Outliers

Review: the t-test

When does it crop up?

To understand the effect of a dirty air filter on gasoline mileage, suppose we take a random sample of 9 cars and measure the miles per gallon they get when driven with a clean air filter and with a dirty air filter. Suppose that we have properly designed this study: that the order in which the cars receive a clean or dirty air filter is determined by random assignment, that the dirty filters are equally dirty in each car, that the filters are the same brand, and that the cars are driven under the same conditions when the gasoline mileage is measured. The table given below shows the results of the testing.

Car ID	clean filter	dirty filter	difference
1	19.1	17.8	1.3
2	24.5	24.7	-0.2
3	23.2	21.5	1.7
4	25.1	24.9	0.2
5	25.6	24.3	1.3
6	25.5	23.8	2.8
7	26.0	26.9	-1.5
8	28.3	28.1	0.2
9	30.9	29.0	1.9
$\bar{x}$	25.35	24.55	0.80
s	3.24	3.42	1.00

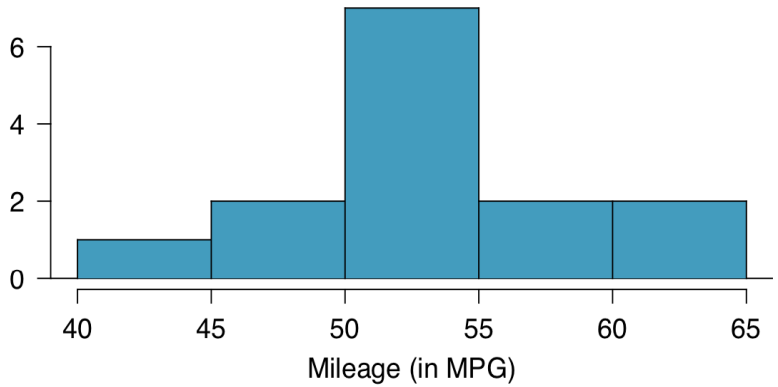
12. How many *independent* observations are in this dataset?

13. Does there appear to be a difference between the gasoline mileage obtained when cars have clean air filters and when they have dirty air filters at the 10% level? Give complete statistical evidence to support your answer, including the hypotheses, test statistic, degrees of freedom, p-value, an interpretation of your results, and an indication of the conditions needed to ensure a reasonably accurate p-value (suggestion: sketch a picture on scratch paper to be sure you compute the p-value correctly.)

14. Give a 95% confidence interval for the magnitude of the average difference and provide an interpretation of the interval.

Practice

4.6 Fuel efficiency of Prius. Fueleconomy.gov, the official US government source for fuel economy information, allows users to share gas mileage information on their vehicles. The histogram below shows the distribution of gas mileage in miles per gallon (MPG) from 14 users who drive a 2012 Toyota Prius. The sample mean is 53.3 MPG and the standard deviation is 5.2 MPG. Note that these data are user estimates and since the source data cannot be verified, the accuracy of these estimates are not guaranteed.[29]

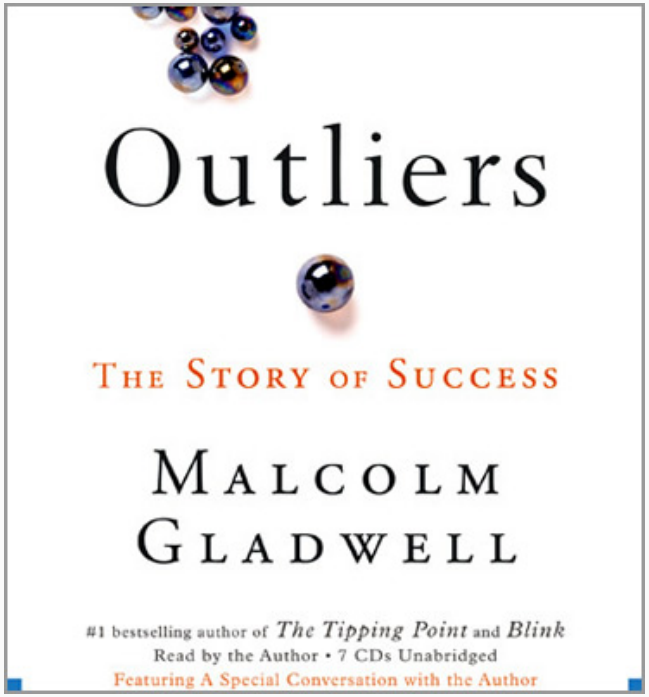


(a) We would like to use these data to evaluate the average gas mileage of all 2012 Prius drivers. Do you think this is reasonable? Why or why not?

(b) The EPA claims that a 2012 Prius gets 50 MPG (city and highway mileage combined). Do these data provide strong evidence against this estimate for drivers who participate on fueleconomy.gov? Note any assumptions you must make as you proceed with the test.

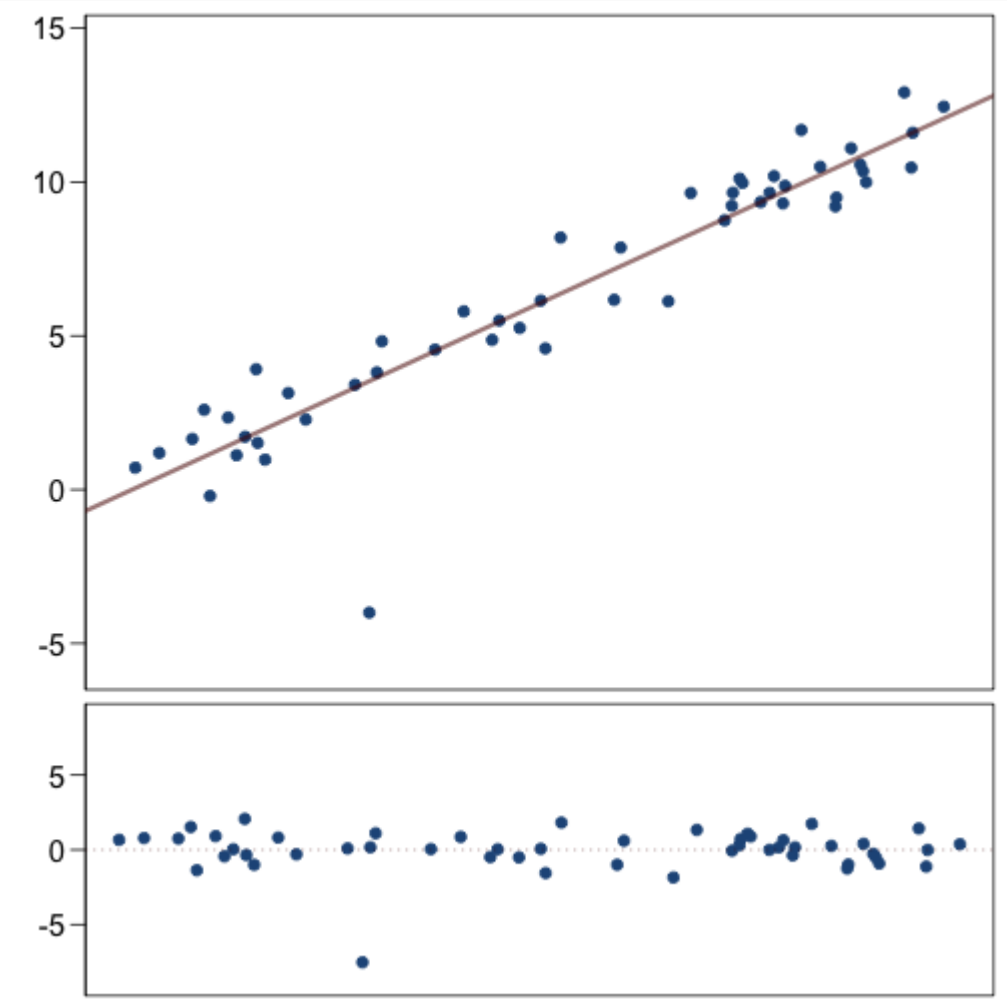
(c) Calculate a 95% confidence interval for the average gas mileage of a 2012 Prius by drivers who participate on fueleconomy.gov.

What is an outlier?

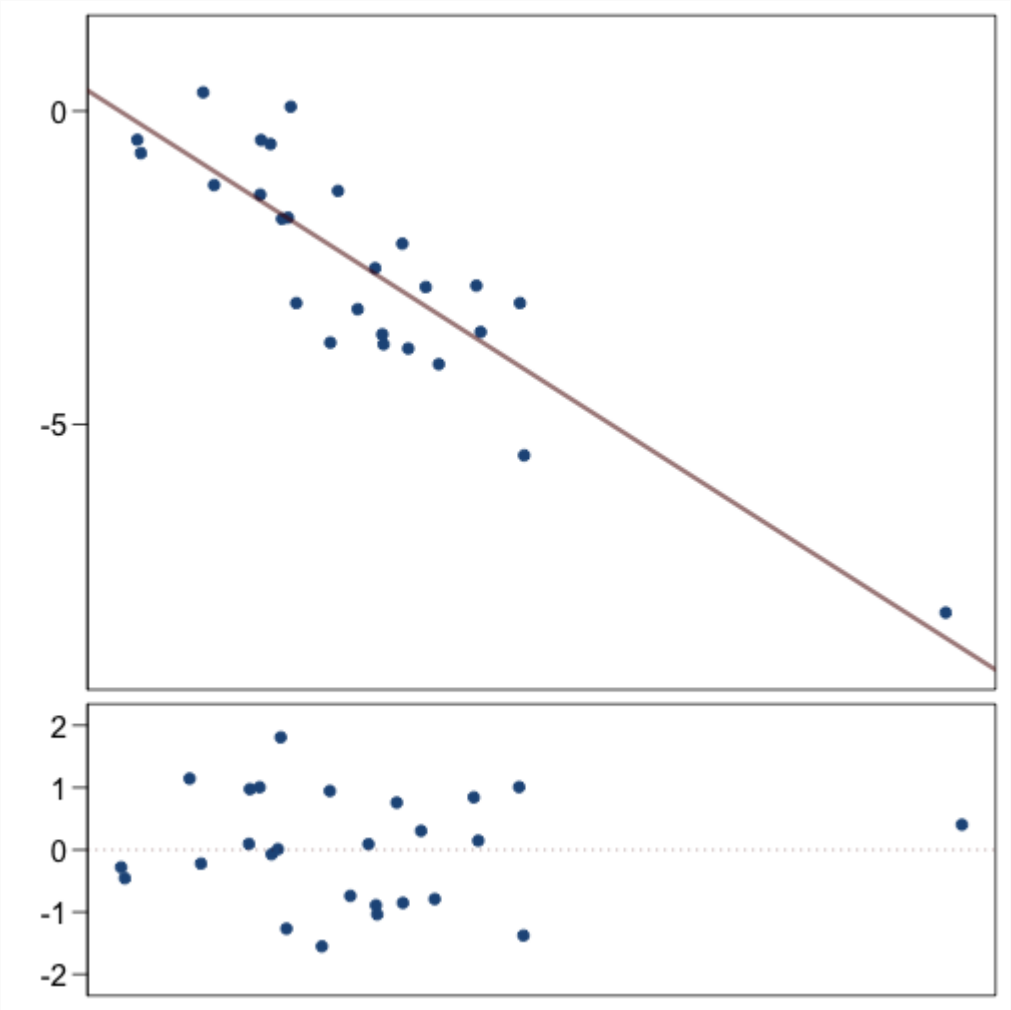


Outlier is a general term to describe a data point that doesn't follow the pattern set by the bulk of the data when one takes into account the model.

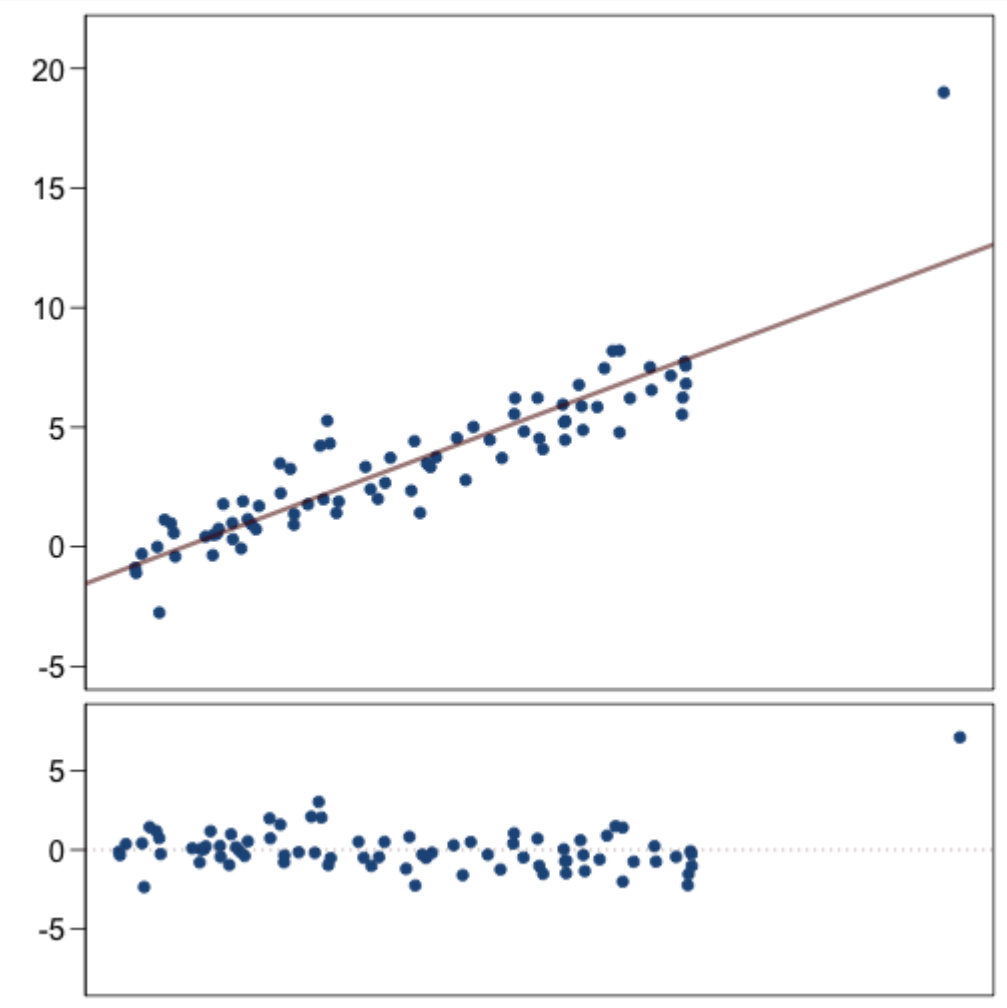
Outlier Example One



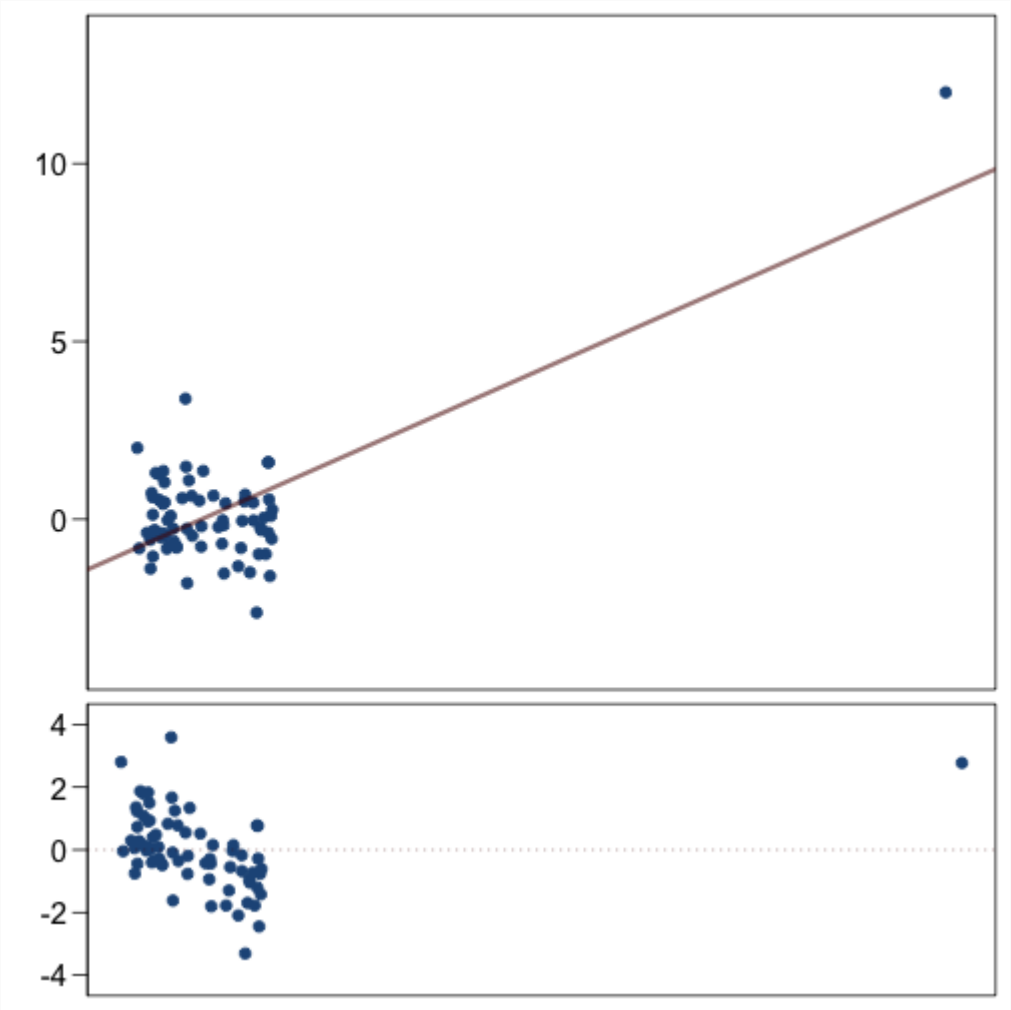
Outlier Example Two



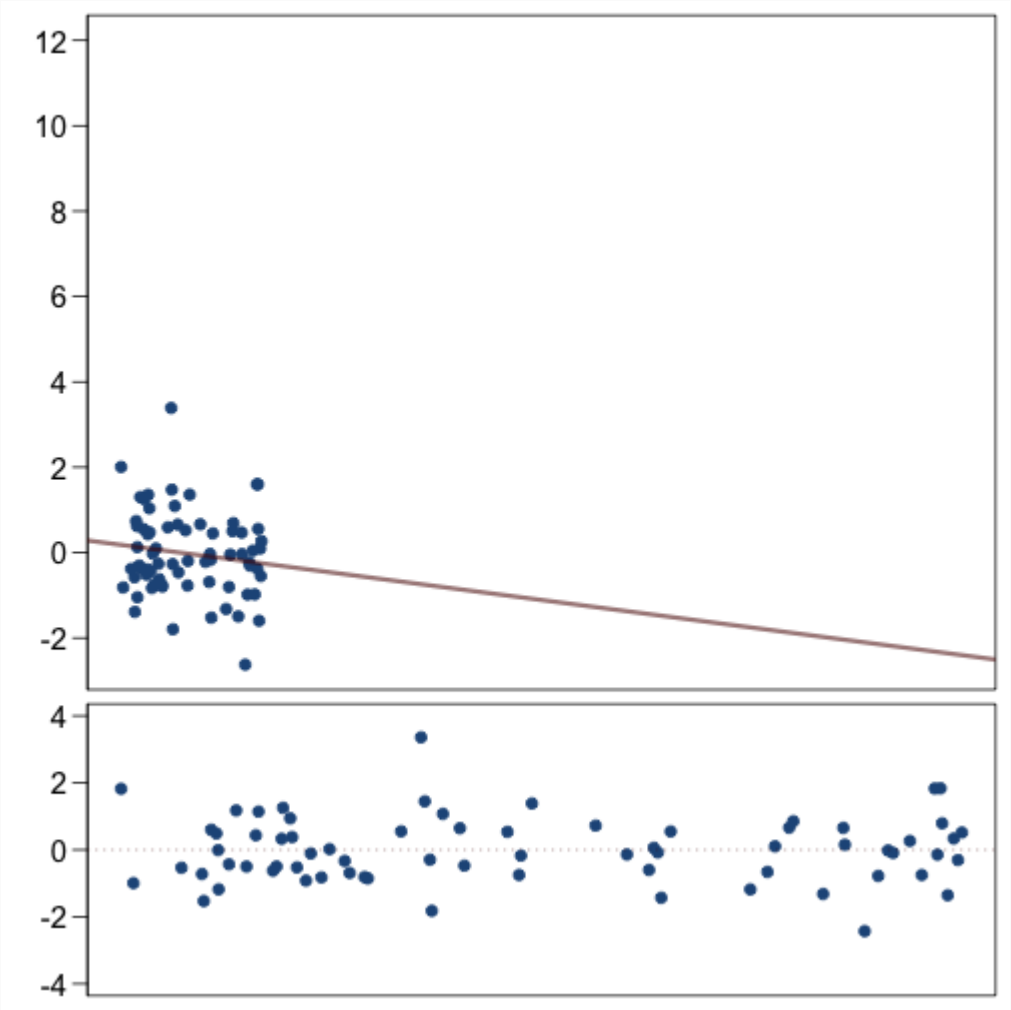
Outlier Example Three



Outlier Example Four



Outlier Example Four



Outliers, leverage, influence

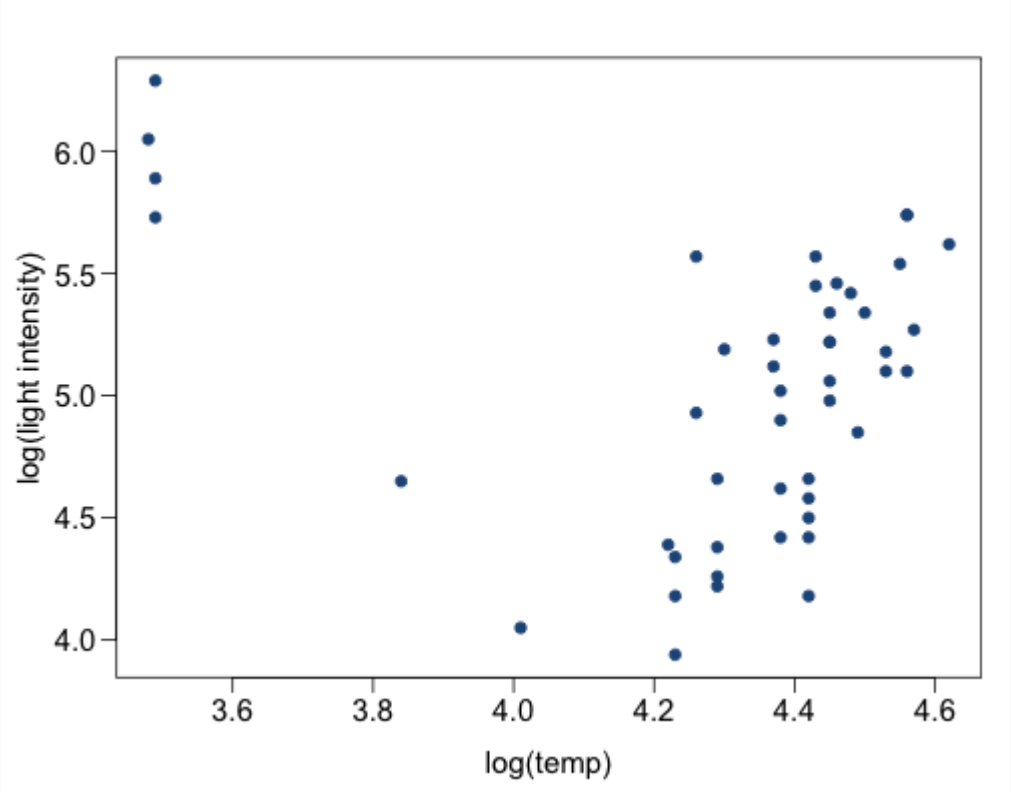
Outliers are points that don't fit the trend in the rest of the data.

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.

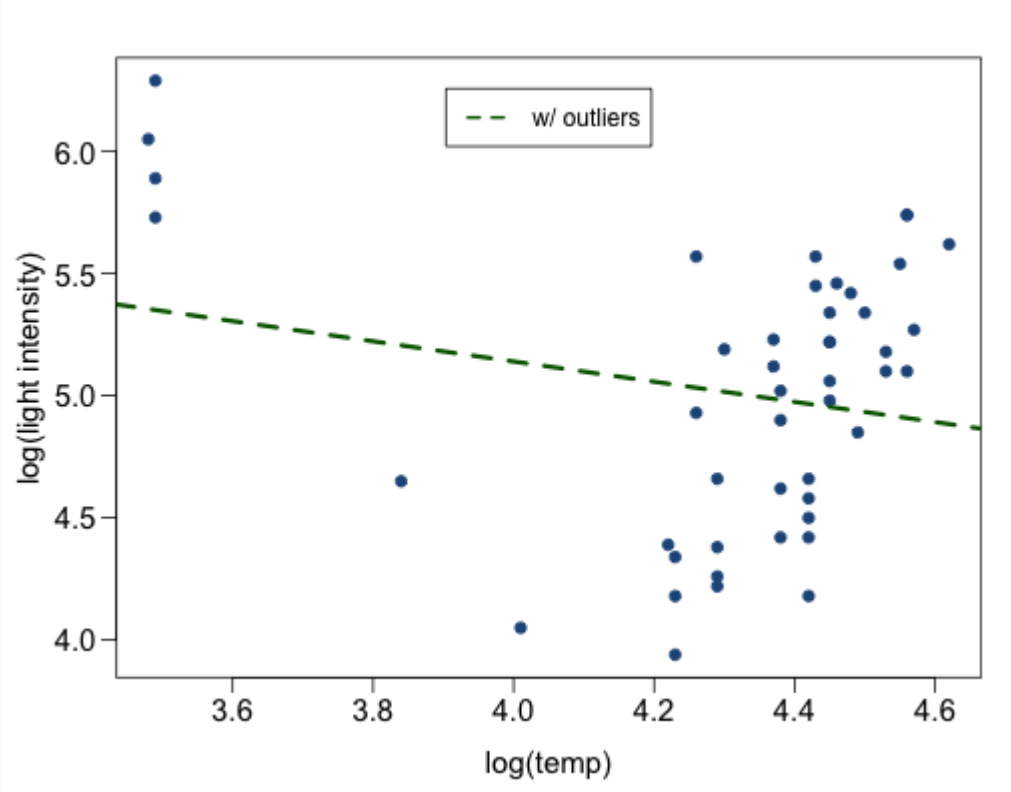
Example of high lev, high influence

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



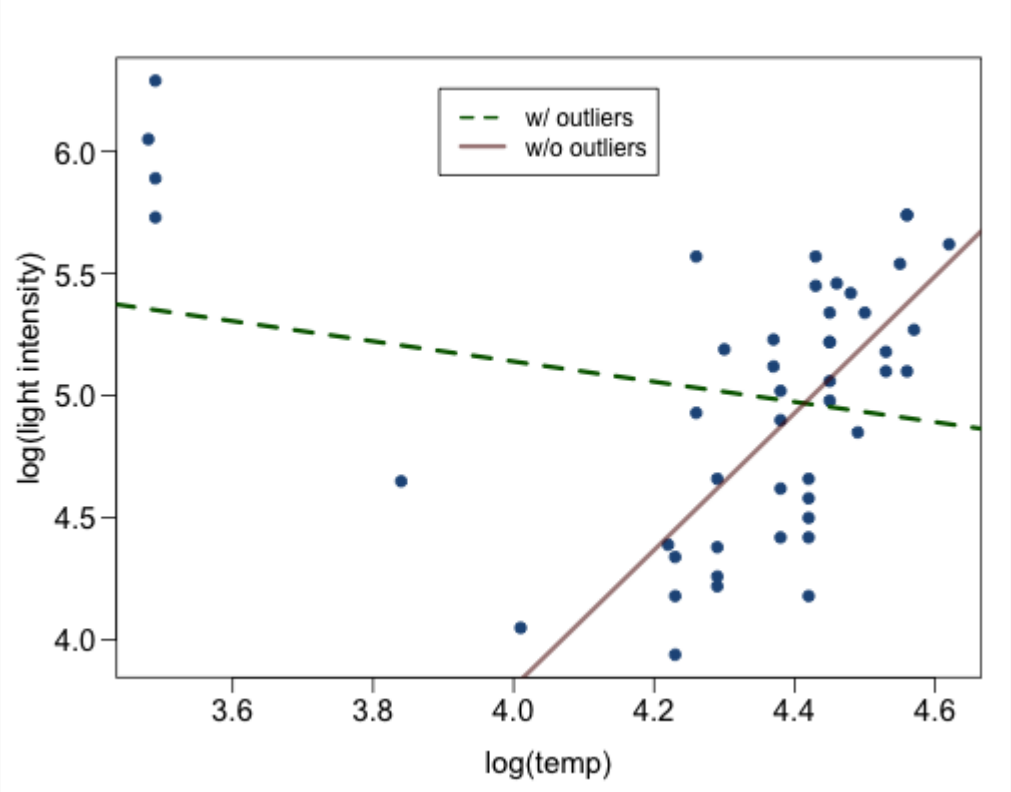
Example of high lev, high influence

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

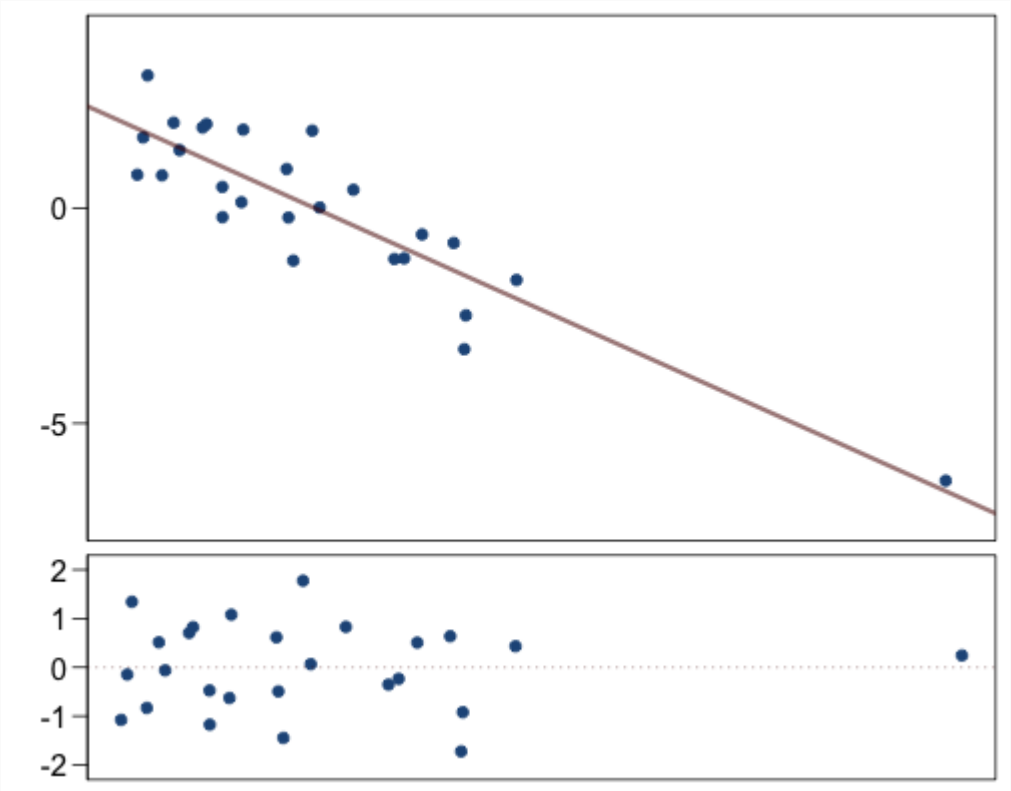


Example of high lev, high influence

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



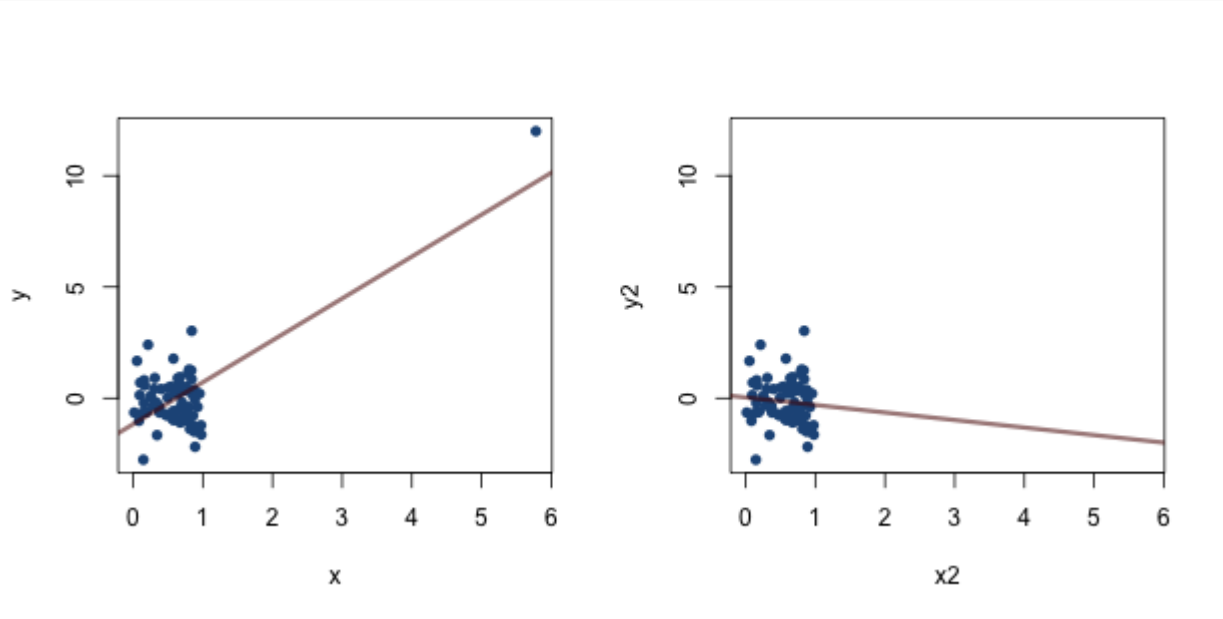
Example of high lev, low influence



From leverage to influence

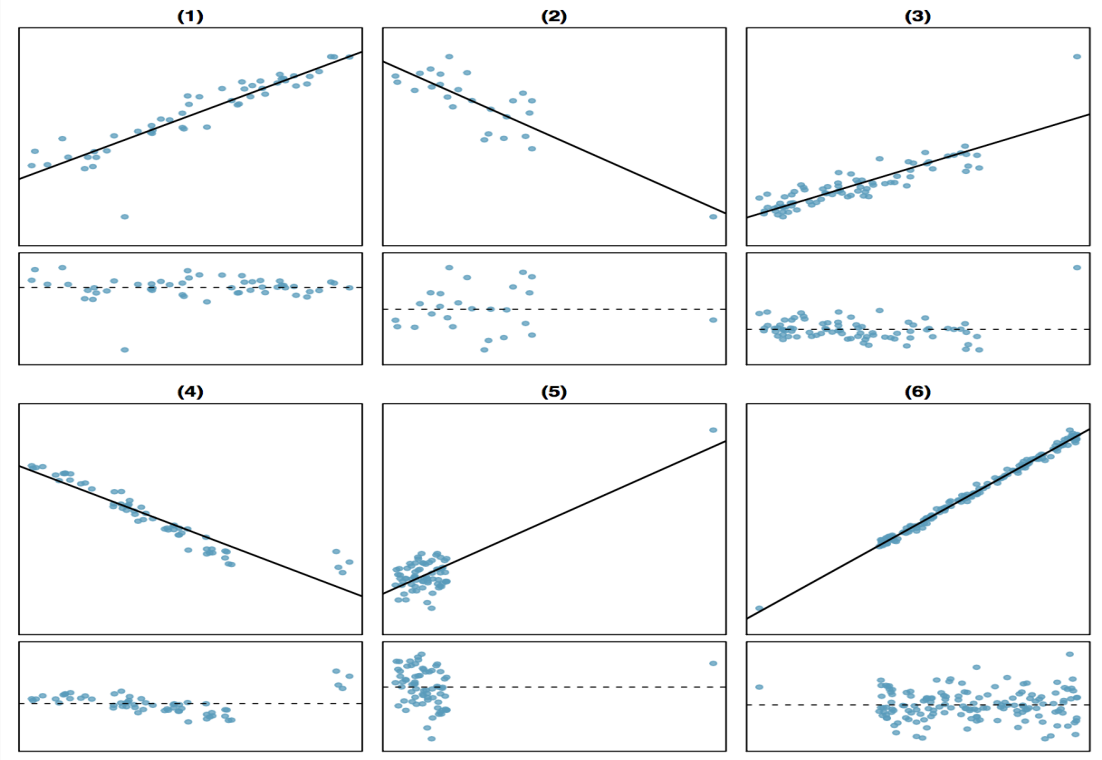
Leverage measures the weight given to each point in determining the regression line.

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

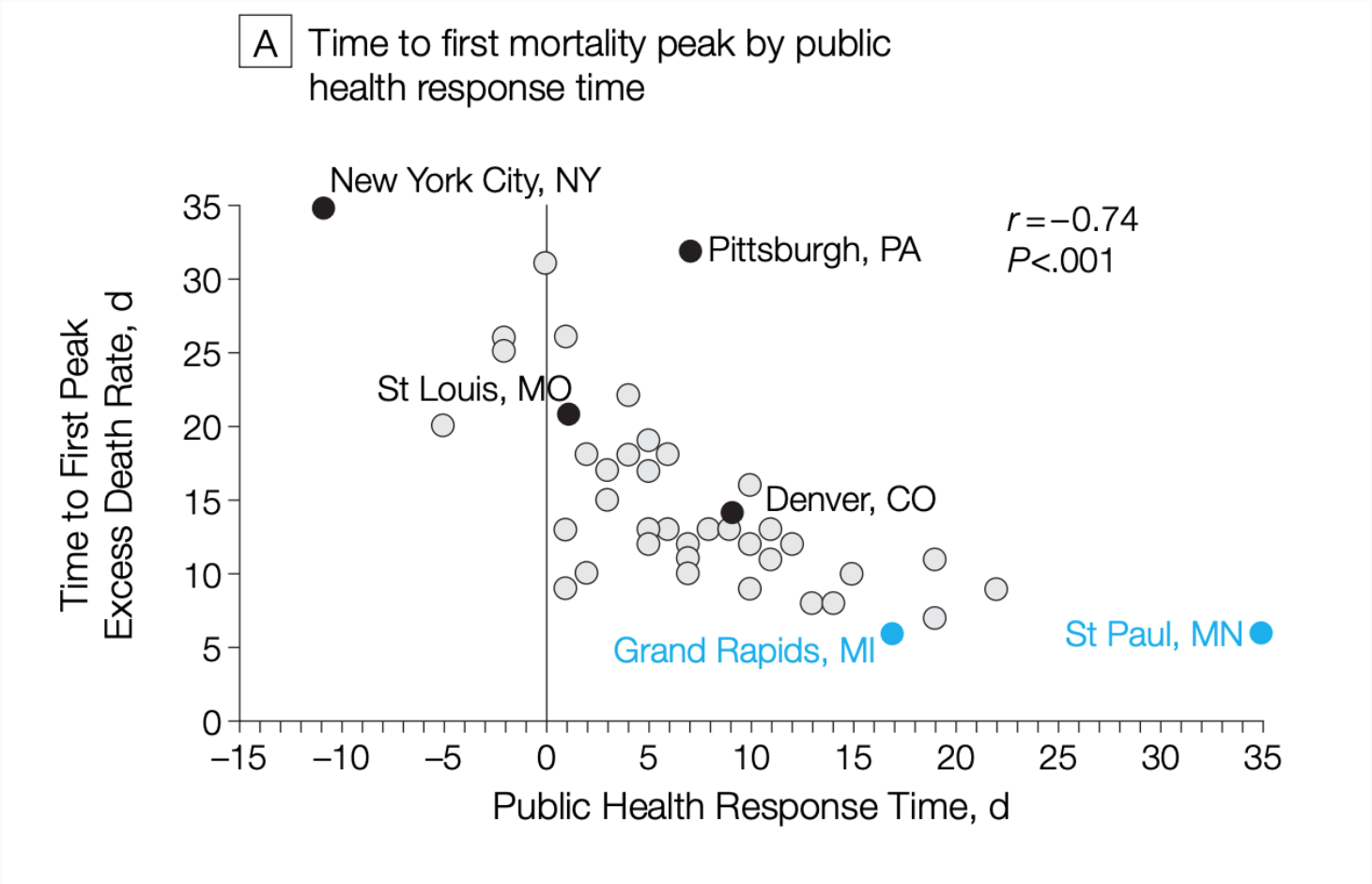


Your Turn

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



From the problem set



From the problem set