Sta 111 - Summer II 2017
Probability and Statistical Inference

19. Conditions for least squares regression and types of outliers

Lu Wang

Duke University, Department of Statistical Science

August 1, 2017

Outline

1. Conditions for least squares regression

2. Using R^2 to assess model fit

3. Types of outliers in linear regression

Conditions for the least squares line

When fitting a least squares line, we generally require

1. *Linearity:* The data should show a linear trend.

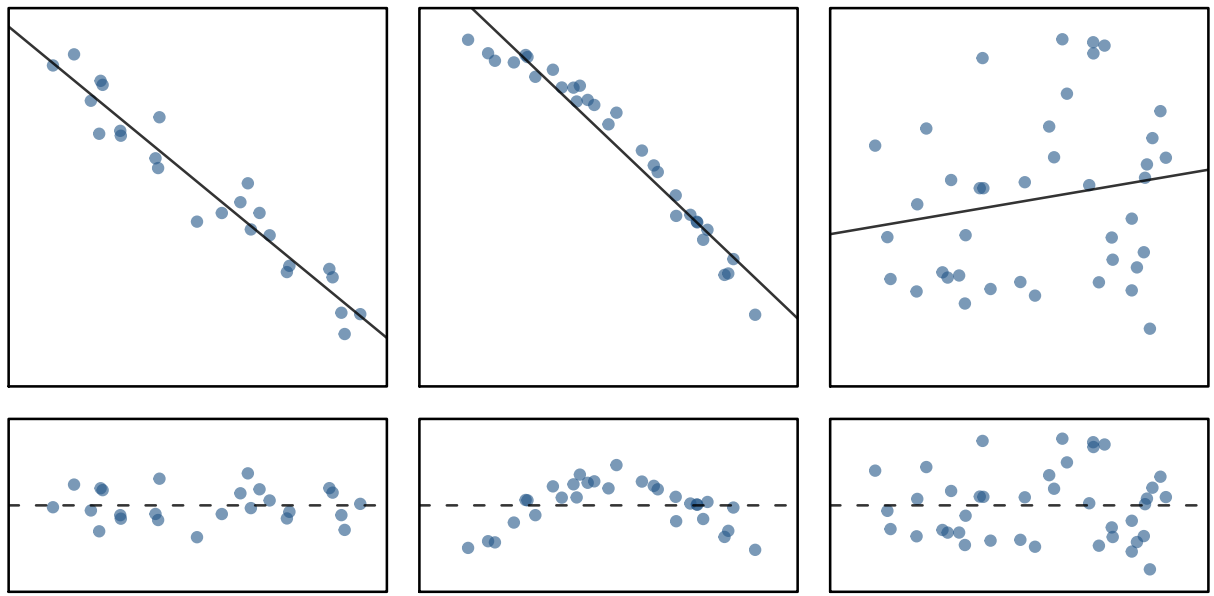
2. *Nearly normal residuals:* The residuals must be nearly normally distributed.

3. *Constant variability:* The variability of points around the least squares line remains roughly constant.

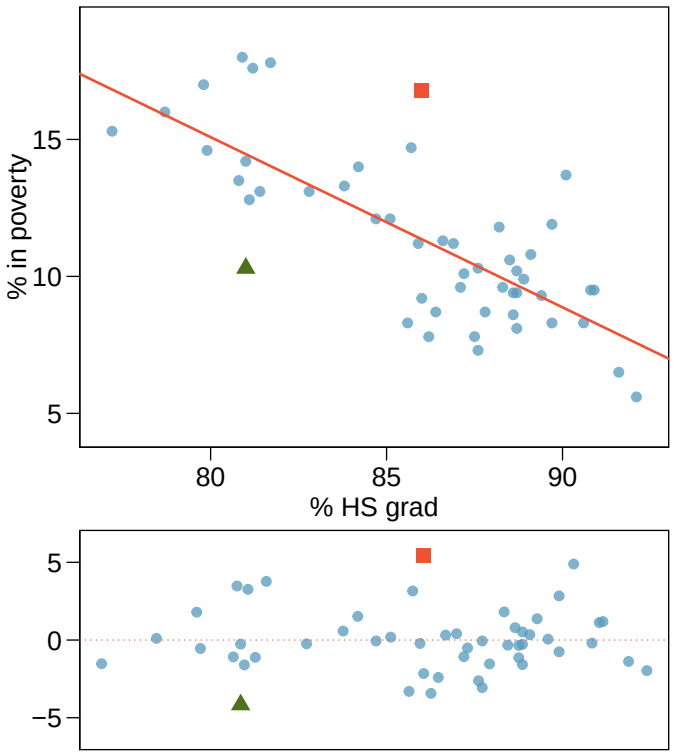
4. *Independent observations:* depends on data collection method, often violated for time-series data

Conditions: (1) Linearity

- The relationship between the explanatory and the response variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class. If this topic is of interest, an Online Extra is available on openintro.org covering new techniques.
- Check using a scatterplot of the data, or a *residuals plot* where the residuals should be scattered around 0.



Anatomy of a residuals plot



▲ *RI:*

$$\%\ \textit{HS grad} = 81 \qquad \%\ \textit{in poverty} = 10.3$$

$$\widehat{\%\ \textit{in poverty}} = 64.68 - 0.62 * 81 = 14.46$$

$$e = \%\ \textit{in poverty} - \widehat{\%\ \textit{in poverty}}$$

$$= 10.3 - 14.46 = -4.16$$

■ *DC:*

$$\%\ \textit{HS grad} = 86 \qquad \%\ \textit{in poverty} = 16.8$$

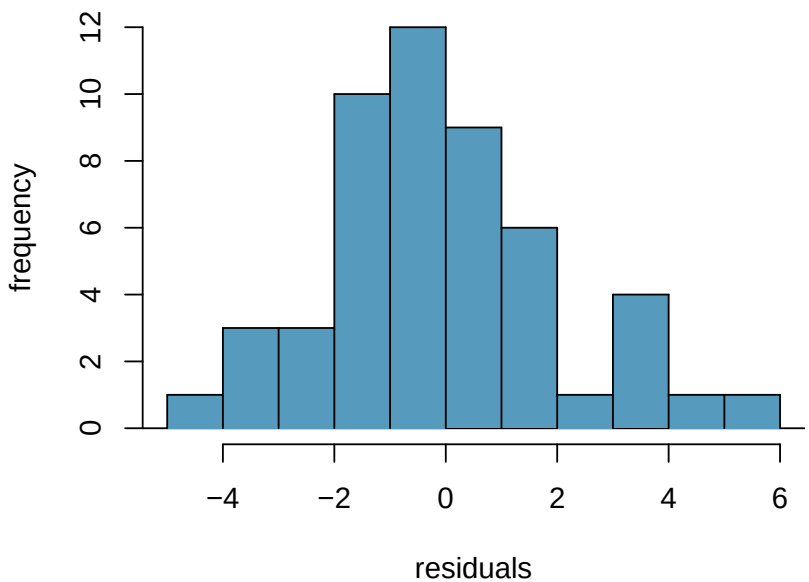
$$\widehat{\%\ \textit{in poverty}} = 64.68 - 0.62 * 86 = 11.36$$

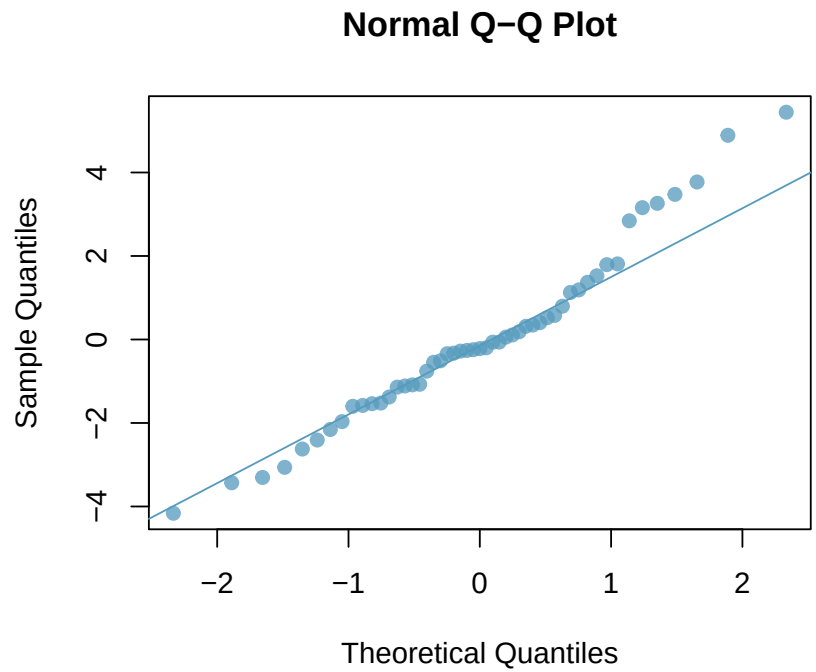
$$e = \%\ \textit{in poverty} - \widehat{\%\ \textit{in poverty}}$$

$$= 16.8 - 11.36 = 5.44$$

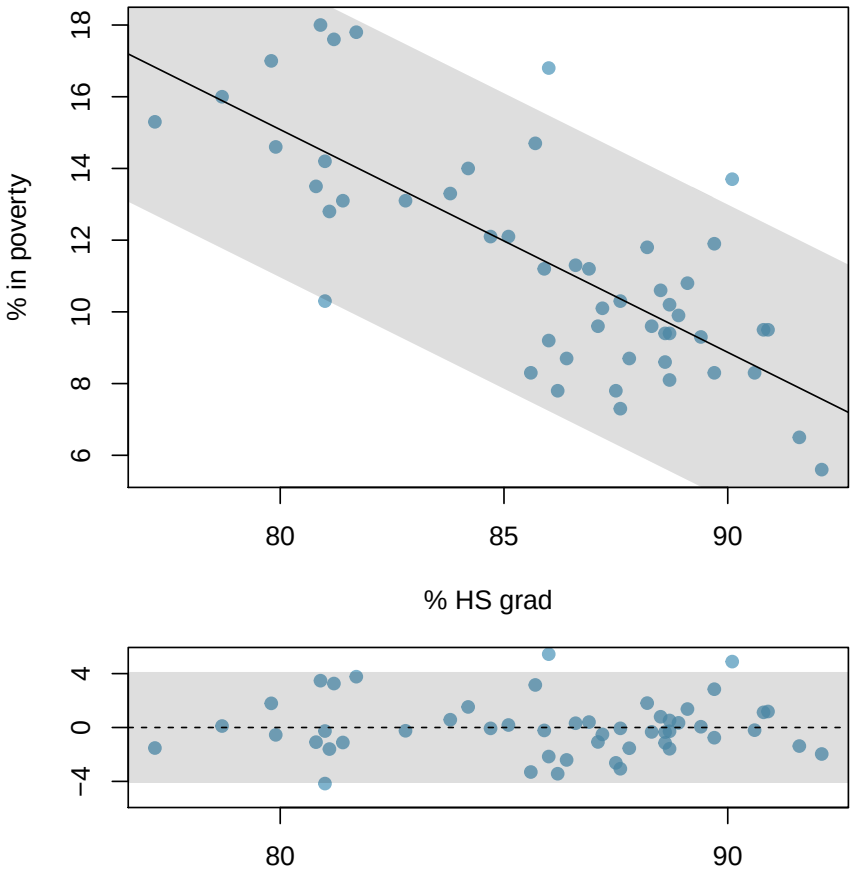
Conditions: (2) Nearly normal residuals

- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Check using a histogram or Q-Q plot of residuals.





Conditions: (3) Constant variability



- The variability of points around the least squares line should be roughly constant.
- Check using the residuals plot. The variability of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.

Checking conditions

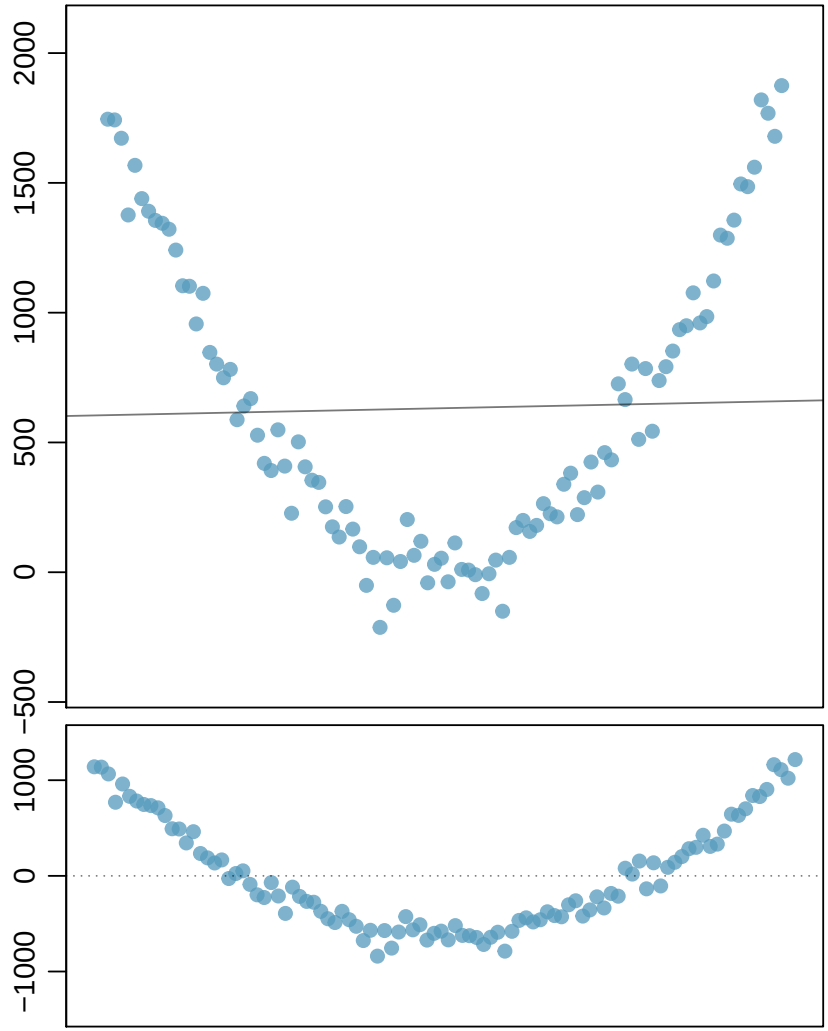
What condition is this linear model obviously violating?

(a) Constant variability

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



Checking conditions

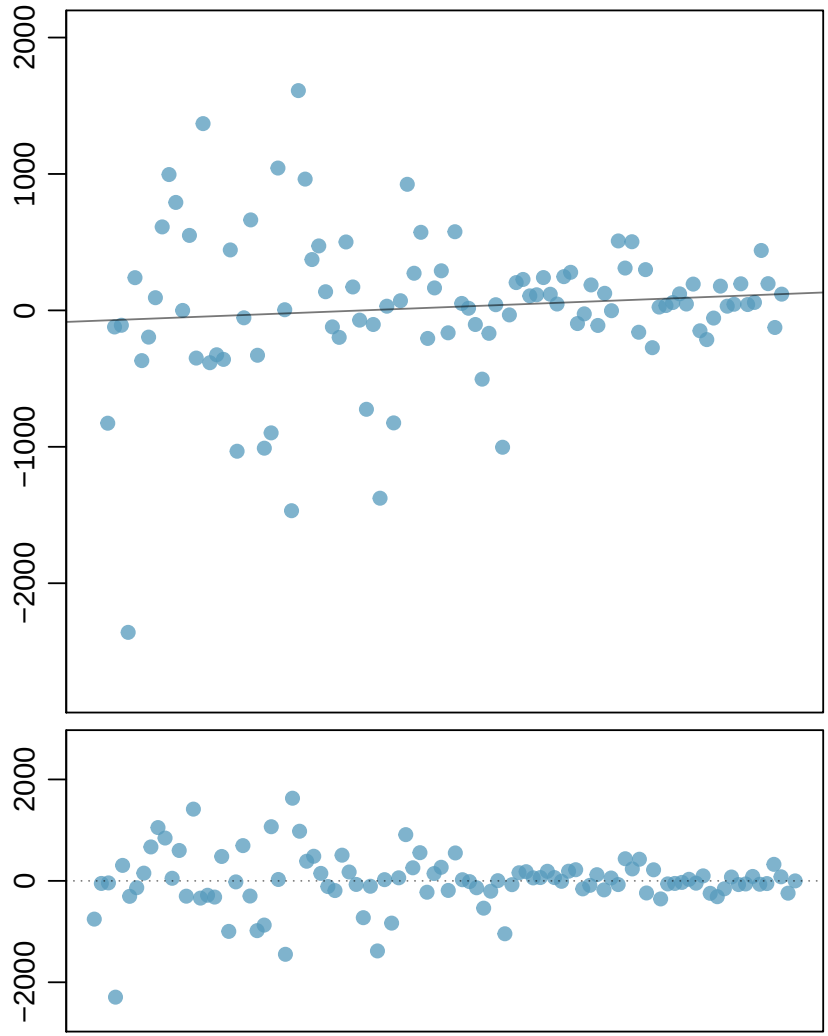
What condition is this linear model obviously violating?

(a) Constant variability

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 tells us what percent of variability in the response variable is explained by the model.

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y} = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2}$$

- For single predictor regression: R^2 equals the square of the correlation coefficient.
  - For the model we've been working with, $R^2 = (-0.62)^2 = 0.38$.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.

Interpretation of R^2

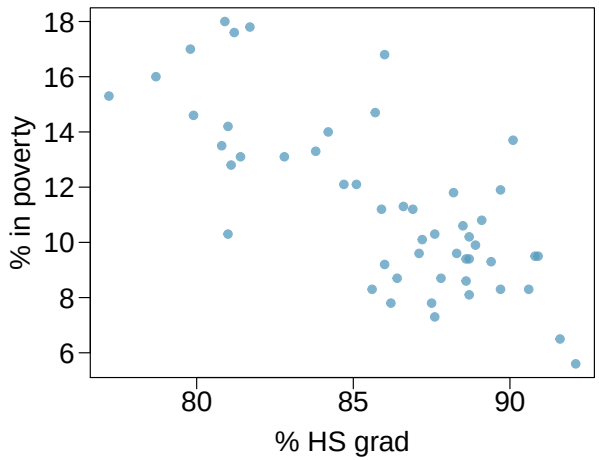
Which of the below is the correct interpretation of *corr* $= -0.62$, $R^2 = 0.38$?

(a) 38% of the variability in the % of HS graduates among the 51 states is explained by the model.

(b) 38% of the variability in the % of residents living in poverty among the 51 states is explained by the model.

(c) 38% of the time % HS graduates predict % living in poverty correctly.

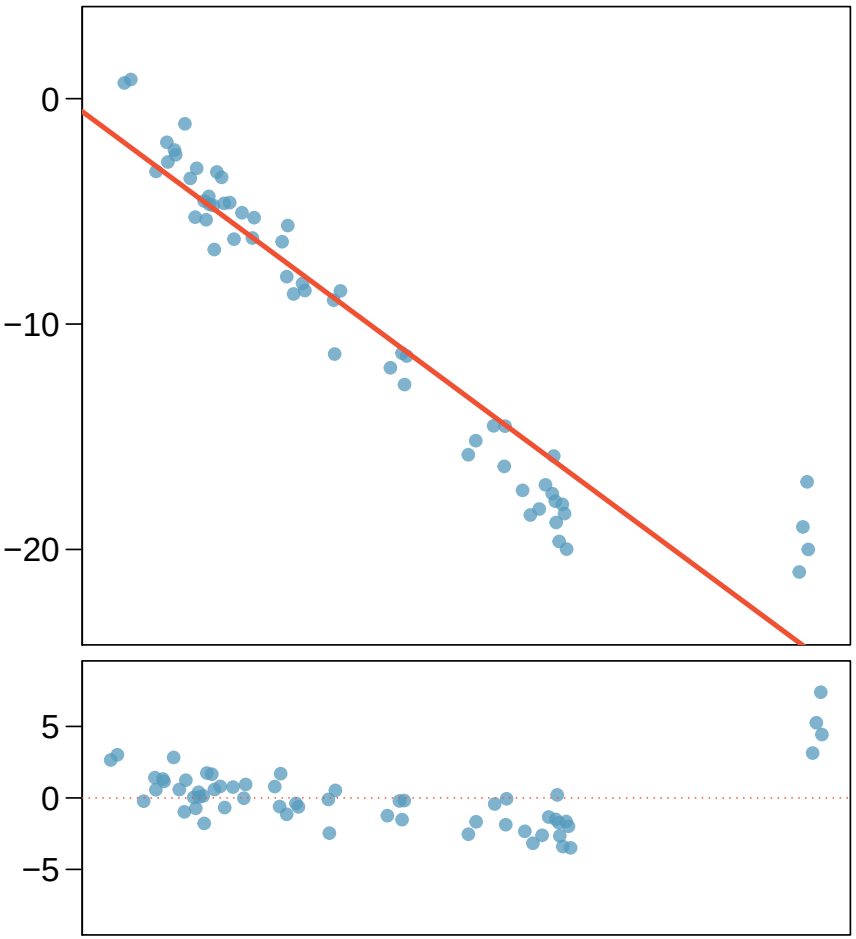
(d) 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



Types of outliers

- *Outliers* in regression are observations that fall far from the ”cloud” of points.
- These points are especially important because they can have a strong influence on the least squares line.
- We identify criteria for determining which outliers are important and influential.
- Type of outlier determines how it should be handled.

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

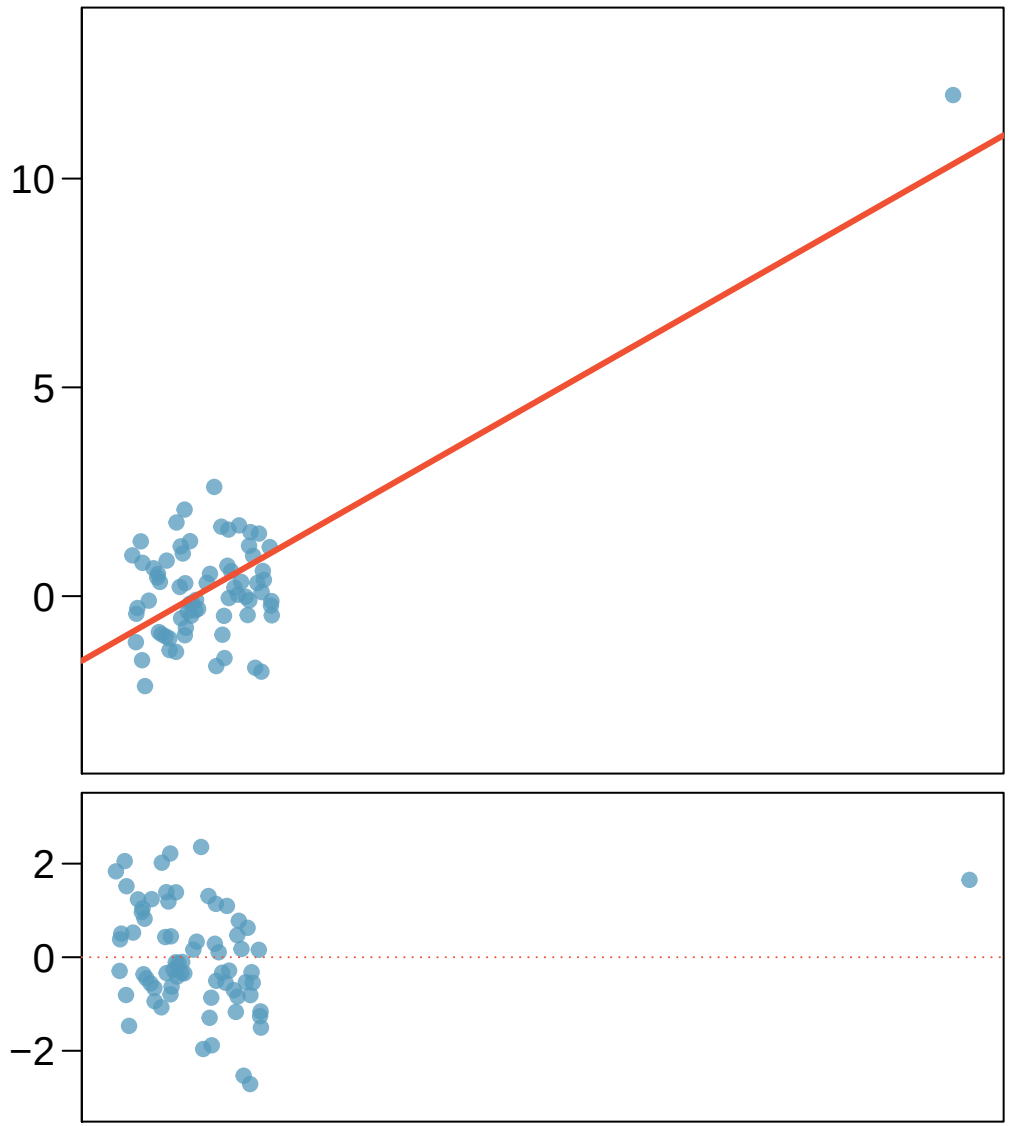


- To answer this question think of where the regression line would be with and without the outlier(s).
- Without the outliers the regression line would be steeper, and lie closer to the larger group of observations.
- With the outliers the line is pulled up and away from some of the observations in the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

Types of outliers

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

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

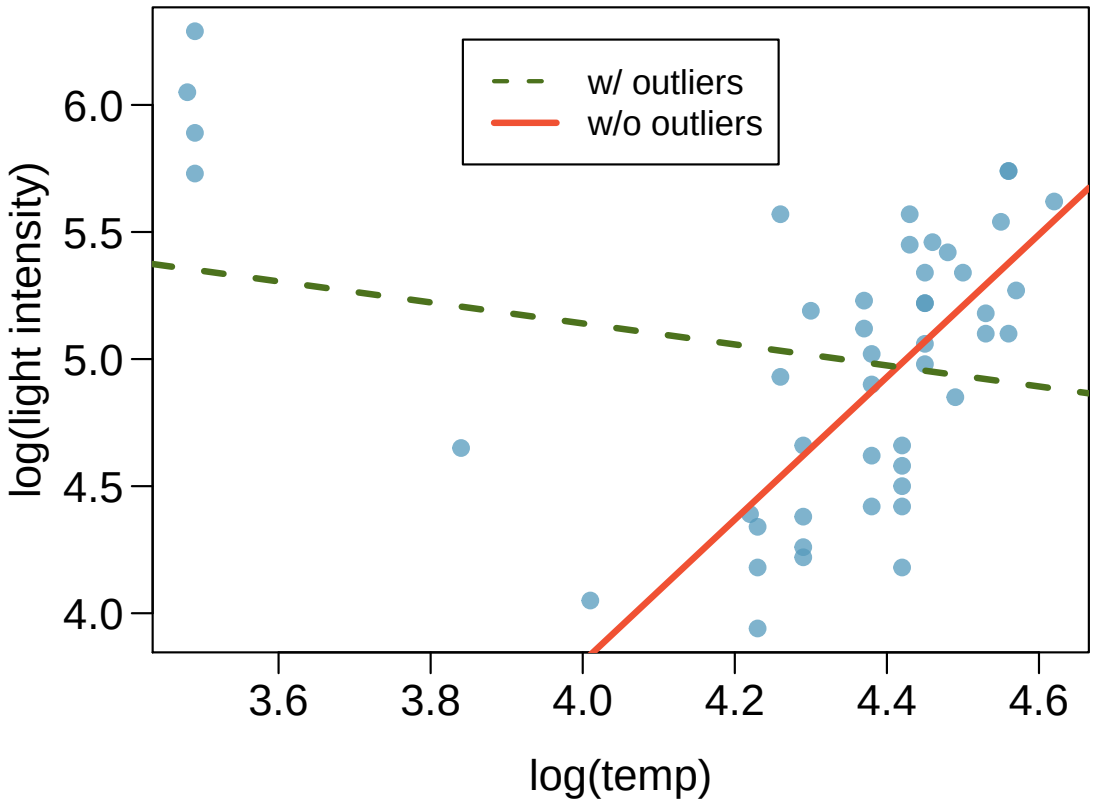


Some terminology

- 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 is 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.



Practice

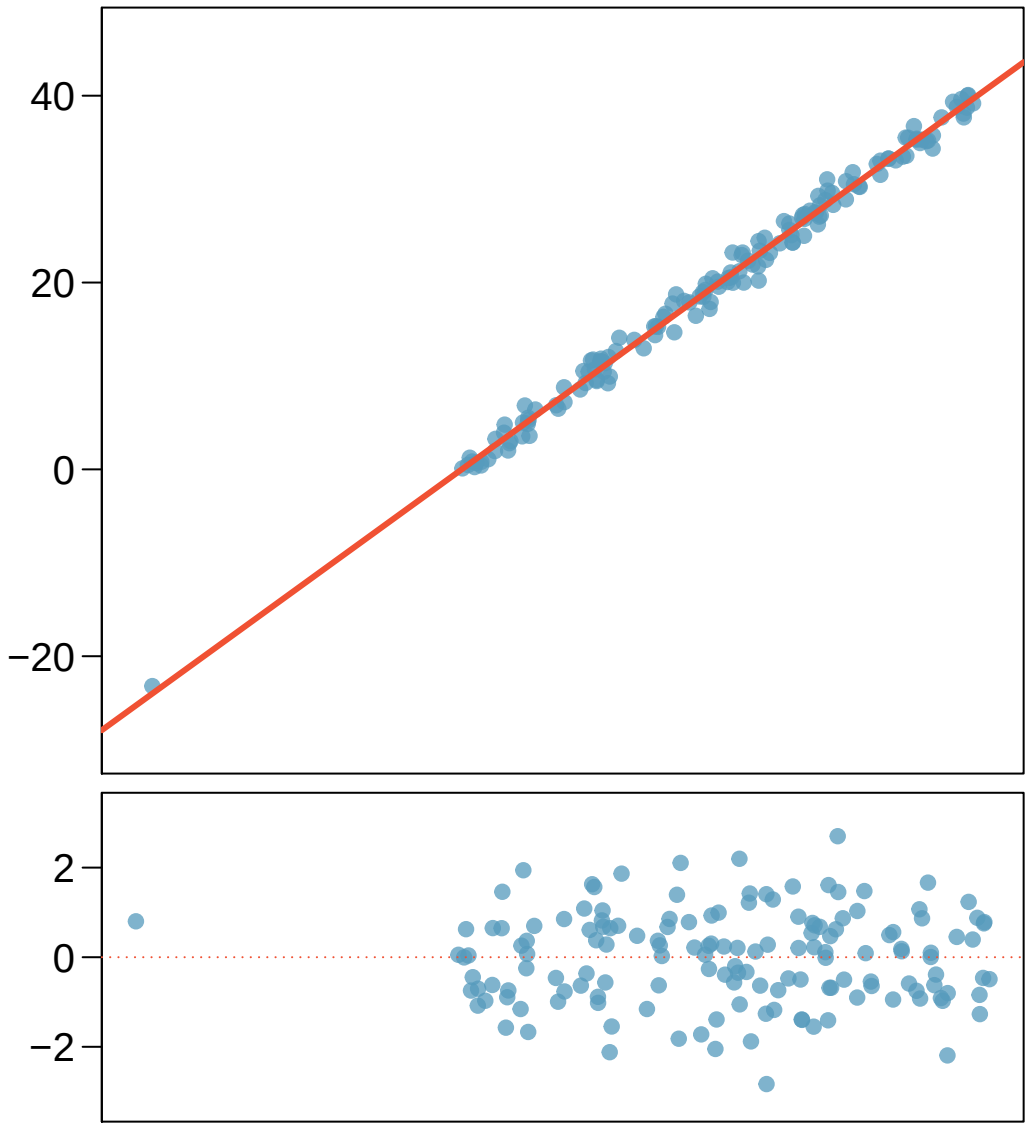
Which of the below best describes the outlier?

(a) influential

(b) high leverage

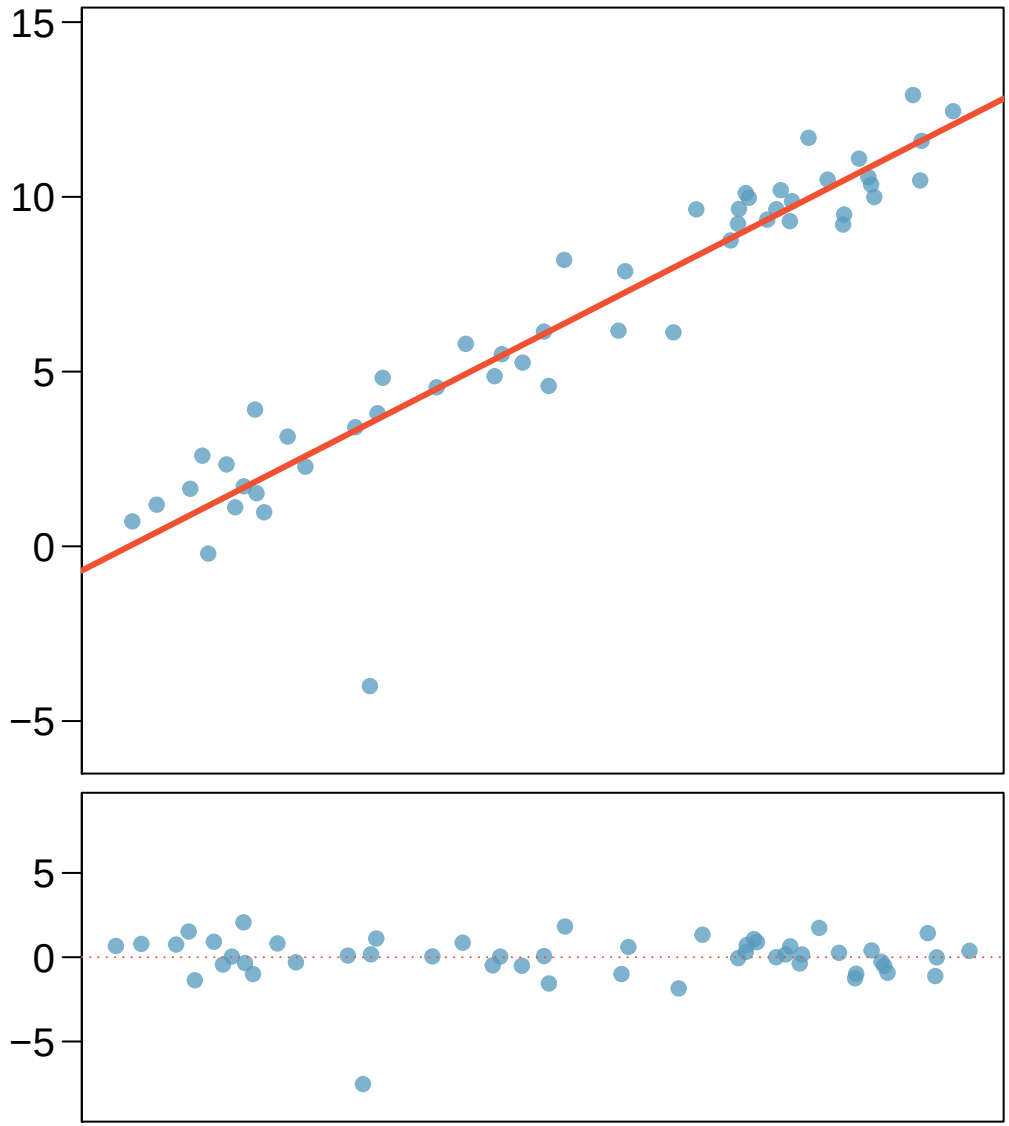
(c) none of the above

(d) there are no outliers



Practice

Does this outlier influence the slope of the regression line?



Which of following is true?

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

(b) Influential points always reduce R^2.

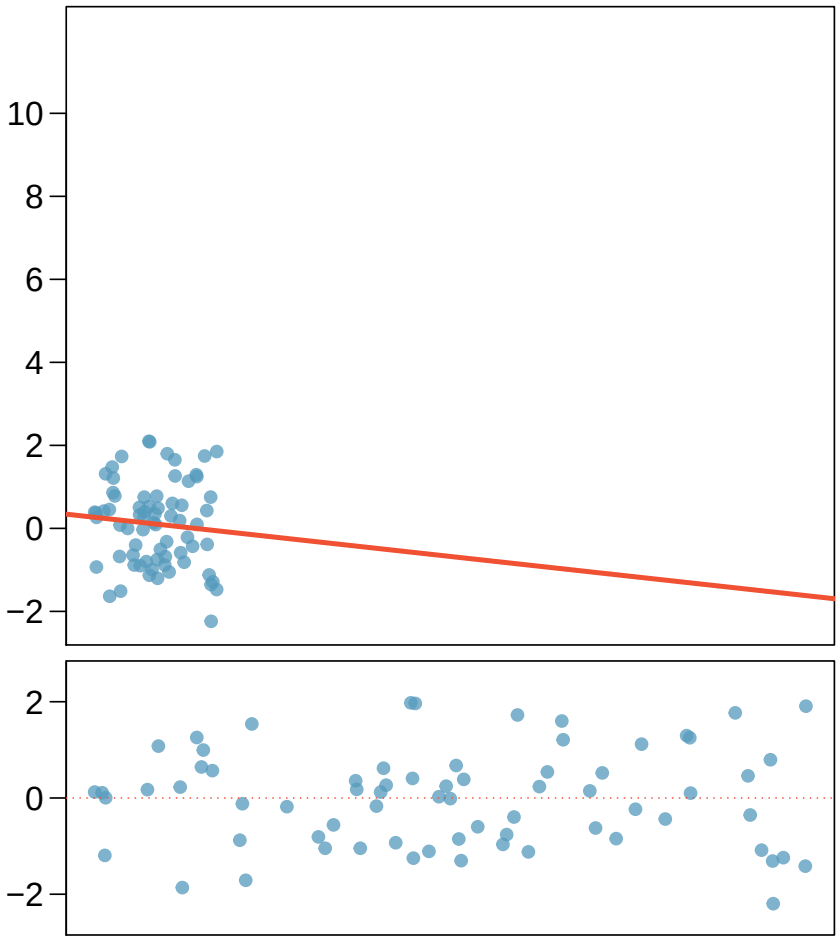
(c) It is much more likely for a low leverage point to be influential, than a high leverage point.

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

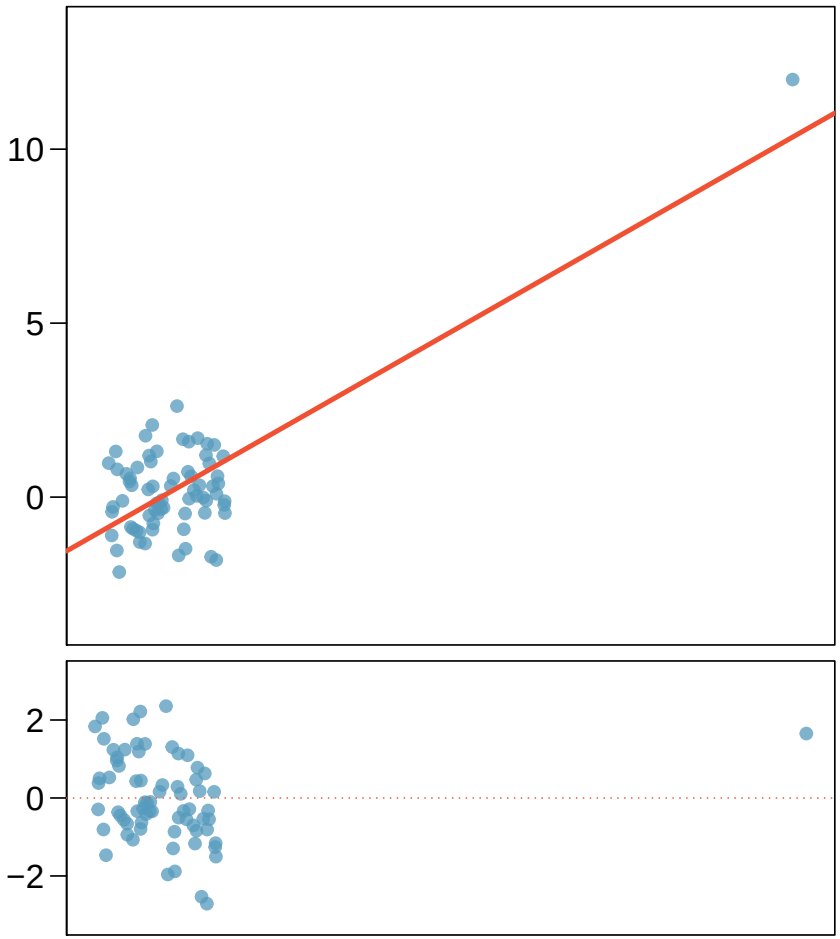
(e) None of the above.

Practice (cont.)

$corr = 0.08, R^2 = 0.0064$

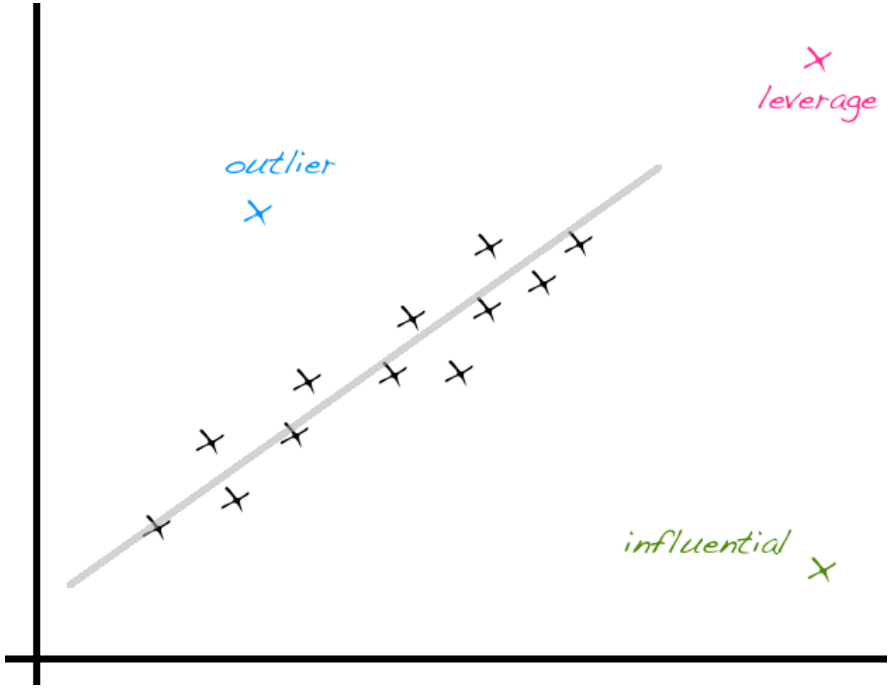


$corr = 0.79, R^2 = 0.6241$



Recap: Types of outliers

- *Leverage* point is away from the cloud of points horizontally, does not necessarily change the slope
- *Influential* point changes the slope (most likely also has high leverage) – run the regression with and without that point to determine



- *Outlier* is an unusual point without these special characteristics (this one likely affects the intercept only)
- Don't remove outliers without a good reason. Models that ignore exceptional cases often perform poorly.
- If clusters (groups of points) are apparent in the data, it might be worthwhile to model the groups separately.