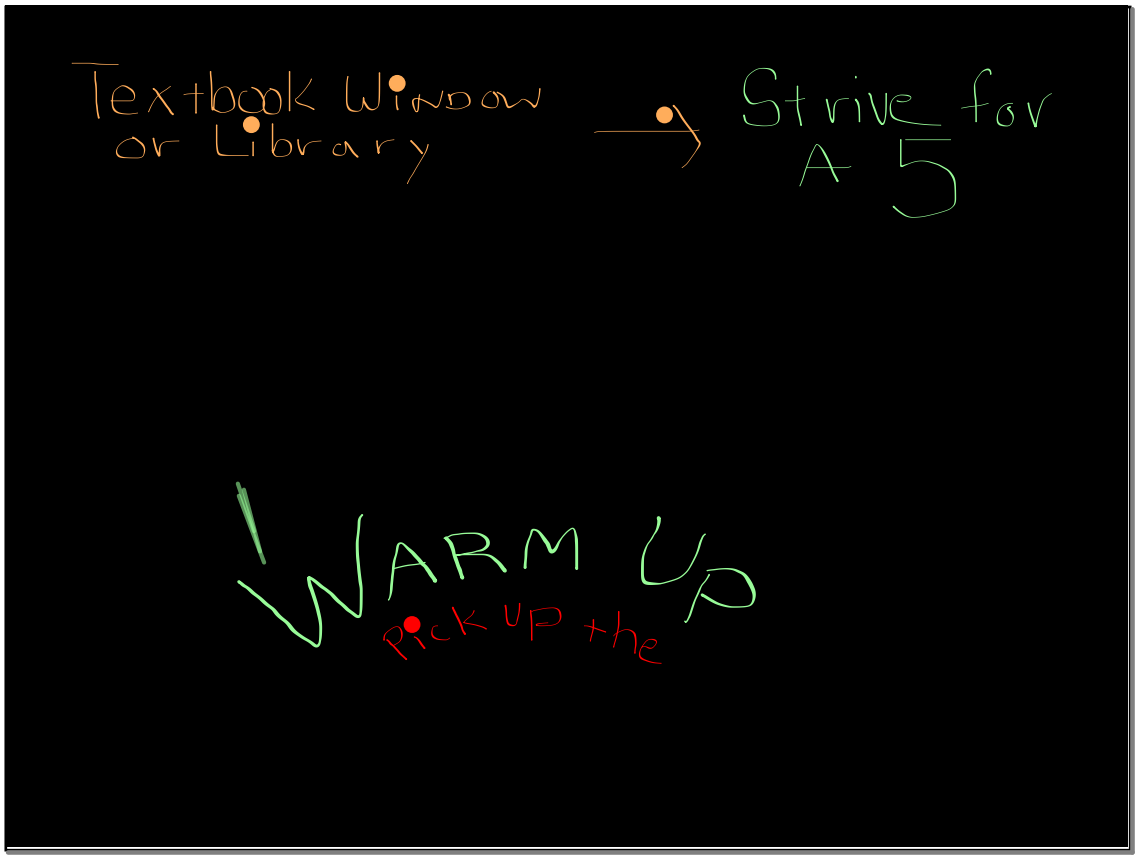


Using technology is often the most convenient way to find the equation of a least-squares regression line.

It is also possible to calculate the equation of the least-squares regression line using only the means and standard deviations of the two variables and their correlation.

$\bar{x}$ $\bar{y}$ r s_x s_y

How to Calculate the Least-squares Regression Line Using Summary Statistics

We have data on an explanatory variable *x* and a response variable *y* for *n* individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y of the two variables and their correlation *r*.

The least-squares regression line is the line $\hat{y} = b_0 + b_1 x$

with **slope** $b_1 = r \cdot \frac{s_y}{s_x}$

and ***y* intercept** $b_0 = \bar{y} - b_1 \bar{x}$

2. We collect data from a random sample of 15 high school students to investigate the relationship between foot length (in centimeters) and height (in centimeters).
   - The mean and standard deviation of the foot lengths are $\bar{x} = 24.76$ and $s_x = 2.71$.
   - The mean and standard deviation of the heights are $\bar{x} = 171.43$ and $s_y = 10.69$.
   - The correlation between foot length and height is $r = 0.697$.

   Find the equation of the least-squares regression line for predicting height from foot length.

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

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

2. We collect data from a random sample of 15 high school students to investigate the relationship between foot length (in centimeters) and height (in centimeters).
   - The mean and standard deviation of the foot lengths are $\bar{x} = 24.76$ and $s_x = 2.71$.
   - The mean and standard deviation of the heights are $\bar{x} = 171.43$ and $s_y = 10.69$.
   - The correlation between foot length and height is $r = 0.697$.

   Find the equation of the least-squares regression line for predicting height from foot length.

slope $b = r \cdot \frac{s_y}{s_x}$ $b = 0.697 \cdot \frac{10.69}{2.71} = 2.75$

y-int $a = \bar{y} - b\bar{x}$ $a = 171.43 - 2.75(24.76) = 103.34$

$\hat{y} = 103.34 + 2.75x$

New terms
univariate & bivariate

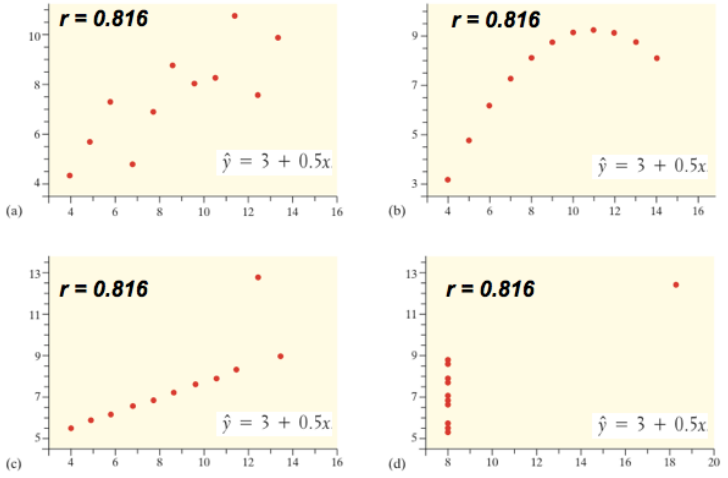
We have been working with bivariate quantitative data

Correlation and Regression Wisdom

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, you should be aware of their limitations.

What form do you visualize in a scatterplot if you were told the linear correlation coefficient is **0.86** *?*

CORRELATION AND REGRESSION LINES DESCRIBE ONLY LINEAR RELATIONSHIPS



ASSOCIATION DOES NOT IMPLY CAUSATION

When we study the relationship between two variables, we often hope to show that changes in the explanatory variable *cause* changes in the response variable.

CAUTION:

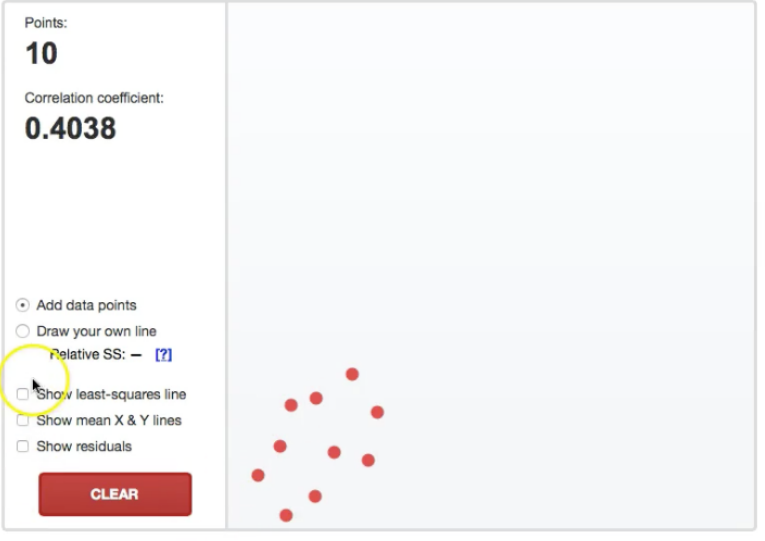
A strong association between two variables is not enough to draw conclusions about cause and effect.

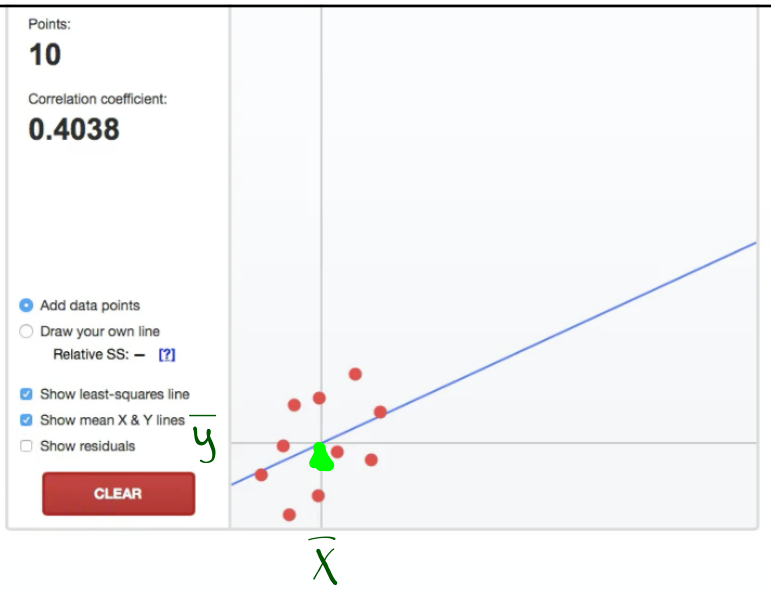
Aim

To understand how Unusual Points influence the LSRL

Pick Up

1) Laptops 2) Instructions



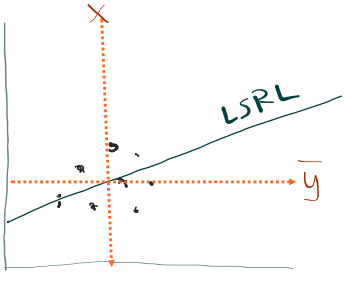


Lesson 3.2: Day 4

How do Unusual Points affect the LSRL?

1. Go to stapplet.com and find the link for ***Correlation and Regression.***

- Click on the graphing area to add 10 points in the lower-left corner so that the correlation is about r = 0.40.
- Check the boxes to show the LSRL and the mean X and Y lines.
- Sketch it..



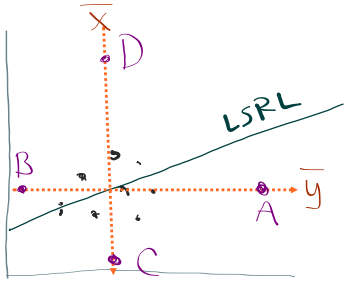
Lesson 3.2: Day 4

How do Unusual Points affect the LSRL?

1. Go to stapplet.com and find the link for ***Correlation and Regression.***

- Click on the graphing area to add 10 points in the lower-left corner so that the correlation is about r = 0.40.
- Check the boxes to show the LSRL and the mean X and Y lines.
- Sketch it..

sample



2. For each of the following situations add the point to the scatterplot and decide if the slope, *y*-intercept and correlation will increase or decrease.

a. If a point is added on the far-right side of the graph on the horizontal line for the mean of Y.

Slope: y-intercept: Correlation:

b. If a point is added on the far-left side of the graph on the horizontal line for the mean of Y.

Slope: y-intercept: Correlation:

c. If a point is added below the LSRL on the vertical line for the mean of X.

Slope: y-intercept: Correlation:

d. If a point is added above the LSRL on the vertical line for the mean of X.

Slope: y-intercept: Correlation:

2. For each of the following situations add the point to the scatterplot and decide if the slope, *y*-intercept and correlation will increase or decrease.

a. If a point is added on the far-right side of the graph on the horizontal line for the mean of Y.

Slope: decreases y-intercept: increases Correlation: decreases

b. If a point is added on the far-left side of the graph on the horizontal line for the mean of Y.

Slope: decreases y-intercept: increases Correlation: decreases

c. If a point is added below the LSRL on the vertical line for the mean of X.

Slope: Same y-intercept: decreases Correlation: decreases

d. If a point is added above the LSRL on the vertical line for the mean of X.

Slope: Same y-intercept: increases Correlation: decreases

2. For each of the following situations add the point to the scatterplot and decide if the slope, *y*-intercept and correlation will increase or decrease.

a. If a point is added on the far-right side of the graph on the horizontal line for the mean of Y.

Slope: decreases y-intercept: increases Correlation: decreases

b. If a point is added on the far-left side of the graph on the horizontal line for the mean of Y.

Slope: decreases y-intercept: increases Correlation: decreases

c. If a point is added below the LSRL on the vertical line for the mean of X.

Slope: Same y-intercept: decreases Correlation: decreases

d. If a point is added above the LSRL on the vertical line for the mean of X.

Slope: Same y-intercept: increases Correlation: decreases

Tilts line from (x̄, ȳ)

2. For each of the following situations add the point to the scatterplot and decide if the slope, *y*-intercept and correlation will increase or decrease.

Tilts line from $(\bar{x},\bar{y})$

a. If a point is added on the far-right side of the graph on the horizontal line for the mean of Y.

Slope: decreases y-intercept: increases Correlation: decreases

b. If a point is added on the far-left side of the graph on the horizontal line for the mean of Y.

Slope: decreases y-intercept: increases Correlation: decreases

Lifts line from $(\bar{x},\bar{y})$

c. If a point is added below the LSRL on the vertical line for the mean of X.

Slope: Same y-intercept: decreases Correlation: decreases

d. If a point is added above the LSRL on the vertical line for the mean of X.

Slope: Same y-intercept: increases Correlation: decreases

2. For each of the following situations add the point to the scatterplot and decide if the slope, *y*-intercept and correlation will increase or decrease.

Tilts line from $(\bar{x},\bar{y})$

a. If a point is added on the far-right side of the graph on the horizontal line for the mean of Y.

Slope: decreases y-intercept: increases Correlation: decreases

b. If a point is added on the far-left side of the graph on the horizontal line for the mean of Y.

Slope: decreases y-intercept: increases Correlation: decreases

Lifts line from $(\bar{x},\bar{y})$

c. If a point is added below the LSRL on the vertical line for the mean of X.

Slope: Same y-intercept: decreases Correlation: decreases

d. If a point is added above the LSRL on the vertical line for the mean of X.

Slope: Same y-intercept: increases Correlation: decreases

Out of pattern
Outliers

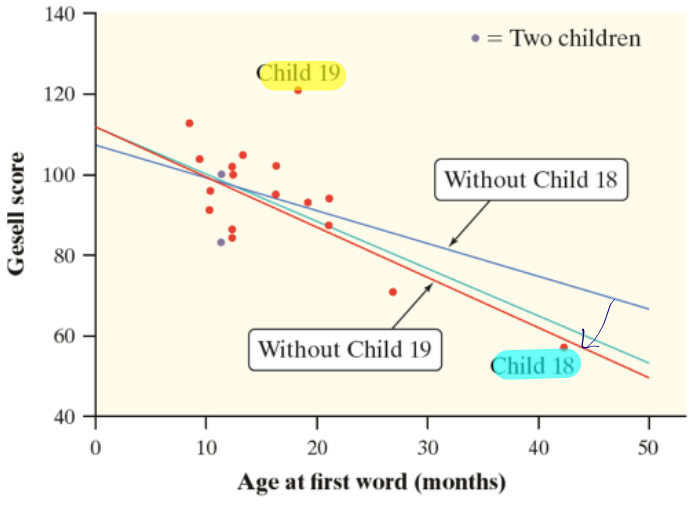
3. Which outliers had the greatest impact on the LSRL, vertical or horizontal outliers?

HORIZONTAL

Those outliers have more "leverage" since they are to the left and right of the center $(\bar{x}, \bar{y})$ of the plot

CORRELATION AND LEAST-SQUARES REGRESSION LINES ARE NOT RESISTANT

ORRELATION AND LEAST-SQUARES REGRESSION LINES ARE NOT RESISTANT



With all 19 points:
$\hat{y} = 109.9 - 1.127x$
$s = 11.02$ and $r^2 = 0.410$

Without Child 18:
$\hat{y} = 105.6 - 0.779x$
$s = 11.11$ and $r^2 = 0.112$

Without Child 19:
$\hat{y} = 109.3 - 1.193x$
$s = 8.63$ and $r^2 = 0.572$

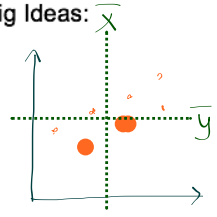
Child 18 gives more leverage than Child 19

Unusual Points and the LSRL

Big Ideas:

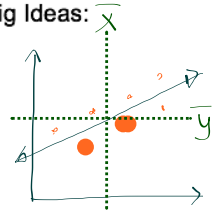
Unusual Points and the LSRL

Big Ideas:



Unusual Points and the LSRL

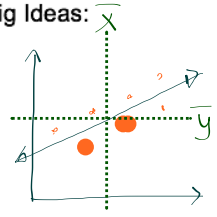
Big Ideas:



LSRL passes thru $(\bar{x}, \bar{y})$

Unusual Points and the LSRL

Big Ideas:

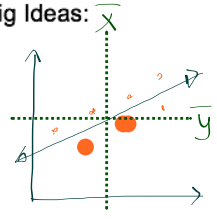


LSRL passes thru $(\bar{x}, \bar{y})$

GOOD LSRL has
- Low s
- high r^2

Unusual Points and the LSRL

Big Ideas:



LSRL passes thru $(\bar{x}, \bar{y})$

GOOD LSRL has
- Low s
- high r^2

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

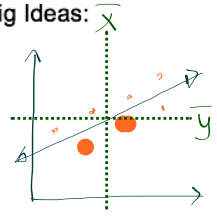
a: y-int, b: slope

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

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

Unusual Points and the LSRL

Big Ideas:



LSRL passes thru $(\bar{x}, \bar{y})$

GOOD LSRL has
- Low S
- high r^2

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

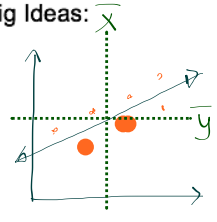
a: y-int, b: slope

$b = r \cdot \frac{s_y}{s_x}$ $a = \bar{y} - b\bar{x}$

Points with high leverage in regression have much larger, or much smaller, x-values than other points in data.

Unusual Points and the LSRL

Big Ideas:



LSRL passes thru $(\bar{x}, \bar{y})$

GOOD LSRL has
- Low S
- high r^2

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

a: y-int, b: slope

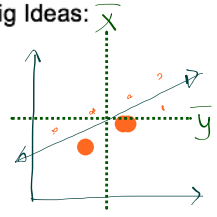
$b = r \cdot \frac{s_y}{s_x}$ $a = \bar{y} - b\bar{x}$

Points with high leverage in regression have much larger, or much smaller, x-values than other points in data.

An influential point is any point, if removed, substantially changes the slope, y-int, r, r^2, or S

Unusual Points and the LSRL

Big Ideas:



LSRL passes thru $(\bar{x}, \bar{y})$

GOOD LSRL has
- Low S
- high r^2

LSRL $\hat{y} = a + bx$ (a: y-int, b: slope)

$b = r \cdot \frac{s_y}{s_x}$ $a = \bar{y} - b\bar{x}$

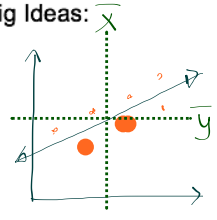
Points with high leverage in regression have much larger, or much smaller, x-values than other points in data.

An influential point is any point, if removed, substantially changes the slope, y-int, r, r^2, or S

I know it doesn't fit, but...

Unusual Points and the LSRL

Big Ideas:



LSRL passes thru $(\bar{x}, \bar{y})$

GOOD LSRL has
- Low S
- high r^2

LSRL $\hat{y} = a + bx$ (a: y-int, b: slope)

$b = r \cdot \frac{s_y}{s_x}$ $a = \bar{y} - b\bar{x}$

Points with high leverage in regression have much larger, or much smaller, x-values than other points in data.

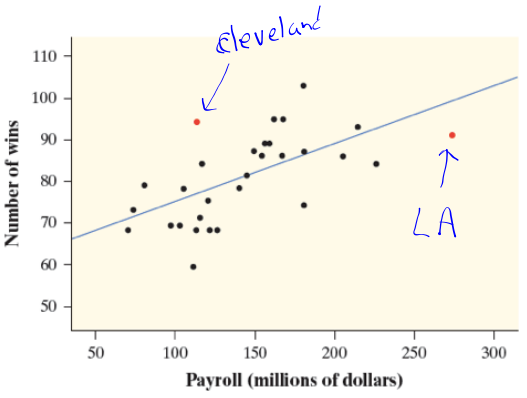
An influential point is any point, if removed, substantially changes the slope, y-int, r, r^2, or S

I know it doesn't fit, but...

High leverage points have more influence on the LSRL than outliers in the vertical direction.

4. The scatterplot shows the payroll (in millions of dollars) and number of wins for Major League Baseball teams in 2016, along with the least-squares regression line. The points highlighted in red represent the Los Angeles Dodgers (far right) and the Cleveland Indians (upper left).

a. Describe what influence the point representing the Los Angeles Dodgers has on the **equation of the least-squares** regression line. Explain your reasoning.

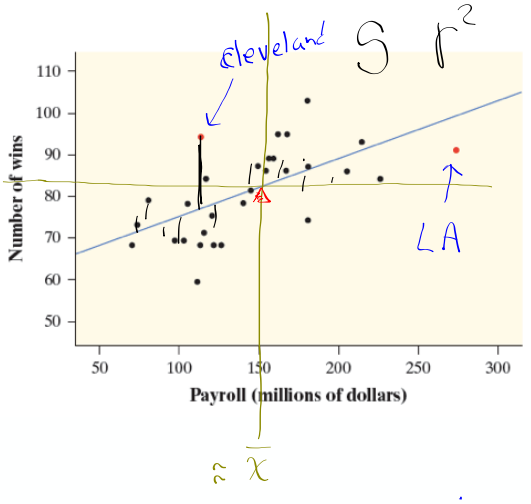


4. The scatterplot shows the payroll (in millions of dollars) and number of wins for Major League Baseball teams in 2016, along with the least-squares regression line. The points highlighted in red represent the Los Angeles Dodgers (far right) and the Cleveland Indians (upper left).

Discuss in your group

a. Describe what influence the point representing the Los Angeles Dodgers has on the **equation of the least-squares** regression line. Explain your reasoning.

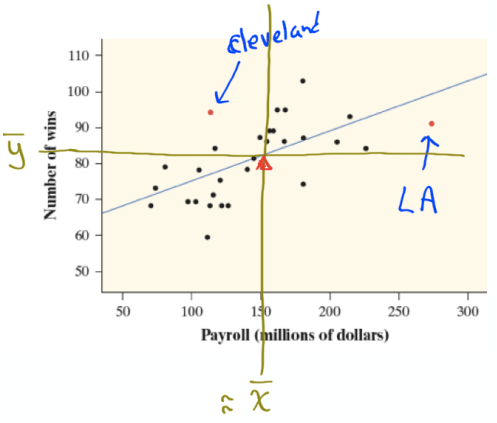
Since LA is on the far right and below the LSRL, it will decrease the slope and ~~decrease~~ increase the y-int.



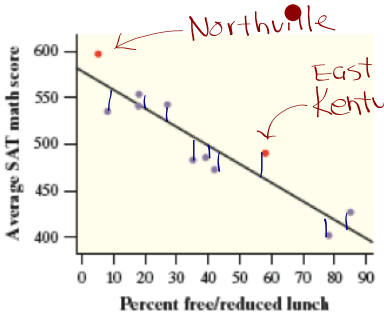
b. Describe what influence the point representing the Cleveland Indians has on the **standard deviation** of the residuals and r^2. Explain your reasoning.

Because Cleveland has a large residual it makes the standard deviation of the residuals (s) larger and r^2 smaller.

↑ less correlation.



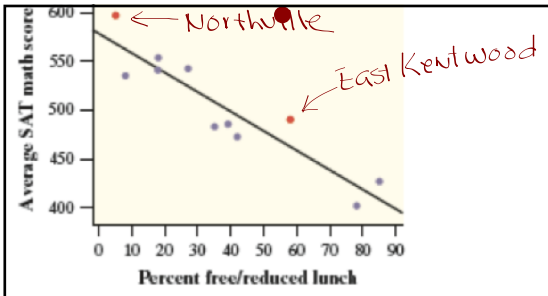
5. The scatterplot below shows the percent of students who are eligible for free/reduced lunch and the average SAT math score for 11 randomly selected high schools in Michigan in 2016, along with the least-squares regression line. The points highlighted in red are Northville High School (upper left) and East Kentwood High School (right middle).



r

r^2

(a) Describe the influence the point representing Northville High School has on the equation of the least-squares regression line. Explain your reasoning.



(a) Describe the influence the point representing Northville High School has on the equation of the least-squares regression line. Explain your reasoning.

Because the point for Northville is is on the far left and above the LSRL, it is making the line steeper (further from 0) and the y-int greater. If it was removed, the line would be less steep.

(b) Describe the influence the point representing East Kentwood High School has on the standard deviation of the residuals and r^2. Explain your reasoning.

Because the point for E. Kentwood has a large residual it is making the std deviation of the residuals, s, greater and the value of r^2 smaller.

Assignment

3.2.... 63, 65, 71-78

and study pp.194-200

Also, be sure you are making progress on the assignment posted for you on "My AP". Needs to be completed by Monday.