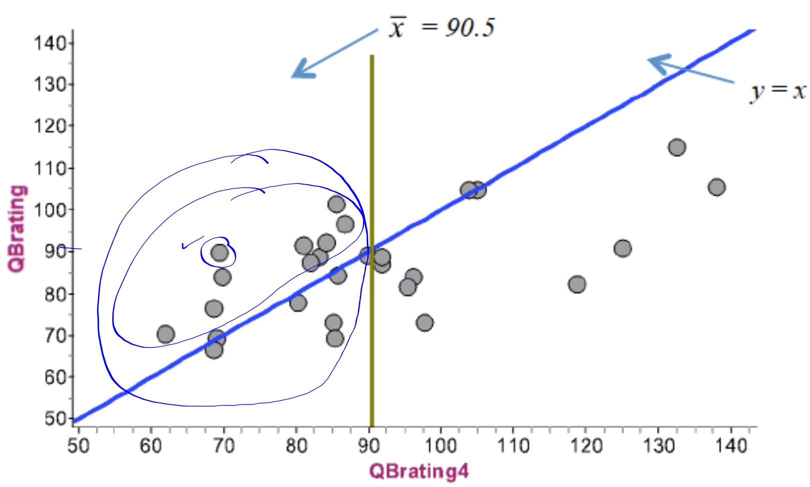
do as much as you can in the first 10 minutes.

on the Warm Up

Regression to the Mean in Sports

•



(a) What would it mean if a point was exactly on the line $y = x$?

The QB's rating is the same at 4 weeks and at the end of the season.

(b) The average QB rating after 4 weeks was $\overline{x}$ = 90.5. Of the 28 quarterbacks, 11 had QB ratings greater than $\overline{x}$ = 90.5 in the first 4 weeks. What percentage of these 11 ended up with a smaller QB rating by the end of the season?

9/11 = 82% (possibly 10/11 = 91%)

(c) Of the 28 quarterbacks, 17 had QB ratings less than $\overline{x}$ = 90.5 in the first 4 weeks. What percentage of these 17 ended up with a larger QB rating by the end of the season?

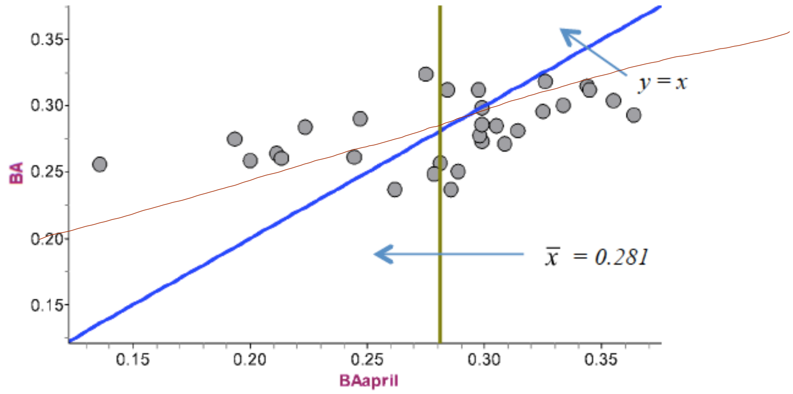
10/17 = 59% (possibly 11/17 = 65%)

(d) If you calculated the least-squares regression line for these data, how do you think it would compare to the line $y = x$?

It would be flatter.

Now
Baseball

where x = batting average in April (the first month of the season) and y = batting average at the end of the seasor average is the proportion at at-bats where the player gets a hit. The line $y = x$ and a vertical line showing the mea average in April are also shown on the scatterplot.



(a) What would it mean if a point was exactly on the line $y = x$?

The player's batting average is the same in April and at the end of the season.

(b) The mean batting average in April was $\bar{x} = 0.281$. Of the 30 players, 19 had batting averages greater than $\bar{x} = 0.281$ in April. What percentage of these 19 ended up with a smaller batting average by the end of the season?

16/19 = 84% (possibly 17/19 = 89%)

(c) Of the 30 players, 11 had batting averages less than $\bar{x} = 0.281$ in April. What percentage of these 11 ended up with a larger batting average by the end of the season?

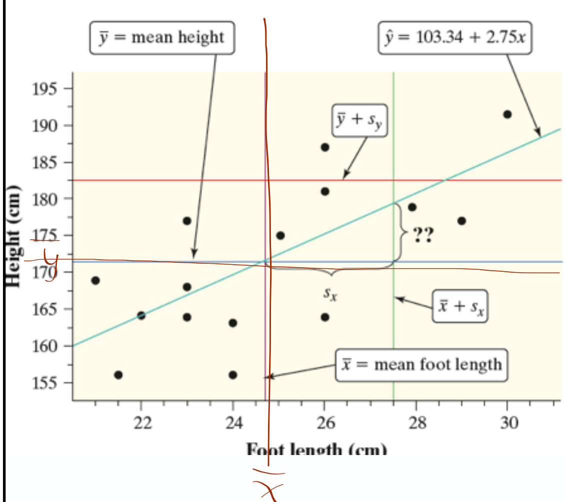
9/11 = 82%

(d) If you calculated the least-squares regression line for these data, how do you think it would compare to the line $y = x$?

It would be flatter.

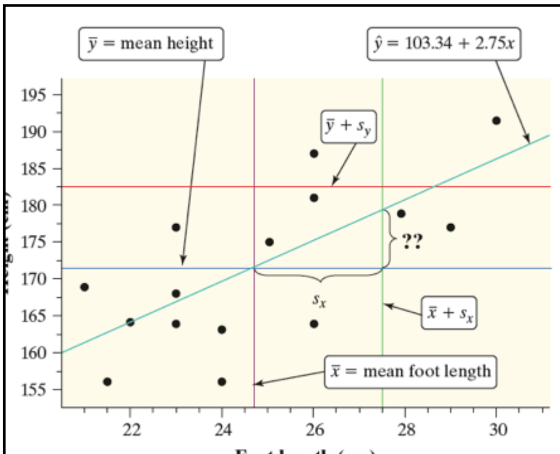
3. Based on these two examples, what typically happens to players who get off to a really good start? What typically happens to players who get off to a really poor start?

When players get off to a good start, they usually don't continue to perform that well (i.e. they perform worse). When players get off to a poor start, they usually don't continue to perform that badly (i.e. they perform better).



The scatterplot shows height versus foot length and the regression equation
$\hat{y} = 103.34 + 2.75x$.
We have added four more lines to the graph:

- a vertical line at the mean foot length $\bar{x}$
- a vertical line at $\bar{x} + s_x$
- a horizontal line at the mean height $\bar{y}$
- a horizontal line at $\bar{y} + s_y$

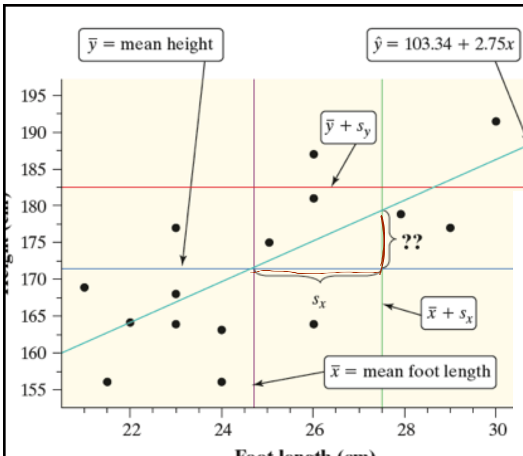


The scatterplot shows height versus foot length and the regression equation
$\hat{y} = 103.34 + 2.75x$.
We have added four more lines to the graph:

- a vertical line at the mean foot length x
- a vertical line at $x + s_x$
- a horizontal line at the mean height y
- a horizontal line at $y + s_y$

For an increase of 1 standard deviation in the value of the explanatory variable x, the least-squares regression line predicts an increase of r standard deviations in the response variable y.

r = .90 for example



The scatterplot shows height versus foot length and the regression equation
$\hat{y} = 103.34 + 2.75x$.
We have added four more lines to the graph:

- a vertical line at the mean foot length x

This is called regression to the mean, because the values of y "regress" to their mean.

For an increase of 1 standard deviation in the value of the explanatory variable x, the least-squares regression line predicts an increase of r standard deviations in the response variable y.

Regression to the Mean
(pages 194-197)

We can use the means and SDs of *x* and *y*, along with their correlation, to calculate the equation of the LSRL.

Regression to the Mean refers to the fact that predicted values of *y* tend to be less extreme (in terms of SDs) than their corresponding values of *x*.

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 $\cancel{\bar{x}}\ \bar{y} = 171.43$ 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 $b_1 = r \cdot \frac{S_y}{S_x}$

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

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 $b_1 = r \cdot \frac{S_y}{S_x}$ $\quad b_1 = 0.697 \cdot \frac{10.69}{2.71} = 2.75$

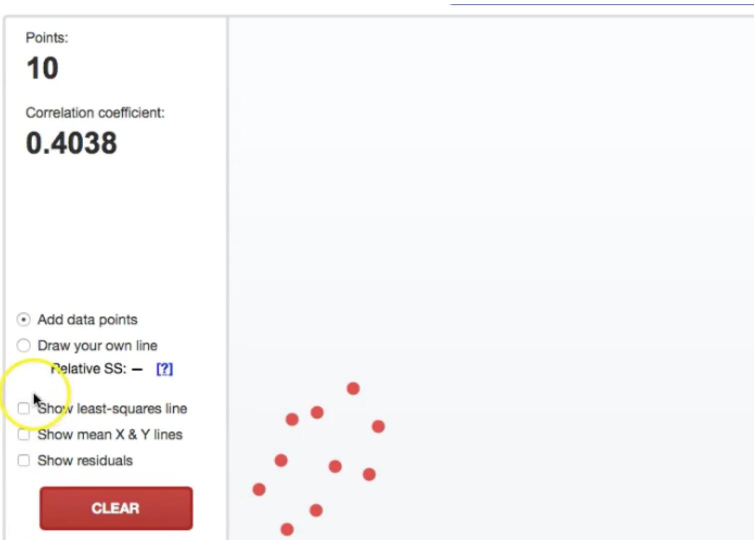
y-int $b_0 = \bar{y} - b_1\bar{x}$ $\quad b_0 = 171.43 - 2.75(24.76) = 103.34$

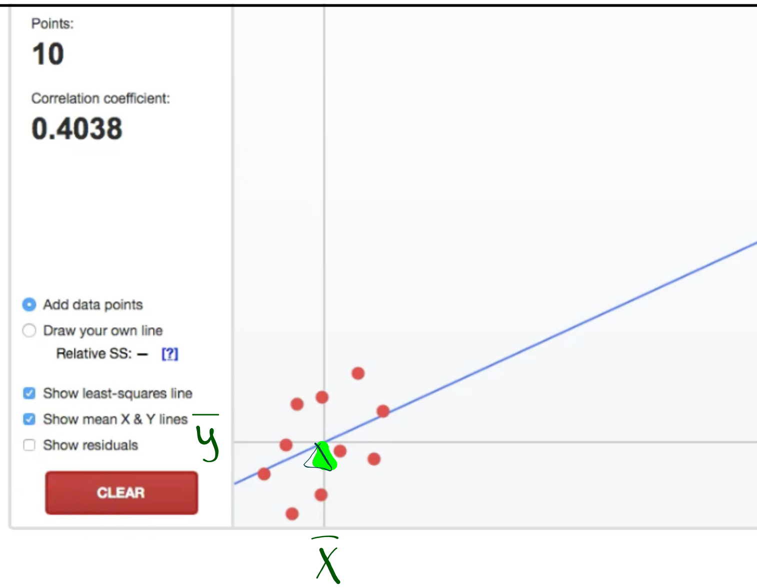
$\hat{y} = 103.34 + 2.75x$

LAptops

p. 198 Activity

write your answers on the class notes





3. Page 198 Activity. Answer the questions below for :

#3- The least-squares regression line did not change

#4 The least-squares regression line is dragged toward the point.

#5 The LSRL changes very slightly. Outliers in the vertical direction have less influence than the than the outliers in the horizontal direction

#6 Outliers can greatly affect the LSRL, with outliers in the horizontal direction having more influence than the vertical direction.

CORRELATION AND LEAST-SQUARES REGRESSION LINES ARE NOT RESISTANT

CORRELATION AND LEAST-SQUARES REGRESSION LINES ARE NOT RESISTANT

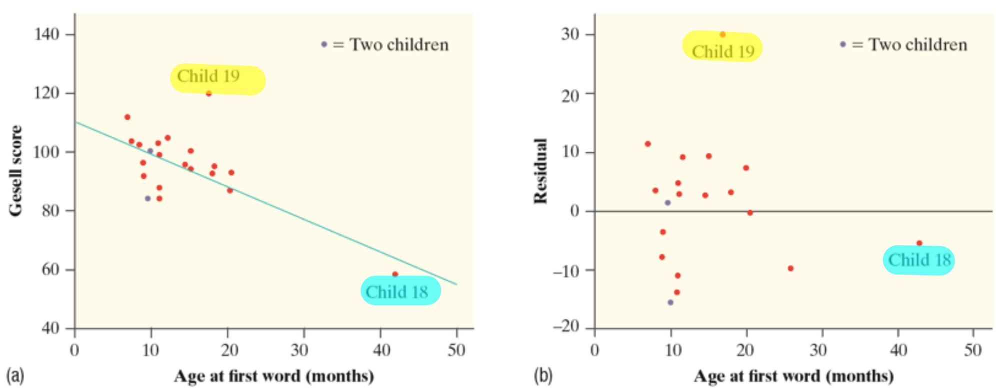
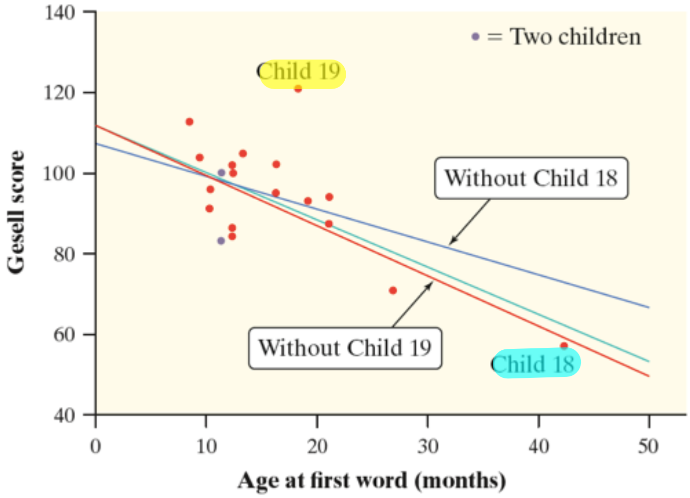


FIGURE 3.16 (a) Scatterplot of Gesell Adaptive Scores versus the age at first word for 21 children, along with the least-squares regression line. (b) Residual plot for the linear model. Child 18 and Child 19 are outliers. Each purple point in the graphs stands for two individuals.

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 $r^2 = 0.112$

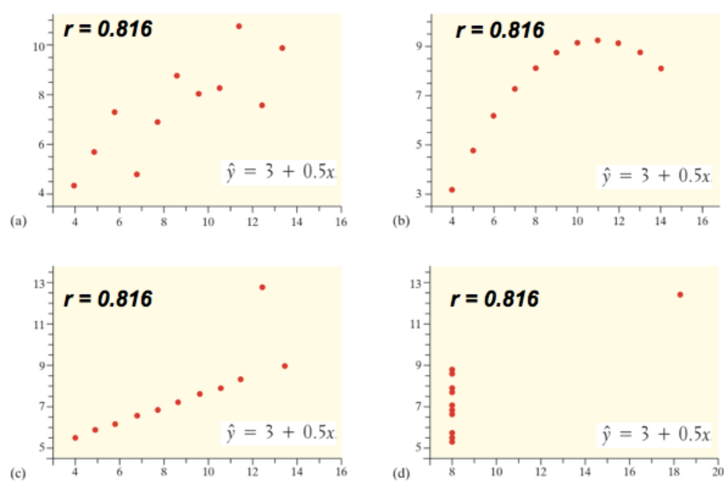
Without Child 19:
$\hat{y} = 109.3 - 1.193x$
$s = 8.63$ and $r^2 = 0.572$

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.

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:

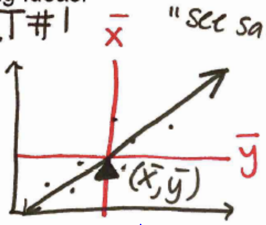
A strong association between two variables is not enough to draw conclusions about cause and effect.

Lesson 3.2 – Outliers and the LSRL

Big Ideas:

Big Ideas:

LT#1 "see saw"



the mean point

Good LSRL has
- Low s
- high r^2

LT#2

LSRL: $\hat{y} = b_0 + b_1 x$

b_0: y-int

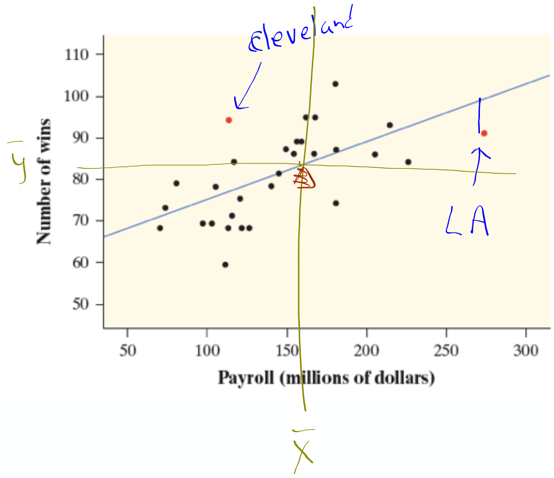
b_1: slope

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

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

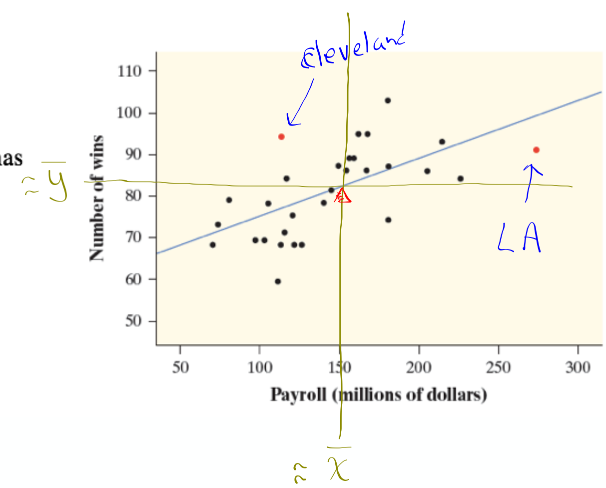
4. The scatterplot shows the payroll (in millions of dollars) and number of wins for Major League Baseball teams in 2016, along with the least-squares regression line. The points highlighted in red represent the Los Angeles Dodgers (far right) and the Cleveland Indians (upper left).

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



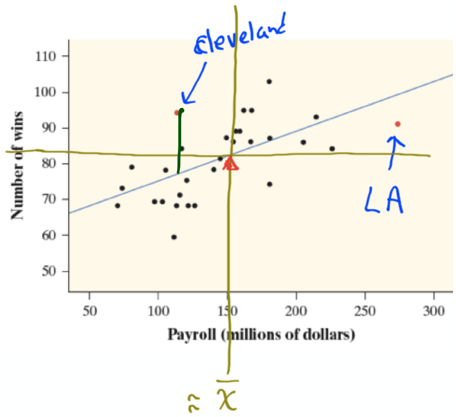
4. The scatterplot shows the payroll (in millions of dollars) and number of wins for Major League Baseball teams in 2016, along with the least-squares regression line. The points highlighted in red represent the Los Angeles Dodgers (far right) and the Cleveland Indians (upper left).

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

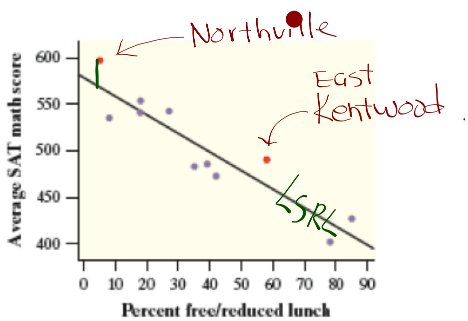
b. Describe what influence the point representing the Cleveland Indians has on the **standard deviation** of the residuals and r^2. Explain your reasoning.



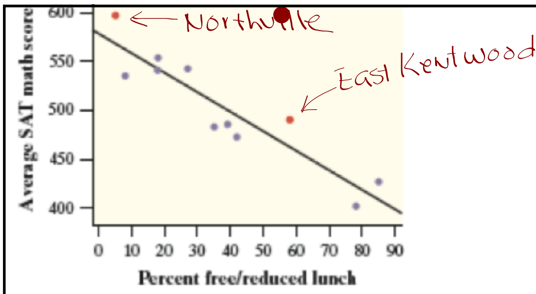
Because Cleveland has a large residual it makes the standard deviation of the residuals (s) larger and r^2 smaller.

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).



(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.

3. **SAT math scores again (Extra Practice)**

In the preceding alternate example, we used data from a random sample of 11 high schools in Michigan to investigate the relationship between the percent of students who are eligible for free/reduced lunch and the average SAT math score. The mean and standard deviation of the percent of students on free/reduced lunch are $\bar{x} = 37.55$ and $s_x = 26.37$. The mean and standard deviation of the average SAT math scores are $\bar{y} = 503.04$ and $s_y = 57.68$. The correlation between percent free/reduced lunch and average SAT math score is $r = -0.9236$. Find the equation of the least-squares regression line for predicting average SAT math scores from percent free/reduced lunch. Show your work.

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

$b_1 = -.9236 \frac{57.68}{26.73}$

$= -1.993$

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

$b_0 = 503.04 - (-1.993)(37.55)$

$= 577.9$

$\boxed{\hat{y} = 577.9 - 1.993x}$

Assignment

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