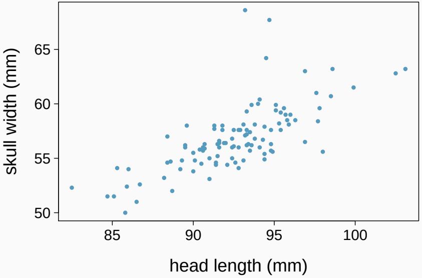
Announcements

- HW3 due tonight
- Lab 3 write-up due tonight as well
- I'm behind in grading here, and trying to get caught up. My apologies for the delay
- In-class lab on Wednesday!
- Read Ch 6.1 for Friday

Warm Up Question



	Mean	sd
head length	92.6	3.57
skull width	56.9	3.11

R=0.71

Given the data to the left, determine the best least squares linear model for how possum head lengths compare to their skull widths.

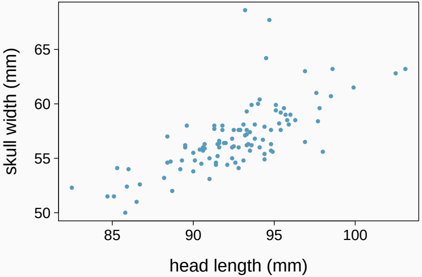
A) $\hat{y} = -0.42 + 0.619x$

B) $\hat{y} = 114.2 - 0.818x$

C) $\hat{y} = -0.43 + 0.815x$

D) $\hat{y} = 0.42 + 0.871x$

Warm Up Question



	Mean	sd
head length	92.6	3.57
skull width	56.9	3.11

R=0.71

Given the data to the left, determine the best least squares linear model for how possum head lengths compare to their skull widths.

A) $\hat{y} = -0.42 + 0.619x$

B) $\hat{y} = 114.2 - 0.818x$

C) $\hat{y} = -0.43 + 0.815x$

D) $\hat{y} = 0.42 + 0.871x$

Some extra on R^2

- Tells us what percent of variability in the response variable is explained by the model.
  - Variability in the response variable is s_y
  - Variability in the model is the standard deviation of the residuals s_{res}
- The remainder variability is explained by variables not in the model or by inherent randomness

Example

For the possum heads:

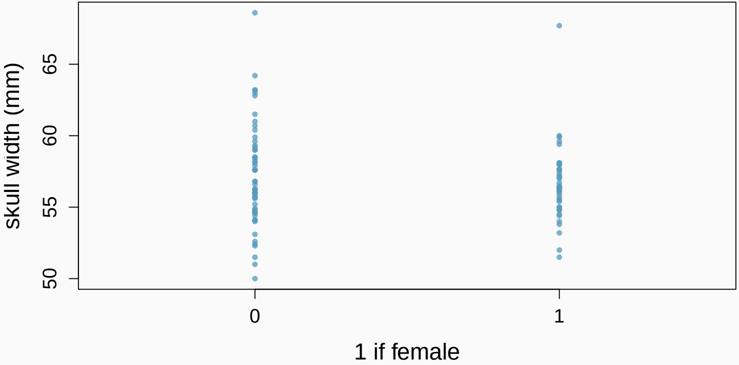
$$R = 0.71 \quad \Rightarrow \quad R^2 = 0.504$$

In terms of the variance:

$$\frac{s_y^2 - s_{res}^2}{s_y^2} = \frac{3.11^2 - 2.19^2}{3.11^2} = 0.504$$

Categorical Regression

- Can also use regression to look at relationships between two categorical values
- Use an ***indicator variable*** that takes a value of 1 for one category and 0 for the other.
- Calculating the least square model parameters then remains the same!



Fat Head Possums

- Summarizing the data:

	Mean	Sd
gender	0.413	0.495
skull width	56.9	3.11

R=-0.08

Fat Head Possums

- Summarizing the data:

	Mean	Sd
gender	0.413	0.495
skull width	56.9	3.11

R=-0.08

- Gives us:

$$\hat{y} = 57.09 – 0.503x$$

Fat Head Possums

- Summarizing the data:

	Mean	Sd
gender	0.413	0.495
skull width	56.9	3.11

R=-0.08

- Gives us:

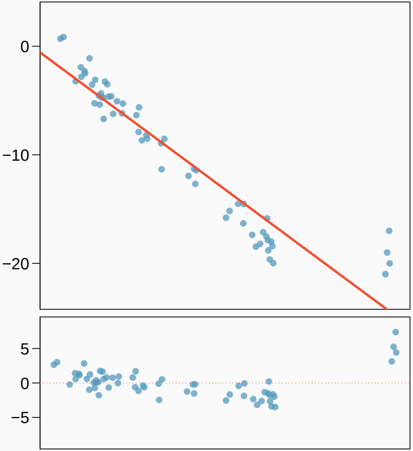
$$\hat{y} = 57.09 - 0.503x$$

Interpretation

- Males have an average skull size of 57.09 mm
- Females have an average skull size that is 0.503 mm smaller than males
- Our model is pretty terrible at describing the variation

Types of outliers

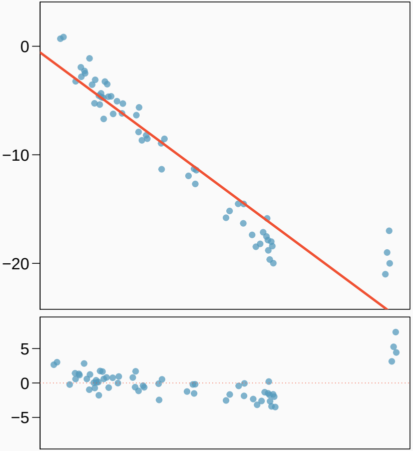
How do outliers influence the least squares line in this plot?



Types of outliers

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

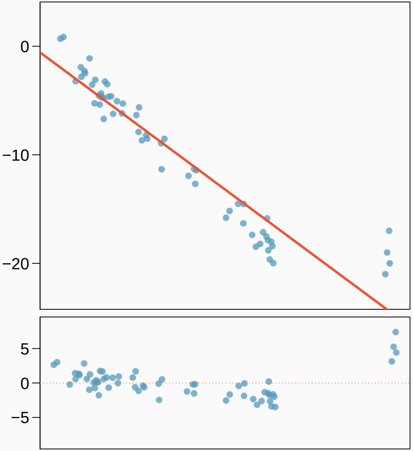
- How would the regression line change if you removed the outliers?



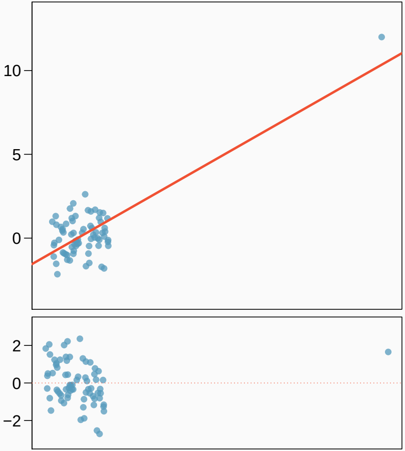
Types of outliers

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

- How would the regression line change if you removed the outliers?
- Here the slope is pulled up near the outliers more than it would be otherwise.

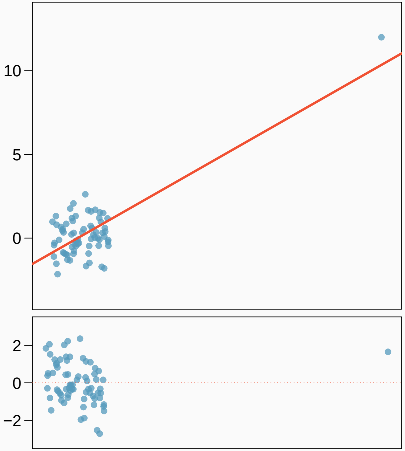


Exerting Leverage



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

Exerting Leverage



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

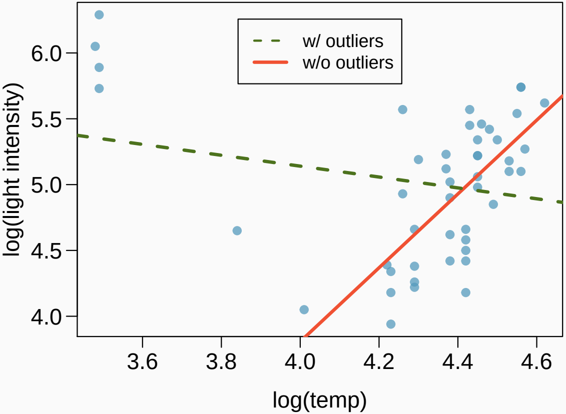
- There is not real clear linear relationship (or any relationship at all) without the outlier

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***.
- To determine if a point is influential, visualize the regression line with and without the point.
  - Does the slope change considerably? Then influential

Influential Stars

Stars generally show strong trends between temperature and light intensity. Here are 47 stars in the cluster CYG OB1.



Understanding Check

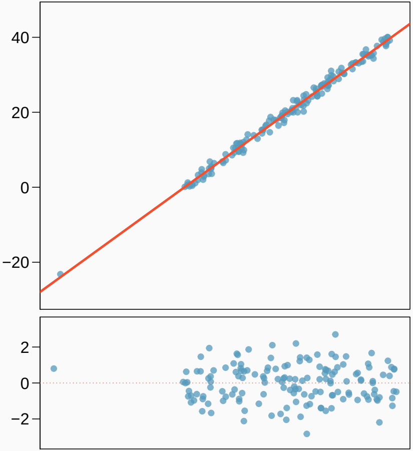
Which of the below best describes the outlier?

A) influential

B) high leverage

C) none of the above

D) there are no outliers



Understanding Check

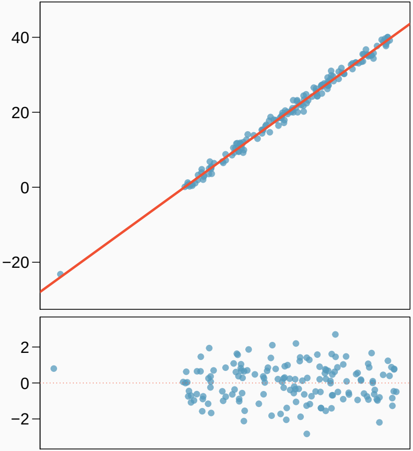
Which of the below best describes the outlier?

A) influential

B) high leverage

C) none of the above

D) there are no outliers



Summary

- You can have outliers that are neither high leverage nor influential
- An outlier won't always reduce R^2
- High leverage points are more likely to be influential than low leverage points