Coursera Statistical Inference Project: Part II

Basically Inferential Data Analysis

Giuseppe Di Bernardo

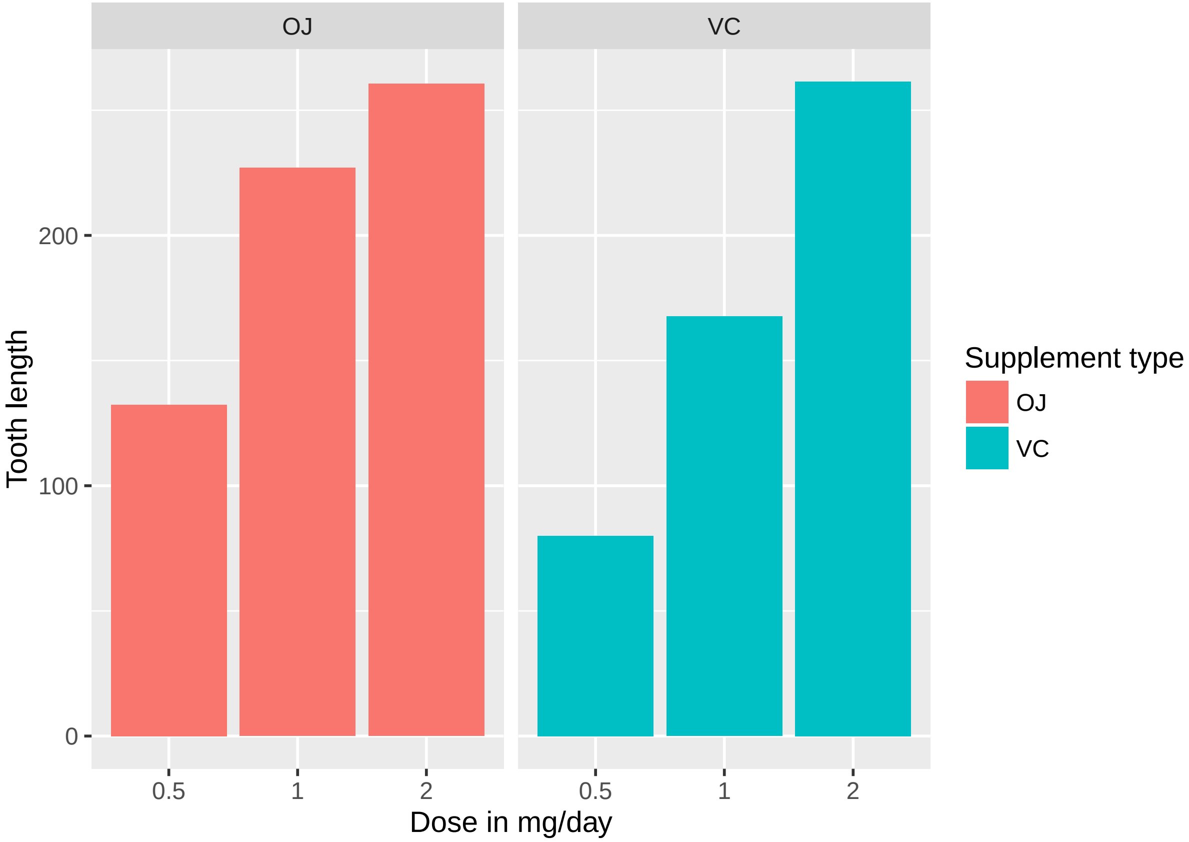
February 29, 2016

This is the Assignment Project part II of the Coursera Statistical Inference course. We are going to analyze the `ToothGrowth` data provided in the `R` dataset packages, and statistically infer conclusions about the relationship between the length of odontoblasts (cells responsible for tooth growth) and the `dose` levels of vitamin C (in milligrams/day), by one of two delivery methods (`supp` factor, e.g. orange juice `OJ`, or ascorbic acid coded as `VC`).

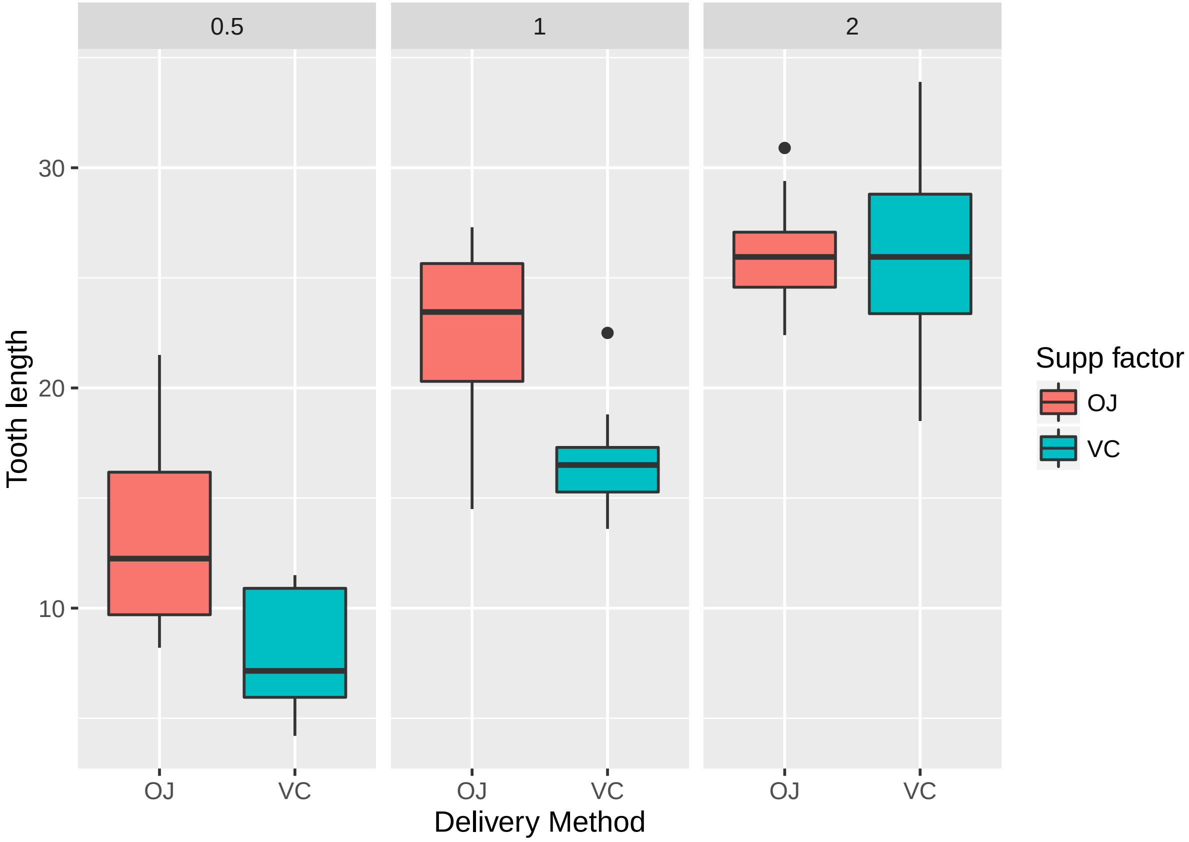
1. Data Analysis

Preliminarly, we perform a basic exploratory data analysis.

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```



A clear correlation between the tooth length and the dose levels of vitamin C - for both delivery methods - can be oserved in the above histogram plot: we may conclude that the larger is the dosage, the longer is the tooth. However, the relation between the length and the supplement type is not so immediate, at least at this stage. In fact, at low dosage, orange juice seems to positively correlate with longer teeth, more than the ascorbic acid, but at higher dosages this difference is not significant anymore, as can be observed in the below boxplot:



2. Statistical Inference

Let us now quantify the effect of the dose on the length of the teeth. In particular, we may be interested in answering the following question: how much of the variability in the tooth length, if any, can be explained by the supplement type (i.e., orange juice or ascorbic juice)? We claim to address this issue through a linear regression analysis, by using the method `lm` of the `R` linear regression fit:

```r
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
```

According to the `R-squared` value, we infer that the 70% of the variation in the data are explained by the adopted model. Regarding the correlation with the amount of the dose, we notice that the corresponding coefficient is 9.764, which means that increasing the delivered dose of $1mg$, all else equal (i.e. no change in the supplement type), would increase the length of the teeth of corresponding 9.764 units. The intercept is equal to 9.273, i.e. without supplement of vitamic C we get the average tooth lenght of 9.2725 units. The `suppVC` coefficient is for the supplement type categorical variable. The computed coefficient has a value of -3.7, which implies a decrease of 3.7 units in the tooth length by delivering a given dose as ascorbic acid, without changing the total dose. Then we conclude that the lenght of the teeth will increase of a same amount of 3.7 units if the delivering method of dosage is the orange juice `OJ`. Finally, to test the statistical significance of linear regression coeffcients (`intercept`, `dose` and `suppVC`) we show the results of the 95% of confidence intervals for the variables of dose and delivery methods, and the intercept as well.

```
confint(fit)
```

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

The null hypothesis H_0 corresponds to all null coefficients, and the hope is to reject H_0, i.e. variation of tooth length cannot be explained by the variable used. From the above computation values, we conclude that our coefficient estimated are statistically significant at the 5% of level (i.e. we cannot reject the alternative hypothesis H_a, that coeffiecients $\neq 0$).