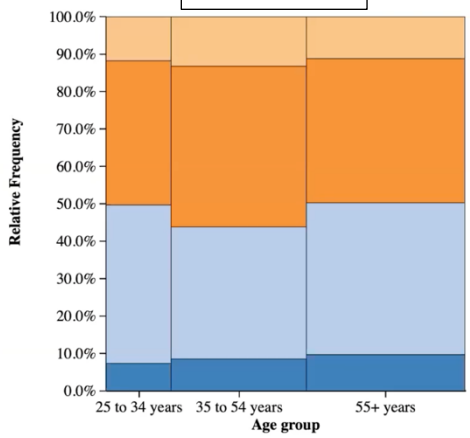
How do we determine if there is a relationship? _____

Graphical Representations

Segmented Bar Chart



Mosaic Plot



Numerical representations

		Age group			
		25 to 34 years old	35 to 54 years old	55+ years old	**Total**
Educational Attainment	No high school diploma	3264 (7.2%)	7982 (8.5%)	10729 (9.8%)	21975 (8.8%)
	High school diploma	19169 (42.4%)	33082 (35.4%)	44696 (40.7%)	96947 (39%)
	Associate's or Bachelor's degree	17468 (38.6%)	40035 (42.8%)	42333 (38.5%)	99836 (40.2%)
	Master's degree or higher	5311 (11.7%)	12407 (13.3%)	12163 (11.1%)	29881 (12%)
	Total	45212 (100%)	93506 (100%)	109921 (100%)	**248639 (100%)**

People age _____ are less likely to drop out of high school than other age groups.

People age _____ are more likely to have earned a master's degree or high than other age groups.

Name_____

Example: Are school attendance and math score related?

Random sample of 11 students:

- _____ of school days attended
- _____ of question the correctly answered during the Texas end-of-year Algebra 1 assessment.

Categorical or quantitative variable? _____

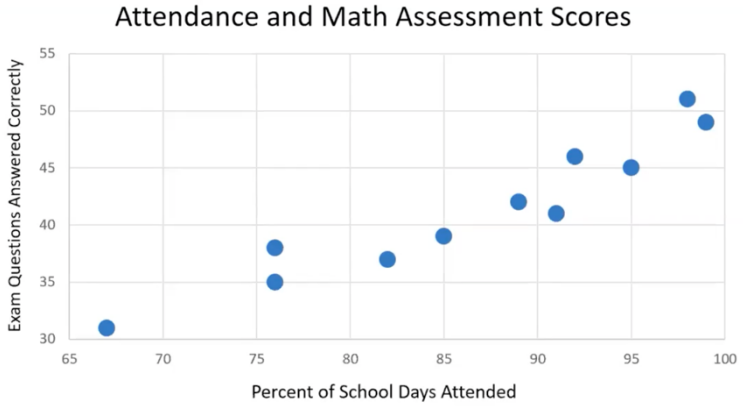
How do we determine if there is a relationship?

_____

Percent Attendance	Questions Correct
95	45
89	42
67	31
98	51
99	49
76	38
92	46
91	41
76	35
85	39
82	37

Graphical representations

This will be represented with a _____.



Numerical representation

Correlation value: **r = 0.95** Equation for a line of best fit: $\hat{y} = -7.69 + 0.56692x$

Coefficient of determination: **r^2 = 0.903**

But even at first glance of the scatterplot above it is clearly evident that,
As the attendance level _____, exam performance tends to _____.

What Should We Take Away?

How will we decide if there is a relationship between two categorical variables?

Graphical representations: _____

Numerical representations: _____, _____

How will we decide if there is a relationship between two quantitative variables?

Graphical representation: _____

Numerical representation: _____, _____ **and** _____.

Name_____

AP Statistics CED 2.2 Daily Video 1 (Skill 4.E)

Representing Two Categorical Variables

What Will We Learn?

How do we construct graphical displays to show the relationship between two categorical variables? How do we use graphical displays to determine if there is an association between two categorical variables?

Examples: Are age group and educational attainment related?

	25 to 34 years old	35 to 54 years old	55+ years old
No high school diploma	3,264	7,982	10,729
High school diploma	19,169	33,082	44,696
Associate's or bachelor's degree	17,468	40,035	42,333
Master's degree or higher	5,311	12,407	12,163

Source: 2019 Census data. Numbers are in thousands.

This table is called a _____ because it shows the two categorical variables.

Make a side-by-side bar graph, segmented bar graph, and mosaic plot. Then, determine if there is an association between age group and educational attainment using the above graphs.

Calculating percents (Complete the two-way table as you watch the video.)

	25 to 34 years old	35 to 54 years old	55+ years old
No high school diploma	3,264	7,982	10,729
High school diploma	19,169	33,082	44,696
Associate's or bachelor's degree	17,468	40,035	42,333
Master's degree or higher	5,311	12,407	12,163
TOTAL			

Start by finding the _____ for each column.

Find the _____ of educational attainment within each group.

For 25-34 years old:

$$\frac{3,264}{45,212}\times 100 = 7.2\% \qquad \frac{19,169}{45,212}\times 100 = 42.4\% \qquad \frac{17,468}{45,212}\times 100 = 38.6\% \qquad \frac{5,311}{45,212}\times 100 = 11.7\%$$

For 35-54 years old: (copy these calculations

_____ _____ _____ _____

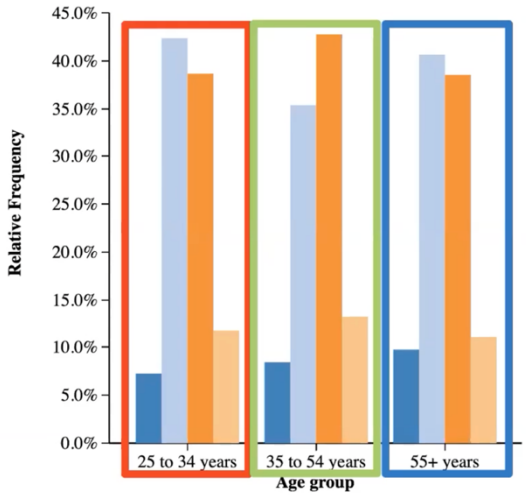
For 55+ years old:

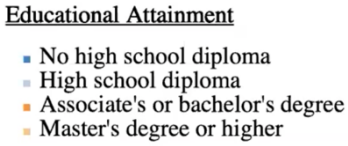
$$\frac{10,729}{109,921}\times 100 = 9.8\% \qquad \frac{44,696}{109,921}\times 100 = 40.7\% \qquad \frac{42,333}{109,921}\times 100 = 38.5\% \qquad \frac{12,163}{109,921}\times 100 = 11.1\%$$

Name_____

Side-by-side bar graph

	25 to 34 years old	35 to 54 years old	55+ years old
No high school diploma	3,264 (7.2%)	7,982 (8.5%)	10,729 (9.8%)
High school diploma	19,169 (42.4%)	33,082 (35.4%)	44,696 (40.7%)
Associate's or bachelor's degree	17,468 (38.6%)	40,035 (42.8%)	42,333 (38.5%)
Master's degree or higher	5,311 (11.7%)	12,407 (13.3%)	12,163 (11.1%)
TOTAL	45,212 (100%)	93,506 (100%)	109,921 (100%)





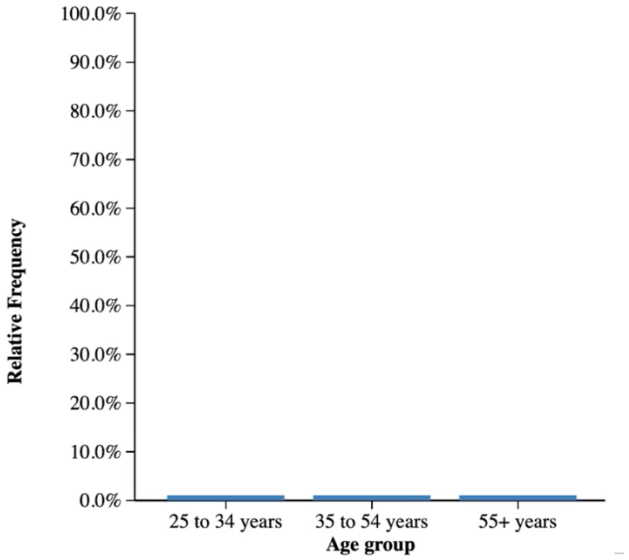
Use the side-by-side bar graph to create a segmented bar graph.

Segmented bar graph (Using the side-by-side bar graphs above sketch the segmented bar graph.)

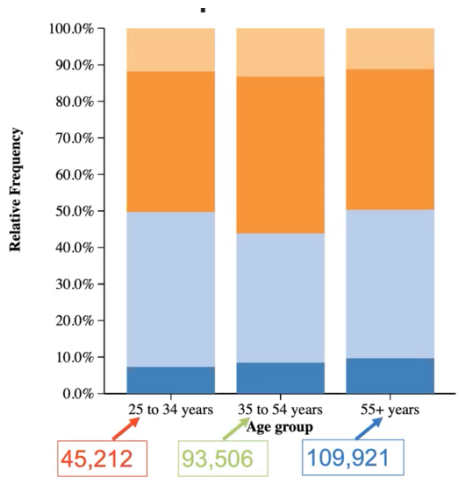
When making the segmented bar graph stack the bars on top of each other. The first bar (7%) should go from 0 – 7%, the second bar (42%) from 7% to (7% + 42%) or 49%, the third bar (39%) from 49% to (49% + 39%) or 88% and the fourth bar (12%) from 88% to *88% +12%) or 100%.

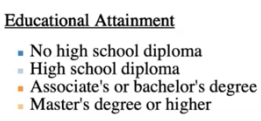
Repeat for each age group.

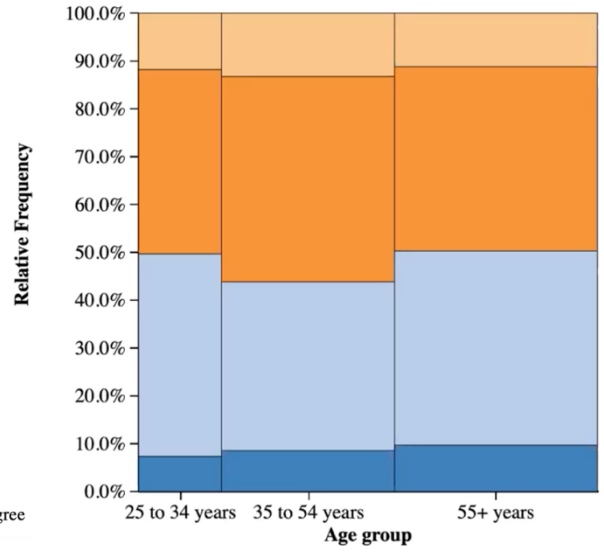
Make sure you understand how to properly "stake" the bars!!



Mosaic Plot







STATS MEDIC

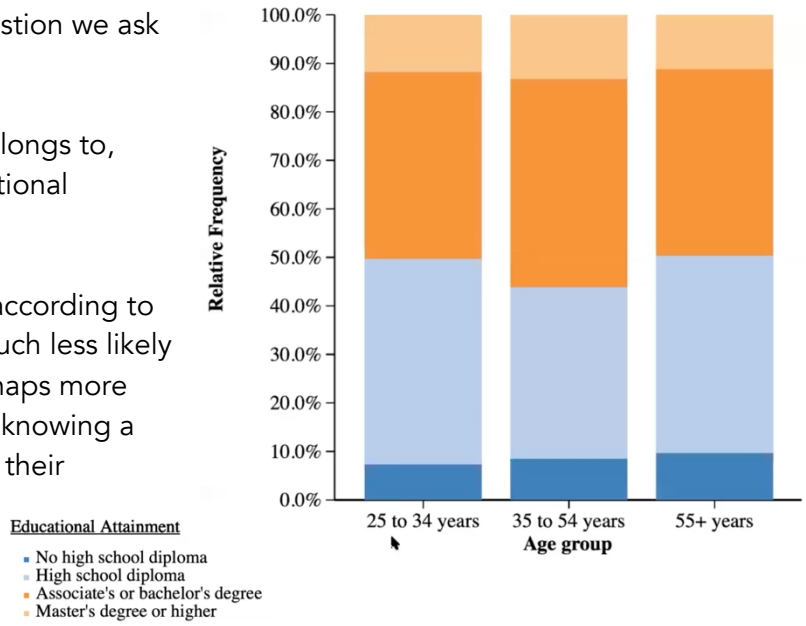
Name_____

Are age group and educational attainment related?

Using the segmented bar graph, the question we ask is,

"If we know what age group a person belongs to, would that help us predicate their educational attainment level?"

Looking at the 25 to 34 year old group, according to the graph, we would see that they are much less likely to have no high school diploma and perhaps more likely to have a high school diploma. So, knowing a person's age group does help us predict their educational attainment level.



Because the _____ of educational attainment is not the _____ for each age group, these two variables are _____.

What Should We Take Away?

How do we construct graphical displays to show the relationship between two categorical variables?

_____, _____, and _____

How do we use graphical displays to determine if there is an association between two categorical variables?

If the _____ are not the same for _____, the there _____ an association between the _____ variables.

Name_____

AP Statistics CED 2.3 Daily Video 1 (Skill 4.E)

Statistics for Two Categorical Variables

What Will We Learn?

How do we calculate summary statistics for two categorical variables?

How do we use summary statistics to determine if there is an association between two categorical variables?

Examples: Are age group and educational attainment related?

Use summary statistics to determine if there is an association between age group and educational attainments?

	25 to 34 years old	35 to 54 years old	55+ years old
No high school diploma	3,264	7,982	10,729
High school diploma	19,169	33,082	44,696
Associate's or bachelor's degree	17,468	40,035	42,333
Master's degree or higher	5,311	12,407	12,163

Source: 2019 Census data. Numbers are in thousands.

Table with row totals and table total.

	25 to 34 years old	35 to 54 years old	55+ years old	TOTAL
No high school diploma	3,264	7,982	10,729	**21,975**
High school diploma	19,169	33,082	44,696	**96,947**
Associate's or bachelor's degree	17,468	40,035	42,333	**99,836**
Master's degree or higher	5,311	12,407	12,163	**29,881**
TOTAL	**45,212**	**93,506**	**109,921**	**248,639**

(Underline and solve as you watch the video!)

What percentage of the people in the survey are 25 to 34 years old with a Master's degree or higher?

Joint relative frequency: A _____ frequency divided by the _____ for the entire table.

Table with row totals and table total.

	25 to 34 years old	35 to 54 years old	55+ years old	TOTAL
No high school diploma	3,264	7,982	10,729	**21,975**
High school diploma	19,169	33,082	44,696	**96,947**
Associate's or bachelor's degree	17,468	40,035	42,333	**99,836**
Master's degree or higher	5,311	12,407	12,163	**29,881**
TOTAL	**45,212**	**93,506**	**109,921**	**248,639**

(Underline and solve as you watch the video!)

What percent of the people in the survey have only a high school diploma?

What percent of people in the survey are 35 to 54 years old?

Marginal relative frequency: _____ and _____ totals in the _____ table divided by the _____ for the entire table.

Name_____

Table with row totals and table total

	25 to 34 years old	35 to 54 years old	55+ years old	TOTAL
No high school diploma	3,264	7,982	10,729	**21,975**
High school diploma	19,169	33,082	44,696	**96,947**
Associate's or bachelor's degree	17,468	40,035	42,333	**99,836**
Master's degree or higher	5,311	12,407	12,163	**29,881**
TOTAL	**45,212**	**93,506**	**109,921**	**248,639**

(Underline and solve as you watch the video!)

What percent of those with only a high school diploma are 35 to 54 years old?

What percent of those 25 to 34 years old have no high school diploma?

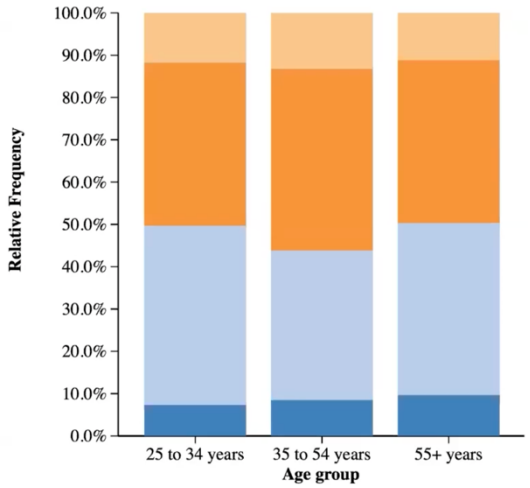
Conditional relative frequency: A relative frequency for a _____ of a two-way table.
*Note: **"specific part"** generally means within _____ row or within _____ column of the two-way table.

More on conditional relative frequency

(Watch the video to review how to calculate distribution for educational attainment within each group.)

Now turn the table to the left into a segmented bar chart.

	25 to 34 years old	35 to 54 years old	55+ years old
No high school diploma	3,264 (7.2%)	7,982 (8.5%)	10,729 (9.8%)
High school diploma	19,169 (42.4%)	33,082 (35.4%)	44,696 (40.7%)
Associate's or bachelor's degree	17,468 (38.6%)	40,035 (42.8%)	42,333 (38.5%)
Master's degree or higher	5,311 (11.7%)	12,407 (13.3%)	12,163 (11.1%)
TOTAL	**45,212** (100%)	**93,506** (100%)	**109,921** (100%)



Interpretation: Because the distribution of _____ is different for each age group, these two variables are _____.

What Should We Take Away?

How do we calculate summary statistics for two categorical variables?

_____, _____, _____

How do we use summary statistics to determine if there is an association between two categorical variables?

If the _____ of conditional relative frequencies are _____ the same for _____ groups, then there _____ an _____ between the two variables.

STATS MEDIC

Name_____

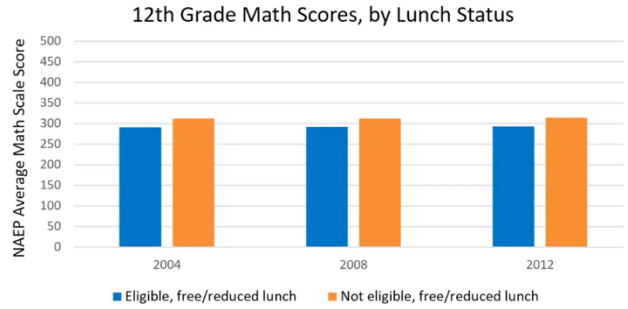
AP Statistics CED 2.4 Daily Video 1 (Skill 2.B)

Relationship Between Two Quantitative Variables

What Will We Learn?

For bivariate data, how do we determine which variable is the explanatory variable?
For bivariate data, how do we determine which variable is the response variable?
How do we construct a scatterplot?

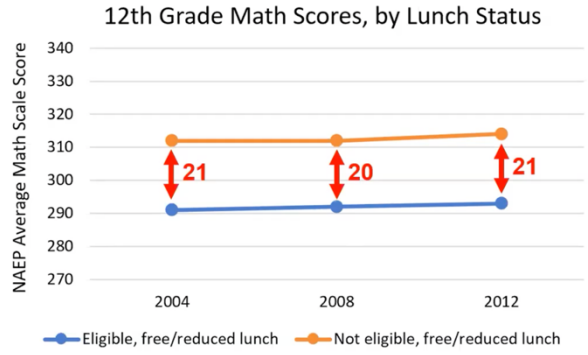
Example: The Income Achievement Gap



Source: NAEP (National Assessment of Education Progress) Long-Term Trend Mathematics scores. According to NAEP, differences are statistically significant. Data taken from nces.ed.gov

Nationally, high/middle income students tend to perform better on math exams than low-income students, on average.

Lesson adapted from skewthescript.org



Source: NAEP (National Assessment of Education Progress) Long-Term Trend Mathematics scores. Image scale based on NAEP online chart tool. According to NAEP, differences are statistically significant. Data taken from nces.ed.gov

Zooming in on these gaps…

These so-called "achievement gaps" have stayed consistent over multiple years.

***Note:**
This data says nothing about individual performance (it's merely a trend of averages).

This data says nothing about innate "intelligence".

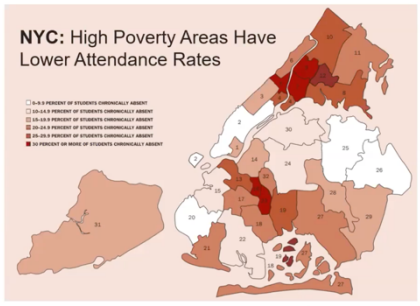
Schools and "Equal Opportunity"

Illustration by Ameena Golding, *The Daily Californian*, 2018

Middle/upper income student have wealth privilege. Often, they don't face as many educational barriers as their lower income peers.

Can education systems equalize opportunities for lower income students?

The Income Attendance Gap



Graphic from Nauer et al., "A Better Picture of Poverty," *Center for New York City Affairs*, Nov. 2014: https://www.attendanceworks.org/wp-content/uploads/2017/06/BetterPictureofPoverty_PA_FINAL_001.pdf

Nationally, higher income areas tend to have fewer chronically absent student. Possible reasons:

- Transportation Issues
- Work to support family

Is attendance the solution?

Poverty
↓
High ~~Low~~ Attendance
↓
High ~~Low~~ Scores

Some school systems have targeted attendance as the key to raining test scores for lower income students.

STATS MEDIC

Name_____

Let's look at the data

Random sample of 11 students:

- Percent of school days attended
- Number of questions they correctly answered during the Texas end-of-year Algebra 1 assessment.

Explanatory (x): _____

Response (y): _____

Percent Attendance	Questions Correct
95	45
89	42
67	31
98	51
99	49
76	38
92	46
91	41
76	35
85	39
82	37

Let's visualize it!

Hard to see _____ from raw data.
We should make a plot.

1. One or two variables?

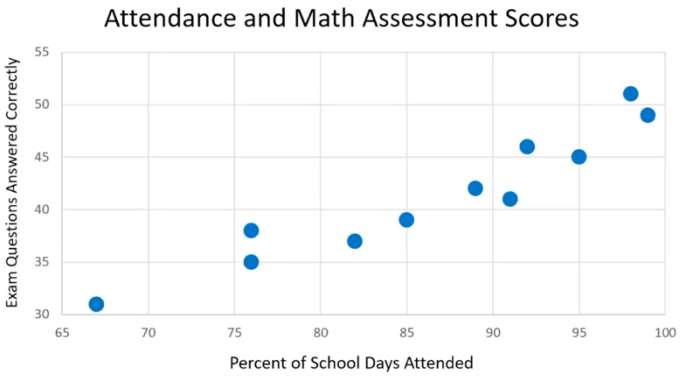
2. Categorical or quantitative data?

explains → responds

(x) (y)

Percent Attendance	Questions Correct
95	45
89	42
67	31
98	51
99	49
76	38
92	46
91	41
76	35
85	39
82	37

Constructing a Scatterplot



Each individual has and _____ (attendance) and a _____ (questions correct).

* Graph has a _____
* Axes, are _____ (with units, if applicable)
* Scales are shown with _____

(Be sure to highlight scatterplot as you watch the video.)

To think about: How would you describe trends you see in this scatterplot??

What Should We Take Away?

Explanatory variables _____ or "explain" trends in the _____ variable.

Scatterplots _____ trends between two _____ variables.

When making a scatterplot, include a _____, _____ (with units), and properly show _____,

Name_____

AP Statistics CED 2.4 Daily Video 2 (Skill 2.A)

Relationships Between Two Quantitative Variables

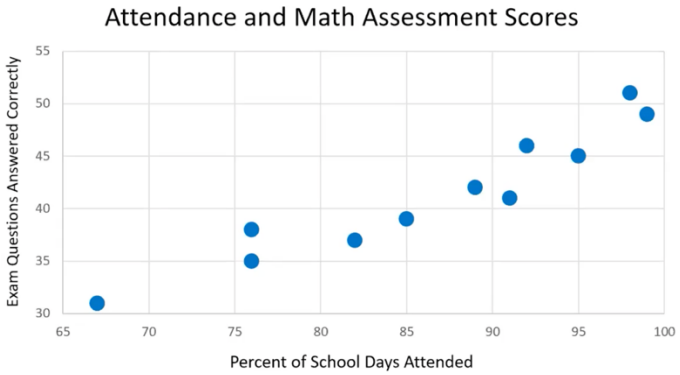
What Will We Learn?

How do we describe direction in scatterplot?
How do we describe form and unusual features in a scatterplot?
How do we describe strength in a scatterplot?

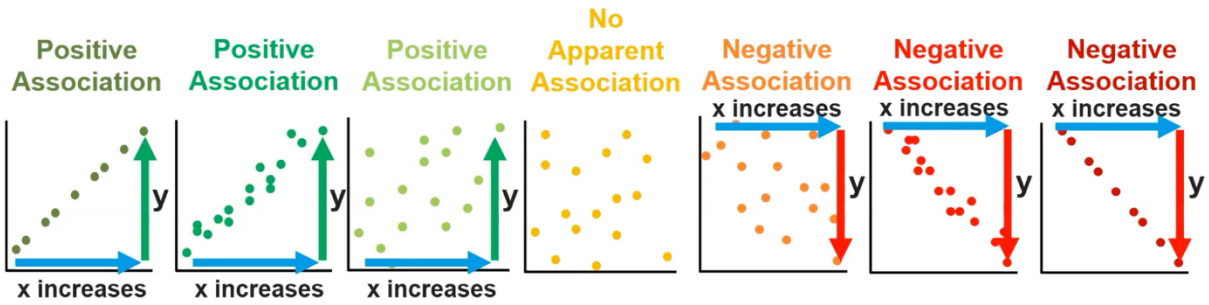
Example: Is Attendance the Solution?

We collected data from a random sample of 11 students. We found the percent of school days they attended and the number of questions they answered correctly on the end-of-year Texas Algebra 1 assessment.

Take a pause: How would you describe any trends you see in this scatterplot?



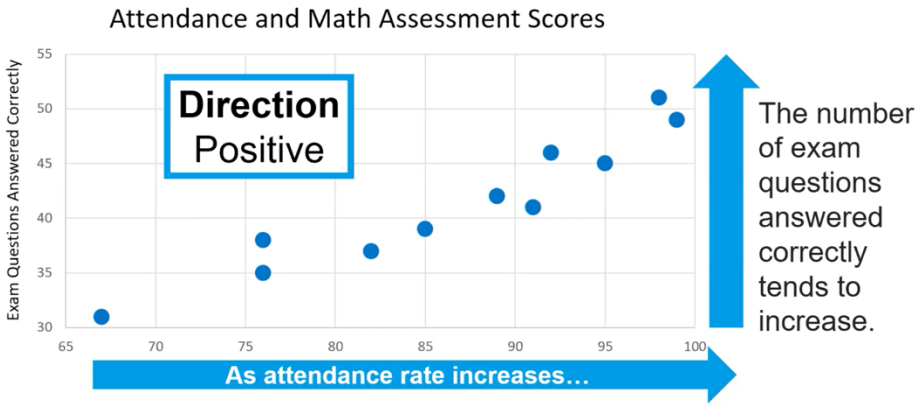
Describing Direction



Directions

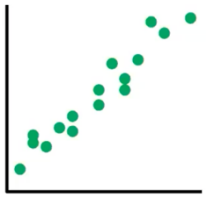
Positive: as x values _____, the y values also tend to _____.
Negative: as x values _____, the y values tend to _____.

Direction

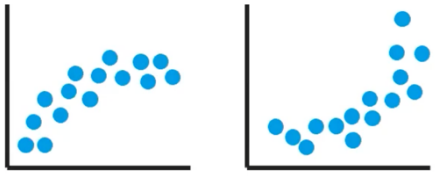


Name_____

Form and Unusual Feature



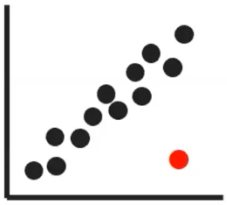
Form: _____



Form: _____

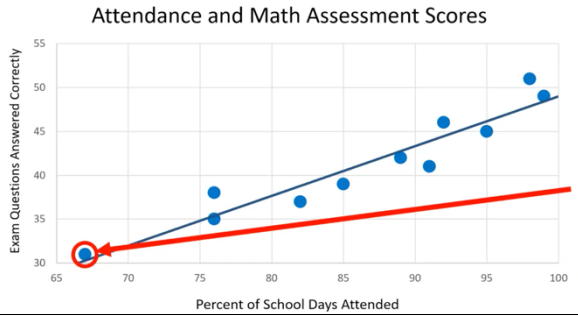
Unusual Features: _____

Unusual Feature: _____



Unusual Feature: _____

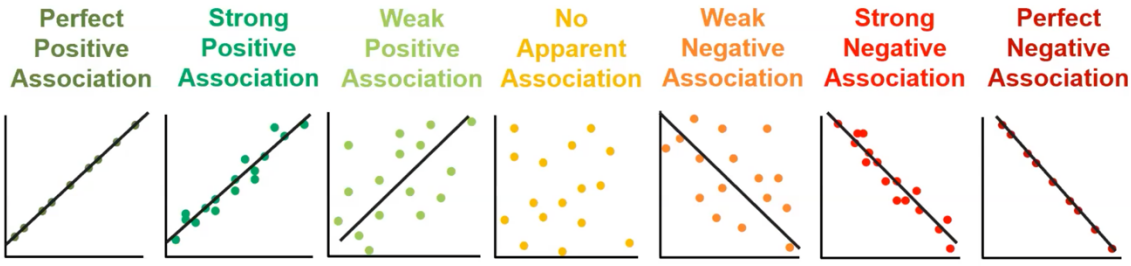
Form and Unusual Features



Form: **Linear**

Unusual: **one student had unusually low attendance.**

Describing Strength



Strong: data _____ the pattern (e.g. linear)

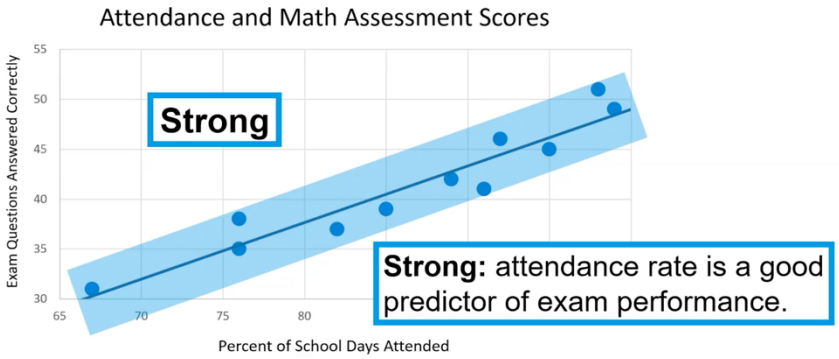
* In the above cases, if you used the linear model to _____ new data, you would tend to make _____.

Weak: data _____ the patter (e.g. linear)

*. In the above cases, if you used a linear model to _____ new data, you may be _____.

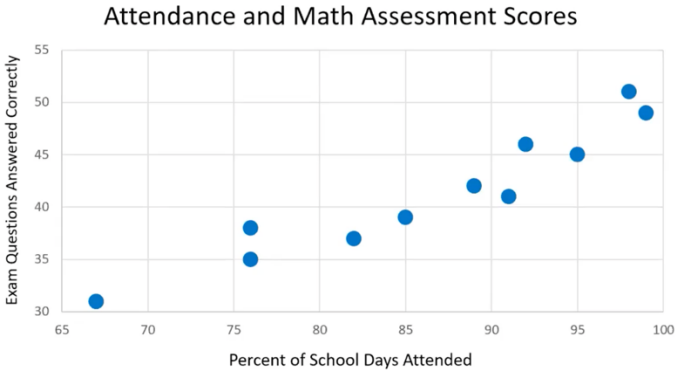
Name_____

Strength



Putting it all together

Describe the relationship between attendance rate and exam performance:



Direction: _____

Form: _____

Strength: _____

Unusual Features: _____

Context: _____

The relationship between _____ and _____ appears to be _____, _____ and _____. There appears to be one student with _____.

Lingering Question:

Given the strong and positive relationship, if a school starts a new policy that raises attendance, what do you think will happen to test scores? Explain.

What Should We Take Away?

_____ describes the overall trend in the _____ as the _____ increase.

_____ can be categorized as _____ or _____.

_____ describes how closely the data follow a _____.

_____ include _____ and apparent _____.

Name_____

AP Statistics CED 2.5 Daily Video 1 (Skill 2.C)

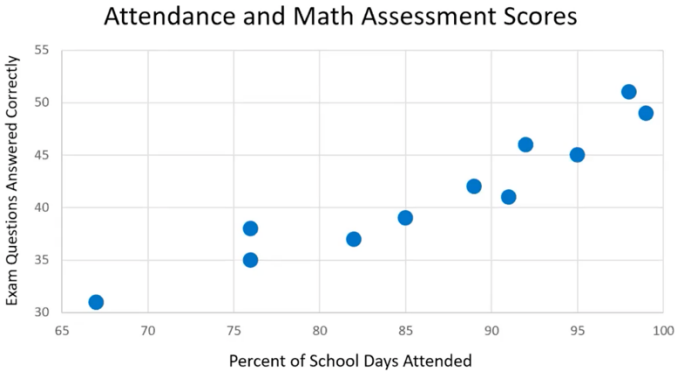
Correlation

What Will We Learn?

How do we calculate the correlation (r)?
How do we interpret the correlation (r)?
What properties of a scatterplot can we learn from the r-value alone?

Example: Is Attendance the Solution? The Data Looks Promising!

We collected data from a random sample of 11 students. We found the percent of school days they attended and the number of questions they answered correctly on the end-of-year Texas Algebra 1 assessment.



Strong, positive linear relationship.

As attendance rates rise, exam performance also tends to rise!

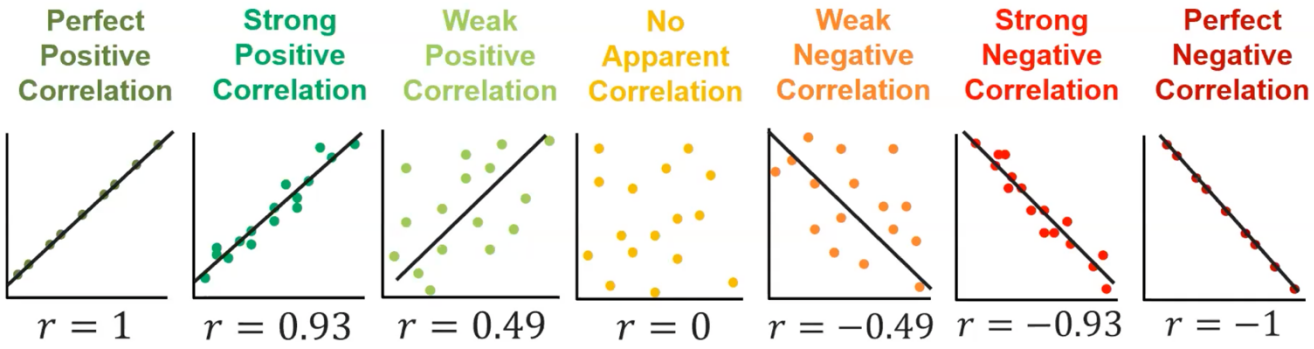
Can we quantify how strongly attendance predicts exam performance?

Lesson adapted from skewthescript.org

The Correlation Coefficient (r)

- Gives the _____ and _____ the strength of a _____ relationship.
- Fancy Formula: $r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
- Technology does the calculations of this formula for us.

Correlation (r): **r-values span between ____ ≤ r ≤ ____**



Take a pause: Do you see any pattern(s) in these r-values?

Direction:

Negative r-value → _____ correlation
Positive r-value → _____ correlation

Strength:

r closer to 0 → _____ correlation
r closer to -1, 1 → _____ correlation

STATS MEDIC

Name_____

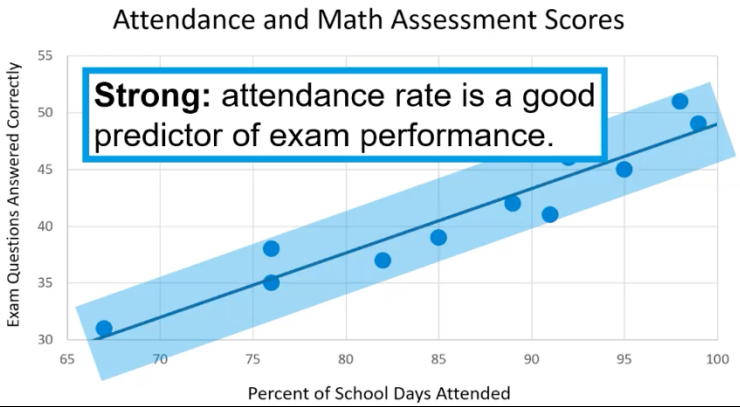
Correlation in Our Context

Attendance rate is a good predictor of exam performance, but how strong is the correlation.

Using a calculator, you would find that:

r = _____

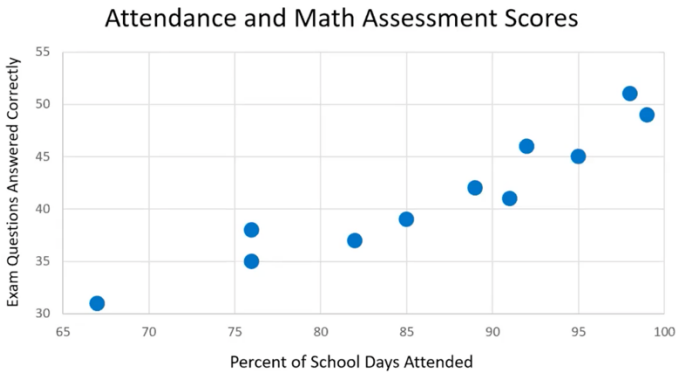
This is a very strong, positive relationship.



Question from last time….

Describe the relationship between attendance rate and exam performance:

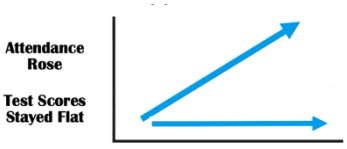
From last video: Given the _____, _____, and _____ relationship, if a school starts a new policy that raises attendance, what do you think will happen to test scores? Explain.



Here's what they did

In the past several years, superintendents have piloted large-scale (and sometimes quite expensive) initiatives to improve student attendance. These included:

1. Call programs for chronically absent students
2. Hiring attendance case managers and coordinators
3. Using Uber/Lyft for students with transportation issues



WHAAAAAT????

For next video:

Many datasets with thousands of student data points show a strong, positive correlation between attendance rate and exam performance. Given this fact, how could a new attendance initiative still not succeed in boosting test scores?

What Should We Take Away?

The _____ of the correlation coefficient (r) tell you the _____ of the linear relationship.

The _____ of the correlation coefficient (r) quantifies the _____ of the linear relationship

The correlation coefficient (r) _____ does not provide enough information to make claims about _____ or _____ in the relationship

Name_____

AP Statistics CED 2.5 Daily Video 2 (Skill 4.B)

Correlation

What Will We Learn?

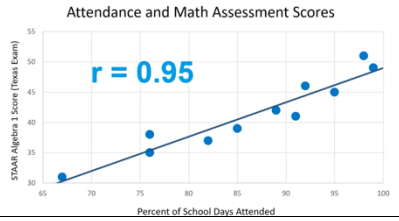
What is the difference between correlation and causation?

How do we factor alternative explanations into our data analysis?

What kinds of inferences can we make from data that show a strong correlation?

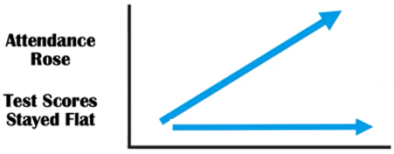
Watch the review of The Income Achievement Gap. Is attendance the solution?

LOOK: ATTENDANCE HELPS!



Positive, linear, and very strong relationship

Here's what happened



After piloting initiative to improve attendance we found attendance rose but test scores remained flat.

One possible explanation

Poverty → High ~~Low~~ Attendance → High ~~Low~~ Scores

What if instead →

Poverty → High ~~Low~~ Attendance → Low Scores (Cause)

-Hunger
-Worse Schools
-Less Study Time (job, taking care of family members)

One of our fundamental concepts:

Correlation ≠ Causation

- Just because attendance and test performance _____, it doesn't mean they have a _____ relationship.

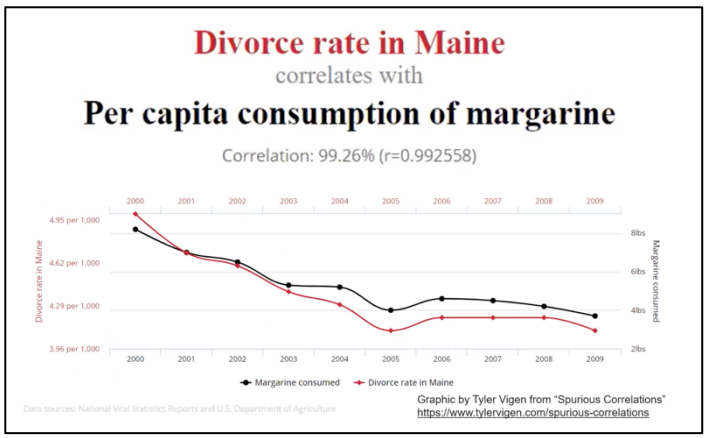
If it is not attendance, then….

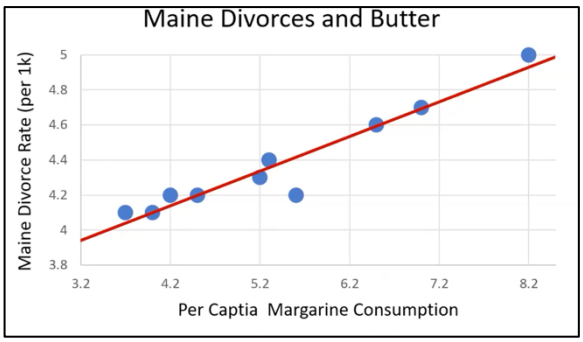
- How can we identify the true causal mechanisms that will help us fight educational inequity?

The Causation Fallacies

- Sometime there are _____ causal variables _____ in the background (last example).
- Other times, correlations are completely _____!

Example:





High Correlation ≠ Causation!

STATS MEDIC

Name_____

What Should We Take Away?

Correlation ≠ Causation

Beware of other variables or coincidental correlations!

Think about what **you** can do to investigate and, ultimately, fight educational inequity.

Name_____

AP Statistics CED 2.6 Daily Video 1 (Skill 2.C)

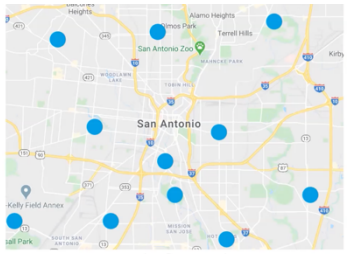
Linear Regression Models

What Will We Learn?

How do we construct a linear regression model?
How do we make predictions using a linear model?
How can we gauge the reliability of our predictions?

Food Access: Is neighborhood a good predictor of access to healthy foods?

Supermarket locations in San Antonio, TX.



Former student, Linda Saucedo, noticed that her local supermarket offered fewer organic fruits/vegetables than another location from the same company in a wealthier part of town. She wondered if this is a broader pattern throughout the city?

The Data

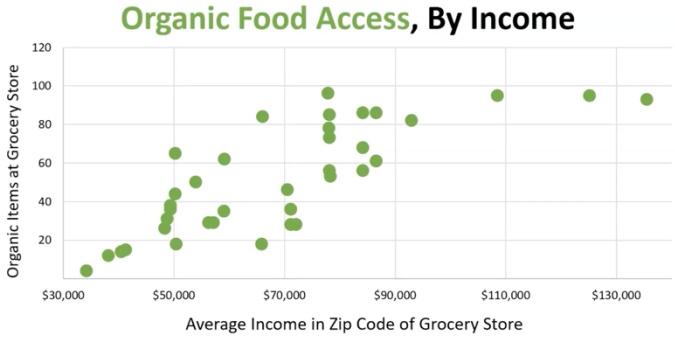
Zip Code	Mean Household Income	Number of Organic Vegetables Offered
78204	$71,186	36
78207	$34,234	4
78204	$71,186	28
78201	$48,760	31
78212	$78,096	78
78202	$40,506	14
78237	$38,166	12
78228	$50,398	18
...	...	...

Data collected in November, 2019

Linda's Dataset

n = 37
Income data from census aggregator
Food data from stores' listings

Scatterplot of Linda's Dataset



Explanatory Variable: _____
Response Variable: _____

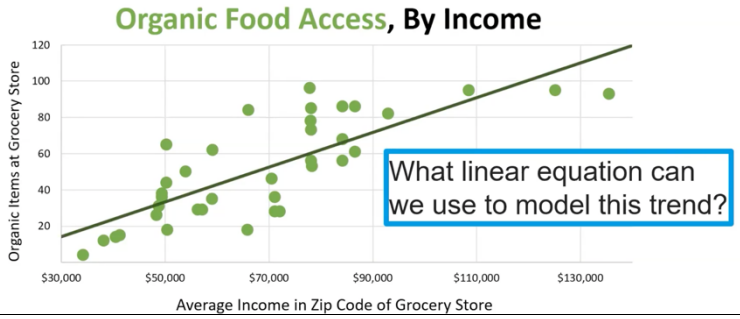
Linear Model

Algebra: Linear Equation

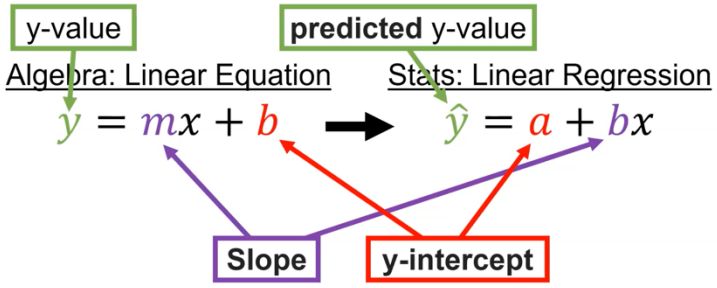
$$y = mx + b$$

Stats: Linear Regression

$$\hat{y} = a + bx$$



Linear Regression Model



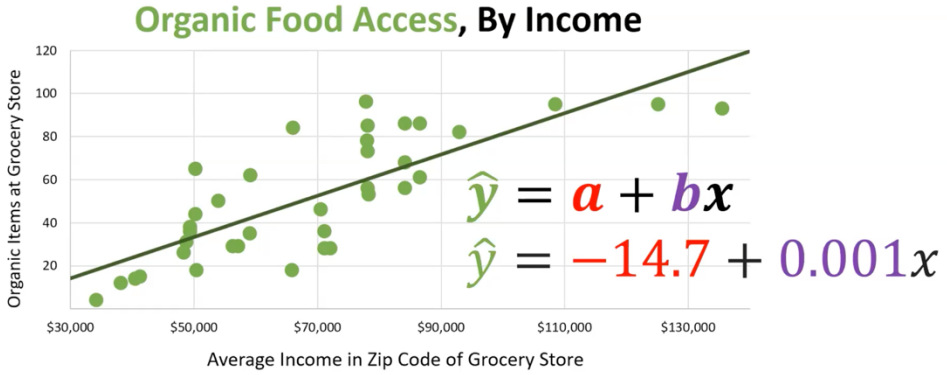
STATS MEDIC

Name_____

Linear Regression Model

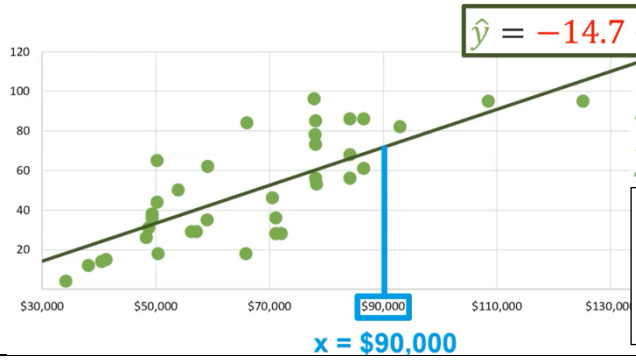
y-intercept =_____

slope = _____



Making Predictions

You move to a zip code in San Antonio in which the average household income is $90,000. Use the linear regression model to predict the average number of organic vegetable items offered at supermarkets in your zip code. (Calculate the predicted value as you watch the video.)



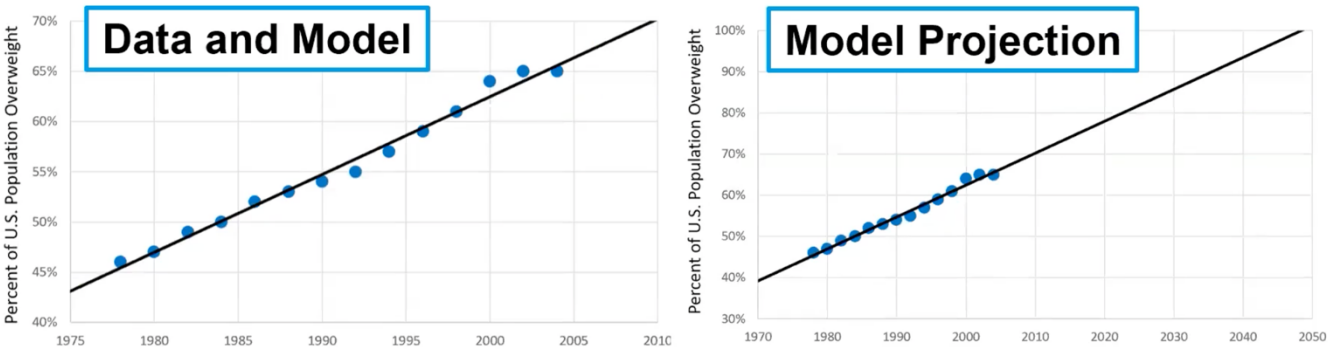
$\hat{y}$ = _____

$\hat{y}$ = _____

How could we get a decimal value?
75.3 is the _____ number of organic items offered per store in zip codes of _____ wealth. This is not the actual offering at a specific store. (not a data value).

Lingering Question…

For next video: Using linear regression, a widely-cited study* concluded that by 2048, if trends continue, **100% if Americans would be overweight.** Using the graphs below, do you believe 100% of Americans will be overweight by 2048? Why or why not?



*Wang, Beydoun, et al., "Will all Americans become overweight or obese? estimating the progression and cost of the US obesity epidemic." *Obesity* (Silver Spring). 2008;16(10):2323-2330. doi:10.1038/oby.2008.351. Graphs provided are representative approximations of analyses from the paper

Example inspired by Ellenberg, J. *How Not to be Wrong*, pg. 50-61

What Should We Take Away?

The linear regression model is composed of a _____ and a _____.

We can use the linear regression model to make _____ about the response variable.

Predictions made by the linear regression model do not represent _____ data values.

Name_____

AP Statistics CED 2.6 Daily Video 2 (Skill 2.C)

Linear Regression Models

What Will We Learn?

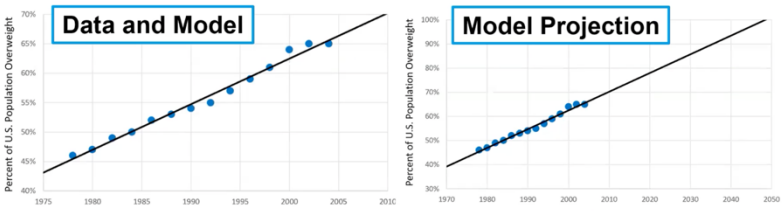
Why is it dangerous to extrapolate using linear models?

How can we use the linear regression tools we've learned so far to answer AP Exam questions?

What strategies can we employ to earn maximum credit on AP Exam free-response questions?

Question from last video…

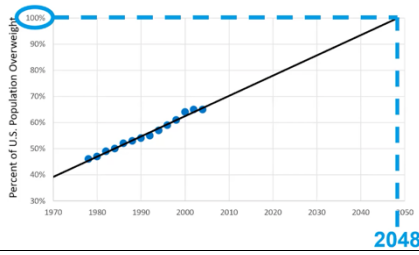
Using linear regression, a widely-cited study* concluded that 2048, if trends continue, **100% of Americans would be overweight.** Using the graphs below, do you believe 100% of Americans will be overweight by 2048? Why or why not?



*Wang, Beydoun, et al., "Will all Americans become overweight or obese? estimating the progression and cost of the US obesity epidemic." *Obesity* (Silver Spring). 2008;16(10):2323-2330. doi:10.1038/oby.2008.351. Graphs provided are representative approximations of analyses from the paper

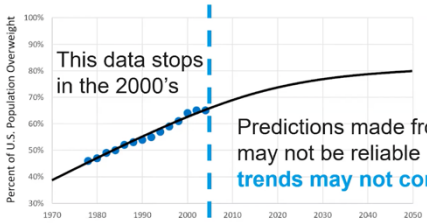
Example inspired by Ellenberg, J. *How Not to be Wrong*, pg. 50-61

Extrapolation



According to this model: yes, **all** Americans are predicted to be overweight by 2048

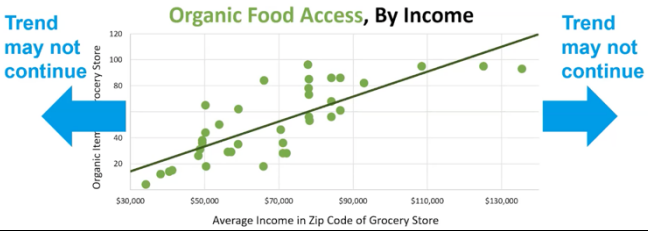
Extrapolation is dangerous…



Predictions made from extrapolation may not be reliable because **current trends may not continue.**

Important Note

_____ is also dangerous when working with variables that aren't time-related.



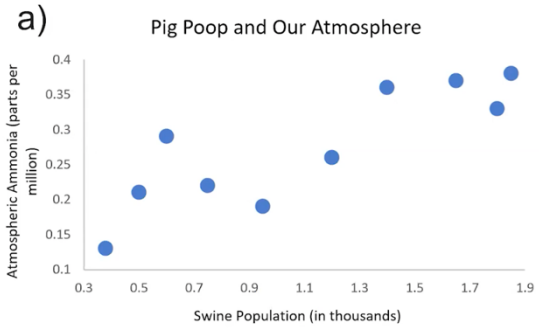
Let's Practice: Free-Response Question 2002, Form B, Question #1, a

Animal-waste lagoons and spray fields near aquatic environments may significantly degrade water quality and engender health. The National Atmospheric Deposition Program has monitored the atmospheric ammonia at swine farms since 1978. The data on the swine population size (in thousands) and atmospheric ammonia (in parts per million) for one decade are given here.

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
Swine Population	0.38	0.50	0.60	0.75	0.95	1.20	1.40	1.65	1.80	1.85
Atmospheric Ammonia	0.13	0.21	0.29	0.22	0.19	0.26	0.36	0.37	0.33	0.38

a) Construct a scatterplot for these data:
(Annotate the problem as you watch the video.)

a)



Identify:

Explanatory (x): _____

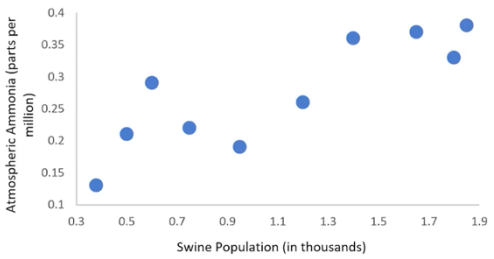
Response (y): _____

STATS MEDIC

Name_____

Scoring

a)



E = Essentially Correct - scatterplot has _____ points, includes _____ and _____.

P = Partially Correct – scatterplot has _____ points, but only one of _____ or _____ is included.

I = Incorrect – Any of these: Neither axis labels or scale is included OR not a scatterplot OR has missing points.

Let's Practice: Free-Response Question 2002, Form B, Question #1, b

b) The value for the correlation coefficient for these data is 0.85. Interpret this value.

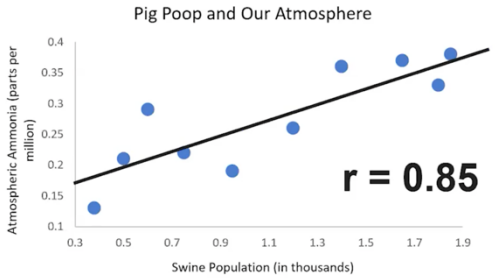
There is a _____, _____, _____ relationship between _____ population size and _____ ammonia concentration.

Here scorers are looking for three things: i) _____, ii) _____, iii)_____

Scoring: **E –** 3/3; **P –** 2/3; **I -** less

Let's Practice: Free-Response Question 2002, Form B, Question #1, c

c) Based on the scatterplot and the value of the correlation coefficient, does it appear that the amount of atmospheric ammonia is linearly related to the swine population size?



Because the data in the scatterplot appear to follow an approximately _____ pattern and the magnitude of the _____ (which describes the _____ of the linear relationship) is relatively _____ (0.85), the relationship appears to be _____.

Looking for: i) Correct comment, based on scatterplot; ii) Correct comment, based on r-value.

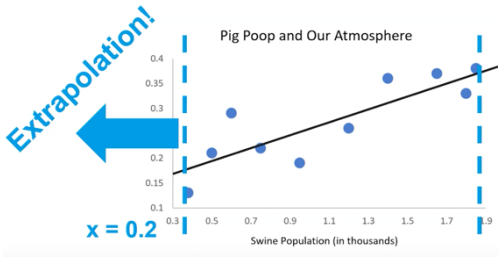
Scoring: **E –** 2/2; **P –** 1/2; **I –** 0/2

Modified from original FRQ, d

d) A scientist constructs a linear regression model for this relationship: $\hat{y} = 0.127 + 1.33x$, where $\hat{y}$ is the predicted atmospheric ammonia concentration and x is the swine population size. Predict the atmospheric ammonia concentration is the swine population size is 200 and comment on whether the prediction is reliable. (Recall: Swine population size is in thousands. (200 is 0.2 of a thousand).

$\hat{y} = 0.127 + 1.33x = 0.127 + 1.33(0.2) =$ _____

Because 0.2 falls outside of the data, this is _____!!



The predication is _____ due to extrapolation: the prediction was made _____ the interval of _____ data x-values.

Trends seen in the scatterplot _____ at this new x-value.

Looking for: i) correct prediction; ii) showing plugging into formula; iii) state unreliable: extrapolation

Scoring: **E –** 3/3; **P –** 2/3; **I -** less

Name_____

Free-Response Question: Score yourself!
Note: total scores between whole values (e.g. 2.5 points) are rounded up or down based on a holistic approach.

E's = 1 point each
P's = ½ point each
I's = 0 points each

___ / 4

Note: AP Exams are graded on their own scale! For example, scoring 2.4 isn't an "F". Actually, depending on the problem, 2/4 may be a pretty solid score!

What Should We Take Away?

_____ is dangerous because trends may not continue.

When tackling an FRQ, _____ the question, include _____ and show _____ work.

Practice makes _____ (there's no such thing a '_____').

Name_____

AP Statistics CED 2.7 Daily Video 1 (Skill 2.B)

Residuals

What Will We Learn?

How do we calculate a residual?

How do we interpret a residual?

How do we construct a residual plot?

Example: Food Access (Watch as the video reviews this topic.)

Supermarket locations in San Antonio, TX.

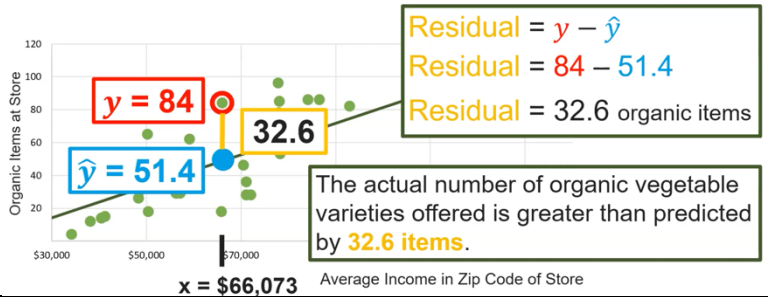
Former student, Linda Saucedo, noticed that her local supermarket offered fewer organic fruits/vegetables than another location from the same company in a wealthier part of town.

Residuals

Residual: The difference between the _____ response value and the model's _____ response value.

Residual = _____ - _____; Residual = $y - \hat{y}$

Residuals



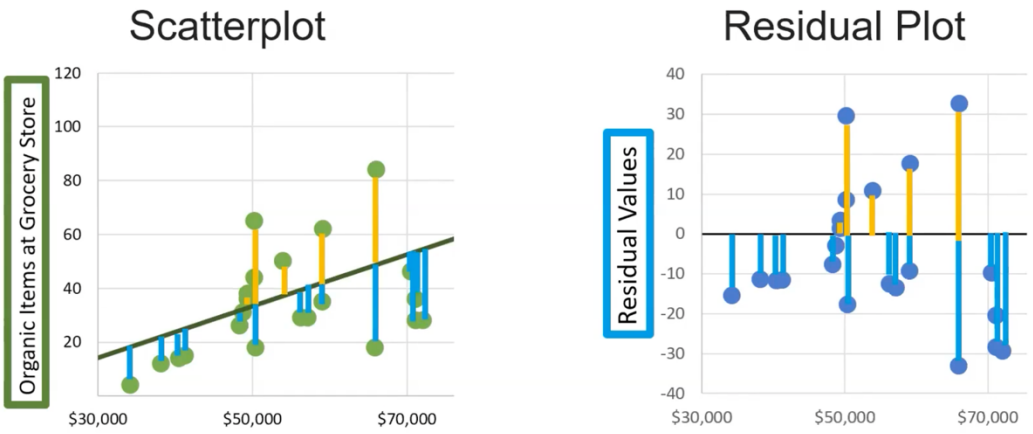
Residuals

$$\text{Residual} = y - \hat{y}$$
(actual) (predicted)

Residual is positive: model _____ the actual response value.

Residual is negative: model _____ the actual response value.

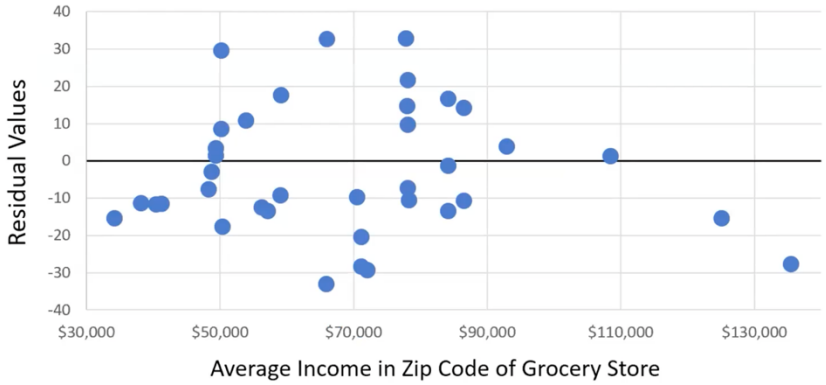
All Residuals - (Watch the video as the residual plot is created.)



STATS MEDIC

Name_____

Residual Plot



What's the point?

A residual plot:

_____ and _____ the residuals, allowing us to assess our _____ fit.

Lingering Question…

For the next video: What does the residual plot indicate about the fit of our model?

What Should We Take Away?

_____ measure the _____ between _____ and _____ response values.

_____ residual values indicate model _____. Negatives indicate _____.

Residual plots can be used to _____ on a model's _____ and assess _____.

Name_____

AP Statistics CED 2.7 Daily Video 2 (Skill 2.A)

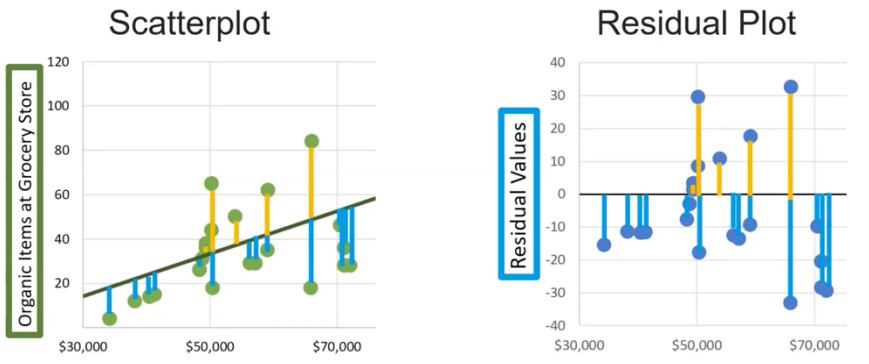
Residuals

What Will We Learn?

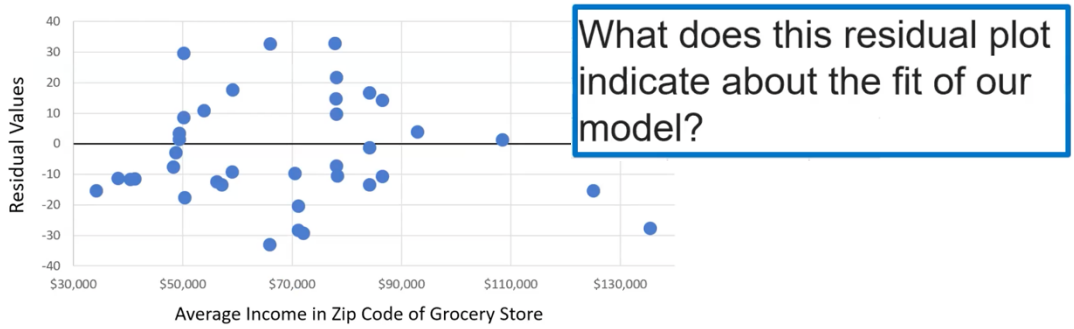
If a linear model is a good fit, what will the residual plot look like?
If a linear model is a bad fit, what might the residual plot look like?
What do we do when we encounter a bad fit?

Residual Plot

The way that the residual plot differs from a scatterplot is that instead of the _____ y-values on the y-axis we chart the _____ values on the y-axis.



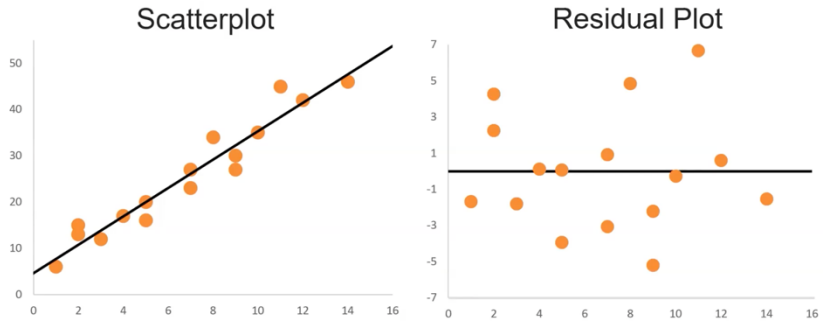
From video… **last**



A Good Fit

- Apparent _____
- Centered at _____
- No clear _____

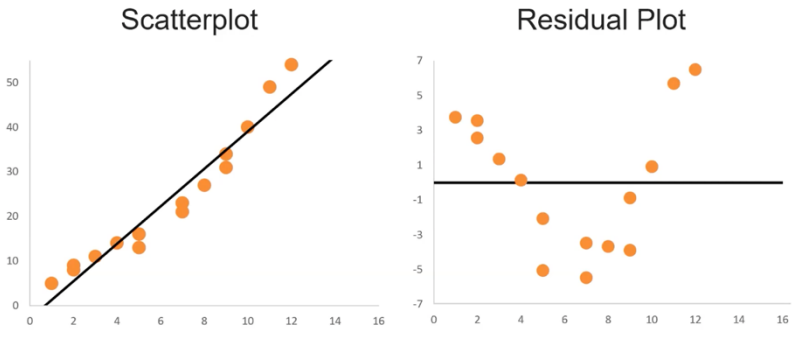
Our model captured the linear _____ of the data, without capturing the random residual _____ around it.



A Bad Fit

- Curved pattern

These residuals are _____ random noise. There is a _____ that our model failed to capture. Linear model may _____ be the best fit.



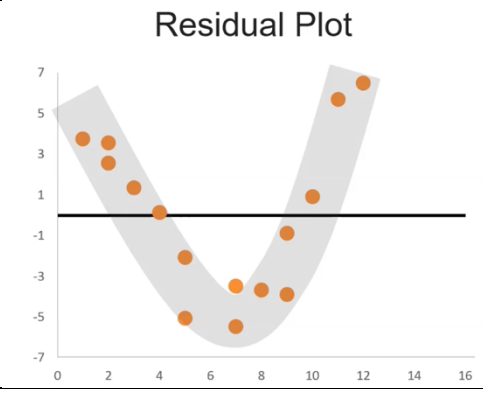
STATS MEDIC

Name_____

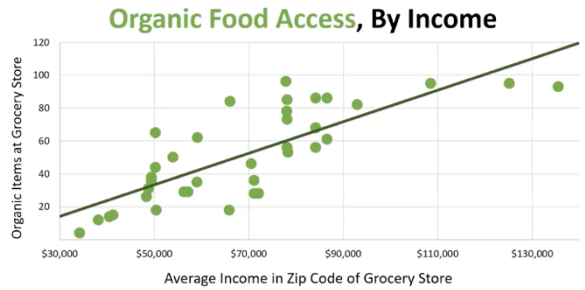
A Bad Fit

- Curved pattern

Residual plots _____ possible trends in residuals, allowing us to _____ assess model fit.



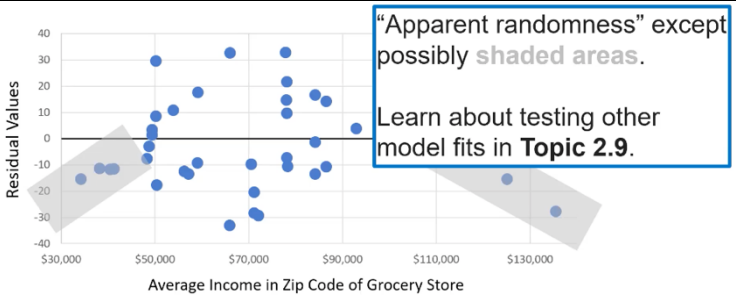
Original Data



Residual Plot

Residual Plot

We will explore how to make more complex models in Topic 2.9.



What Should We Take Away?

Residual plots _____ on residuals, allowing us to better assess _____.

"_____," centering at _____, is a good sign for model fit.

_____ among residual values tend to be _____ for linear model fit.

Name_____

AP Statistics CED 2.8 Daily Video 1 (Skill 2.C)

Least Squares Regression

What Will We Learn?

How do we determine the least-squares regression line?

What are the important properties of the least-squares regression line?

How can we determine the slope of the least-squares regression line using the correlation?

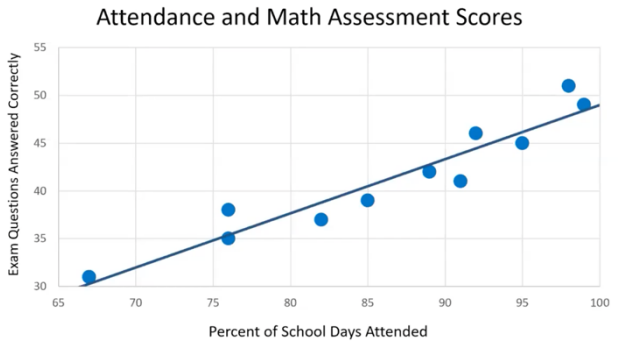
Schools and "Equal Opportunity" – Watch as the video reviews this context.

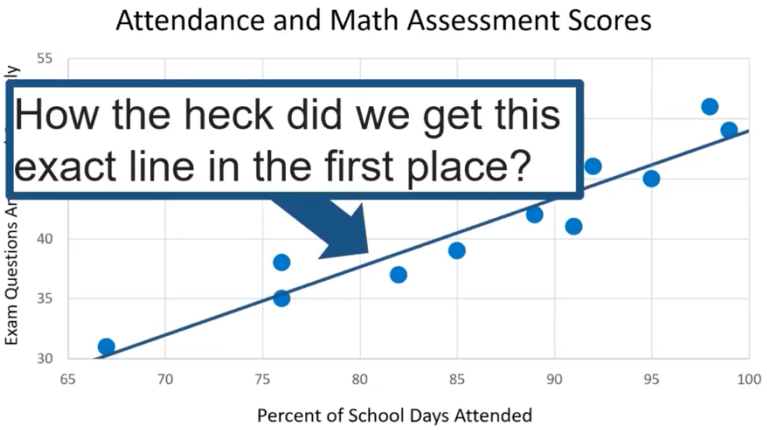
Why focus on attendance?

Random sample of _____ in Texas

x: _____ of school days attended

y: _____ of questions answered correctly on state Algebra 1 test

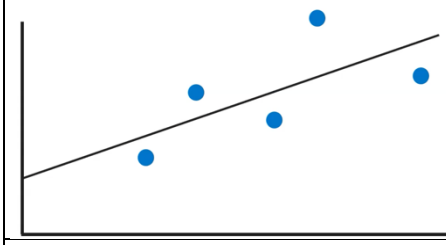


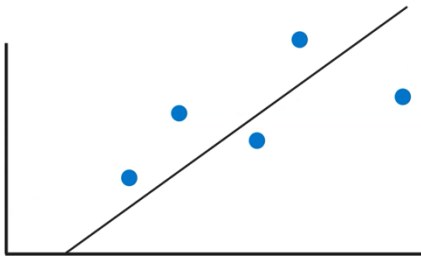


Many datasets, like this one, show a _____, _____, _____ relationship between attendance rates and exam performance.

But how the heck did we get this exact line in the first place????

Finding the best line

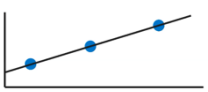




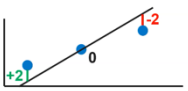
Which model is better?

A look at the residuals may help.

First thought: The line of best fit is the line that minimizes the sum of the residuals. BUT…



Sum of residuals: 0 + 0 + 0 = 0
Good fit.



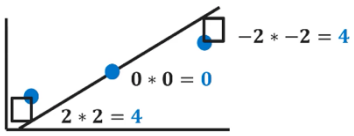
Sum of residuals: 2 + 0 + (-2) = 0
Good fit?!

How can we get rid of negative residuals???
SQUARE IT!!

STATS MEDIC

Name_____

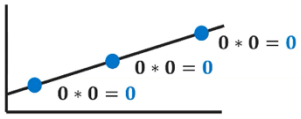
Finding the Best Line!



Sum of squared residuals:

_____ = _____

Not a great fit!



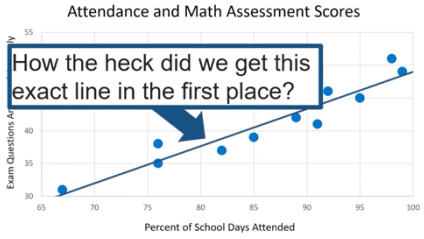
Sum of squared residuals:

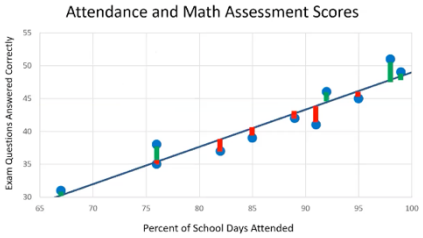
_____ = _____

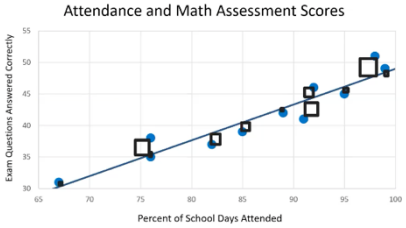
Not a great fit!

Least-Square Regression

Lease Squares Regression Line (LSRL): a _____ model that _____ the sum of the _____.







This was the line that _____ the sum of the _____.

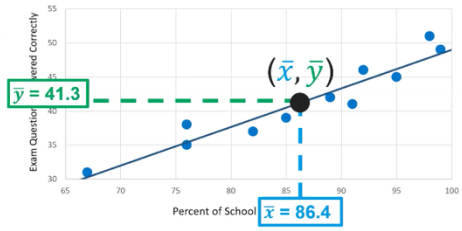
Technology helps us get the exact equation!!

A couple of LSRL properties…

1. **The LSRL contains the MEAN** $(\bar{x}, \bar{y})$ p

$\bar{x}$ = _____ attendance

$\bar{y}$ = _____ correct answers



$(\bar{x}, \bar{y})$

(86.4, 41.3)

explains → responds

(x) (y)

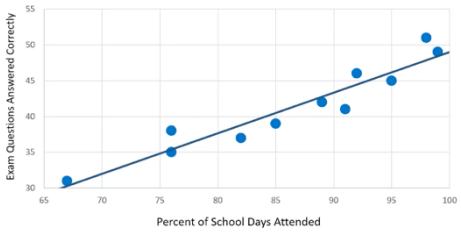
Percent Attendance	Questions Correct
95	45
89	42
67	31
98	51
99	49
76	38
92	46
91	41
76	35
85	39
82	37

Lesson adapted from skewthescript.org 48

2. **Slope and Correlation** $b = r\left(\frac{s_y}{s_x}\right)$

$b = r\left(\frac{s_y}{s_x}\right)$

Slope | Correlation | Standard Deviation of y values | Standard Deviation of x values



Given:

r = 0.95

s_y = 6.08

s_x = 10.2

Find: Slope

(Calculate the slope)

What Should We Take Away?

The LSRL is the _____ model that _____ the sum of the _____ residuals.

The LSRL contains the MEAN point (_____).

The _____ can be determined from the _____ and the standard _____ of _____ variables.

STATS MEDIC

Name_____

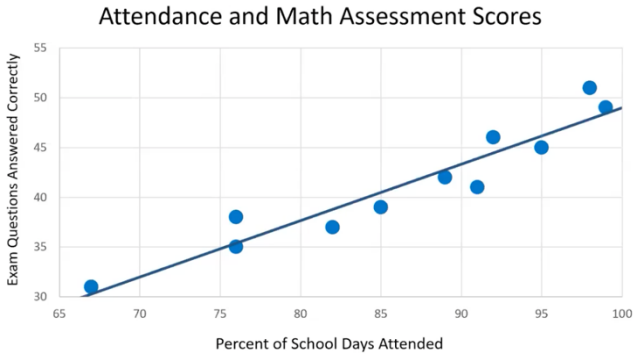
AP Statistics CED 2.8 Daily Video 2 (Skill 4.B)

Least Squares Regression

What Will We Learn?

How do we interpret the slope of the LSRL?
How do we interpret the y-intercept of the LSRL?
How do we determine if the y-intercept has a meaningful interpretation in context?

The LSRL Equation



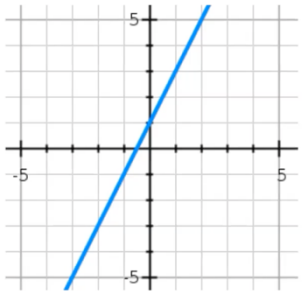
Random sample of 11 student in Texas.

x: percent of school days attended
y: number of questions answered correctly on state Algebra 1 test.

Our Model: $\hat{y} = -7.69 + 0.57x$ where,
$\hat{y}$: Predicted number of questions correct
x : Percent of school days attended

Algebra's Linear Equation

Slope intercept form of a line: $\boldsymbol{y = mx + b}$ for our examples let's look at: $y = 2x + 1$



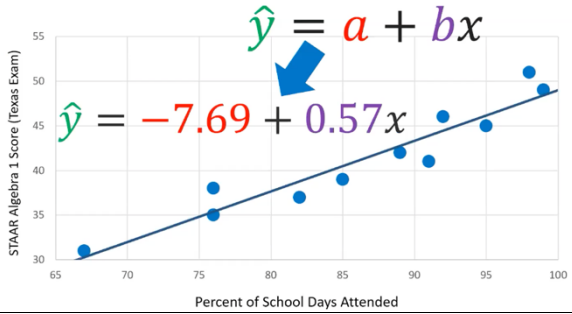
1) Slope? What does it mean? (Plot finding slope on the graph.)
Slope: _____.
Means: for every increase in _____ by _____ unit, the _____ valued increases by _____.
2) Y intercept? What does it mean? (Plot the y-intercept on the graph.)
Y-intercept: _____
When x = _____, the y-intercept value is _____.

Linear Regression Model

Algebra: Linear Equation → Stats: Linear Regression
(Label similarities as you watch the video.)

$$\boldsymbol{y = mx + b} \qquad \boldsymbol{\hat{y} = a + bx}$$

The LSRL Equation

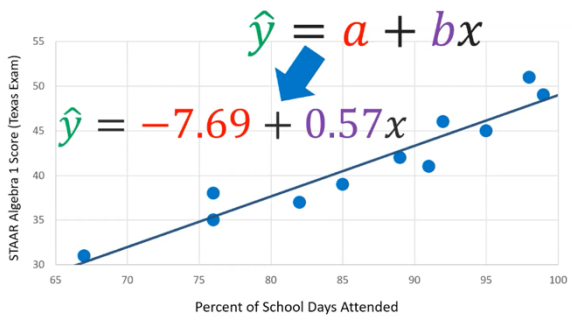


Interpret the **slope** value…
For every 1 unit increase in explanatory variable, our model predicts an average increase/decrease of slope in response variable.
For every 1 _____ increase in _____, our model predicts an average _____ of _____ _____.

STATS MEDIC

Name_____

The LSRL Equation



Interpret the **y-intercept…**

$$\hat{y} = -7.69 + 0.57x$$

When the explanatory variable is zero units, our model predicts that the response variable would be y-intercept.

When _____ is zero _____, our model predicts that students would score _____.

Take a pause: Is the y-intercept _____ in this context? Why or why not?
The y-intercept value is _____, since anyone with _____ attendance doesn't really go to the school OR take the exam. The y-intercept gives a _____ predicted number of correct answers, which is _____.

What Should We Take Away?

The _____ value tells you the _____ change in _____for every _____ increase in _____.

The _____ value tells you the _____ y-value when _____.

The _____ is only _____ in contexts where the _____ can reasonably take on the value of _____.

Name_____

AP Statistics CED 2.8 Daily Video 3 (Skill 2.C)

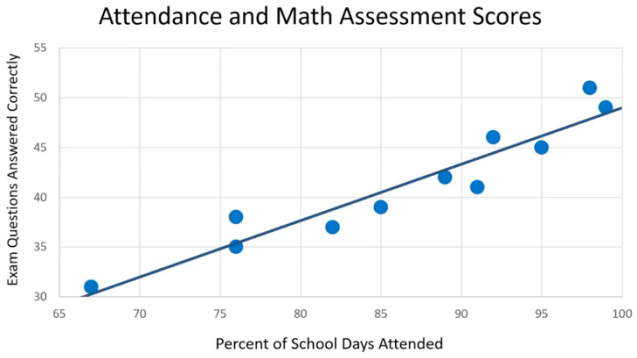
What Will We Learn?

How do we determine r^2?

How do we interpret r^2?

How do we interpret computer output for linear regression?

Coefficient of Determination (r^2)



Random sample of 11 student in Texas.

x: percent of school days attended

y: number of questions answered correctly on state Algebra 1 test.

Our Model: $\hat{y} = -7.69 + 0.57x$ where,

$\hat{y}$: Predicted number of questions correct

x : Percent of school days attended

Our questions:

How well does attendance predict test scores?

How much better are the LSRL predictions than a model that can't use attendance *at all*?

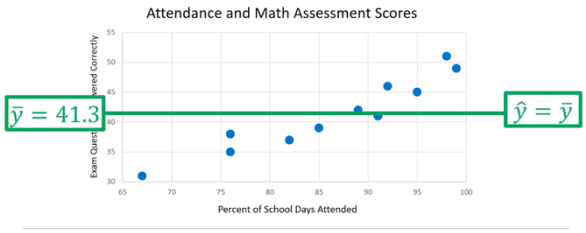
Mean of y

$\bar{y} = 41.3$ **correct answers**

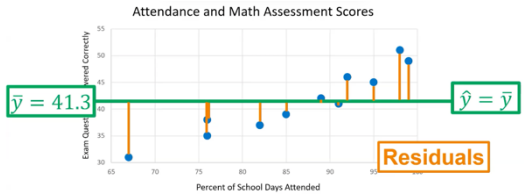
This is our model when we're not allowed to use attendance (x) to make predictions.

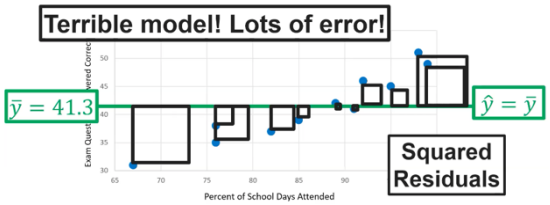
explains (x) → responds (y)

Percent Attendance	Questions Correct
95	45
89	42
67	31
98	51
99	49
76	38
92	46
91	41
76	35
85	39
82	37

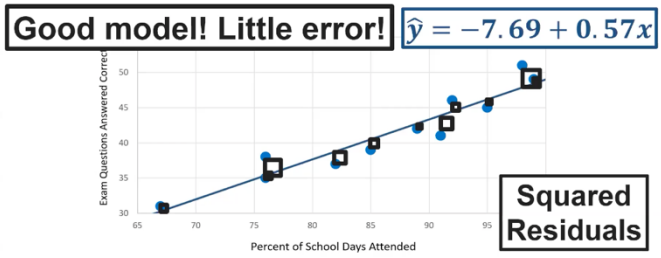


Coefficient of Determination (r^2) – So we get the residuals (prediction errors)





Therefore, when we create the LSRL we use a model that actually uses x to predict y! So, we know this s a better model, but how much better now becomes our question.

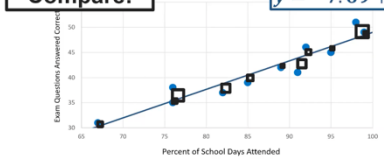


Name_____

Coefficient of Determination (r^2)

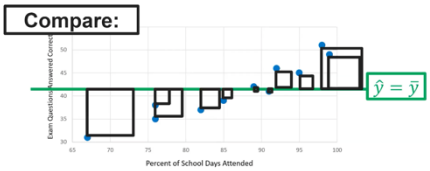
LSRL Model

Compare: $\hat{y} = -7.69 + 0.57x$



Simple Mean Model

Compare: $\hat{y} = \bar{y}$

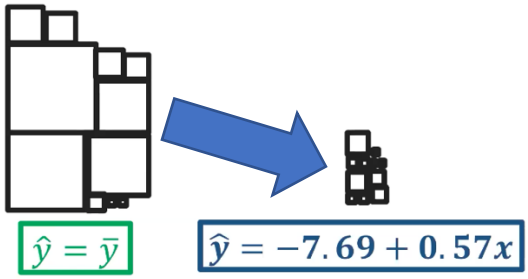


When we use x we get much better predictions!

Sum of Square Residuals

Using x in our model resulted in a _____ in the sum of squared residuals.

That number _____ is what we call the coefficient of determination or r^2.



Coefficient of Determination (r^2)

The _____ of variation in the _____ that is explained by the _____ in the model.

Interpretation of the LSRL:

r^2% of the variation in response variable can be explained by the linear relationship with explanatory variable.

In our case: _____ of the variable in _____ can be explained by the linear relationship with _____.

Fun Fact:

Coefficient of Determination = (Correlation)^2 = $r^2 = (r)^2$

r^2 properties

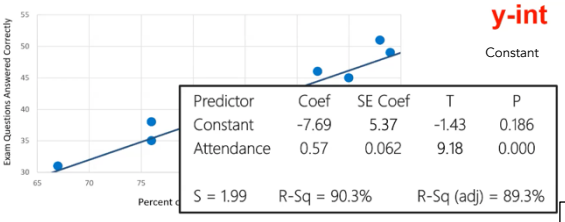
_____; _____ → stronger; _____ → weaker

Computer Output

$\hat{y} = a + bx$

y-int: Constant

slope: Coefficient of x



$\hat{y} = -7.69 + 0.57x$

Predictor	Coef	SE Coef	T	P
y-int Constant	-7.69	5.37	-1.43	0.186
slope Attendance (x)	0.57	0.062	9.18	0.000
S = 1.99	R-Sq = 90.3%		R-Sq (adj) = 89.3%	

r^2

Note slope determines whether correlation is positive or

Note: 90.3% changed to .903! Percent was changed to a decimal!

You can calculate r (correlation) by taking the square root of r^2. So, $r = \pm\sqrt{r^2} = \pm\sqrt{.903} = \pm.95$

What Should We Take Away?

____ is the proportion of variation in the ____ variable that is explained by the ____ variable in the model.

____ is a measure of _____ and can be obtained by squaring the _____.

When reading computer output, recall that the _____ is the x-value's coefficient. The _____ is a constant.

STATS MEDIC

Name_____

AP Statistics CED 2.9 Daily Video 1 (Skill 2.A)

Analyzing Departures from Linearity

What Will We Learn?

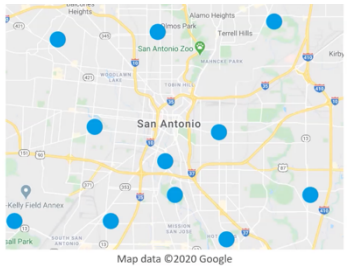
What is a high-leverage point in linear regression?

What is an outlier in linear regression?

What effects do different types of influential points have on the LSRL, correlation and r^2?

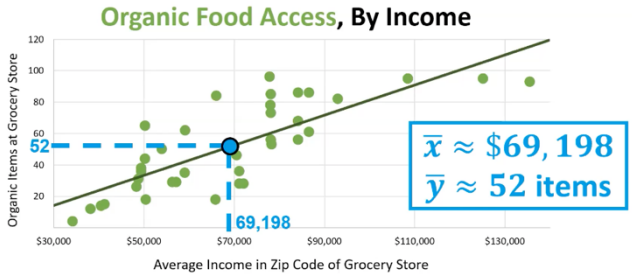
Food Access: Is neighborhood a good predictor of access to healthy foods?

Supermarket locations in San Antonio, TX.

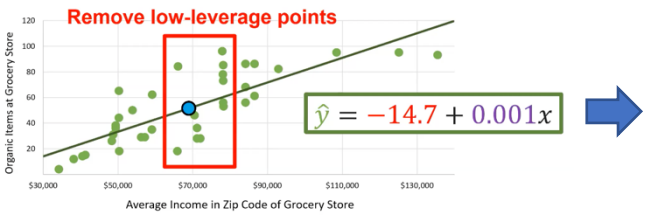


Former student, Linda Saucedo, noticed that her local supermarket offered fewer organic fruits/vegetables than another location from the same company in a wealthier part of town. She wondered if this is a broader pattern throughout the city?

Linear Regression Model

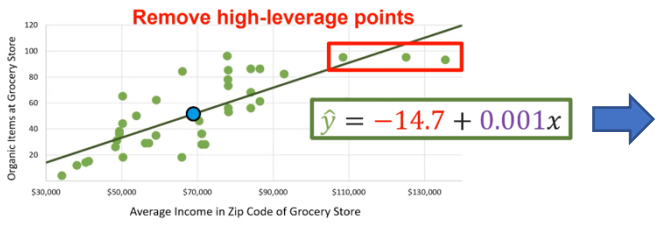


Removing Low-Leverage Points



When the points which, on the x-axis, are _____ to the mean point, the LSRL does not change much.

Removing High-Leverage Points



When the points which, on the x-axis, are _____ above the mean point, the LSRL does have a substantial shift.

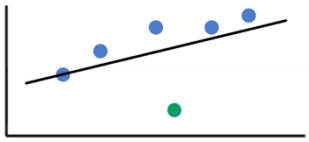
High-Leverage Points

High-Leverage Points: point with unusually _____ or _____ x-values (far away from $\bar{x}$. If removed, usually have a _____ effect on _____ of LSRL.

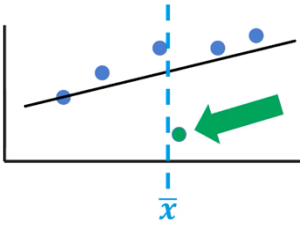
STATS MEDIC

Name_____

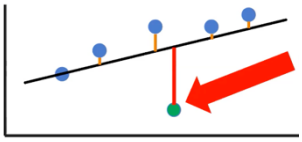
What about this point?



What about this point below the LSRL?



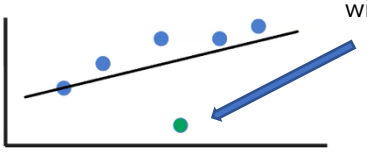
It appears to be near $\bar{x}$,so not necessarily a high leverage point. But it does appear unusual…



Unusually high magnitude residuals are:

* Called an Outlier
* Big effect on strength of relationship

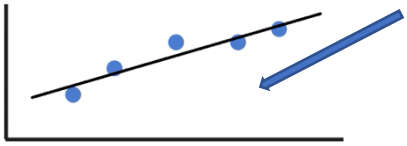
Outliers and Strength



with point!

$r = 0.463$

$r^2 = 0.214$



without point!

$r = 0.940$

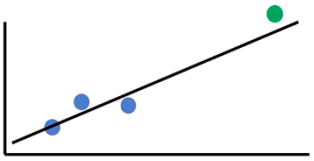
$r^2 = 0.883$

Influential Points

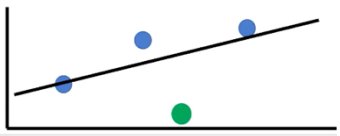
Points that, _____, change the _____, _____, and/or _____ substantially.

Three types of Influential Points:

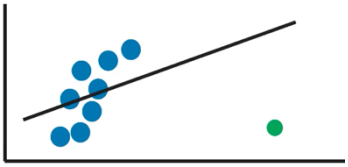
Outliers (change _____) High-Leverage (change _____) Both of the above.



Influential Point
- High-Leverage
- If removed, may change slope/y-int substantially



Influential Point
- Outlier
- If removed, may change correlation substantially



Influential Point
- High-Leverage **and** Outlier
- If removed, may change slope/y-int and correlation substantially

What Should We Take Away?

High-leverage points have unusually _____ or _____ x-values.
Outliers have unusually high magnitude _____ (don't follow the modeled trend as closely).
Points are considered "_____" if, when removed, they affect the _____, _____, and/or _____.

Name_____

AP Statistics CED 2.9 Daily Video 2 (Skill 2.C)

Analyzing Departures from Linearity

What Will We Learn?

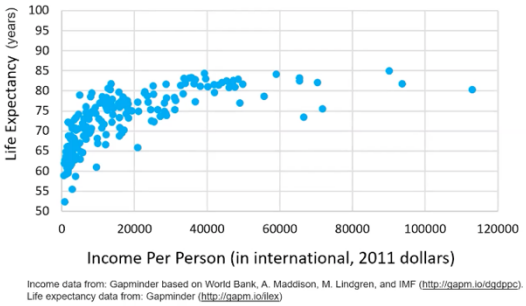
What do we do when we encounter bivariate data that has a non-linear form?
What effects do transformations have on different datasets?
How do we assess if transforming our data improved model fit?

Income and Life Expectancy

What is the relationship between income per person in a country and the average life expectancy?

Notes:

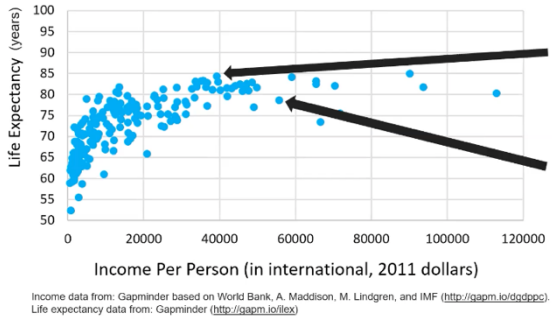
- Income per person = GDP per person adjusted for difference in purchasing power (in international dollars, based on 2011 prices).
- Life expectancy is the average number of years a newborn child would live if current mortality patterns were the same.



2018 data on countries' income per person and life expectancy (each data point is a country).

Direction:

_____ association; as income rises, life expectancy tends to rise (Not a perfect trend).



Japan
x = $39,300 (int.)
y = 84.4 years

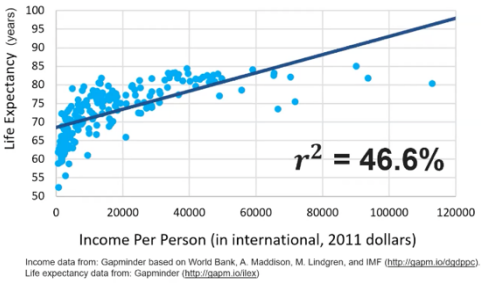
United States
x = $55,700 (int.)
y = 78.6 years

These points show us that there is some _____ from the trend.

But the overall trend does show a _____ trend between income and life expectancy.

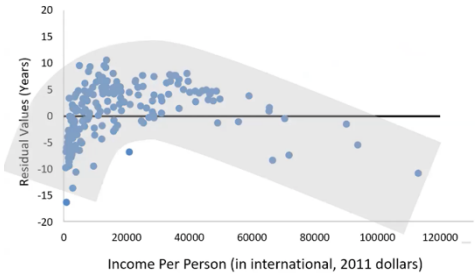
Form

Fitting with the LSRL



The r^2 value of 46.7% is pretty _____ value and does not seem to fit the trend of this data.

Looking at Residual Plot



There looks like there is a curved pattern here.

LSRL is NOT a Good Fit for this Model!

STATS MEDIC

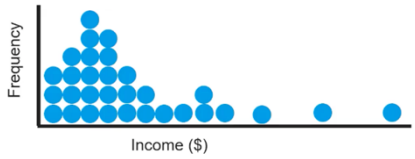
Name_____

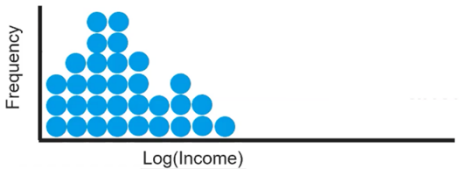
Transforming for Linearity

Problem: sometimes data show _____ relationships.

Solution: transform the data to make the form more _____ (and model with LSRL).

Let's try _____ the income data.

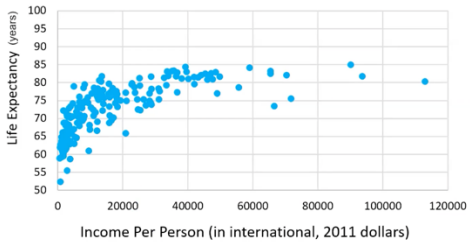




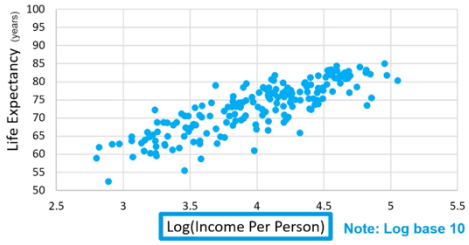
Income data tends to be right-skewed. So, take logarithm of every income (log transformation)

The logarithm makes _____ values less extreme, while _____ order. This reduces _____, which can turn _____ relationships into more _____ ones when income is variable.

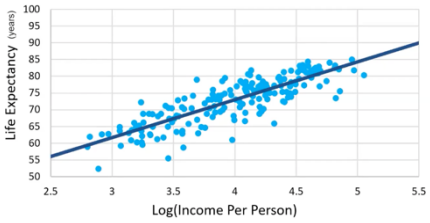
Untransformed (Original)



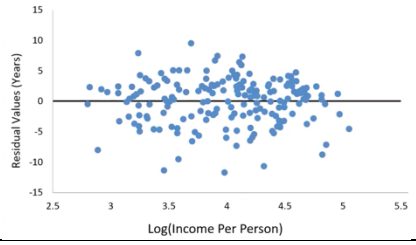
Income Log Transformed



LSRL Income Log Transformed



Residual Plot (Apparent Randomness)



Assessing Fit Using r^2

Untransformed

Predictor	Coef	SE Coef	T	P
Constant	68.5	.518	132.3	0.000
Income	.0002	.00002	12.6	0.000
S = 5.12	R-Sq = 46.6%		R-Sq (adj) = 46.3%	

Transformed

Predictor	Coef	SE Coef	T	P
Constant	27.7	2.16	12.8	0.000
LogIncome	11.3	.54	21.1	0.000
S = 3.76	R-Sq = 71.1%		R-Sq (adj) = 70.9%	

$r^2 = 46.6\%$ < $r^2 = 71.1\%$ ⟸ Stronger Fit!

Other Transformations

There are different types of transformations to achieve linearity. Most common:

-Taking **logarithm** of each data value -**Squaring** each data value -**Exponentiating** each data value

What Should We Take Away?

_____ data that exhibit _____ patterns can often be _____ to achieve _____.

Assess the effectiveness of transformations by seeing if patterns in the _____ were reduced.

Assess the effectiveness of transformations by seeing if the _____ increased.