Homework 5: Dimensions of Anomaly

36-350, Statistical Computing

Due at 11:59 pm on Thursday, 3 October 2013

Computational agenda: Debugging and testing; making code more general.

Statistical agenda: Finding outliers; thinking about what it means to call an observation an "outlier".

Instructions: You know them by now.

In lab, you wrote a function, `tukey.outlier`, to identify outliers in a data vector according to Tukey's rule. This homework extends and applies that function. You can use your code from lab, or the code from the lab solutions; if you do the latter, clearly indicate what you are borrowing.

As stated in lab, Tukey's rule is that a point is an outlier if it is below $Q_1 - 1.5IQR$, or above $Q_3 + 1.5IQR$, where Q_1 and Q_3 are the first and third quartiles, and $IQR = Q_3 - Q_1$ is the inter-quartile range. This only applies to one-dimensional data. If each observation has several dimensions, Tukey's rule says it is an outlier if it is an outlier on *any* dimension.

1. (25) *Handling missing values* Load the rainfall data set from homework 1 (`http://www.stats.uwo.ca/faculty/braun/data/rnf6080.dat`).
   (a) (5) The entries of `-999` represent missing observations, not hours of negative rainfall. Replace the negative numbers with `NA`.
   (b) (5) Run the 6th column of the cleaned data through your `tukey.outlier` function. What error message do you get? Where is the error happening? Why is it happening?
   (c) (5) Write a test case, based on the `x` vector from lab, which shows how you would like your outlier-detector to handle `NA` values. Add it to your testing function.
   (d) (5) Modify your code for `tukey.outlier` until it passes all your test cases, including the new one with `NA`. What did you have to change? *Hint:* `?quantile`.
   (e) (5) How many observations in the 6th column of the rainfall data are anomalies according to your improved `tukey.outlier`? How many are anomalies in the whole data set?

2. (25) *Outlier words?* The R workspace `http://www.stat.cmu.edu./~cshalizi/statcomp/13/hw/05/hw-05.RData` contains an object, `hd_table`, which gives the counts for the number of times different words were used in a famous novel. (Look at recipes 9.3 and 10.20 in *The R Cookbook* for basics of tables, and section 6.3 in Matloff for more.) The names for each component are the words, and the values are the number of times that word was used.

   (a) (4) How many times did the novel use the word "river"? How many times did it use the word "blood"? How many times did it use the word "bloods"? How many times did it use the word "rivers"?

   (b) (4) What was the length of the novel, in words? How many distinct words did the novel use?

   (c) (4) Create a vector which indicates, for each word in `hd_table`, whether or not it is an outlier, by Tukey's rule.

   (d) (4) Create a vector which gives the outlier words and their counts.

   (e) (4) What are the 20 most common words which the rule says are outliers? How often do they occur?

   (f) (5) In large samples of text, the most common words in English are "the", "of", "a", "and", "to", typically each being a few percent of the total number of words. Is the novel anomalous in how often it uses these words?

3. (45) *Multiple dimensions* Consider the following data:

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

   (a) (5) Which rows of `z` ought to be considered outliers, according to Tukey's rule? Why?

   (b) (5) Here is some code which tries to implement Tukey's rule in multiple dimensions. It has a bug. (It's online at `http://www.stat.cmu.edu/~cshalizi/statcomp/13/hw/05/hw-05.R`.)

```
tukey_multiple <- function(x) {
  outliers <- array(TRUE,dim=dim(x))
  for (j in 1:ncol(x)) {
    outliers[,j] <- outliers[,j] && tukey.outlier(x[,j])
  }
  outlier.vec <- vector(length=nrow(x))
  for (i in 1:nrow(x)) {
     outlier.vec[i] <- all(outliers[i,])
  }
  return(outlier.vec)
}
```

What happens when you run this on `z`? How do you know that this is wrong?

(c) (5) Explain what the bug is, and how you know that.

(d) (5) Fix the bug. Verify that the corrected code works properly on `z`.

(e) (5) Modify this code to get rid of the loops. Make sure it still works on `z`. *Hint:* `apply`.

(f) (10) Modify your `tukey.outlier` function so that it still works on vectors, but if it is given an array, it returns a Boolean vector indicating which rows are outliers. Modify your `test.tukey.outlier` function so that it tests both all the old vector cases, and `z` as an array case. Make sure your new `tukey.outlier` works in all your test cases.

(g) (5) How many days in the rainfall data set are outliers?

4. (5, extra credit) What is the novel `hd_table` comes from?