Statistical Inference - Project 2 Tooth Growth Data

Rajesh Thallam

Overview

In the second part of the project, we analyze the ToothGrowth data in the R datasets package. The data is set of 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).

Exploratory Data Analysis

The ToothGrowth data set explains the relation between the growth of teeth of guinea pigs at each of three dose levels of Vitamin C with each of the two delivery methods of orange juice and ascorbic acid. In this section, I first load the data set and then look at the variables in the data set.

```r
library(datasets)
data(ToothGrowth)
str(ToothGrowth)
```

```
## '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 ...
```

So, the data set has 3 variables (len, supp, and dose which are numeric, factor, and numeric variables, respectively) and 60 observations. For each of the two causal factors supp and dose.

```r
unique(ToothGrowth$supp)
```

```
## [1] VC OJ
## Levels: OJ VC
```

```r
unique(ToothGrowth$dose)
```

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

Let’s convert the variable dose from numeric to factor. The following code takes care of this conversion and presents the new transformed data set.

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
str(ToothGrowth)
```

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

Looking at the basic summary of the data in the data set

```
summary(ToothGrowth)
```

```
##       len        supp     dose   
##  Min.   : 4.20   OJ:30   0.5:20  
##  1st Qu.:13.07   VC:30   1  :20  
##  Median :19.25           2  :20  
##  Mean   :18.81                   
##  3rd Qu.:25.27                   
##  Max.   :33.90                   
```

The above output provides the summary statistics for all of the variables in the data set. In the following section, I provide the summary of the data for each combination of dose level and delivery method.

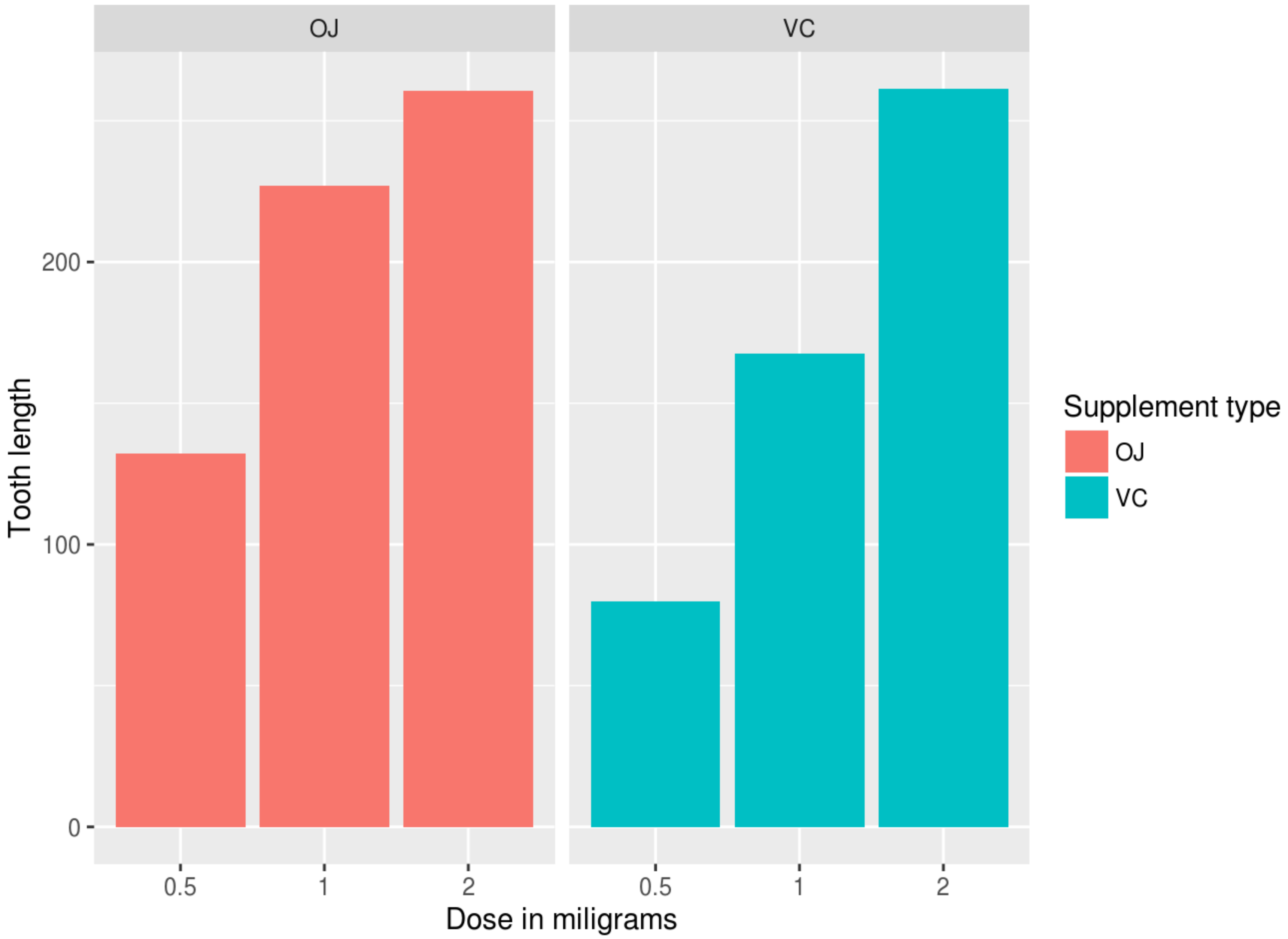
```
by(ToothGrowth$len, INDICES = list(ToothGrowth$supp, ToothGrowth$dose), summary)
```

```
## : OJ
## : 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    8.20    9.70   12.25   13.23   16.18   21.50 
## -------------------------------------------------------- 
## : VC
## : 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.20    5.95    7.15    7.98   10.90   11.50 
## -------------------------------------------------------- 
## : OJ
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   14.50   20.30   23.45   22.70   25.65   27.30 
## -------------------------------------------------------- 
## : VC
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   13.60   15.27   16.50   16.77   17.30   22.50 
## -------------------------------------------------------- 
## : OJ
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   22.40   24.58   25.95   26.06   27.08   30.90 
## -------------------------------------------------------- 
## : VC
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.50   23.38   25.95   26.14   28.80   33.90
```

Hypothesis Testing

Evaluating the effects of delivery methods on tooth growth at different levels of Vitamin C dosage

```r
library(datasets)
library(ggplot2)
ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) +
    geom_bar(stat="identity",) +
    facet_grid(. ~ supp) +
    xlab("Dose in miligrams") +
    ylab("Tooth length") +
    guides(fill=guide_legend(title="Supplement type"))
```



From the plot, there is a clear positive correlation between the tooth length and the dose levels of Vitamin C, for both delivery methods. The effect of the dosage can also be identified using regression analysis. Another question this plot can address is whether the supplement type (i.e. orange juice or ascorbic acid) has any effect on the tooth length. In other words, 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
## -7.085 -2.751 -0.800  2.446  9.650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4550     0.9883  12.603  < 2e-16 ***
## dose1         9.1300     1.2104   7.543 4.38e-10 ***
## dose2        15.4950     1.2104  12.802  < 2e-16 ***
## suppVC       -3.7000     0.9883  -3.744 0.000429 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 56 degrees of freedom
## Multiple R-squared:  0.7623, Adjusted R-squared:  0.7496
## F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16
```

The model explains 70% of the variance in the data.

- The intercept is 12.455, meaning that with no supplement of Vitamin C, the average tooth length is 12.455 units.
- The coefficient of dose is 9.13. It can be interpreted as increasing the delievered dose 1 mg, all else equal (i.e. no change in the supplement type), would increase the tooth length 9.13 units.
- The last coefficient is for the supplement type. Since the supplement type is a categorical variable, dummy variables are used. The computed coefficient is for suppVC and the value is 15.495 meaning that delivering a given dose as ascorbic acid, without changing the dose, would result in 15.495 units of decrease in the tooth length. Since there are only two categories, we can also conclude that on average, delivering the dosage as orange juice would increase the tooth length by 15.495 units.

95% confidence intervals for two variables and the intercept are as follows.

```
confint(fit)
```

```
##                 2.5 %    97.5 %
## (Intercept) 10.475238 14.434762
## dose1        6.705297 11.554703
## dose2       13.070297 17.919703
## suppVC      -5.679762 -1.720238
```

The confidence intervals mean that if we collect a different set of data and estimate parameters of the linear model many times, 95% of the time, the coefficient estimations will be in these ranges.

For each coefficient (i.e. intercept, dose and suppVC), the null hypothesis is that the coefficients are zero, meaning that no tooth length variation is explained by that variable. All p-values are less than 0.05, rejecting the null hypothesis and suggesting that each variable explains a significant portion of variability in tooth length, assuming the significance level is 5%.