JHU 06 Course Project part 2 - ToothGrowth R dataset

Kamran Haroon

June 22, 2015

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

Overview

This report is part submission for the class project in Statistical Inference - a course in the John Hopkins University Data Science specialization on Coursera. As per the project webpage, we are required to:

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)
4. State your conclusions and the assumptions needed for your conclusions.

The data is the result of measuring the affect of different dosage amounts of Vitamin C on the length of odontoblasts (teeth) of ten guinea pigs. There are two supplement types of Vitamin C tested, Orange Juice and Ascorbic Acid, and they are given in three three different milligram dosage amount, 0.5, 1.0, and 2.0.

```
suppressWarnings(library(ggplot2))
suppressMessages(library(data.table))
library(grid)
```

Data Cleansing

From source, load the ToothGrowth data into a data.table object. Rename the columns and write a join key. To make categorizing a little more simple, we add an additional column for Dosage by converting the Dose variable into factors.

```
# Read in the data file and rename the columns
data1 <- data.table(ToothGrowth)
setnames(data1, c('len', 'supp', 'dose'), c('Length', 'Supplement', 'Dose'))

# Add 'Dosage'and write the join key for factors
data1 <- data1[,Dosage:=sapply(as.character(data1$Dose), function(x)
as.factor(switch(x, '0.5'='SM', '1'='MD', '2'='LG')))]
setkey(data1, Supplement, Dosage)
head(data1, 1)
```

```
##    Length Supplement Dose Dosage
## 1:   15.2         OJ  0.5     SM
```

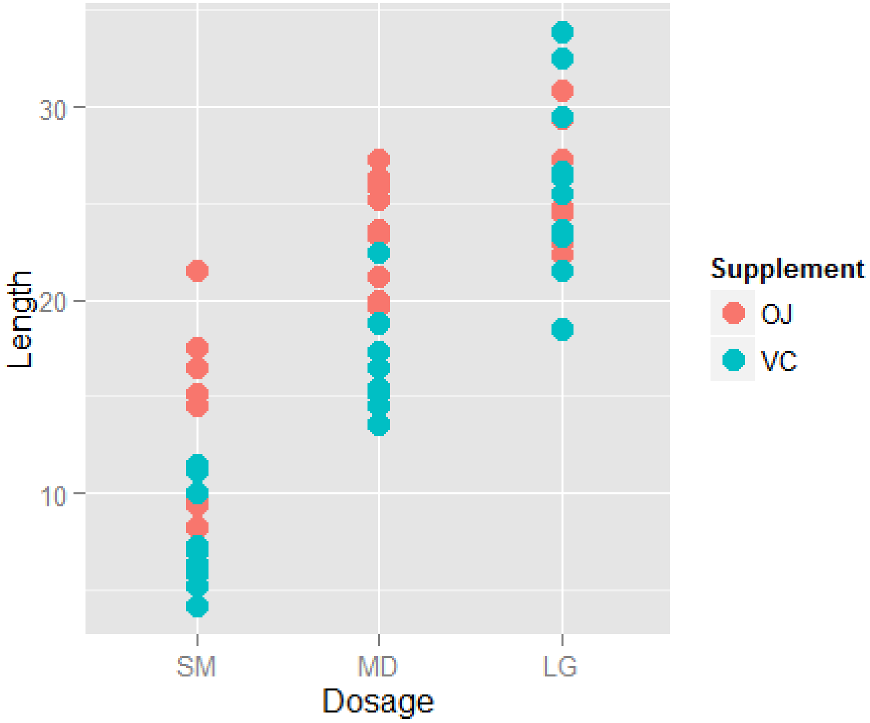
Exploratory Analysis

For a simple exploration of the date to understand the content and the structure of the data.table.

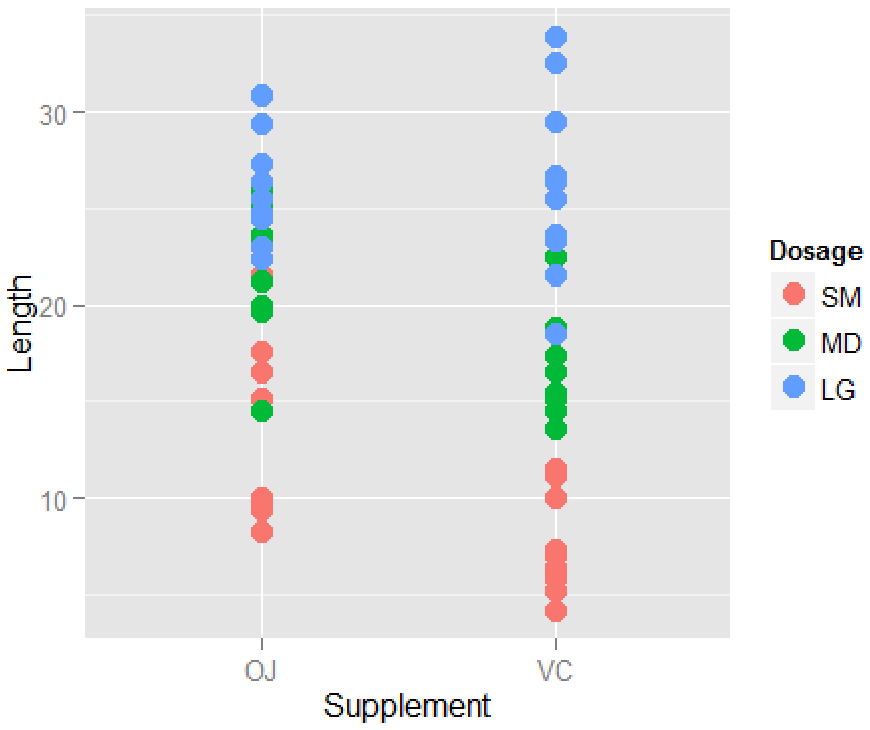
```
summary(data1)

##      Length      Supplement      Dose         Dosage
##  Min.   : 4.20   OJ:30      Min.   :0.500   SM:20
##  1st Qu.:13.07   VC:30      1st Qu.:0.500   MD:20
##  Median :19.25              Median :1.000   LG:20
##  Mean   :18.81              Mean   :1.167
##  3rd Qu.:25.27              3rd Qu.:2.000
##  Max.   :33.90              Max.   :2.000

str(data1)

## Classes 'data.table' and 'data.frame':   60 obs. of  4 variables:
##  $ Length    : num  15.2 21.5 17.6 9.7 14.5 10 8.2 9.4 16.5 9.7 ...
##  $ Supplement: Factor w/ 2 levels "OJ","VC": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Dose      : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
##  $ Dosage    : Factor w/ 3 levels "SM","MD","LG": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
##  - attr(*, "sorted")= chr  "Supplement" "Dosage"
```

Now we plot Length against both Dosage and Supplement. We observe that larger the Dosage, the longer the Length. However, we are not yet certain whether Orange Juice (OJ) or Ascorbic Acid(VC) doses result in greater teeth length.

```
# Plot 1
g1 <- ggplot(data1, aes(x=Dosage, y=Length))
g1 <- g1+geom_point(aes(color=Supplement), size=4)
print(g1)
```



```
# Plot 2
g2 <- ggplot(data1, aes(x=Supplement,y=Length))
g2 <- g2+geom_point(aes(color=Dosage),size=4)
print(g2)
```



Confidence Interval Testing

In order to understand how Vitamin C affects tooth growth, we will conduct the following confidence interval tests. We subset data1 and use the t.test R function to determine the confidence interval, subset means, and p-value for each scenario.

Compare Dosage Alone

```
t1 <- subset(data1, Dosage=='SM')$Length
t2 <- subset(data1, Dosage=='MD')$Length
t <- t.test(t1, t2, paired=FALSE, var.equal=FALSE)
t$conf.int[1:2]

## [1] -11.983781  -6.276219
```

If we increase the Vitamin C dose from 0.5 to 1.0 milligrams, the confidence interval does not contain zero, so we can conclude that dose increase does increase tooth length.

```
t1 <- subset(data1,Dosage=='MD')$Length
t2 <- subset(data1,Dosage=='LG')$Length
t <- t.test(t1,t2,paired=FALSE,var.equal=FALSE)
t$conf.int[1:2]

## [1] -8.996481 -3.733519
```

Now if we increase the Vitamin C dose from 1.0 to 2.0 milligrams, the confidence interval again does not contain zero, so we can conclude that dose increase does increase tooth length.

In both of these scenarios, an increased dose amount leads to an increased tooth length.

Compare Supplement Alone

```
t1<-subset(data1,Supplement=='VC')$Length
t2<-subset(data1,Supplement=='OJ')$Length
t<-t.test(t1,t2,paired=FALSE,var.equal=FALSE)
t$p.value

## [1] 0.06063451

t$conf.int[1:2]

## [1] -7.5710156  0.1710156
```

In this single comparison, the p-value is 0.061 and the confidence interval contains zero so we can conclude that the choice of Vitamin C supplement alone does not affect tooth growth.