Introduction to Outliers

STA 325: Lab 2, Fall 2018

Today's agenda: finding outliers

Programming partners: You should have a programming partner for each lab, and you should switch off who is programming, and use each other for help. We will spend about 30–50 minutes per week on lab exercises and you will be expected to bring your laptops to class to work on these exercises in class. Myself and the TA will be in class to help you.

Background

Identifying outliers in data is an important part of statistical analyses. One simple rule of thumb (due to John Tukey) for finding outliers is based on the quartiles of the data: the first quartile Q_1 is the value $\geq 1/4$ of the data, the second quartile Q_2 or the median is the value $\geq 1/2$ of the data, and the third quartile Q_3 is the value $\geq 3/4$ of the data. The interquartile range, IQR, is $Q_3 - Q_1$.

Tukey's rule says that the outliers are values more than 1.5 times the interquartile range from the quartiles — either below $Q_1 - 1.5IQR$, or above $Q_3 + 1.5IQR$.

In this lab, we will consider the following data

```
x <- c(2.2, 7.8, -4.4, 0.0, -1.2, 3.9, 4.9, 2.0, -5.7, -7.9, -4.9,  28.7,  4.9)
```

We will use these as part of writing a function to identify outliers according to Tukey's rule. Our function will be called `tukey.outlier`, and will take in a data vector, and return a Boolean vector, `TRUE` for the outlier observations and `FALSE` elsewhere.

Lab Tasks

1. (5) Calculate the first quartile, the third quartile, and the inter-quartile range of `x`. Some built-in R functions calculate these; you cannot use them, but you could use other functions, like `sort` and `quantile`.
2. (10) Write a function, `quartiles`, which takes a data vector and returns a vector of three components, the first quartile, the third quartile, and the inter-quartile range. Show that it gives the right answers on `x`. (You do not have to write a formal test for `quartiles`.)
3. (5) Which points in `x` are outliers, according to Tukey's rule, if any?
4. (20) Write `tukey.outlier`, using your `quartiles` function. The function should take a single data vector, and return a Boolean vector, take in a data vector, and return a Boolean vector, `TRUE` for the outlier observations and `FALSE` elsewhere. Show that it passes `test.tukey.outlier`.
5. (20) Write a function, `test.tukey.outlier`, which tests the function `tukey.outlier` against your answer in the previous question. This function should return `TRUE` if `tukey.outlier` works properly; otherwise, it can either return `FALSE`, or an error message, as you prefer.
6. (5) Which data values should be outliers in `-x`?
7. (5) Which data values should be outliers in `100*x`?
8. (10) Modify `test.tukey.outlier` to include tests for these cases.
9. (5) Show that your `tukey.outlier` function passes the new set of tests, or modify it until it does.
10. (15) According to Tukey's rule, which points in the next vector y are outliers? What is the output of your function? If they differ, explain why.

```
y <- c(11.0, 14.0, 3.5, 52.5, 21.5, 12.7, 16.7, 11.7, 10.8, -9.2, 12.3, 13.8, 11.1)
```