Statistical Inference Course Project, Part2 - ToothGrowth analysis

Rodrigo

September 24, 2015

In this second part of the project, we analyze the ToothGrowth data in the R datasets package. The dataset contains 60 observations, length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1 and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

First we load the data and perform some basic exploratory data analyis

```
library(ggplot2)
library(datasets)
data(ToothGrowth)
str(ToothGrowth) # internal strcuct of dataset
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

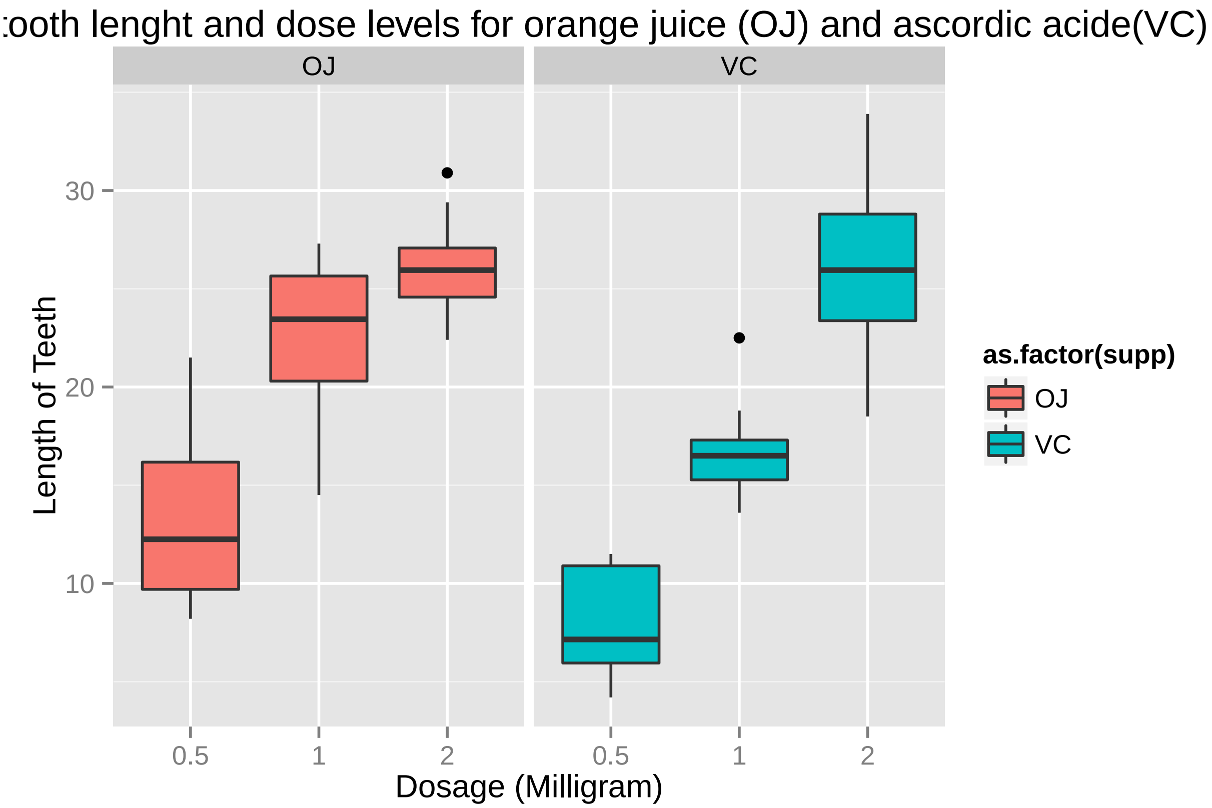
```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
summary(ToothGrowth)
```

```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000  
```

```
plot <- ggplot(ToothGrowth,
               aes(x=as.factor(dose),y=len,fill=as.factor(supp)))
plot + geom_boxplot(notch=F) + facet_grid(.~supp) +
     scale_x_discrete("Dosage (Milligram)") +
     scale_y_continuous("Length of Teeth") +
     ggtitle("Relation tooth lenght and dose levels for orange juice (OJ) and ascordic acide(VC)")
```



The chart aboove shows that there is a rcorrelation between the tooth length and the dose levels of Vitamin C, for both delivery methods: orange juice (OJ) and ascordic acide(VC).

Tabulating the delivery method by the dose levels of vitamin C

```
table(ToothGrowth$supp, ToothGrowth$dose)
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

We can also use regression analysis to identify the effect of the dose. The regression anslysis can help us to answer the following question: "How much of the variance in tooth length, if any, can be explained by the supplement type?"

```
fit <- lm(len ~ dose + supp, data=ToothGrowth)
summary(fit)
```

```
##
## Call:
## lm(formula = len ~ dose + supp, data = ToothGrowth)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -6.600 -3.700  0.373  2.116  8.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
## dose          9.7636     0.8768  11.135 6.31e-16 ***
## suppVC       -3.7000     1.0936  -3.383   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 57 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6934
## F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

One can see that there is 70% of the variance in the data and The intercept 9.2725, means that with no supplement of Vitamin C, the average tooth length is 9.2725 units. The coefficient of `dose` is 9.7635714. One can interpreted this as an increasing the delivering dose 1 mg could increase the tooth length 9.7635714 units.

The last coefficient is for the supplement type. Since the supplement type is a categorical variable, dummy variables are used. Moreover, delivering a given dose as ascorbic acid, without changing the dose, would result in 3.7 units of decrease in the tooth length.

Using Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose:

```
confint(fit)
```

```
##                 2.5 %    97.5 %
## (Intercept)  6.704608 11.840392
## dose         8.007741 11.519402
## suppVC      -5.889905 -1.510095
```

We can conclude that collecting a different set of data and estimate parameters of the linear model over end over again, the coefficient estimations will be in the ranges above in 95% of the time.

Thi null hypothesis in this case is that the coefficients are zero, and the no tooth length variation is explained by that variable. As one can see the p-values are less than 0.5. As a result the null hypothesis is rejecting. In addition, each variable explains a significant portion of variability in tooth length, assuming the significance level is 5%.