Analysis of variance within the R data set ToothGrowth

Michael Hunt

Sunday, June 21, 2015

First, we inspect the data

```
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 ...
```

The dose data is numerical, so we convert it to a factor with three levels, named "low", "medium" and "high".

```
ToothGrowth$dose = factor(ToothGrowth$dose,levels=c(0.5,1.0,2.0),
                          labels=c("low","med","high"))
```

Visualisations of the Data

Boxplots of the data by dose and supp

See Figure 1 in the appendix. This suggests that orange juice leads to longer teeth than vitamin C, at least at low dosages. The variances of the different sets are not the same, which may affect later analyses in which they are presumed to be the same. A later Bartlett test check will investigate this. However we note that the balance of the data, already checked, makes our analysis robust to differences in group variance.

Boxplots by dose type - main effect of dose type

See Figure 2 in the appendix.. The means for different dose types appear to be similar - perhaps there is not much of a main effect due to dose type. The means for different dose strengths appear to be higher for higher doses. Dose strength does appear to have a main effect.

An interaction plot:

A final exploratory visual analysis is an interaction plot. See Figure 3 in the appendix. The slope of the lines is similar, confirming the suggestion that dose strength does have a main effect, but giving no suggestion that dose type does so.

Validity checks for ANOVA analysis- pre analysis

Check balance of the data

For a balanced data set, we need to assure ourselves that there are equal numbers of data with each combination of supp and dose. A table will show if this is so.

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

```
##
##      low med high
##   OJ  10  10   10
##   VC  10  10   10
```

The data set is balanced. This means that the later ANOVA analysis we will carry out will be less sensitive to any difference in variance between the sample groups.

Homogeneity of group variances

```r
with(ToothGrowth, tapply(len, list(supp,dose), var)) # find the variances
```

```
##       low       med      high
## OJ 19.889 15.295556  7.049333
## VC  7.544  6.326778 23.018222
```

{r check for significant difference between variances} # do a Bartlett test bartlett.test(len ~ interaction(supp,dose), data=ToothGrowth) According to ther Bartlett test, there is no significant difference between the group variances ($p > 0.05$), so in fact we can regard variance as homogeneous across the samples in the subsequent ANOVA. Strictly, this is a requirement for the validity of the analysis, as we carry it out here.

Factorial ANOVA to determine whether the differences between the group means for tooth length are significant.

```r
aov.tooth = aov(len ~ supp * dose, data=ToothGrowth)
summary(aov.tooth)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1  205.3   205.3  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *
## Residuals   54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p values here are all less than 0.05, in fact much less, which shows that at the 95% confidence level (at least) both dose and supp have a main effect, that is, there is a difference between at least two of the group means. There also appears to be an interaction between the two effects.

To determine which group means differ, we can conduct a Tukey HSD (Honest Significant Difference) test.

```r
TukeyHSD(aov.tooth)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = len ~ supp * dose, data = ToothGrowth)
##
## $supp
##       diff       lwr       upr     p adj
## VC-OJ -3.7 -5.579828 -1.820172 0.0002312
##
## $dose
##            diff       lwr       upr   p adj
## med-low   9.130  6.362488 11.897512 0.0e+00
## high-low 15.495 12.727488 18.262512 0.0e+00
## high-med  6.365  3.597488  9.132512 2.7e-06
##
## $`supp:dose`
##                   diff        lwr        upr     p adj
## VC:low-OJ:low    -5.25 -10.048124 -0.4518762 0.0242521
## OJ:med-OJ:low     9.47   4.671876 14.2681238 0.0000046
## VC:med-OJ:low     3.54  -1.258124  8.3381238 0.2640208
## OJ:high-OJ:low   12.83   8.031876 17.6281238 0.0000000
## VC:high-OJ:low   12.91   8.111876 17.7081238 0.0000000
## OJ:med-VC:low    14.72   9.921876 19.5181238 0.0000000
## VC:med-VC:low     8.79   3.991876 13.5881238 0.0000210
## OJ:high-VC:low   18.08  13.281876 22.8781238 0.0000000
## VC:high-VC:low   18.16  13.361876 22.9581238 0.0000000
## VC:med-OJ:med    -5.93 -10.728124 -1.1318762 0.0073930
## OJ:high-OJ:med    3.36  -1.438124  8.1581238 0.3187361
## VC:high-OJ:med    3.44  -1.358124  8.2381238 0.2936430
## OJ:high-VC:med    9.29   4.491876 14.0881238 0.0000069
## VC:high-VC:med    9.37   4.571876 14.1681238 0.0000058
## VC:high-OJ:high   0.08  -4.718124  4.8781238 1.0000000
```

This shows that there is indeed a main effect of the way in which the dose was administered, and also of dose strength. The supp:dose table shows that at the 95% confidence level all combinations of dose strength and dose type lead to significantly different tooth lengths, except in 4 cases (those for which p>0.2)

Validity checks for ANOVA analysis- post analysis

A residual plot (residuals vs fitted values) and qq plot are shown in Appendix 4. The residual plot should shown no pattern, and looks fine, while the qq plot should be a straight line if the normality of distribution of differences criterion for the validity of the ANOVA analysis is to be satisfied. It isn't quite, perhaps a consequence of the small sample size.

Appendix

Figure 1: Boxplots by dose and strength

```
boxplot(len ~ supp * dose, data=ToothGrowth,
        ylab="Tooth Length",main="Boxplots of Tooth Growth Data")
```

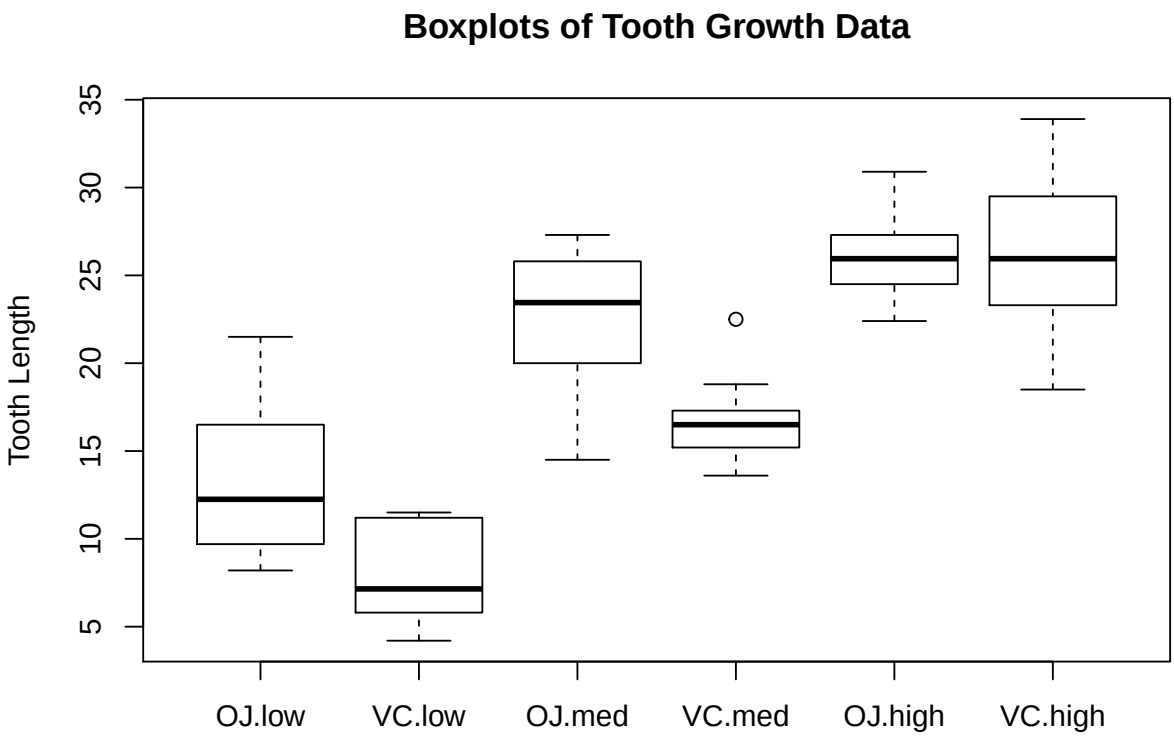


Figure 2: Boxplots of dose and strength main effects.

```
par(mfrow=c(1,2))
boxplot(len ~ supp , data=ToothGrowth,
        ylab="Tooth Length",main="Tooth Growth Data by dose type")
boxplot(len ~ dose , data=ToothGrowth,
        ylab="Tooth Length",main="Tooth Growth Data by dose strength")
```

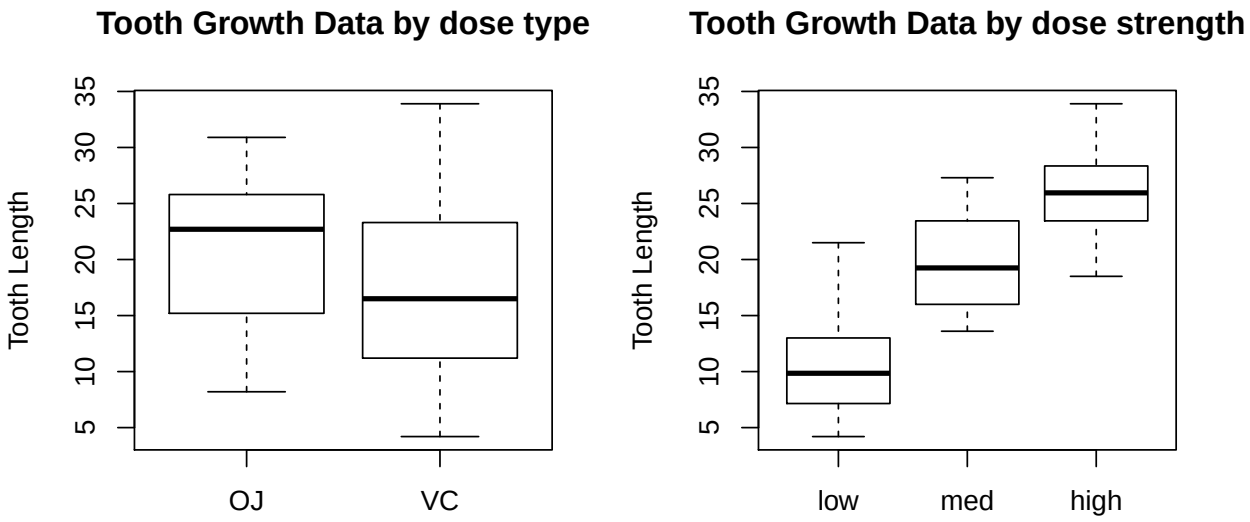


Figure 3: Interaction Plot

```
with(ToothGrowth,interaction.plot(x.factor=dose, trace.factor=supp,
                 response=len,type="b",legend=T,fun=mean,
                 pch=c(1,2),main="Interaction Plot",
                 ylab="Tooth Length",xlab="Dose"))
```

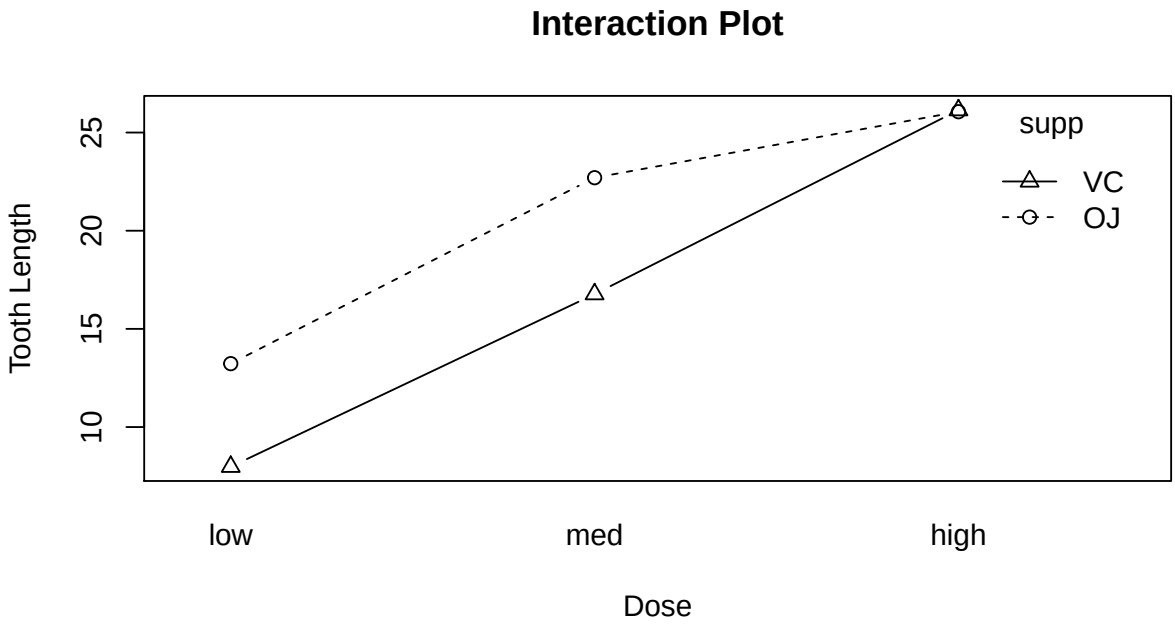
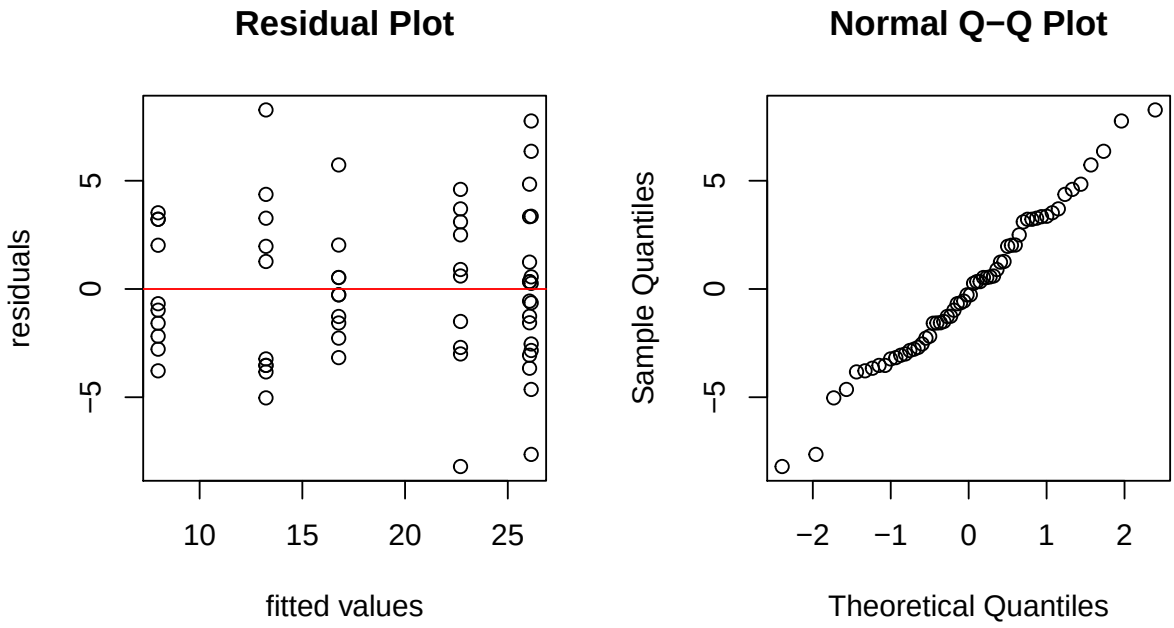


Figure 4: Further check of validity of ANOVA



Source Code

The source code for this document can be found in https://github.com/mbh038/JHU_SI/tree/master/Project

‘‘‘