P2_Basic_Inferential_Data_Analysis

Jerad Acosta

December 21, 2014

```r
library(plyr)
library(ggplot2)
library(datasets)
data(ToothGrowth)
data <- ToothGrowth
```

1 Load Libraries and Data

Then *get to know the data*

```r
# dim(data) tells us there are 3 variables and 60 observations
dim(data)
```

```
## [1] 60  3
```

```r
# summary(data) gives us the name of the variables as well as their range
summary(data)
```

```
##       len        supp         dose     
##  Min.   : 4.2   OJ:30   Min.   :0.50  
##  1st Qu.:13.1   VC:30   1st Qu.:0.50  
##  Median :19.2           Median :1.00  
##  Mean   :18.8           Mean   :1.17  
##  3rd Qu.:25.3           3rd Qu.:2.00  
##  Max.   :33.9           Max.   :2.00  
```

```r
unique(data$dose)
```

```
## [1] 0.5 1.0 2.0
```

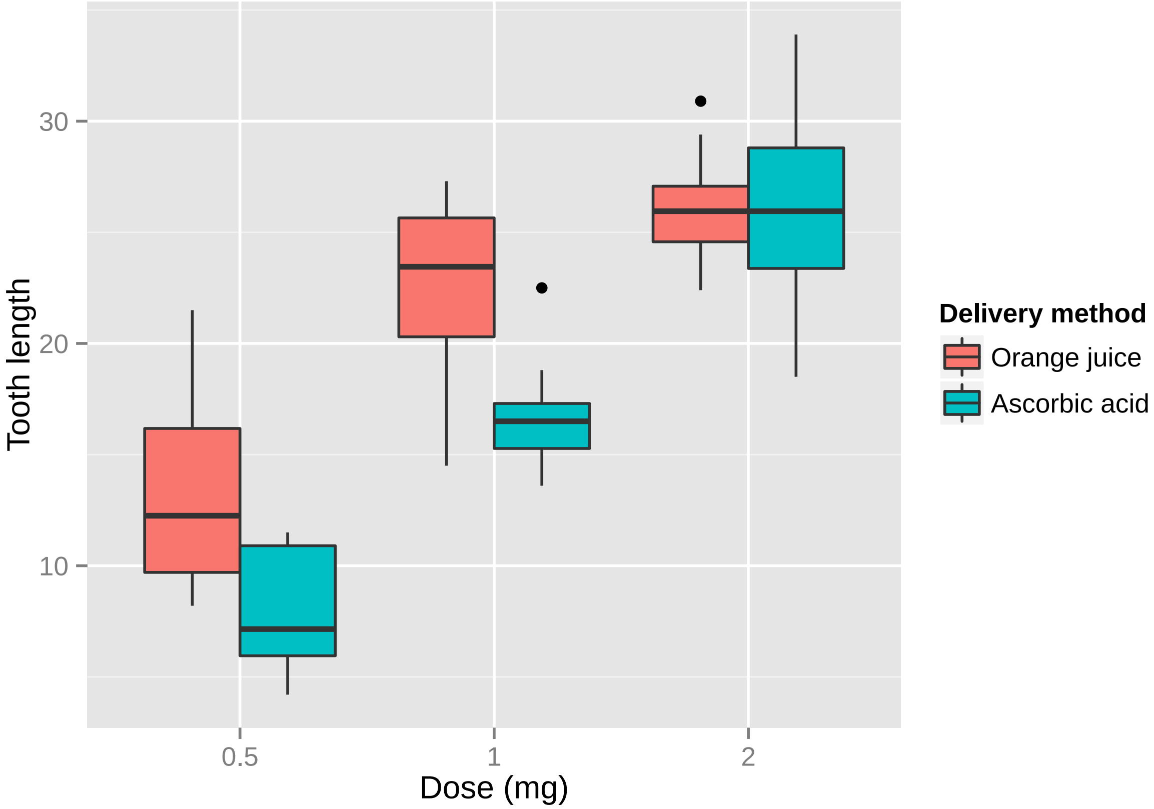
Now we know that of the 3 variables one, **supp** is a factor with two levels which represent different types of delivery methods (Orange Juice or Ascorbic acid) and the other two, **len** and **dose** are numeric but the unique() function tells us that there are only 3 values for **dose**

A little reading into into the dataset with **?ToothGrowth** and we can see that we are measuring tooth length against the two factors **supp** and amount or **dose**

2 basic summary of the data

Thus it seems like the best way to plot out data would be to break it into both **dose** and **supp** values and factors against the tooth length or **len** variable

```
ggplot(ToothGrowth, aes(x = factor(dose), y = len, fill = supp)) +
    xlab("Dose (mg)") +
    ylab("Tooth length") +
    scale_fill_discrete(name="Delivery method",
                breaks=c("OJ", "VC"),
                labels=c("Orange juice", "Ascorbic acid")) + geom_boxplot()
```



The boxplot seems to immediately suggest that a higher **dose** is associated with more **len**. However, while this pattern seems to hold true across the different delivery methods **supp** factors, there doesn't appear to be as significant of a relationship between which *delivery method* or **supp** is used.

To be sure, however, we should run some statistical methods to see how we can quantify this relationship between **dose** and **len** and or lack of one between the factors of **supp** and **len**

3 Confidence intervals using a Two Factor ANOVA test

Since we are trying to find the predictive effect of two regressors **dose** and **supp** on a predictor value **len** it is necessary to remove any linear relationships and confounding coefficient data. Rather that factoring each regressor out from each other one at a time, the ANOVA test is a more efficient and sufficiently detailed methodology

```
data$dose<- as.numeric(as.character(data$dose))
data$dose <- factor(data$dose)
data$supp <- factor(data$supp)
fit <- aov(len ~ dose*supp, data = data)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)    
## dose         2   2426    1213   92.00 < 2e-16 ***
## supp         1    205     205   15.57 0.00023 ***
## dose:supp    2    108      54    4.11 0.02186 *  
## Residuals   54    712      13                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

While the Orange Juice factor of **supp** appears to have a promising positive slope, there is quite a bit of overlap between 1mg and 2mg **dose** so a hypothesis test to see if there is a statistical significance in the increase from 1mg to 2mg **dose**

```r
# Seperate the three subgroups
ToothGrowth.doses_0.5_1.0 <- subset (ToothGrowth, dose %in% c(0.5, 1.0))
ToothGrowth.doses_1.0_2.0 <- subset (ToothGrowth, dose %in% c(1.0, 2.0))
ToothGrowth.doses_0.5_2.0 <- subset (ToothGrowth, dose %in% c(0.5, 2.0))
# Check for group differences due to different dose levels (1.0, 2.0)
t.test(len ~ dose, data = ToothGrowth.doses_1.0_2.0)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len by dose
## t = -4.901, df = 37.1, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996 -3.734
## sample estimates:
## mean in group 1 mean in group 2 
##           19.73           26.10
```

```r
# Check for group differences due to different dose levels (0.5, 1.0)
t.test(len ~ dose, data = ToothGrowth.doses_0.5_1.0)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  len by dose
## t = -6.477, df = 37.99, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.984  -6.276
## sample estimates:
## mean in group 0.5   mean in group 1 
##             10.61             19.73
```

```r
# Check for group differences due to different dose levels (0.5, 2.0)
t.test(len ~ dose, data = ToothGrowth.doses_0.5_2.0)
```

```
## 
##  Welch Two Sample t-test
## 
```

```
## data:  len by dose
## t = -11.8, df = 36.88, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.16 -12.83
## sample estimates:
## mean in group 0.5   mean in group 2
##             10.61             26.10
```

For each dose of the three potential dose combinations , the p-value is less than 0.05 and non of the confidence intervals contain a zero value so we can confidently reject the null hypothesis that increase in **dose** is not correlated with increase in **len**

While delivery method or **supp** does not appear to have a significant affect we should still check for any statistical significance.

```
ddply(ToothGrowth,dose~supp,function(x) c(mean=mean(x$len),confidence.intervall=t.test(x$len)$conf.int)
```

```
##   dose supp  mean confidence.intervall1 confidence.intervall2
## 1  0.5   OJ 13.23                10.040                16.420
## 2  0.5   VC  7.98                 6.015                 9.945
## 3  1.0   OJ 22.70                19.902                25.498
## 4  1.0   VC 16.77                14.971                18.569
## 5  2.0   OJ 26.06                24.161                27.959
## 6  2.0   VC 26.14                22.708                29.572
```

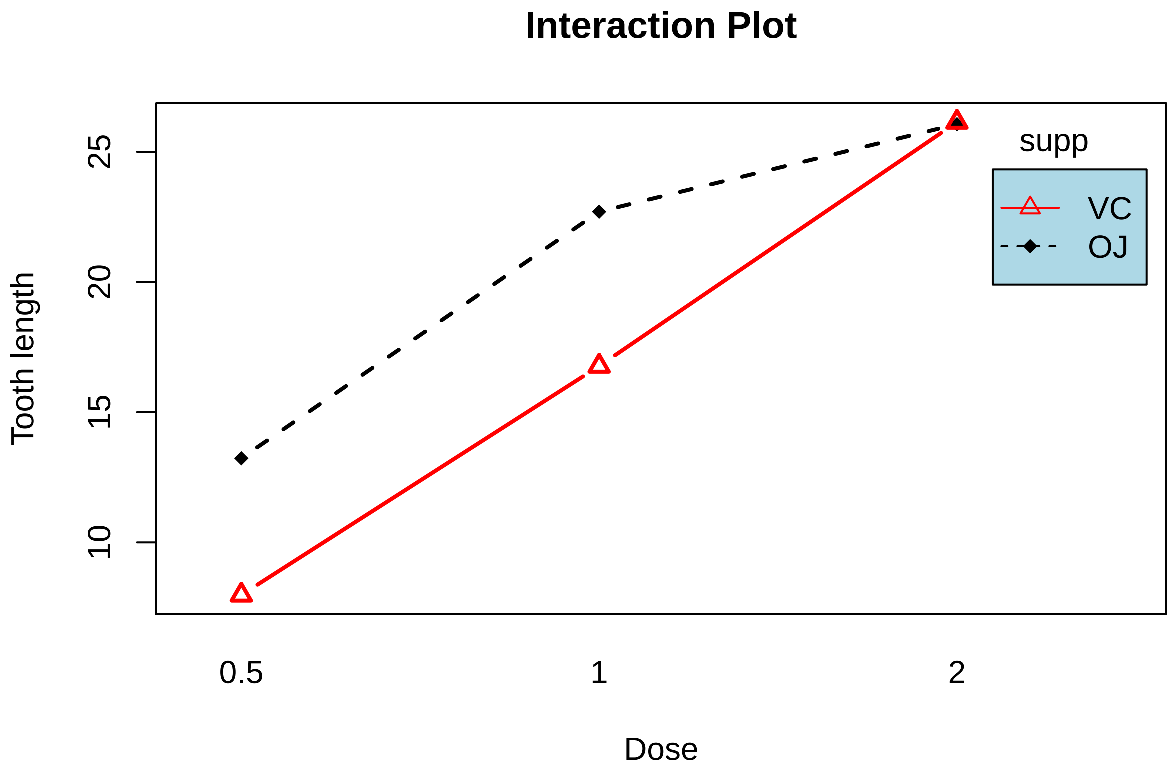
Now that we can see that with these confidence intervals that **supp** does have an effect. Particularly in lower **dose** ranges. This is such that Orange Juice has a significant positive slope over Ascorbic acid in the 0.5mg **dose** range and even more so in the 1mg **dose** range. The 2mg range, however is overlapping and rather inconclusive.

4 Conclusion

Based on removing the develery factor and seeing a significant positive slope with an increase in **dose** to an increase in **len** we could at first conclude at the minimum there was at least a single simple relationship. After controlling for **dose**, however, it became clear that certain relationships were supportable; namely, that at lower doses, Orange Juice has a more positive coefficient with **len** than Ascorbic Acid. Thus, it would be prudent to suspect that there is some mixture of effects going on here. Perhaps both **supp** factors are a positive regressor on **len** but when looking at a single dose - particularly in the lower and even mid ranges - there appears to be a Positive Ceofficient for Orange Juice and a negative or less or an affect correlated with the Ascorbic Acid.

This becomes most evident when we use an interaction plot as such

```
with(data, {
interaction.plot(dose, supp, len, type="b", col=c(1:3),
                 leg.bty="o", leg.bg="light blue", lwd=2, pch=c(18,24,22),
                 xlab="Dose",
                 ylab="Tooth length",
                 main="Interaction Plot")
})
```



And the difference in Coefficient for each regressor is dependent on another regressor, which in this case is the **dose** variable