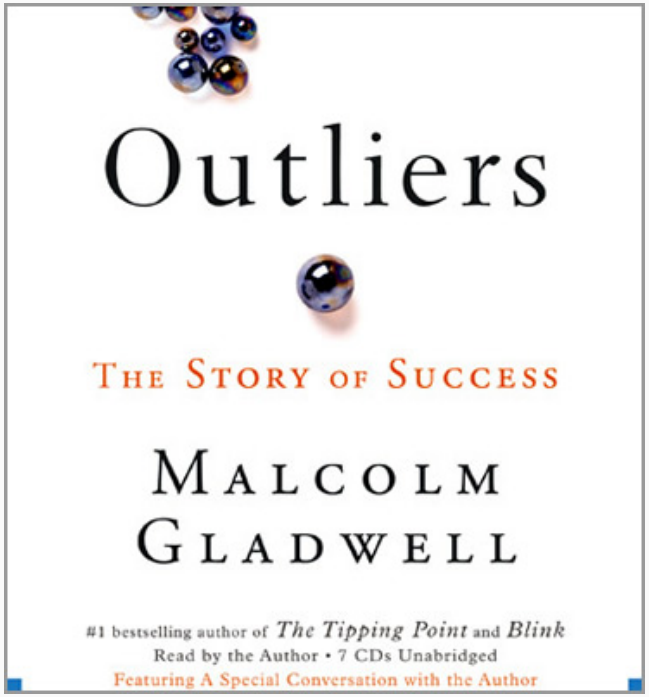
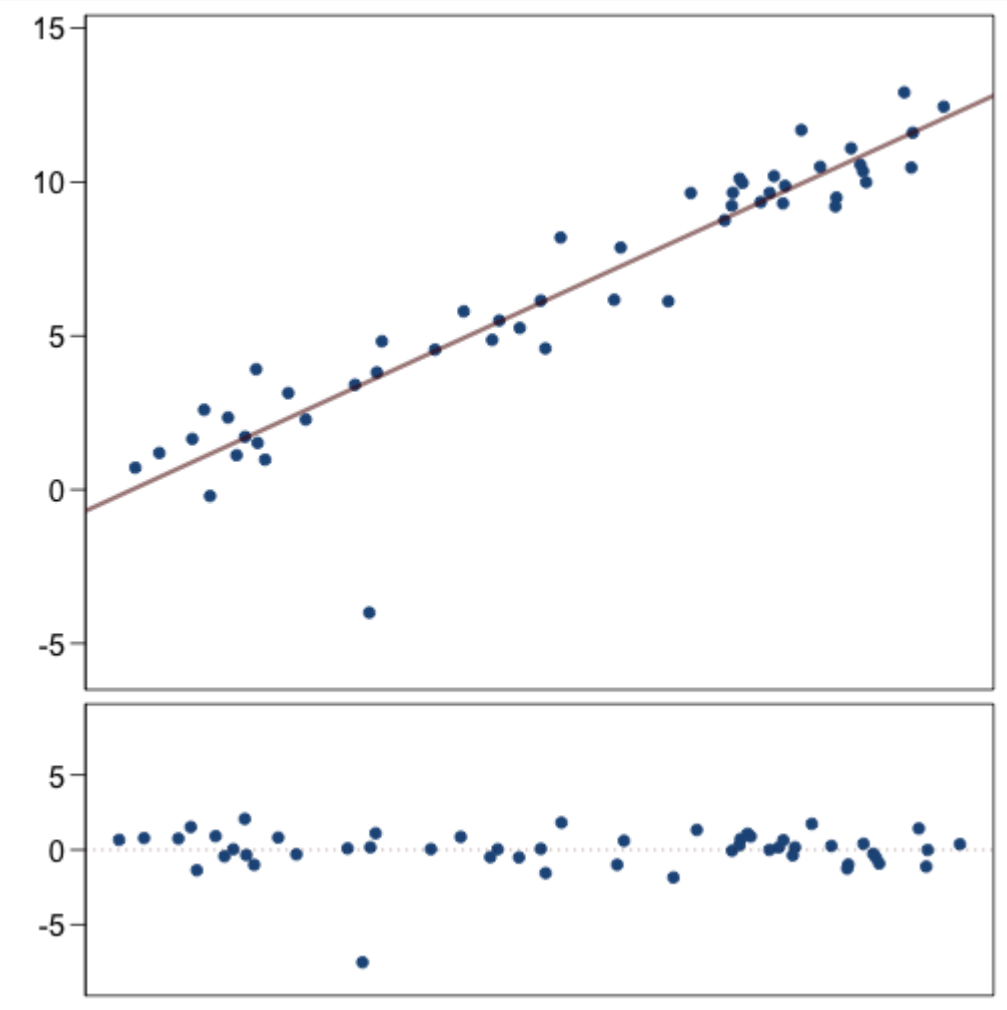
Outliers

What is an outlier?

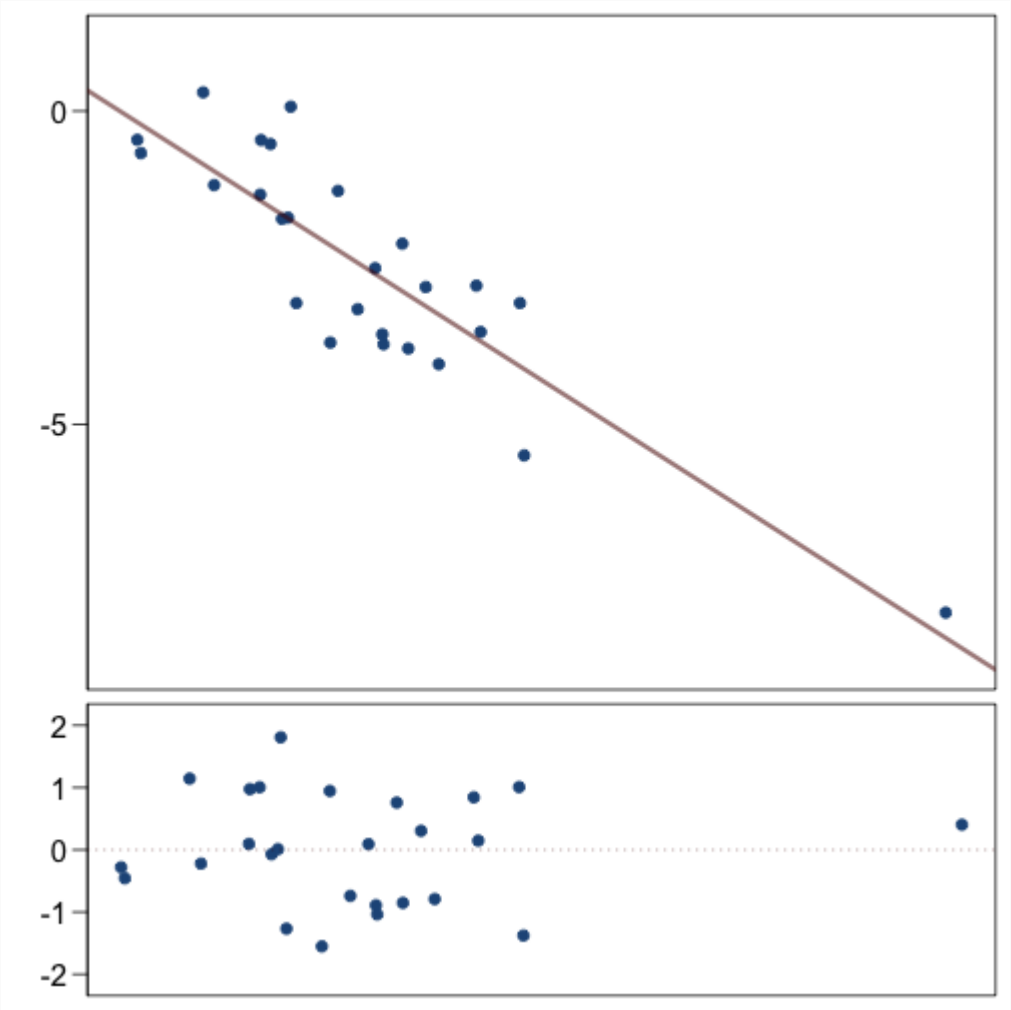


Outlier is a general term to describe a data point that doesn't follow the pattern set by the bulk of the data when one takes into account the model.

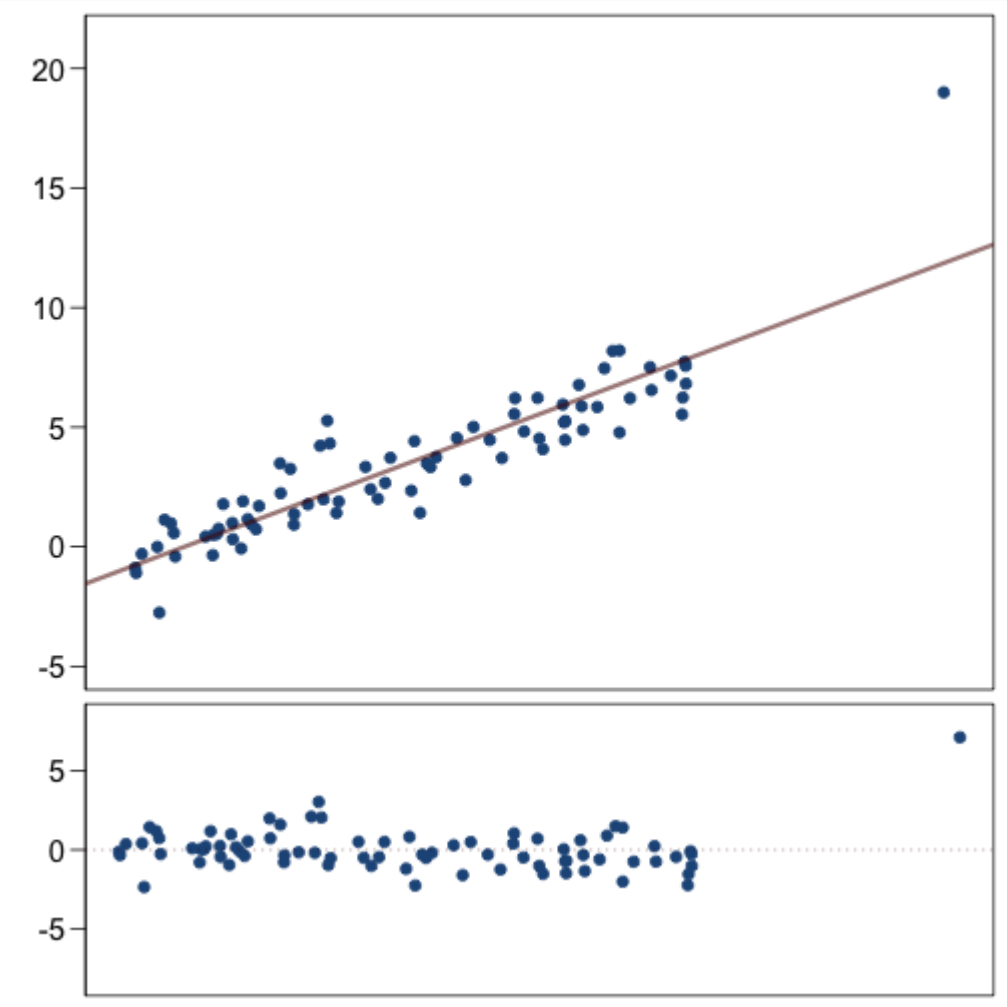
Outlier Example One



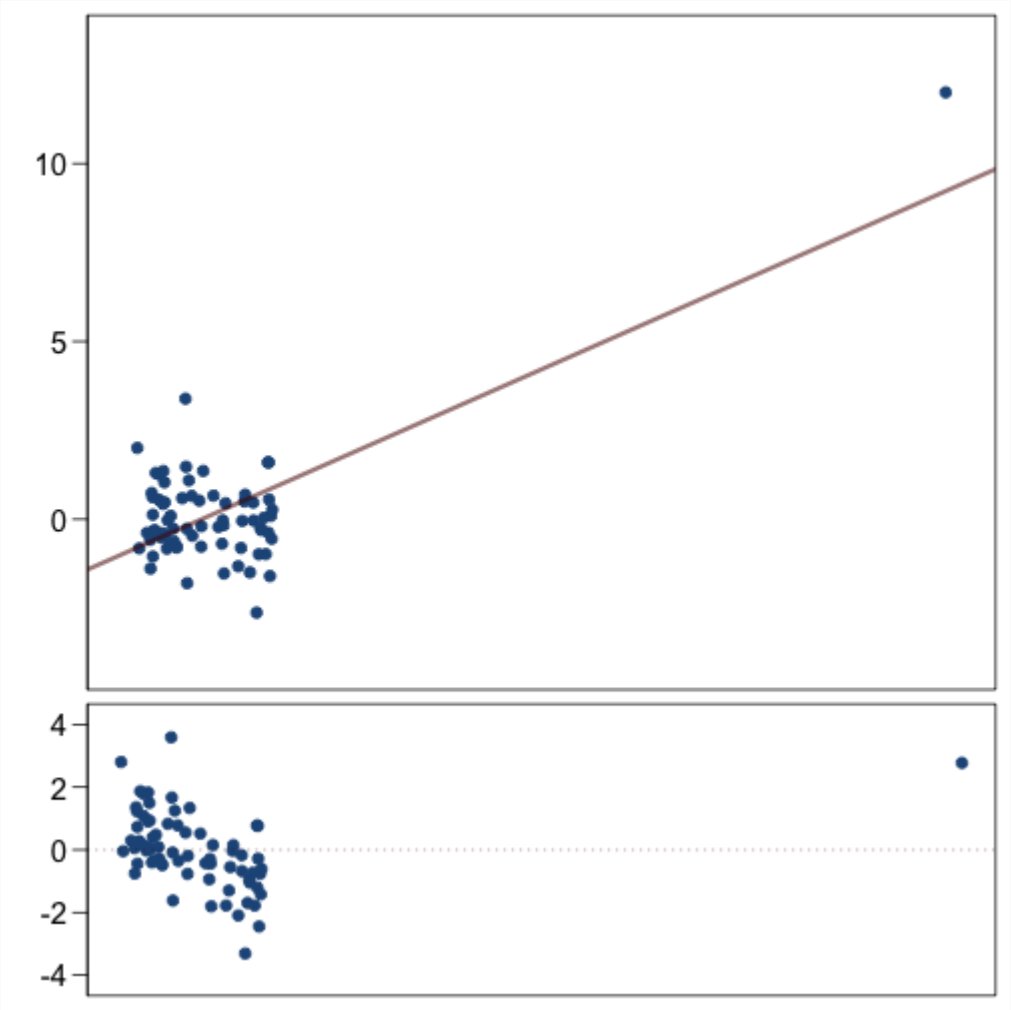
Outlier Example Two



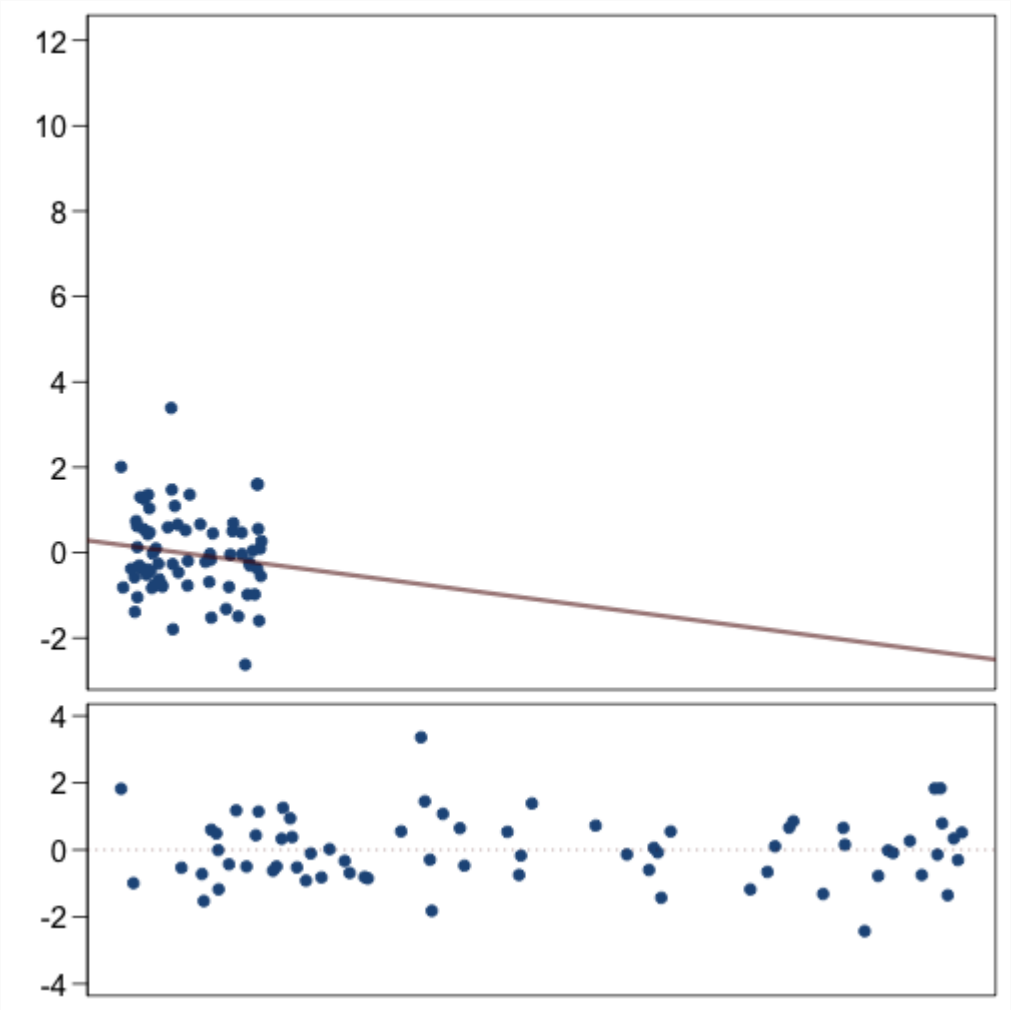
Outlier Example Three



Outlier Example Four



Outlier Example Four



Outliers, leverage, influence

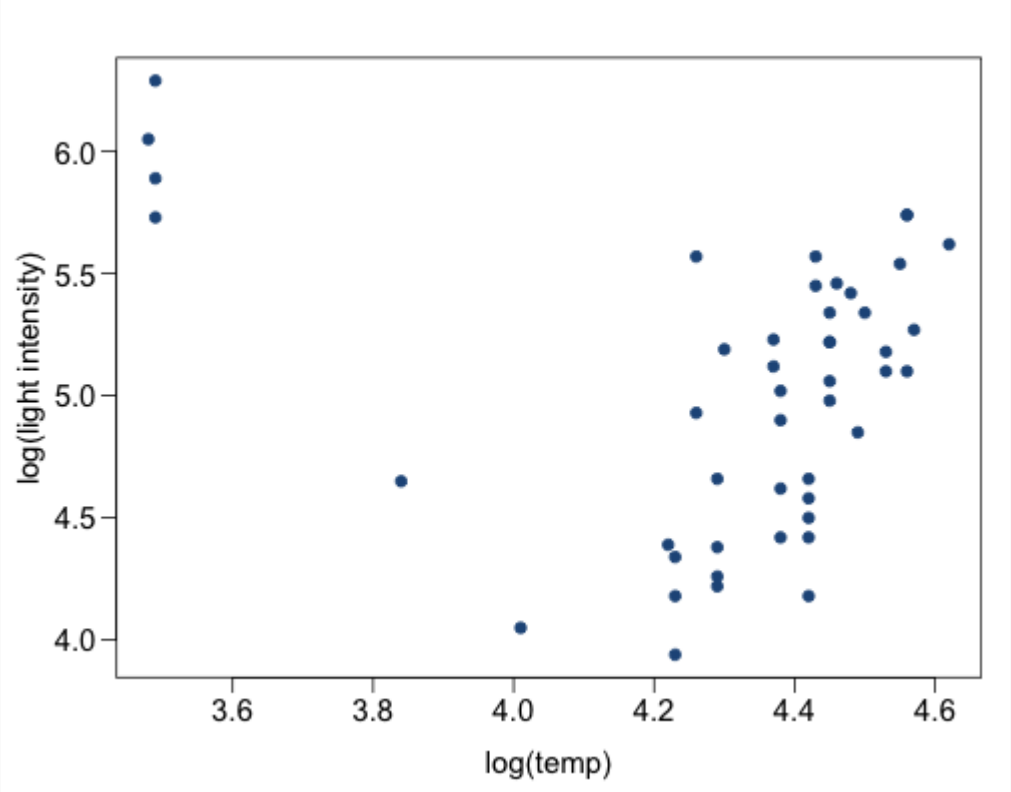
Outliers are points that don't fit the trend in the rest of the data.

High leverage points have the potential to have an unusually large influence on the fitted model.

Influential points are high leverage points that cause a very different line to be fit than would be with that point removed.

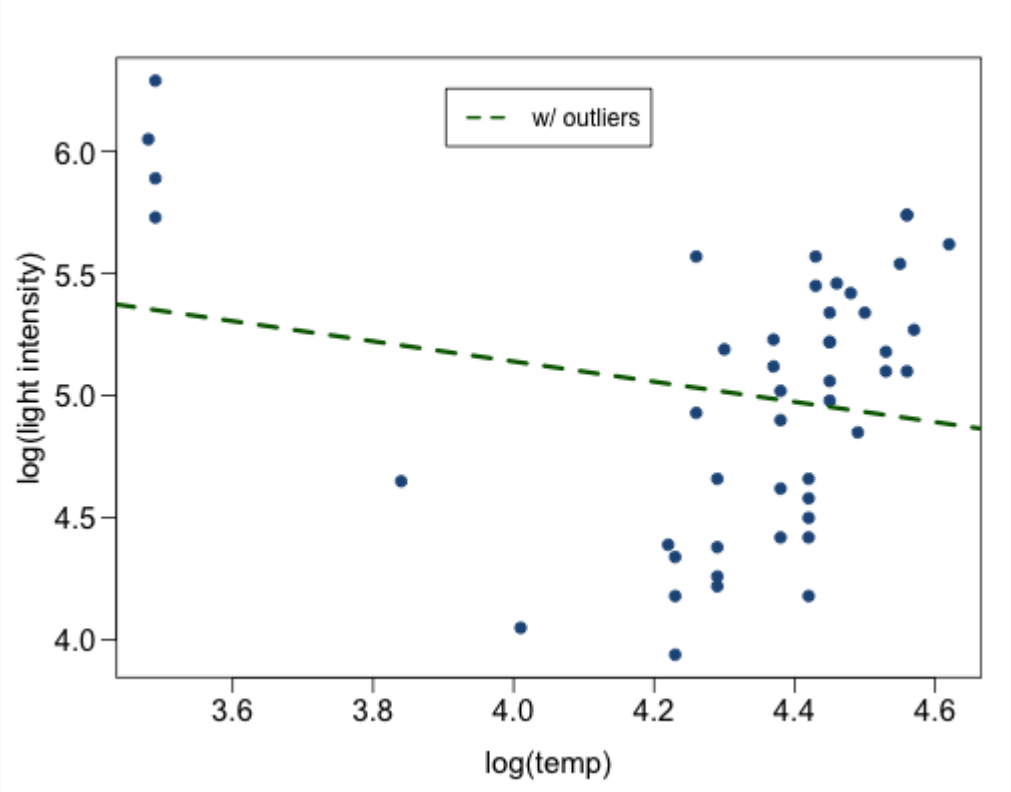
Example of high lev, high influence

We have data on the surface temperature and light intensity of 47 stars in the star cluster CYG OB1, near Cygnus.



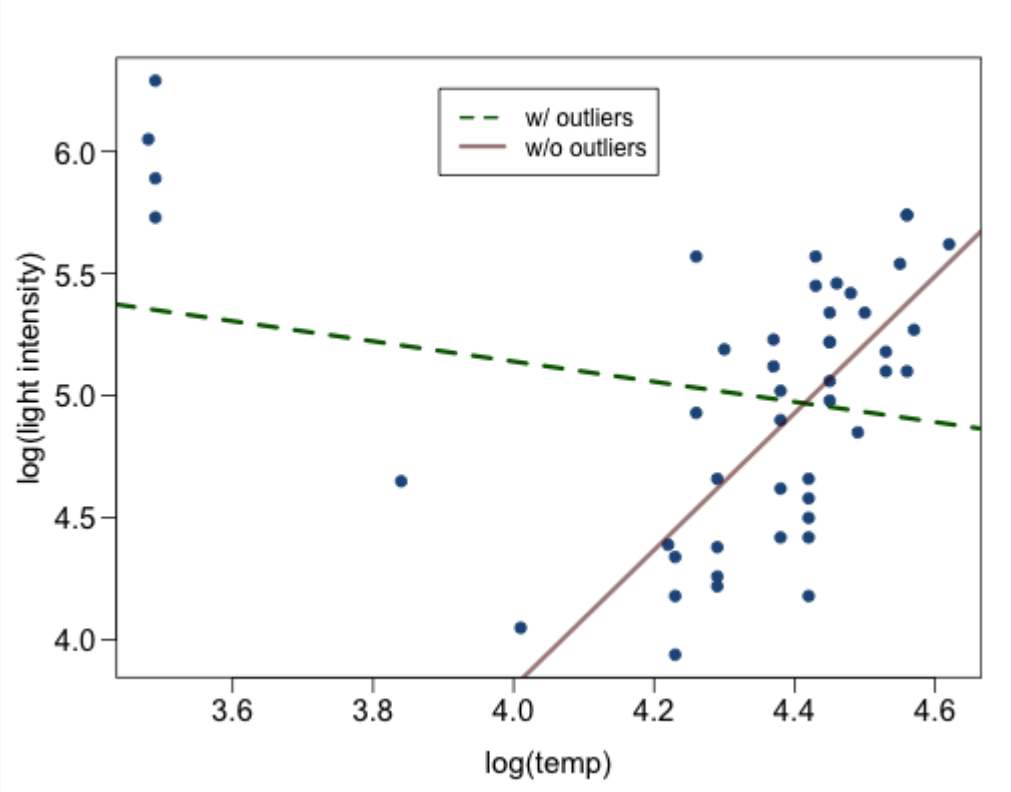
Example of high lev, high influence

We have data on the surface temperature and light intensity of 47 stars in the star cluster CYG OB1, near Cygnus.

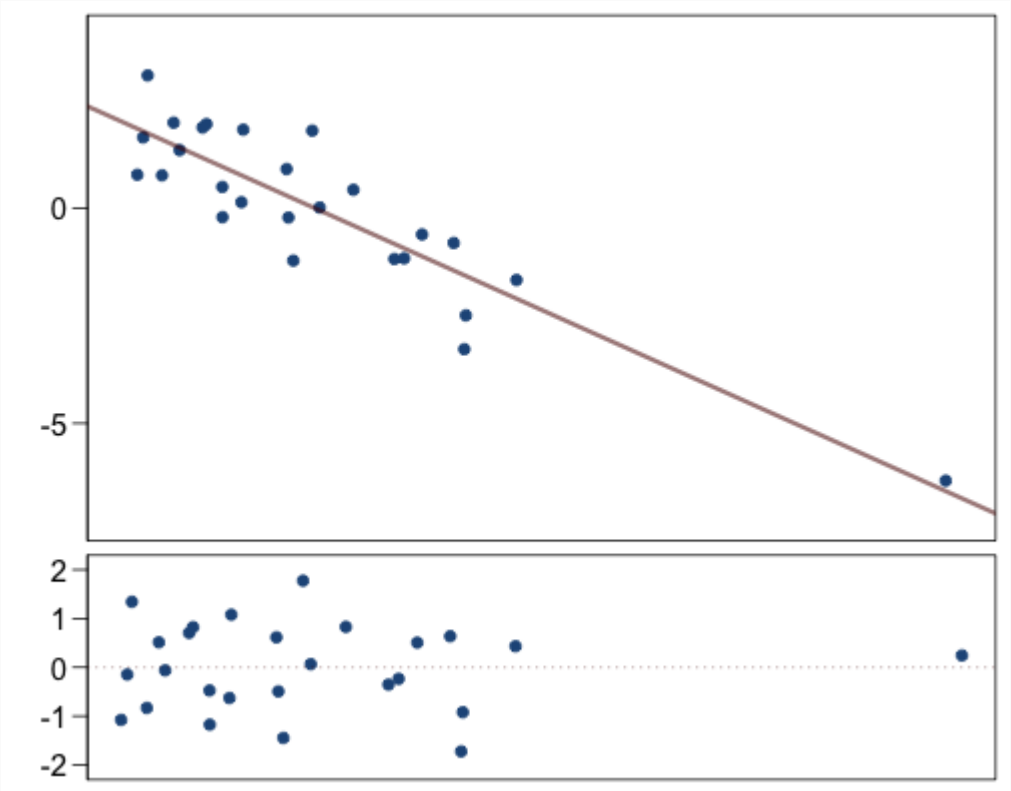


Example of high lev, high influence

We have data on the surface temperature and light intensity of 47 stars in the star cluster CYG OB1, near Cygnus.



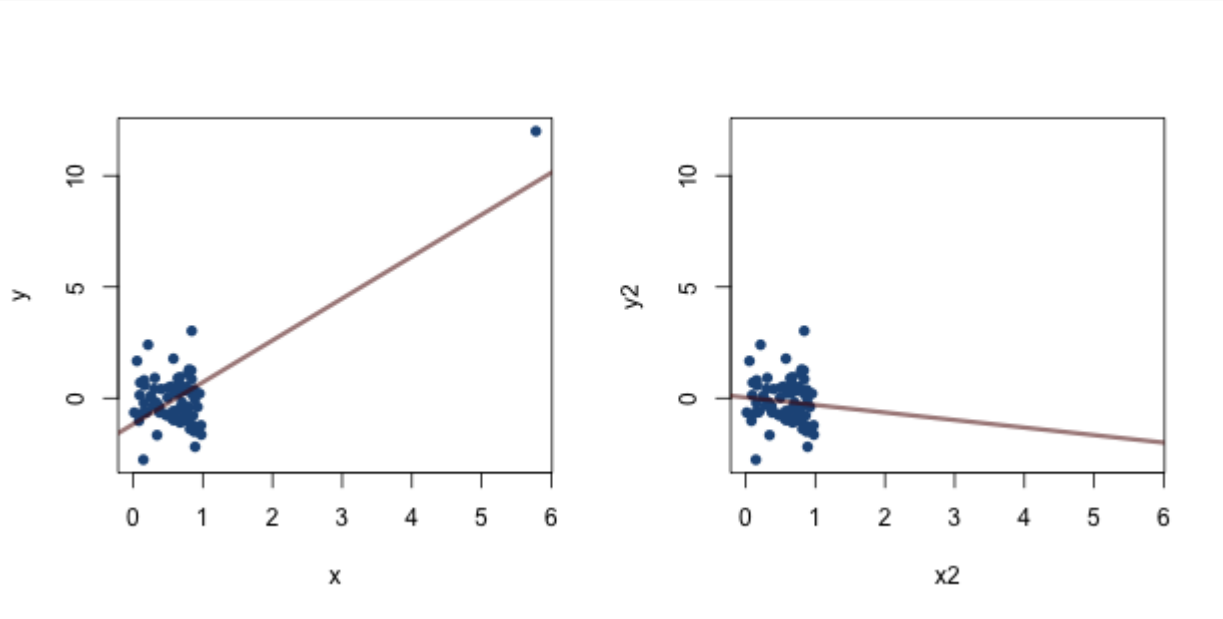
Example of high lev, low influence



From leverage to influence

Leverage measures the weight given to each point in determining the regression line.

Influence measures how different the regression line would be without a given point. Often measured with *Cook's Distance*.



Your Turn

In the following plots are there outliers, leverage pts, or influential pts?