**Outliers:**

All code used is in the folder Task2 on GitHub

How do outliers affect the estimate of the coefficients?

Outliers can have an effect on the linear regression in many way and one of the ways is that it can affect the outcome of the coefficients like the correlation coefficient and variance.

The way it impacts the correlation coefficient is it can make the line steeper of less steep depending on where the outlier lies on the graph. For example if you have an outlier below the regression line it means that it has a negative residual so it brings the correlation coefficient down and therefore makes the line less steep and vice versa for if the outlier is above the line it has a positive residual so makes the coefficient bigger and makes the line steeper.

Example:

I manually plotted some points here of a plot with no outliers and it look like the graph below and has a regression line as shown below. It has the regression coefficient of 1.1583 and intercept 0.825

Now when we add one outlier to this csv file and I will add it below the line first to see the impact it has. So for the first graph below you can see the outlier is below the line at the coordinates (19, 1) and in the second the outlier is above the line with the coordinates (20, 50). In the first graph which the outlier has a negative residual the line becomes less steep and the regression coefficient became 0.229 from the 1.15 before the outlier and the intercept is 7.79 from 0.825. And for the outlier with a positive residual it became 2.276 and the intercept became -7.7 which is an increase also. This shows how much outliers can have an impact on the coefficients.

How can we detect them and remove outliers?

In the version we did above with graph and a regression line you can see that it is quite simple to spot an outlier as I have circled in the graphs because they do not fit a pattern however there are ways to find outliers if you don’t have a graph as well.

There are numerous ways to detect outliers could be with box plots, percentiles, z-score and standard deviation. And for this I will use different data to demonstrate in the csv file outliers2.csv. I will demonstrate box plots and z-score.



