Introduction to Functions, Part II

Rebecca C. Steorts, Duke University

STA 325, Supplemental Material

Agenda

- Multiple functions: Doing different things to the same object
- Sub-functions: Breaking up big jobs into small ones
- Example: Back to Tukey's outlier method

Why Functions?

- Data structures tie related values into one object
- Functions tie related commands into one object
- Both data structures and functions are easier to work with

Defining a function

```
function.name <- function(arguments){
  # computations on the arguments
  # some other code
  # return desired output
}
```

What should be a function?

- Things you're going to re-run, especially if they will be re-run with changes
- Chunks of code you keep highlighting and hitting return on
- Chunks of code that are small parts of bigger analyses
- Chunks that are very similar to other chunks

Why You Have to Write More Than One Function

Meta-problems:

- You've got more than one problem
- Your problem is too hard to solve in one step
- You keep solving the same problems

Meta-solutions:

- Write multiple functions, which rely on each other
- Split your problem, and write functions for the pieces
- Solve the recurring problems once, and re-use the solutions

Writing Multiple Related Functions

- In machine learning, we want to do lots of things with models/algorithms: estimate, predict, visualize, test, compare, simulate, uncertainty, etc.
- Write multiple functions to do these things
- We can then call the functions in order to meet our underlying goal (estimation, prediction, visualization, etc.)

Consistent Interfaces

- Functions for the same kind of object should use the same arguments and presume the same structure
- Functions for the same kind of task should use the same arguments and return the same sort of value

(to the extent possible)

Keep related things together

- Put all the related functions in a single file
- Source them together
- Use comments to note ***dependencies***

Rest of the module

In the rest of the module, we will do the following:

- review Tukey's method for outliers
- review the three functions (from lab 2) needed for Tukey's method
- look at an exercise at how all these functions work together on the rainfall dataset from your first homework assignment

Tukey's method

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 $Q_3 + 1.5IQR$.

Testing for outliers

We have three fuctions:

1. `quartiles`: Outputs the quartiles given a vector
2. `tukey.outlier`: Tells us what indices of a vector are the outlying values
3. `test.tukey.outlier`: Testing function to make sure our tukey.outlier function is working properly

Testing for outliers

```
# Input: data
# Output: quartiles (first, third, iqr)
quartiles <- function(x) {
  q1<-quantile(x,0.25,names=FALSE)  # suppress undesired percentile names
  q3<-quantile(x,0.75,names=FALSE)
  quartiles <- c(first=q1,third=q3,iqr=q3-q1)
  return(quartiles)
}

# Input: data
# Output: outliers according to Tukey's rule
# Dependence: quartiles()
tukey.outlier <- function(x) {
  quartiles <- quartiles(x)
  lower.limit <- quartiles[1]-1.5*quartiles[3]
  upper.limit <- quartiles[2]+1.5*quartiles[3]
  outliers <- ((x < lower.limit) | (x > upper.limit))
  return(outliers)
}

# Input: no arguments
# Output: TRUE if all tests passes,
# o/w error at first failed test
# Dependence: tukey.outlier(), quartiles()
test.tukey.outlier <- function() {
  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)
  x.pattern <- rep(FALSE,length(x)); x.pattern[12] <- TRUE
  stopifnot(all(tukey.outlier(x) == x.pattern))
  return(TRUE)
}
```

Quartiles and Outliers

```
# define a vector of 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)
quartiles(x) # will give the quartiles
```

```
## first third   iqr
##  -4.4   4.9   9.3
```

```
tukey.outlier(x) # output the indices of the outlying data point(s)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FAL
## [12]  TRUE FALSE
```

```
which(tukey.outlier(x)) # outliers
```

```
## [1] 12
```

What is the point of `test.tukey.outlier()`?

- I have a function called `tukey.outlier()`
- How do I know that it works the way that it should?
- A ***testing*** function allow one to perform santity checks to make sure the function behaves as intended

Testing function

```
test.tukey.outlier() # do all the conditions pass? yes!
```

```
## [1] TRUE
```

Rainfall data set

Let's look at a real example now, and one we've seen before with missing values – the rainfall dataset (see homework 1).

```r
# read in rainfall data set
rain <- read.table("data/rnf6080.dat")
```

Our goal will be to walk through five tasks, where we will call the three functions from Tukey's method.

Tasks

1. The entries of `-999` represent missing observations, not hours of negative rainfall. Replace the negative numbers with `NA`.
2. 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?
3. 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.
4. 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: consider the `quantile` function.
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?

Task 1

The entries of `-999` represent missing observations, not hours of negative rainfall. Replace the negative numbers with `NA`.

Task 1

We simply replace all values with a -999 with an NA

```
rain[rain==-999] <- NA
```

Task 2

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?

That is, run the following command in the console (after making sure to load in the functions).

```
rain <- read.table("data/rnf6080.dat")
tukey.outlier(rain[,6])
```

Task 2

we obtain the error `Error in quantile.default(x, 0.25) : missing values and NaN's not allowed if 'na.rm' is FALSE`

Task 2

To understand the error further, we use the traceback command in the console to inspect this further.[1]

This error makes it look like the problem might be happening when we call the built-in function quantile, in our quartile function. Using traceback confirms this.

```
5: stop("missing values and NaN's not allowed
  if 'na.rm' is FALSE")
4: quantile.default(x, 0.25, names = FALSE)
3: quantile(x, 0.25, names = FALSE) at #2
2: quartiles(x) at #2
1: tukey.outlier(rain[, 6])
```

[1] 1 is the command that we ran. 2 – 5 are the issues that came up from running command 1.

Task 2

This makes it seem like `quantile` is refusing to process the data, because it contains `NA` values. Does it?

```
length(rain[,6]) #number of values in column 6
```

```
## [1] 5070
```

```
sum(is.na(rain[,6])) #number of NA values in column 6
```

```
## [1] 111
```

Yes, note that 111 of the 5,070 values in the 6th column are NA.

Task 3

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.

Task 3

When we have NA values, we need to decide whether we are going to say that they are outliers (return `TRUE`), or say that missing values are not outliers (return `FALSE`), or refuse to say either way (return `NA`).

The last two make more sense. We'll consider these two cases.

Task 3 (Case 1)

Case 1: Suppose we want the `NA`s to NOT count as outliers. Our test case might look like the following:

```
# Initialize to data
x.with.nas <- x
# Create a vector which modifies data slightly by
# adding an NA in the middle, at position 7
x.with.nas[7] <- NA
# stop if the pattern is not matched
stopifnot(all(tukey.outlier(x.with.nas)==x.pattern))
```

These three lines can be added to our testing function: `test.tukey.outlier`.

Task 3 (Case 2)

Case 2: Suppose we want to return `NA`. Our test case might look like the following:

```
x.with.nas <- c(x,NA); x.with.nas.pattern <- c(x.pattern,NA)
stopifnot(identical(tukey.outlier(x.with.nas),x.with.nas.pattern))
```

(Annoyingly, `NA==NA` evaluates to `NA`, while `identical(NA,NA)` is `TRUE`.)

Task 3

- The current code for `tukey.outlier` will fail this test case.
- In fact, running `test.tukey.outlier` at this stage should produce the same error message as in the previous question.
- To keep the solutions to a reasonable length, we'll just cover the option of making `NA`s NOT outliers, rather than returning NA for them.

Remark: We will modify the `tukey.outlier` code in the next task section.

Task 4

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: consider the `quantile` function.

Task 4

The `quantile` function has an option for ignoring NA values in the vector we give it, `na.rm`. So we re-define `quartile` to make use of this. Below, we re-define all the functions to incorporate a new test when faced with an NA value (missing is not an outlier).

Task 4

We will adjust each of the three functions to handle NA values (missing is not an outlier).

```
# Redefine quartile to handle NAs
quartiles <- function(x) {
  # na.rm = TRUE removes the NA values
  q1<-quantile(x,0.25,na.rm=TRUE,names=FALSE)
  q3<-quantile(x,0.75,na.rm=TRUE,names=FALSE)
  quartiles <- c(first=q1,third=q3,iqr=q3-q1)
  return(quartiles)
}
```

Task 4

```r
# Redefine tukey.outlier to output FALSE where the input value
# is NA i.e., missing values are never outliers
tukey.outlier <- function(x) {
  quartiles <- quartiles(x)
  lower.limit <- quartiles[1]-1.5*quartiles[3]
  upper.limit <- quartiles[2]+1.5*quartiles[3]
  # Here, we still have NA's that need to be removed
  outliers <- ((x < lower.limit) | (x > upper.limit))
  # Replace all NA values to FALSE
  # We are treating them as not being outliers
  outliers[is.na(outliers)] <- FALSE
  return(outliers)
}
```

Task 4

```r
# Re-defining test.tukey.outlier to incorporate a new test case
  # Return FALSE when faced with an NA value, i.e., missing is
  # not an outlier
test.tukey.outlier <- function() {
  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)
  # Initialize all values to FALSE
  x.pattern <- rep(FALSE,length(x))
  # Put the true outlier to be TRUE
  x.pattern[12] <- TRUE
  stopifnot(all(tukey.outlier(x) == x.pattern))
  stopifnot(all(tukey.outlier(-x) == tukey.outlier(x)))
  stopifnot(all(tukey.outlier(100*x) == tukey.outlier(x)))
  # Initialize the vectors with NA's
  x.with.nas <- x
  # Introduce the NA value
  x.with.nas[7] <- NA
  stopifnot(all(tukey.outlier(x.with.nas)==x.pattern))
  return(TRUE)
}
```

Task 4

Let's test this out before finishing the task.

```
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)
x.with.nas <- x
x.with.nas[7] <- NA
quartiles(x.with.nas)
```

```
##  first  third    iqr
## -4.525  4.150  8.675
```

```
test.tukey.outlier()
```

```
## [1] TRUE
```

```
tukey.outlier(x.with.nas)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FAL
## [10]  TRUE FALSE
```

Task 4

Now, we test this more fully on the rain data set.

```r
summary(tukey.outlier(rain[,6]))
```

```
##    Mode   FALSE    TRUE
## logical    4873     197
```

```r
test.tukey.outlier()
```

```
## [1] TRUE
```

Task 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?

Task 5

In the 6th column, there are

```
sum(tukey.outlier(rain[,6]))
```

```
## [1] 197
```

total observations that are anomalies according to our improved `tukey.outlier`.

In the entire data set, there are

```
sum(tukey.outlier(rain))
```

```
## [1] 20196
```

total observations that are anomalies according to our improved `tukey.outlier`.

Task 5

Remark: you may have noticed that the first three columns actually give the calendar date of the observations, and it doesn't make much sense to ask whether those are outliers or not, so it's also legitimate to give an answer as well for the total observations that are anomalies (for the entire data set):

```
sum(tukey.outlier(rain[,-(1:3)]))
```

```
## [1] 4986
```

Putting everything together

Now that we have completed these tasks, let's consider how we can handle all the final functions efficiently and a few of the tasks.

More specifically, we'll do the following:

1. We will put all our functions that we will call in a tukey.R file
2. We will then `source` this into our markdown file
3. Finally, we can check to make sure that we are able to call the functions and perform analysis on them.
4. This allows for a smoother workflow (and we don't have messy functions running about in our main document).

Putting everything together

Let's source `tukey.R`

The `source` command loads in an R script and all the functions that you have written.

```r
source("scripts/tukey.R")
```

Putting everything together

Let's look back at one of our tasks and see what out put we receive from the sourced file

```
sum(tukey.outlier(rain[,6]))
```

```
## [1] 197
```

```
sum(tukey.outlier(rain))
```

```
## [1] 20196
```

Summary

- We have seen how multiple functions can be used together to perform a task
- We have also seen how conditions can be used to improve these functions
- We have also looked at how to more easily work with markdown by sourcing our scripts once we know that they work, so we can more easily/efficiently work with functions/code