Advanced Statistics
DS2003 (BDS-4A)
Lecture 18

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

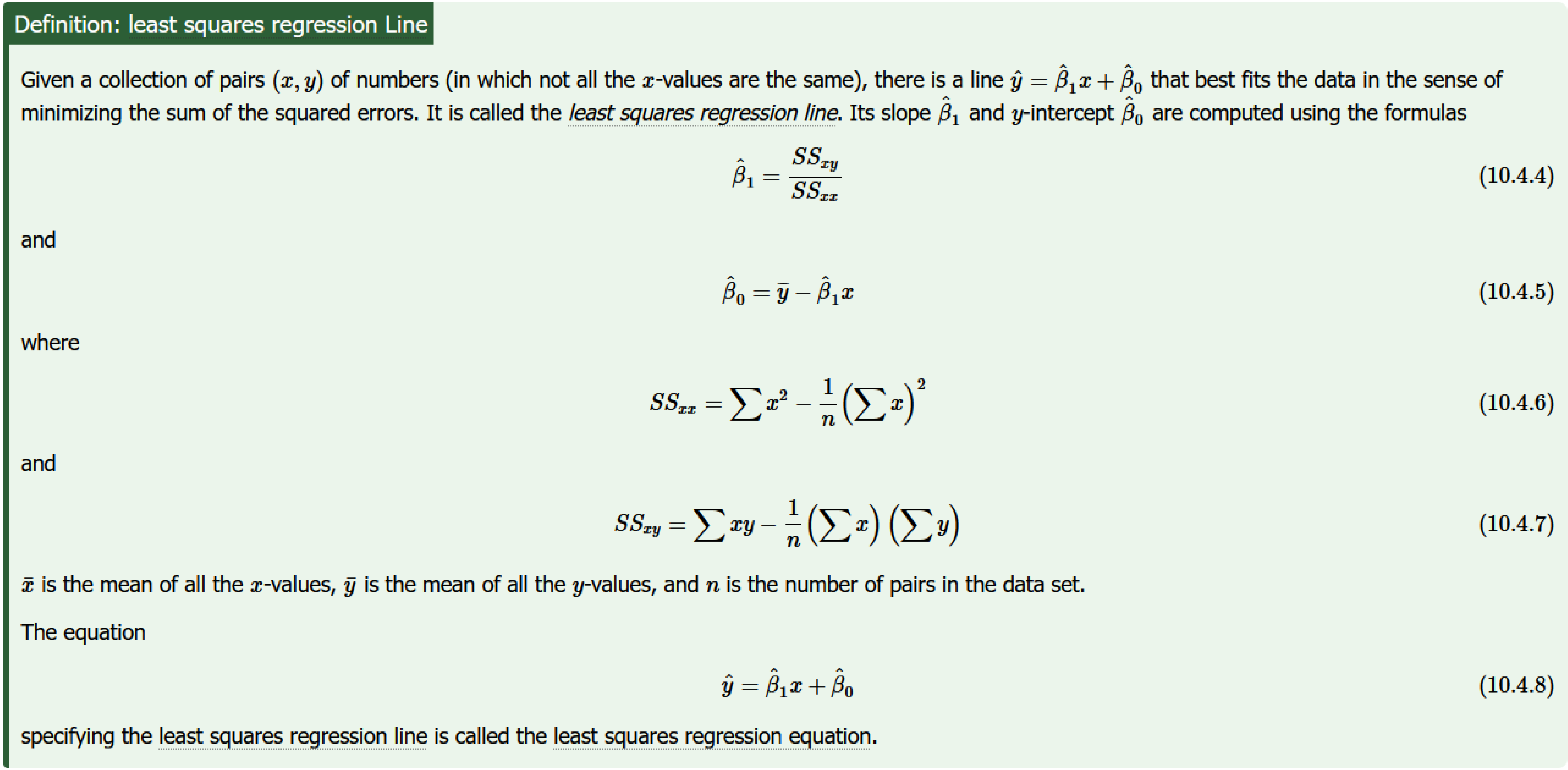
21 April, 2022

Previous Lecture

- Modeling numerical values
  - Example: Poverty vs. High School (HS) graduate rate
- Correlation: Quantifying the relationship
- Fitting a line by least squares regression
- Conditions for a least squares line (Linearity, Nearly normal residuals, Constant variability)
- Slope and Intercept
- Regression Line
- Prediction and Extrapolation

Plan for Today

- Correlation coefficient calculation (r or R)
- Calculating slope using the correlation coefficient and the sample standard deviations



Definition: least squares regression Line

Given a collection of pairs (x, y) of numbers (in which not all the x-values are the same), there is a line $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ that best fits the data in the sense of minimizing the sum of the squared errors. It is called the *least squares regression line*. Its slope $\hat{\beta}_1$ and y-intercept $\hat{\beta}_0$ are computed using the formulas

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \tag{10.4.4}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x \tag{10.4.5}$$

where

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 \tag{10.4.6}$$

and

$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) \tag{10.4.7}$$

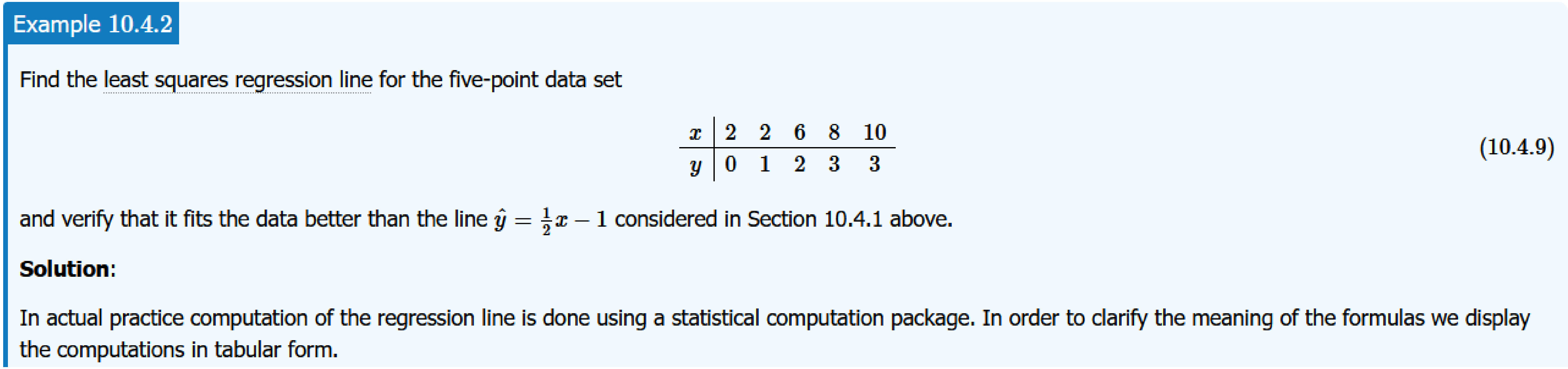
$\bar{x}$ is the mean of all the x-values, $\bar{y}$ is the mean of all the y-values, and n is the number of pairs in the data set.

The equation

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0 \tag{10.4.8}$$

specifying the least squares regression line is called the least squares regression equation.

Example



Example 10.4.2

Find the least squares regression line for the five-point data set

x	2	2	6	8	10
y	0	1	2	3	3

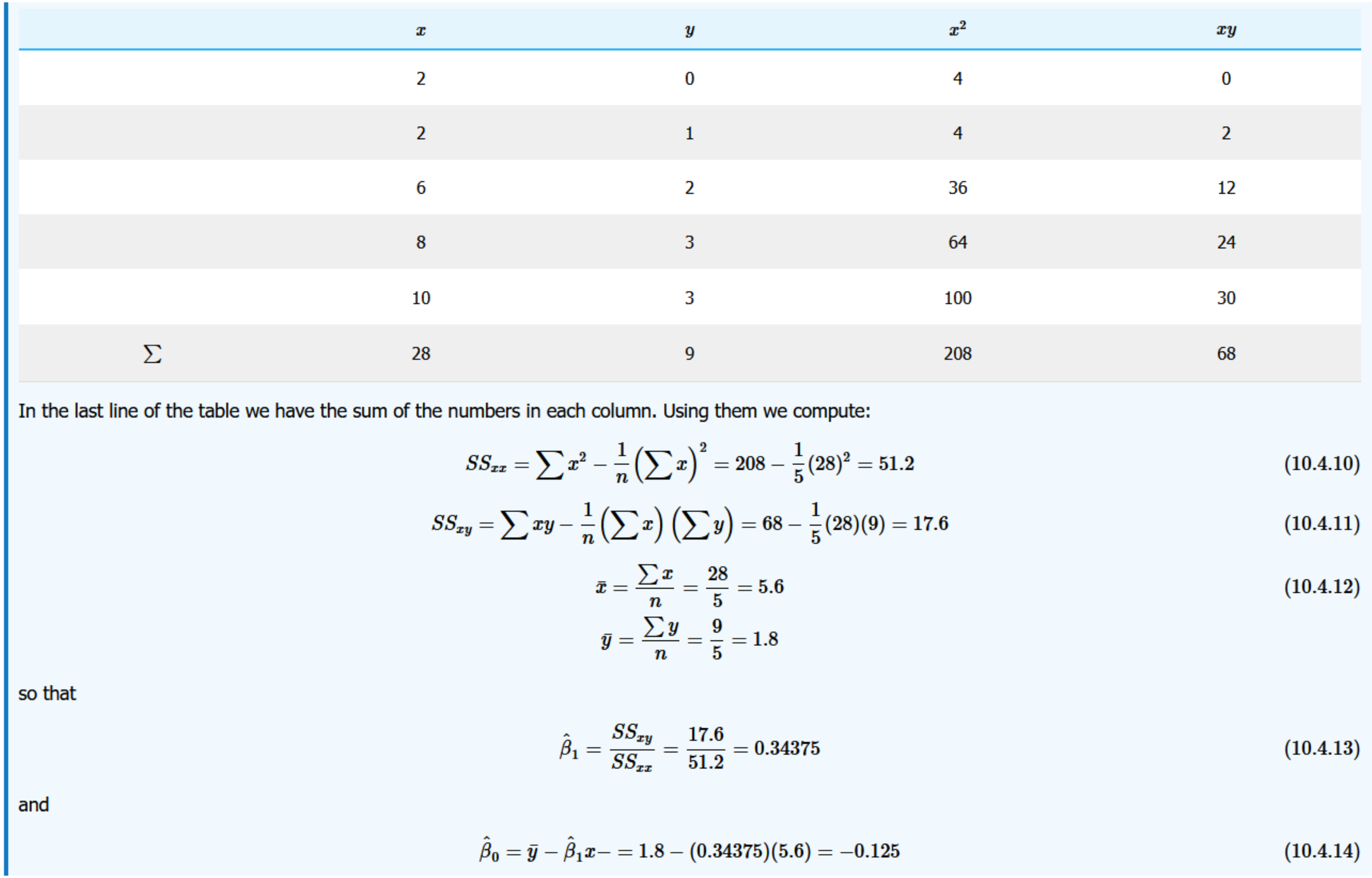
(10.4.9)

and verify that it fits the data better than the line $\hat{y} = \frac{1}{2}x - 1$ considered in Section 10.4.1 above.

Solution:

In actual practice computation of the regression line is done using a statistical computation package. In order to clarify the meaning of the formulas we display the computations in tabular form.

Solution



	x	y	x^2	xy
	2	0	4	0
	2	1	4	2
	6	2	36	12
	8	3	64	24
	10	3	100	30
$\sum$	28	9	208	68

In the last line of the table we have the sum of the numbers in each column. Using them we compute:

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 = 208 - \frac{1}{5}(28)^2 = 51.2 \tag{10.4.10}$$

$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) = 68 - \frac{1}{5}(28)(9) = 17.6 \tag{10.4.11}$$

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{5} = 5.6 \tag{10.4.12}$$

$$\bar{y} = \frac{\sum y}{n} = \frac{9}{5} = 1.8$$

so that

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{17.6}{51.2} = 0.34375 \tag{10.4.13}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x - = 1.8 - (0.34375)(5.6) = -0.125 \tag{10.4.14}$$

Finally

- The least squares regression line for these data is:

$$\hat{y} = 0.34375x \ - 0.12$$

How do we actually calculate the correlation coefficient?

$$r = \frac{\sum\left[\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)\right]}{\sqrt{\Sigma\left(x_i - \overline{x}\right)^2 * \Sigma\left(y_i - \overline{y}\right)^2}}$$

Ice Cream Sales (X)	Temperature °F (Y)
3	70
6	75
9	80

How do we actually calculate the correlation coefficient?

$$r = \frac{\sum \left[(x_i - \overline{x}) \, (y_i - \overline{y}) \right]}{\sqrt{\Sigma(x_i - \overline{x})^2 \, * \, \Sigma(y_i - \overline{y})^2}}$$

Calculate sample means first:

$$\bar{x} = \frac{3 + 6 + 9}{3} = 6$$

$$\bar{y} = \frac{70 + 75 + 80}{3} = 75$$

Ice Cream Sales (X)	Temperature °F (Y)
3	70
6	75
9	80

Calculate the distance of each datapoint from its mean

$$r = \frac{\sum \left[\left(x_i - \overline{x} \right) \left(y_i - \overline{y} \right) \right]}{\sqrt{\Sigma \left(x_i - \overline{x} \right)^2 * \Sigma \left(y_i - \overline{y} \right)^2}}$$

Calculate sample means first:

$$\bar{x} = \frac{3 + 6 + 9}{3} = 6$$

$$\bar{y} = \frac{70 + 75 + 80}{3} = 75$$

Ice Cream Sales (X)	Temperature °F (Y)	$x_i - \bar{x}$	$y_i - \bar{y}$
3	70		
6	75		
9	80		

Calculate the distance of each datapoint from its mean

$$r = \frac{\sum\left[\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)\right]}{\sqrt{\Sigma\left(x_i - \overline{x}\right)^2 * \Sigma\left(y_i - \overline{y}\right)^2}}$$

Calculate sample means first:

$$\bar{x} = \frac{3 + 6 + 9}{3} = 6$$

$$\bar{y} = \frac{70 + 75 + 80}{3} = 75$$

Ice Cream Sales (X)	Temperature °F (Y)	$x_i - \bar{x}$	$y_i - \bar{y}$
3	70	3 − 6 = −3	70 − 75= −5
6	75	6 − 6 = 0	75 − 75= 0
9	80	9 − 6 = 3	80 − 75 = 5

Calculate the distance of each datapoint from its mean

$$r = \frac{\sum\left[\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)\right]}{\sqrt{\Sigma\left(x_i - \overline{x}\right)^2 * \Sigma\left(y_i - \overline{y}\right)^2}}$$

Calculate sample means first:

$$\bar{x} = \frac{3+6+9}{3} = 6$$

$$\bar{y} = \frac{70+75+80}{3} = 75$$

Ice Cream Sales (X)	Temperature °F (Y)	$x_i - \bar{x}$	$y_i - \bar{y}$
3	70	3 – 6 = −3	70 – 75= −5
6	75	6 – 6 = 0	75 – 75= 0
9	80	9 – 6 = 3	80 – 75 = 5

$$r = \frac{30}{\sqrt{18 * 50}} = \frac{30}{30} = 1$$

Calculate the correlation coefficient

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$x_i - \bar{x} * y_i - \bar{y}$
2	0	-3.6	-1.8	6.48
2	1	-3.6	-0.8	2.88
6	2	0.4	0.2	0.08
8	3	2.4	1.2	2.88
10	3	4.4	1.2	5.28
Total				17.6

R = (17.6) / sqrt(((-3.6)^2 + (-3.6)^2 + (0.4)^2 + (2.4)^2 + (4.4)^2) * ((-1.8)^2 + (-0.8)^2 + (0.2)^2 + (1.2)^2 + (1.2)^2))

R = (17.6) / sqrt(348.16) = 0.9432422182838

We can use this R (r) value to calculate the slope

- R = (17.6) / sqrt(348.16) = 0.9432422182838

Slope = $\frac{s_y}{s_x} * R = \frac{\sqrt{\frac{6.76}{5-1}}}{\sqrt{\frac{51.2}{5-1}}} * 0.94324 = 0.34375\ (approx.)$

This is the same value for the slope that we computed in the previous lecture, however we calculated that using SS_{xx} and SS_{xy}.

Note: I left out the calculation procedure for st-deviations → s_x and s_y

Calculate slope and intercept

“x” Hours of Sunshine	"y" Ice Creams Sold
2	4
3	5
5	7
7	10
9	15

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x$$

Slope (b_1) = 1.5183
Intercept (b_0) = 0.3049

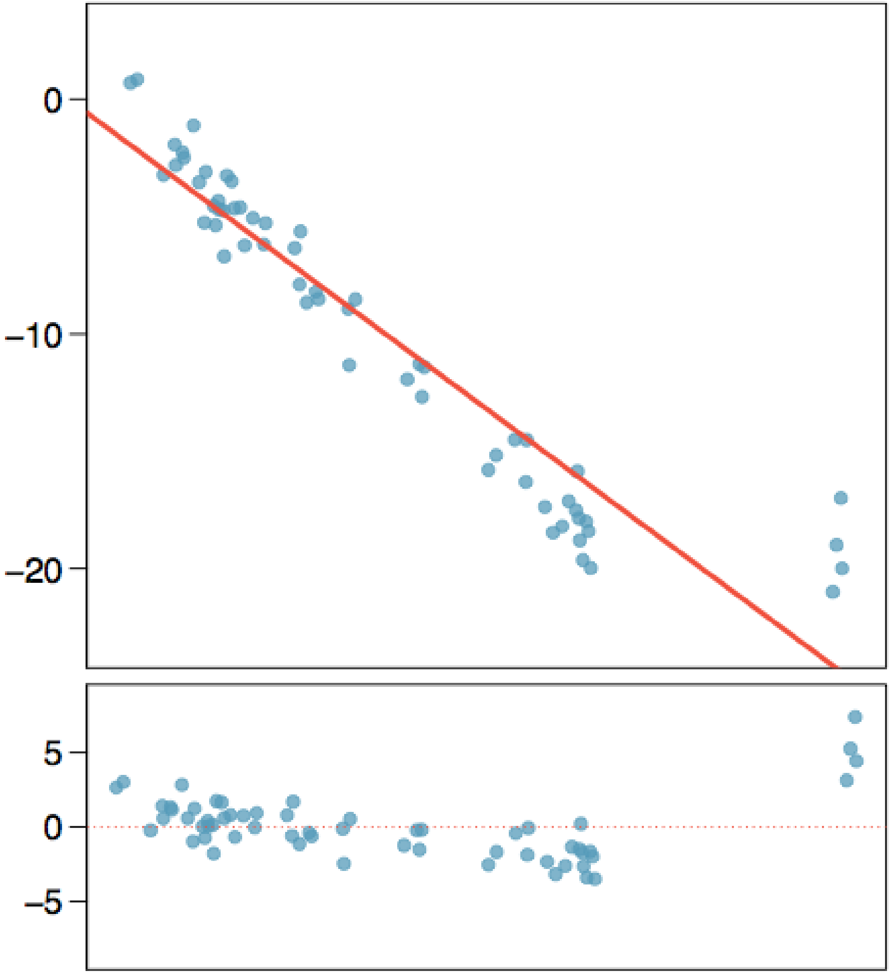
Source: https://www.mathsisfun.com/data/least-squares-regression.html

Types of outliers in linear regression

Types of outliers

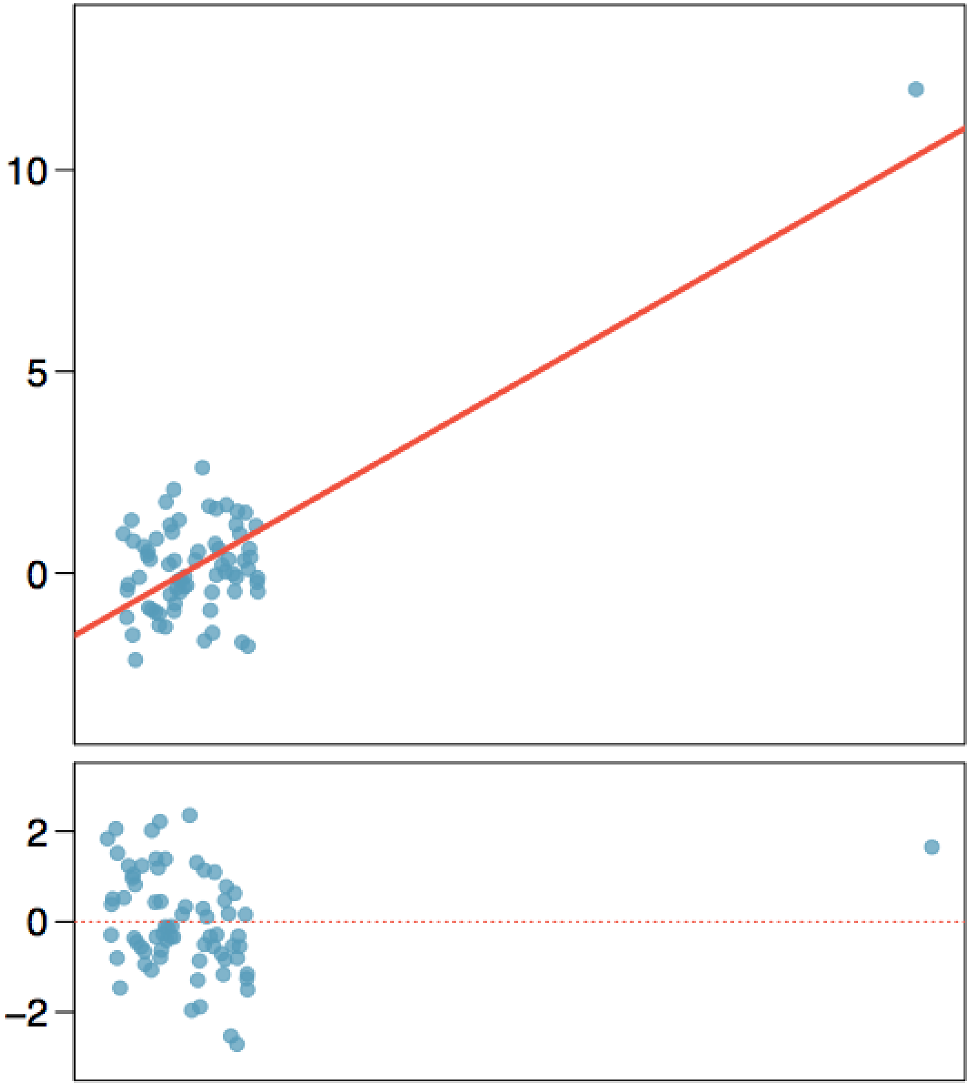
How do outliers influence the least squares line in this plot?

To answer this question think of where the regression line would be with and without the outlier(s). Without the outliers the regression line would be steeper, and lie closer to the larger group of observations. With the outliers the line is pulled up and away from some of the observations in the larger group.



Types of outliers

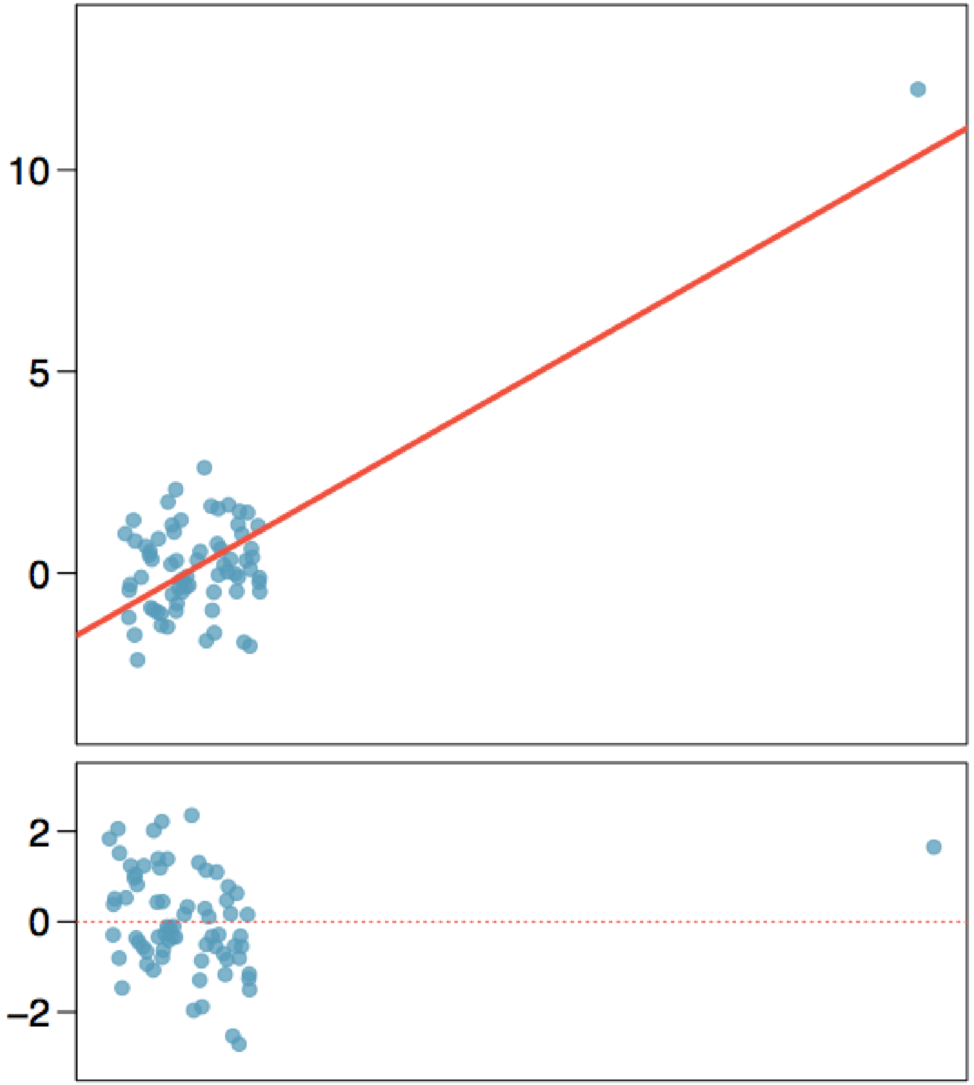
How do outliers influence the least squares line in this plot?



Types of outliers

How do outliers influence the least squares line in this plot?

Without the outlier there is no evident relationship between *x* and *y*.



Some terminology

- *Outliers* are points that lie away from the cloud of points.

Some terminology

- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called *high leverage* points.

Some terminology

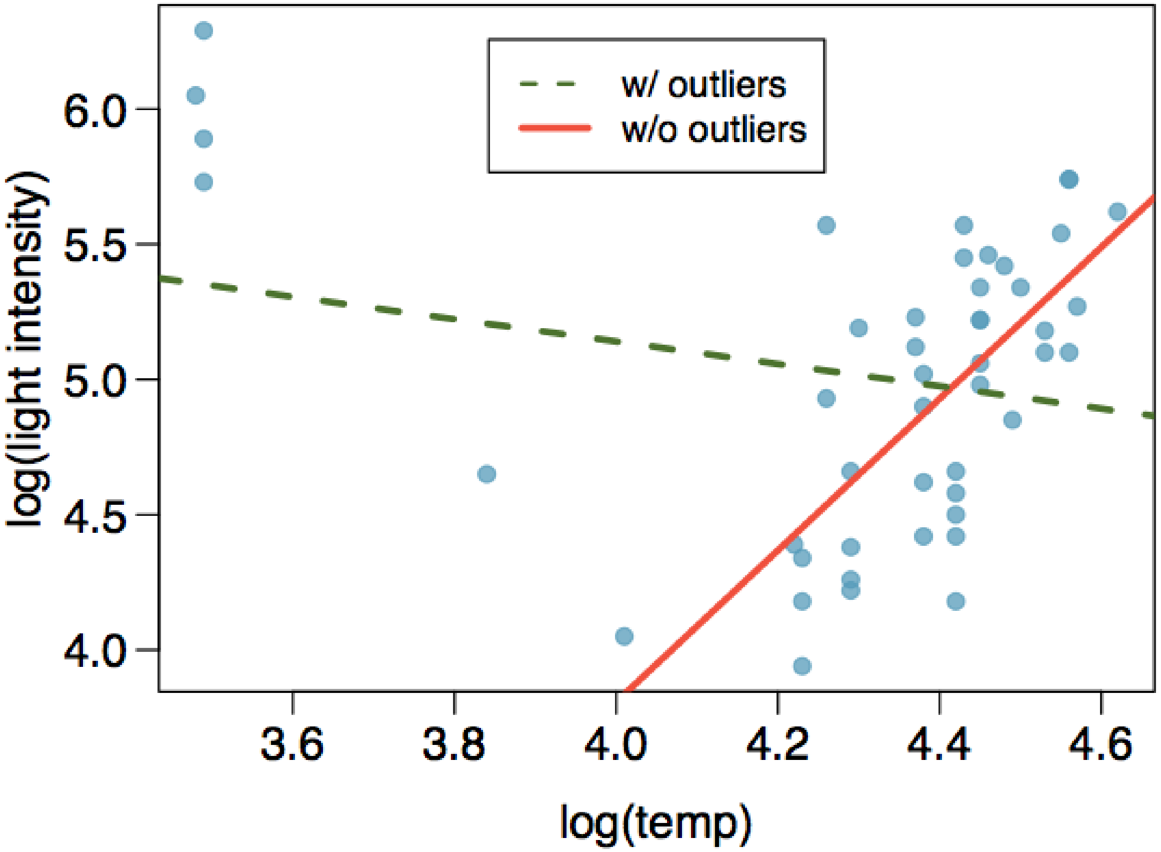
- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called *high leverage* points.
- High leverage points that actually influence the slope of the regression line are called *influential* points.

Some terminology

- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called *high leverage* points.
- High leverage points that actually influence the <u>slope</u> of the regression line are called *influential* points.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then it’s not an influential point.

Influential points

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.



Types of outliers

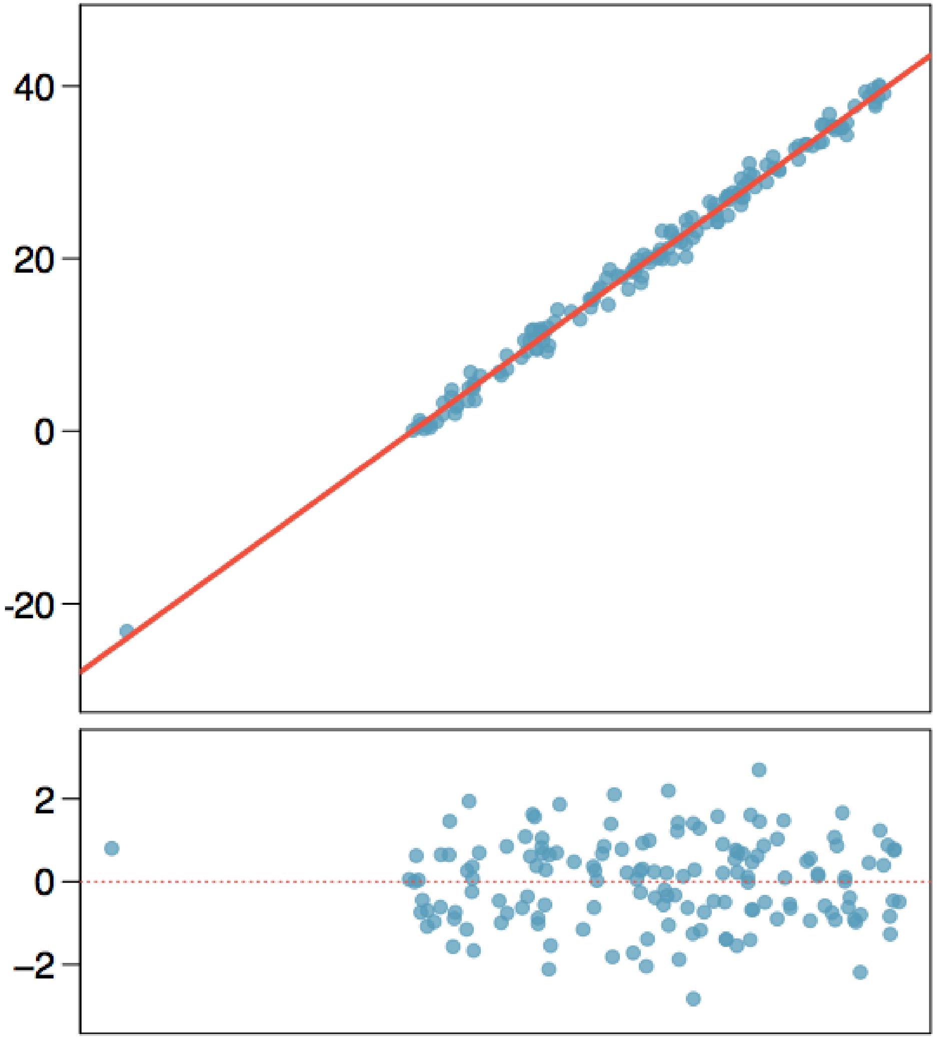
Which of the below best describes the outlier?

(a) influential

(b) high leverage

(c) none of the above

(d) there are no outliers



Types of outliers

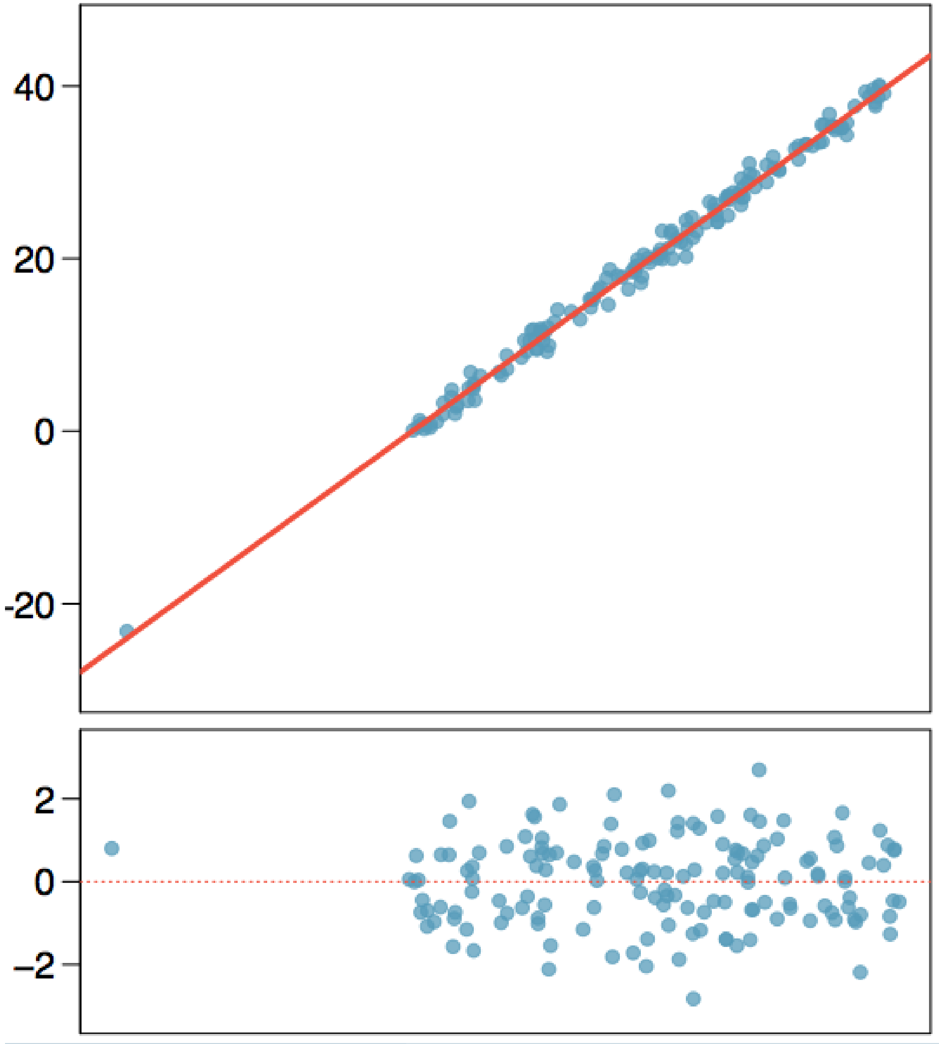
Which of the below best describes the outlier?

(a) influential

(b) high leverage

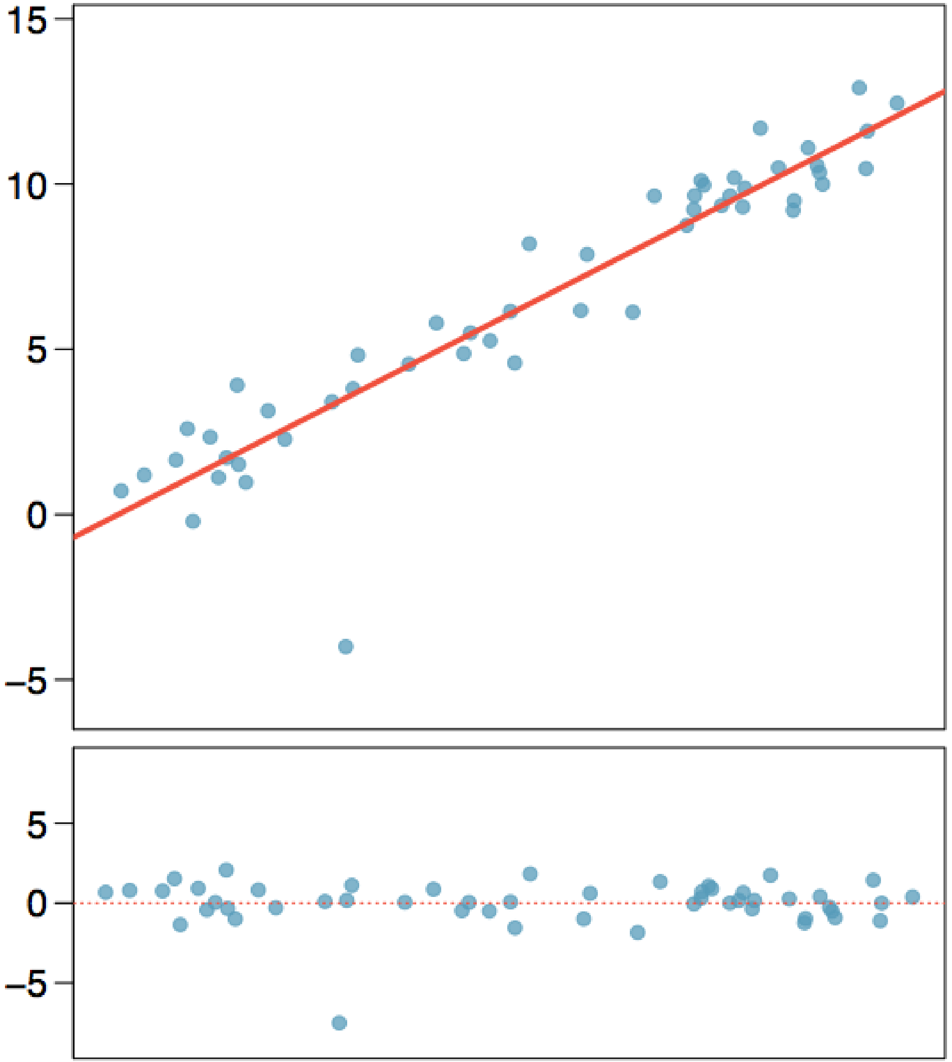
(c) none of the above

(d) there are no outliers



Types of outliers

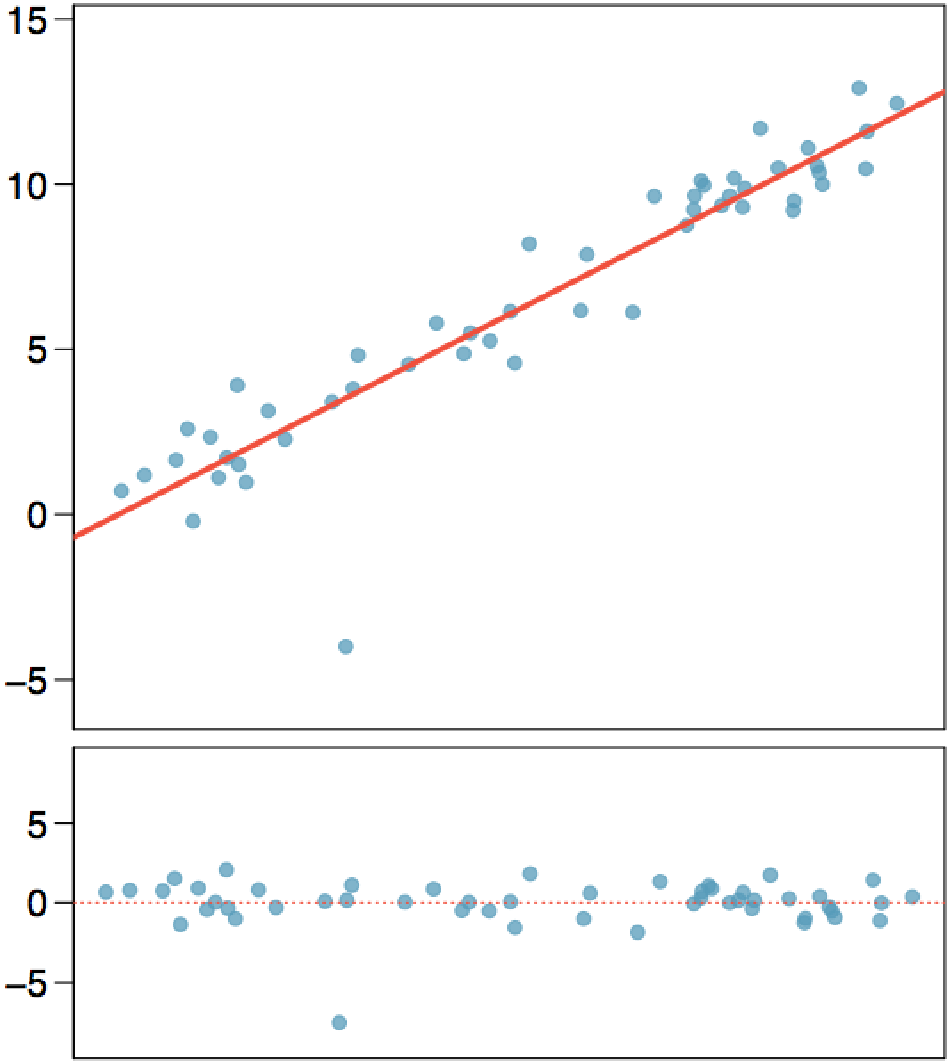
Does this outlier influence the slope of the regression line?



Types of outliers

Does this outlier influence the slope of the regression line?

Not much…



Recap

Which of following is true?

(a) Influential points always change the intercept of the regression line.

(b) Influential points always reduce R^2.

(c) It is much more likely for a low leverage point to be influential, than a high leverage point.

(d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

(e) None of the above.

Recap

Which of following is true?

(a) Influential points always change the intercept of the regression line.

(b) Influential points always reduce R^2.

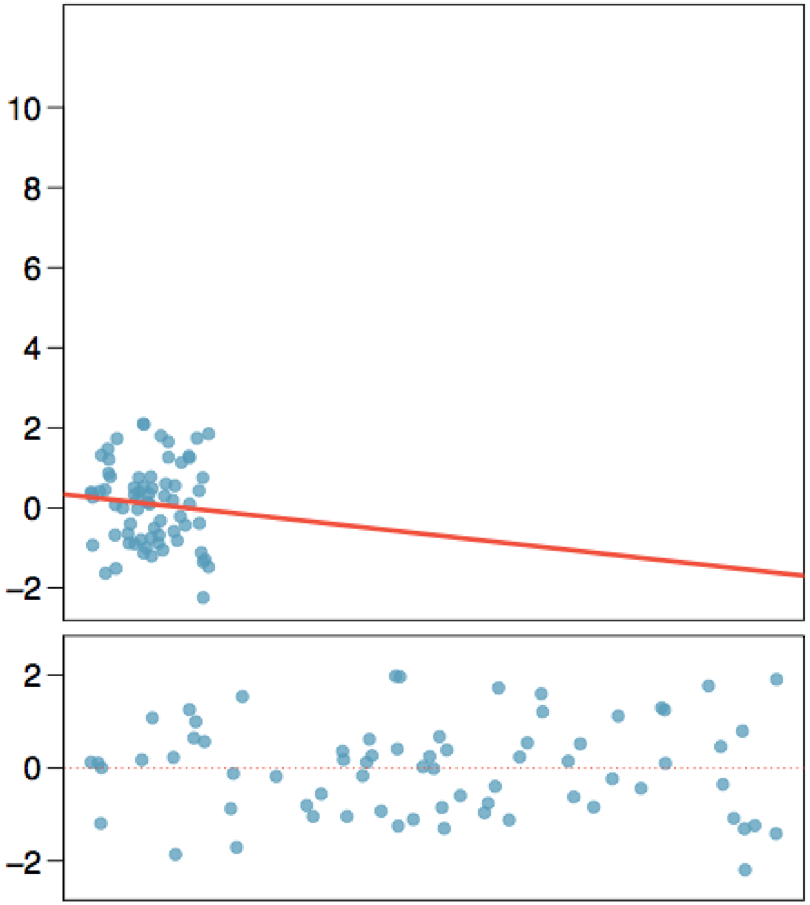
(c) It is much more likely for a low leverage point to be influential, than a high leverage point.

(d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

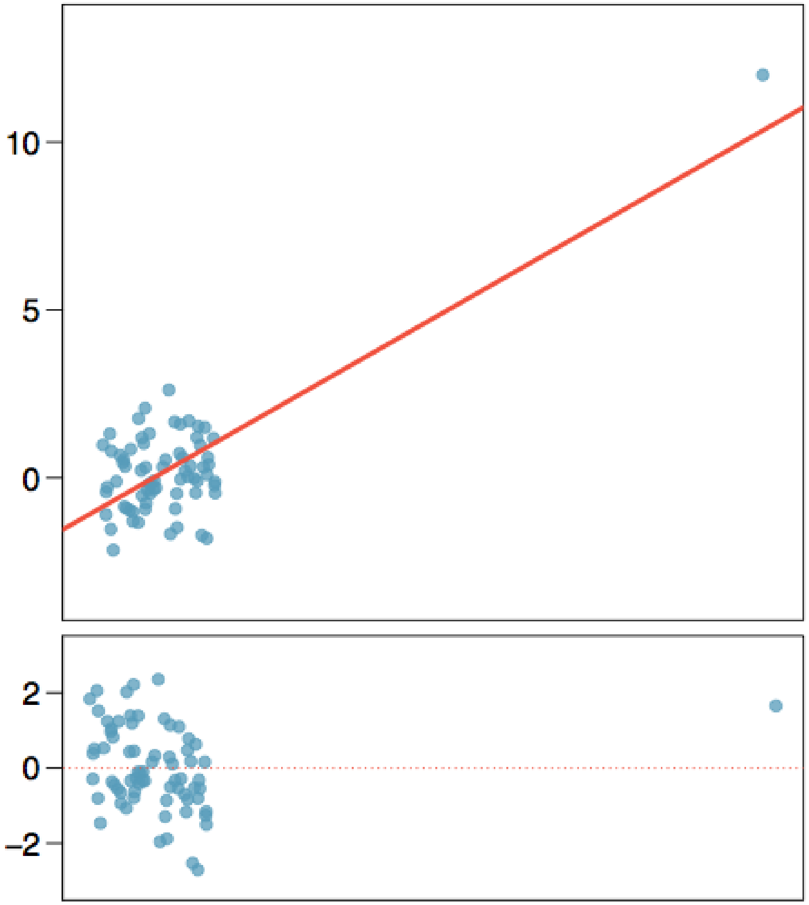
(e) None of the above.

Recap (cont.)

$R = 0.08, R^2 = 0.0064$



$R = 0.79, R^2 = 0.6241$



Sources

- openintro.org/os (Chapter 8, Section 8.3)

Helpful Links:

- https://www.jmp.com/en_sg/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html
- https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/10%3A_Correlation_and_Regression/10.04%3A_The_Least_Squares_Regression_Line