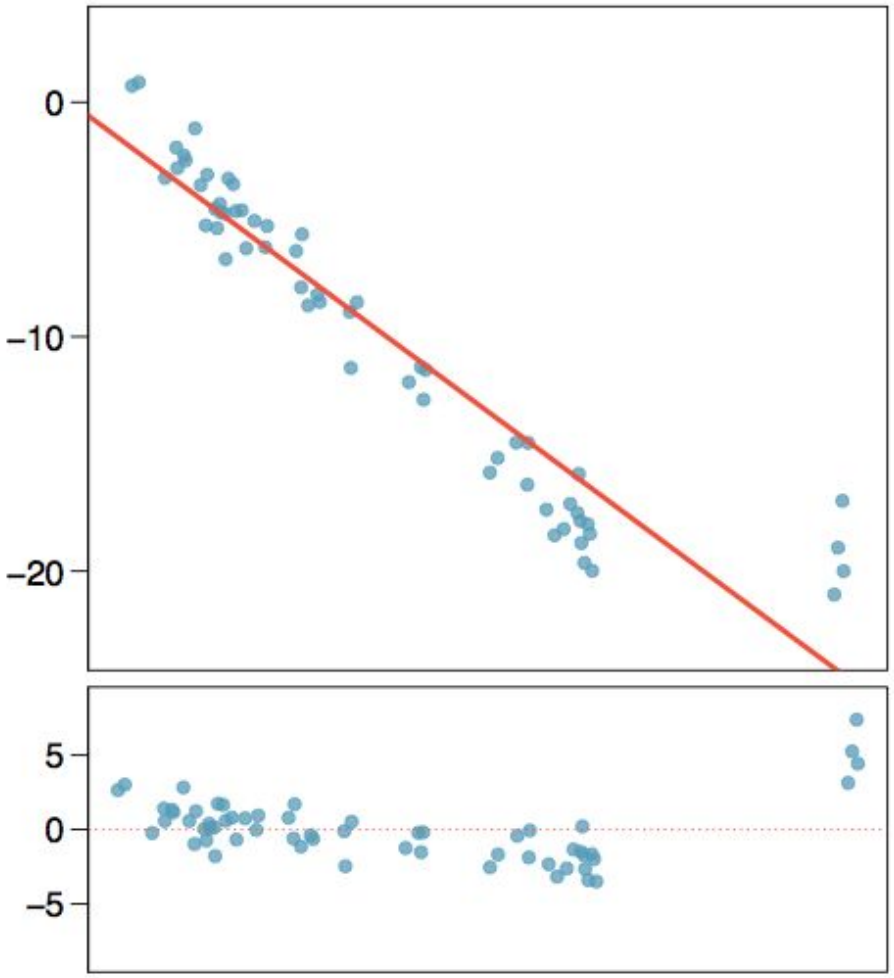
Section 8.3
Types of outliers in linear regression

Stats 7 Summer Session II 2022

Types of outliers

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

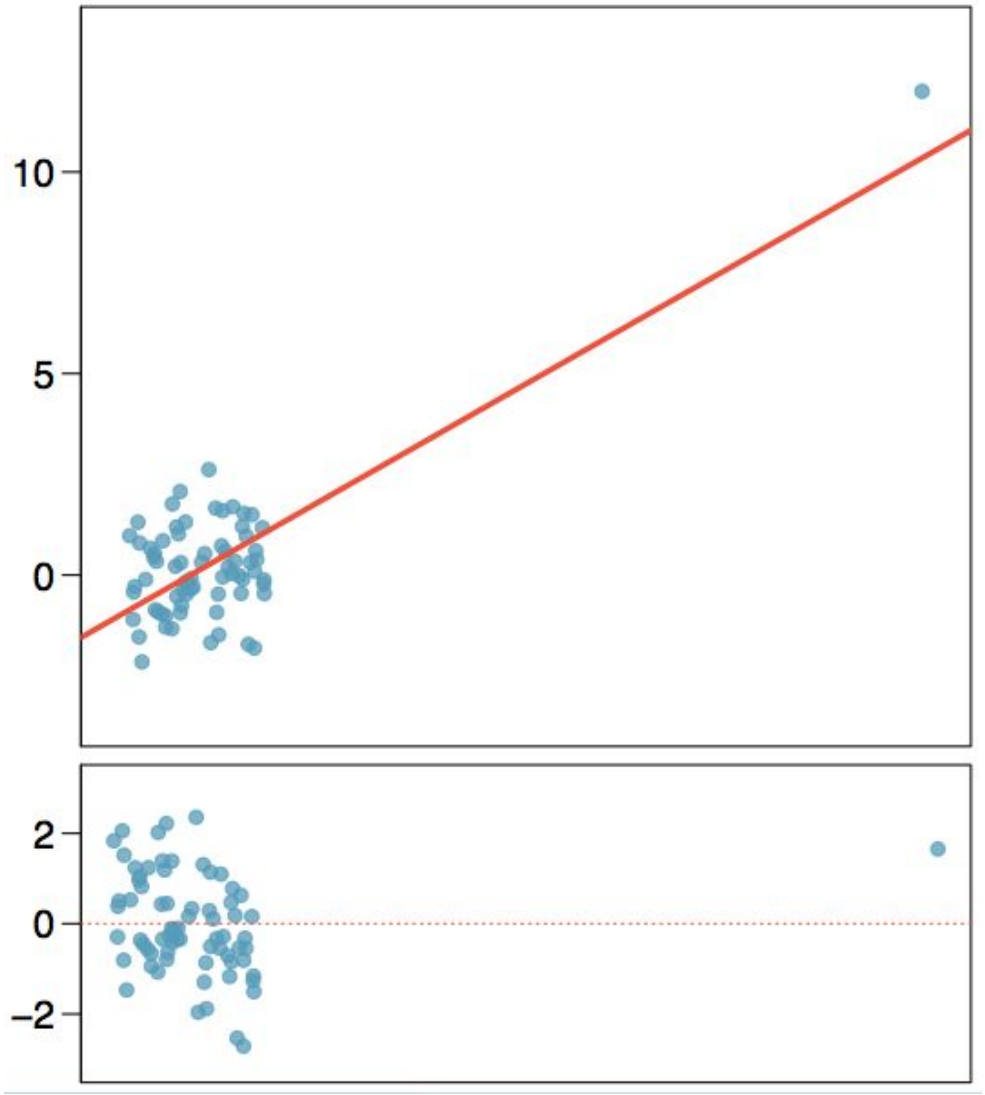
To answer this question think of where the regression line would be with and without the outlier(s). Without the outliers the regression line would be steeper, and lie closer to the larger group of observations. With the outliers the line is pulled up and away from some of the observations in the larger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



Types of outliers

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

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

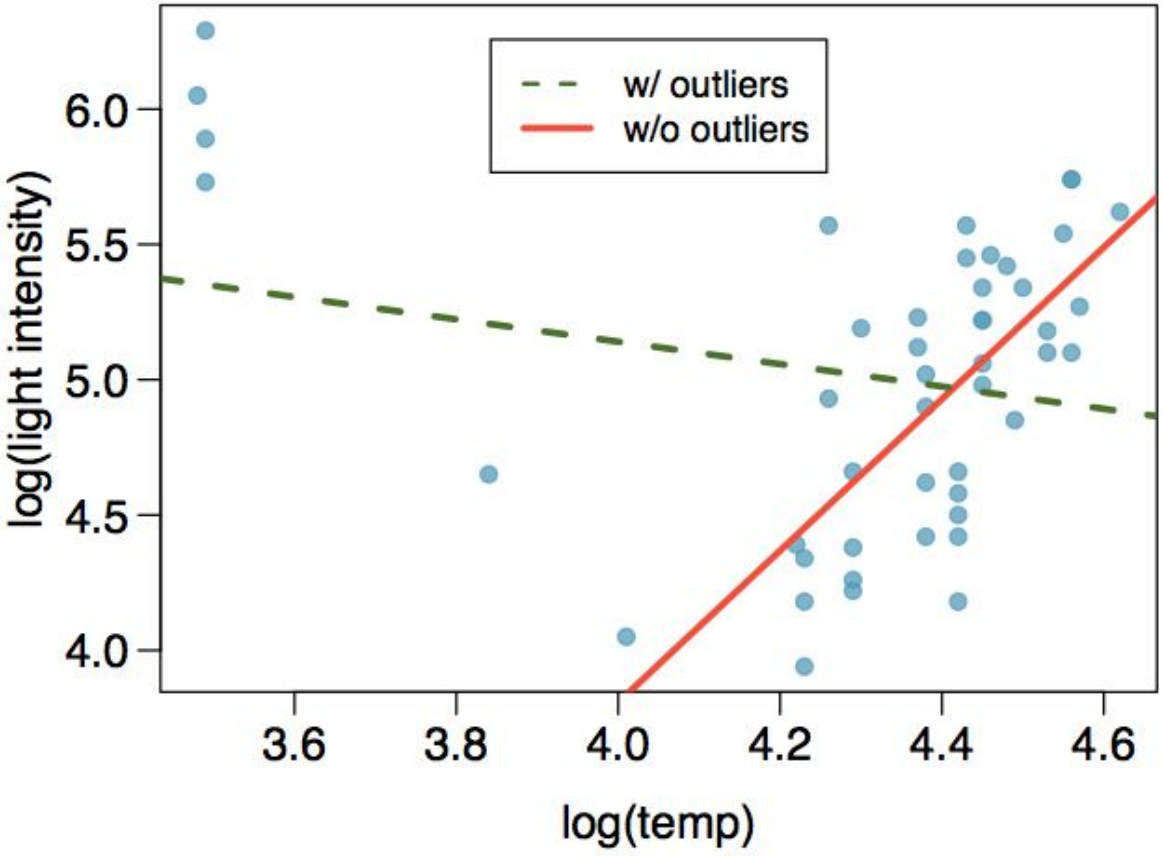


Some terminology

- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie horizontally (x-direction) 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

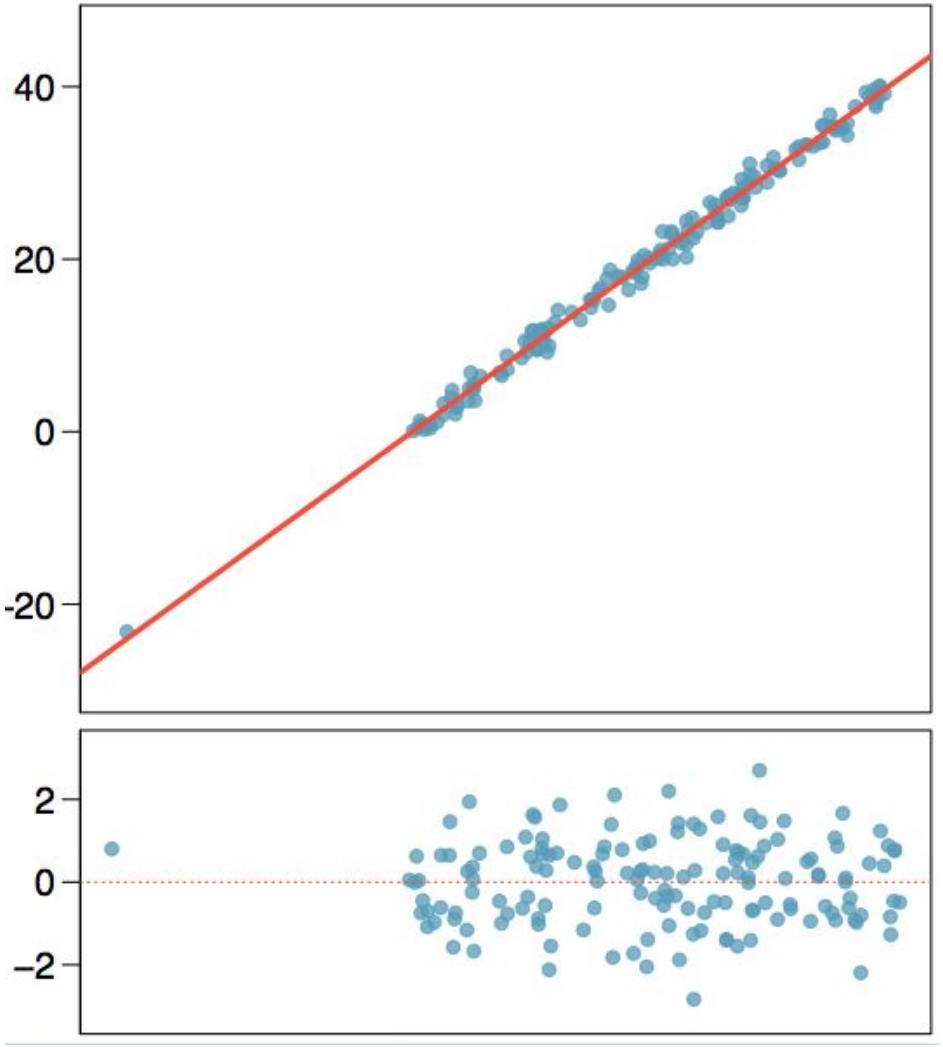
Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.



Types of outliers

Which of the below best describes the outlier?

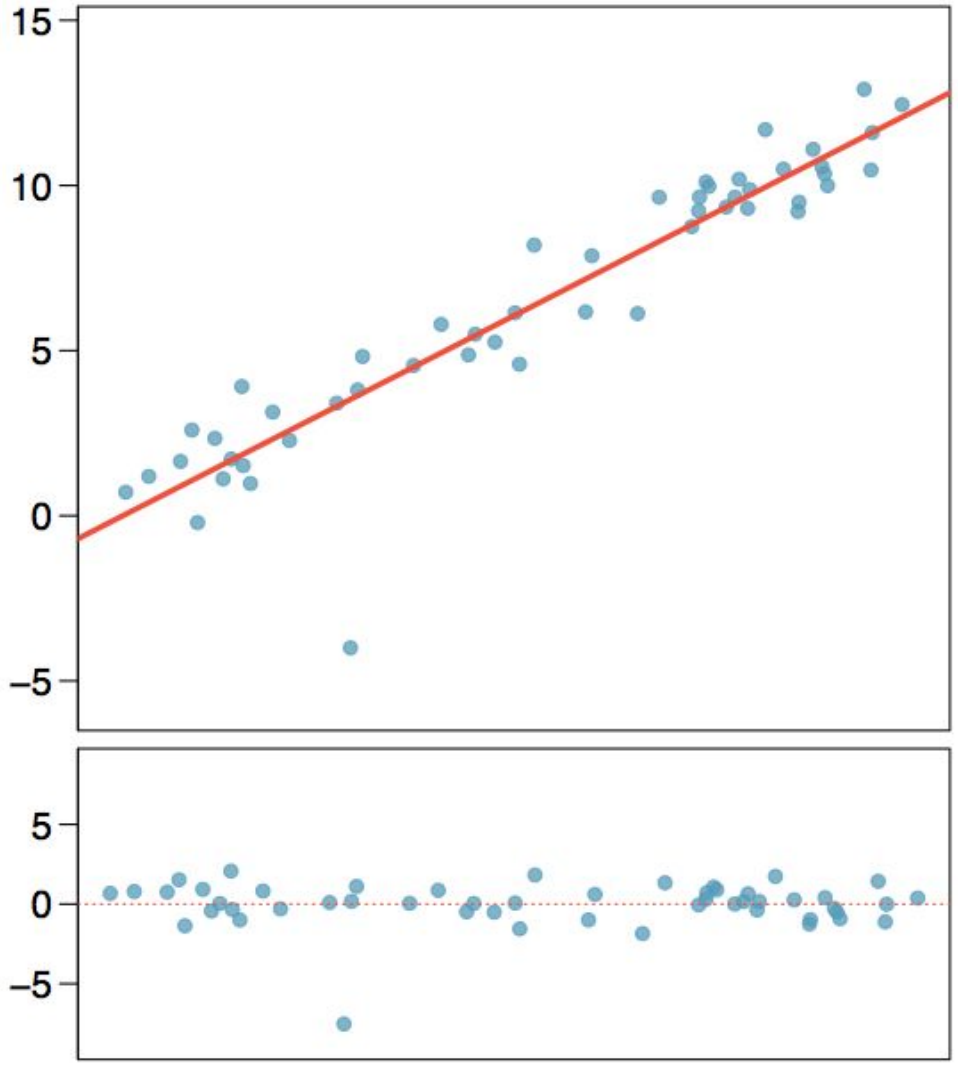
(a) influential
(b) high leverage
(c) none of the above
(d) there are no outliers



Types of outliers

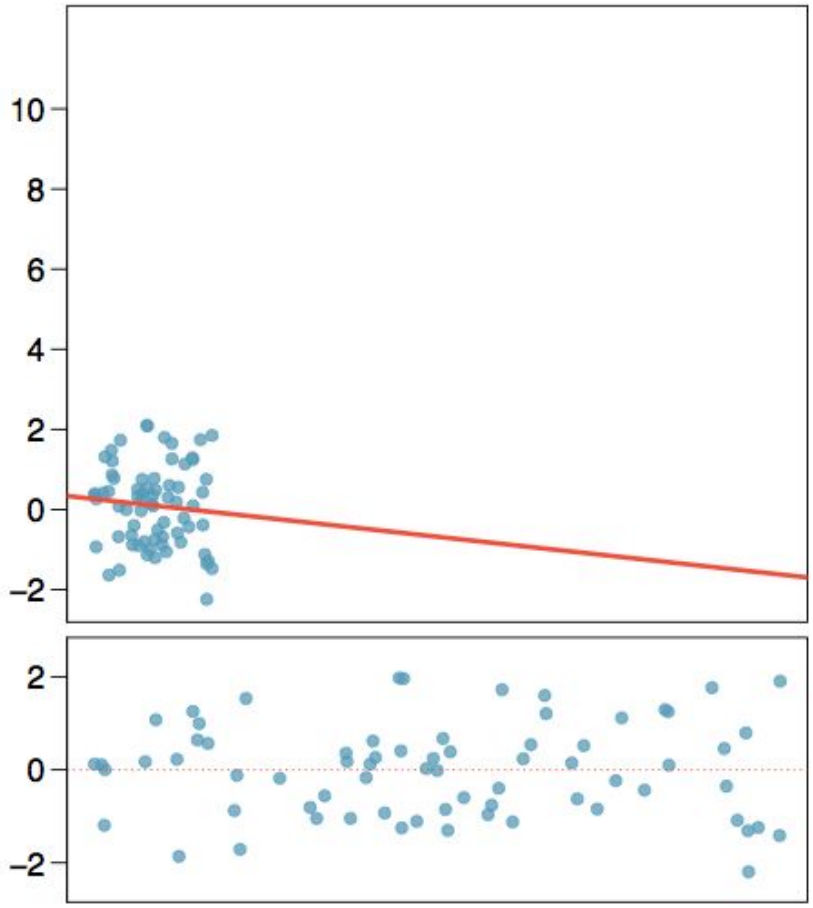
Does this outlier influence the slope of the regression line?

Not much...

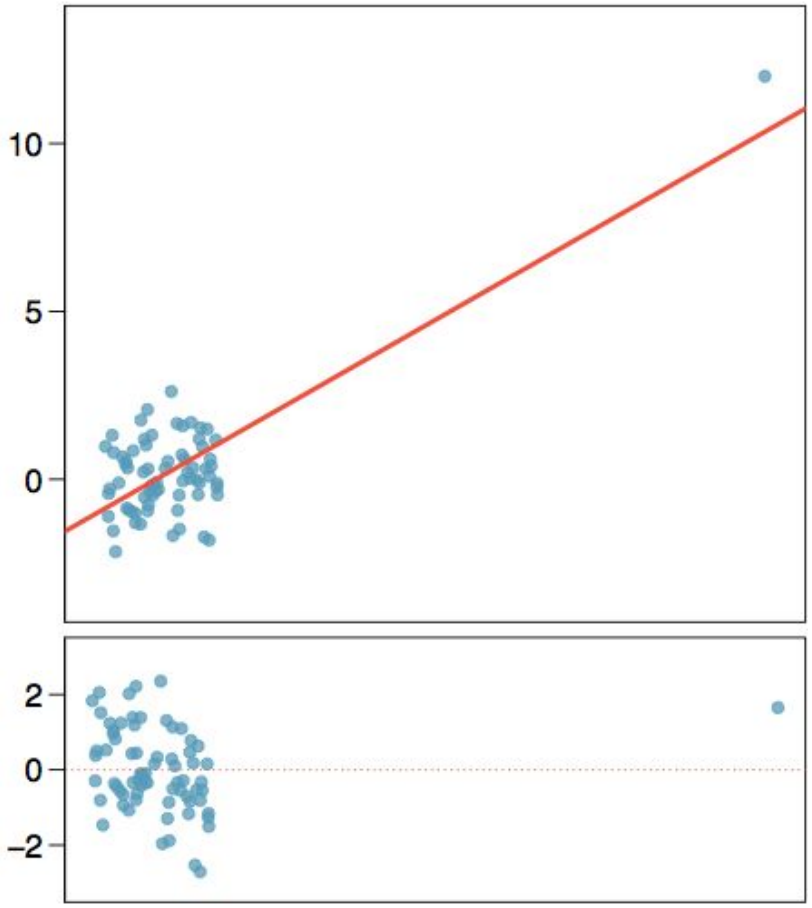


Recap (cont.)

$R = 0.08, R^2 = 0.0064$



$R = 0.79, R^2 = 0.6241$



Should outliers be removed so the line fits better?

- It is tempting to remove outliers. Don’t do this without a very good reason. For example if you:
  - believe that observation was a data entry error
  - refine your population of interest which discludes the outlier
- Models that ignore exceptional (and interesting) cases often perform poorly.
- For instance, if a financial firm ignored the largest market swings – the “outliers” – they would soon go bankrupt by making poorly thought-out investments.