Chapter 8

Introduction to linear regression[1]

Department of Mathematics & Statistics
North Carolina A&T State University

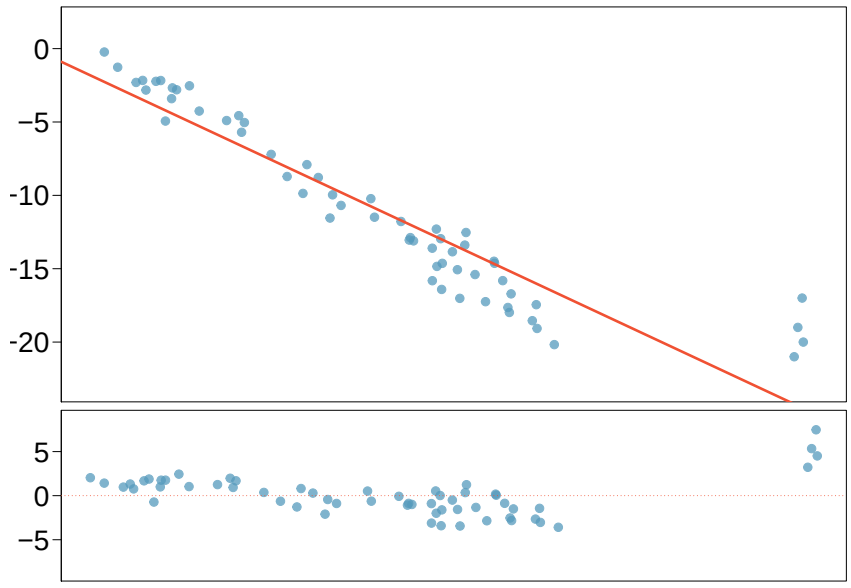
[1] These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

Types of outliers in linear regression

Types of outliers

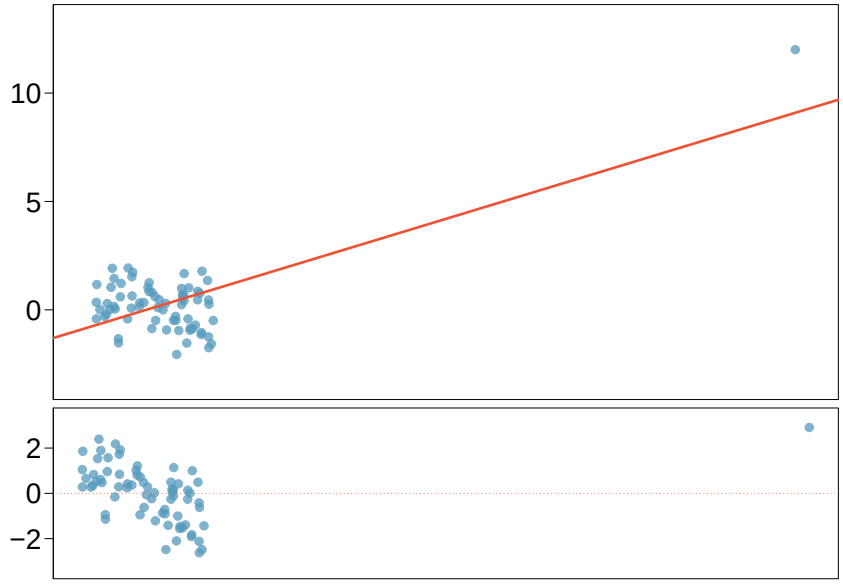
How do outliers influence the least squares in this plot?

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

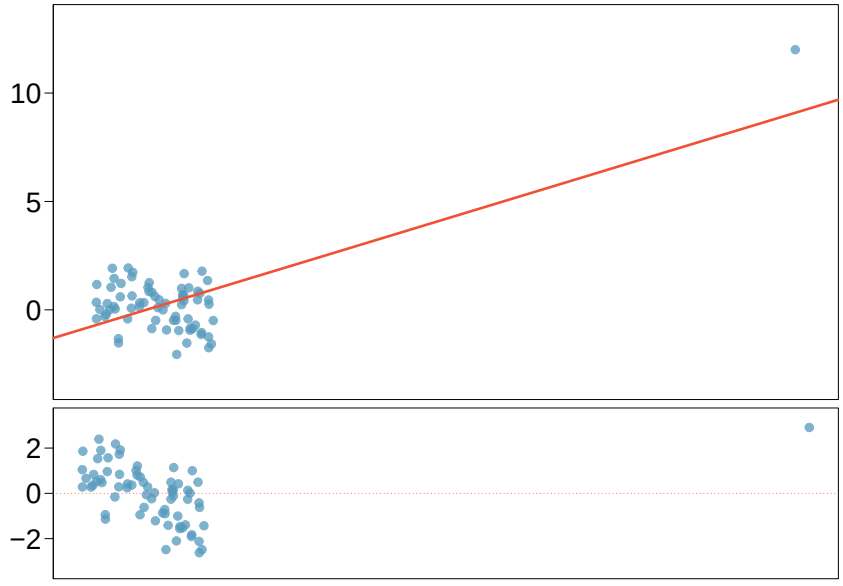
How do outliers influence the least squares in this plot?



Types of outliers

How do outliers influence the least squares 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

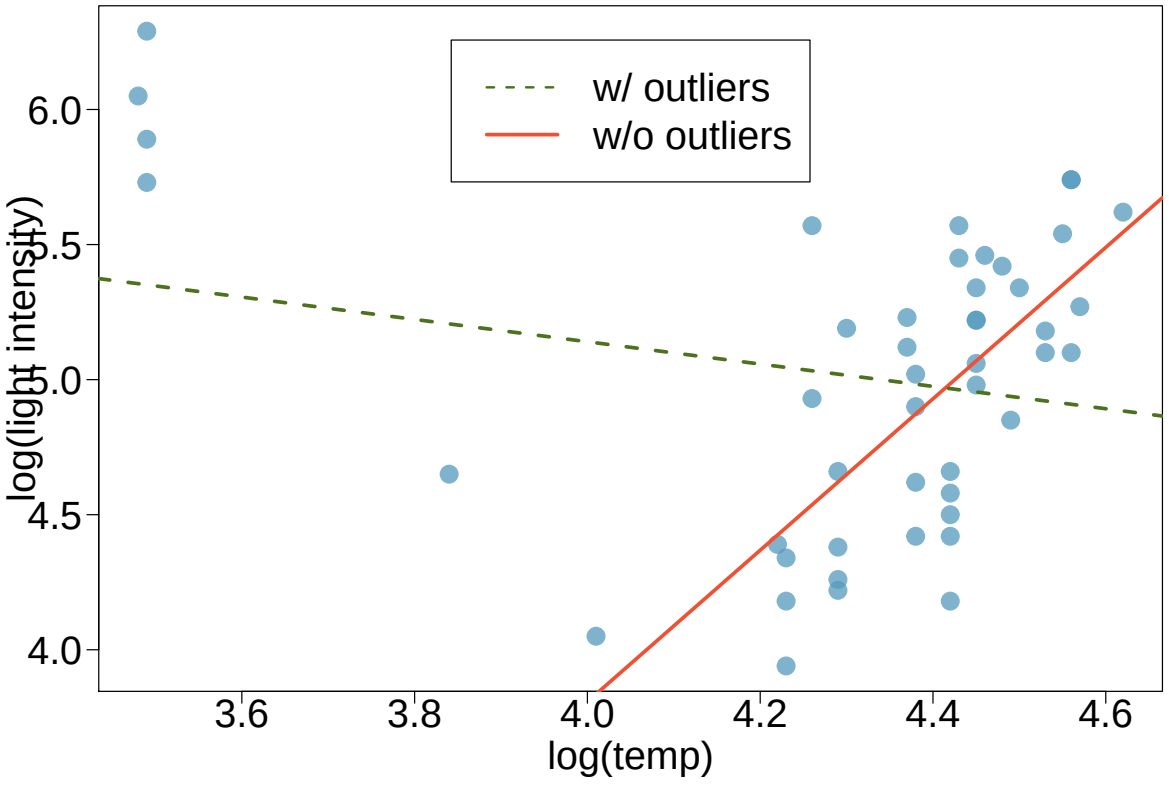
- **Outliers** are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called **high leverage** points.
- High leverage points that actually influence the slope of the regression line are called **influential** points.

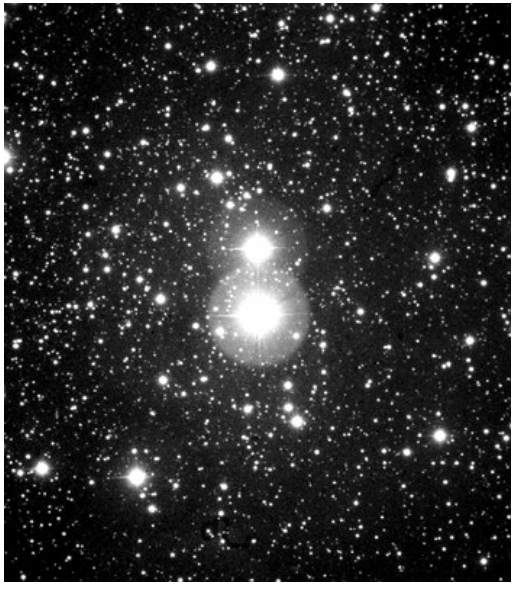
Some terminology

- **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 slope 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.

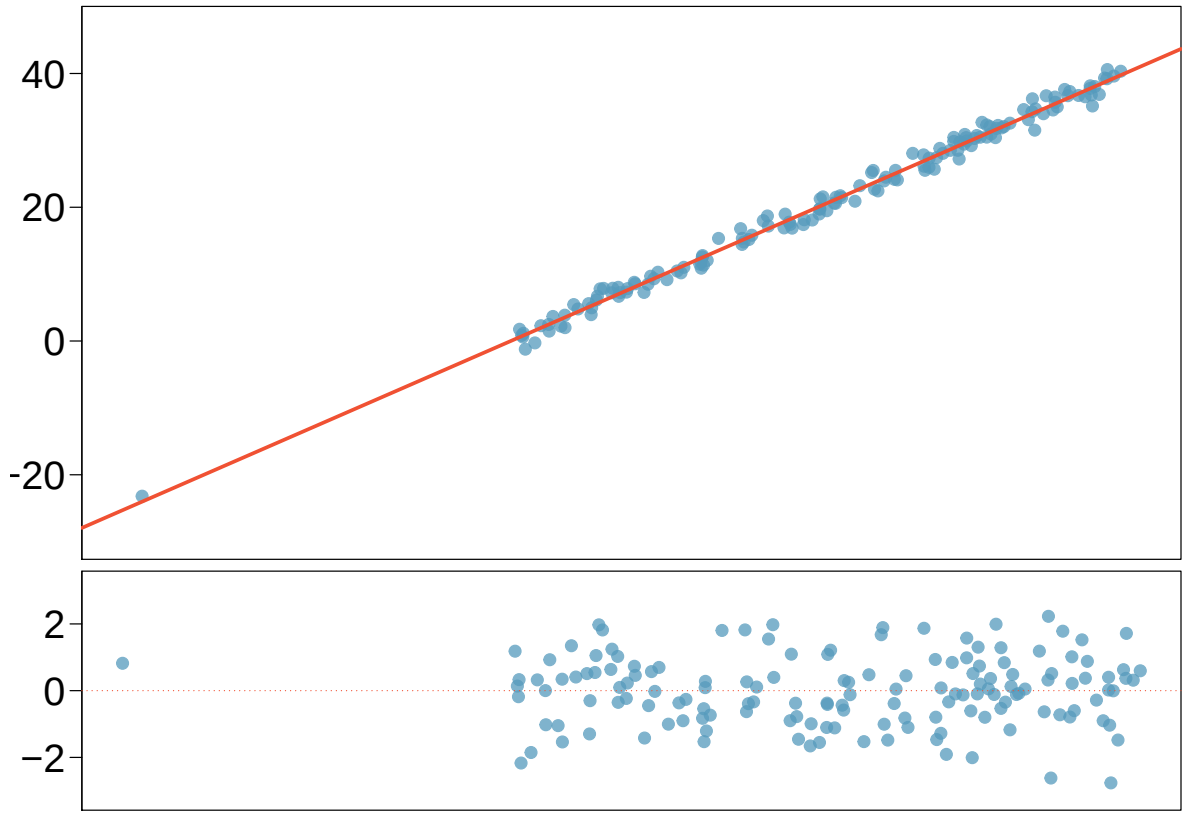




Types of outliers

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

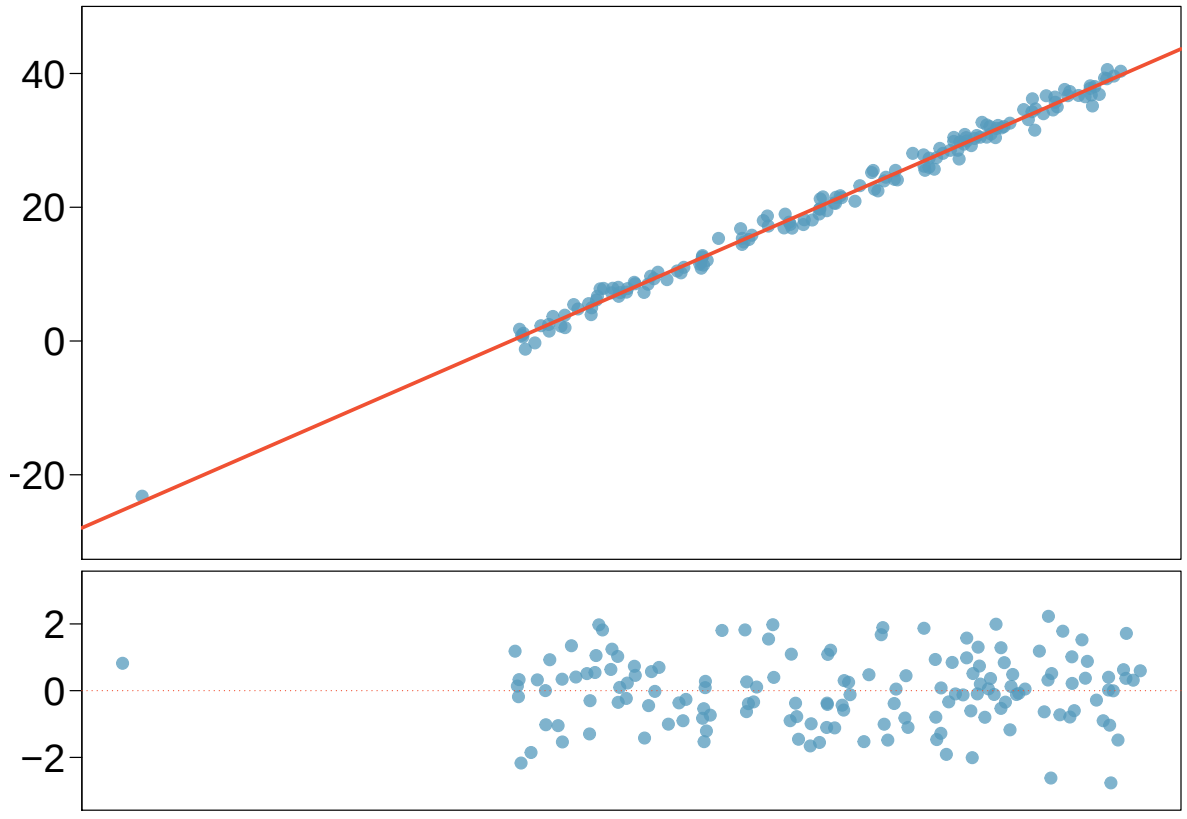
Which of the below best describes the outlier?

A) Influential

B) High Leverage

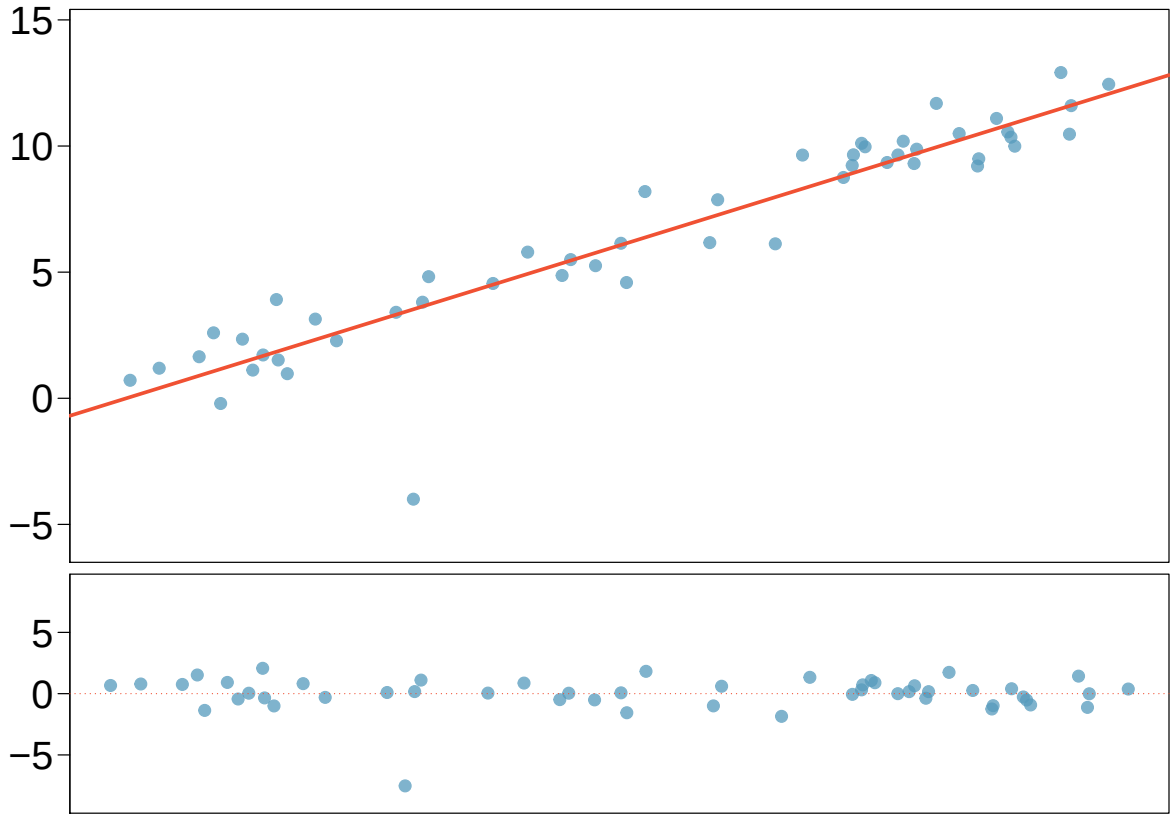
C) None of the above

D) There are no outliers



Types of outliers

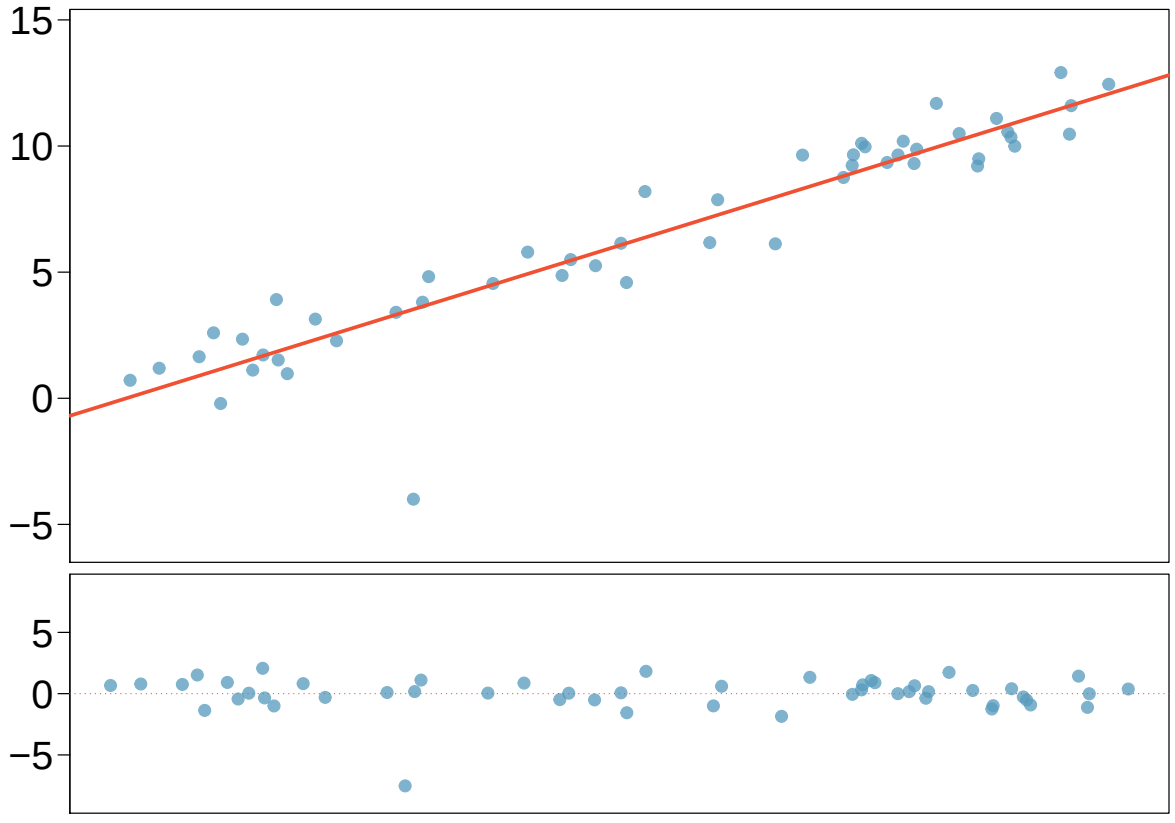
Does this outlier influence the slope of the regression line?



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.

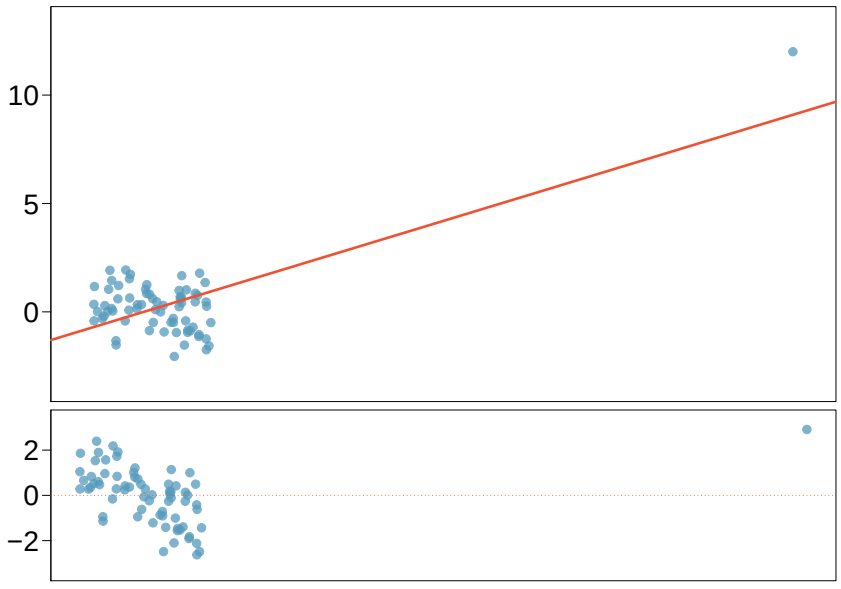
C) It is much more likely for a low leverage point to be influential, than a high leverage point.

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

$R = 0.08, R^2 = 0.0064$



$R = 0.79, R^2 = 0.6241$