Two-way ANOVA and ANCOVA

In this tutorial we discuss fitting two-way analysis of variance (ANOVA), as well as, analysis of covariance (ANCOVA) models in R. As we fit these models using regression methods much of the syntax from the multiple regression tutorial will still be valid. However, we now allow the explanatory variables to be either categorical or quantitative.

A. Two-way Analysis of Variance

Two-way ANOVA is used to compare the means of populations that are classified in two different ways, or the mean responses in an experiment with two factors. We fit two-way ANOVA models in R using the function lm(). For example, the command:

```
>  lm(Response ~ FactorA + FactorB)
```

fits a two-way ANOVA model without interactions. In contrast, the command

```
>  lm(Response ~ FactorA + FactorB + FactorA*FactorB)
```

includes an interaction term. Here both *FactorA* and *FactorB* are categorical variables, while *Response* is quantitative.

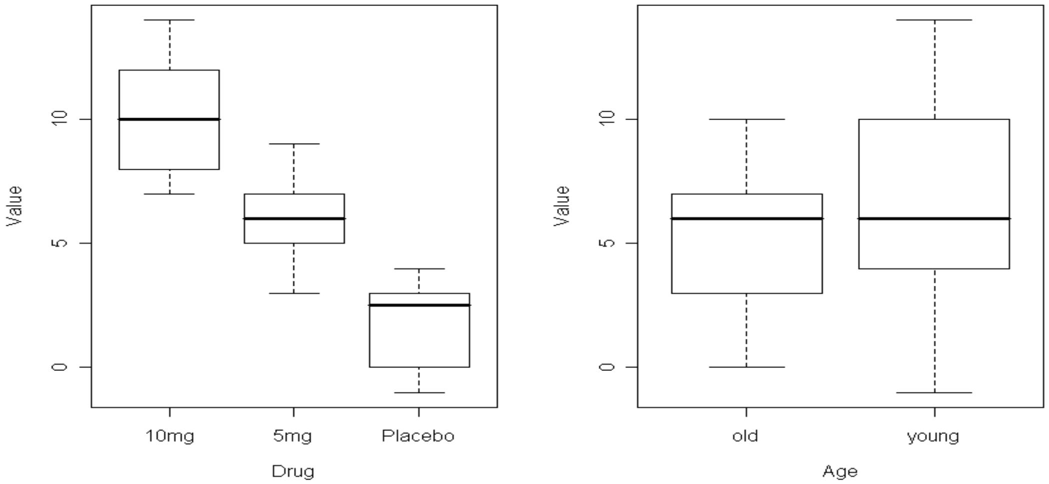
Ex. A study was performed to test the efficiency of a new drug developed to increase high-density lipoprotein (HDL) cholesterol levels in patients. 18 volunteers were split into 3 groups (Placebo/5 mg/10 mg) and the difference in HDL levels before and after the treatment was measured. The 18 volunteers were also categorized into two age groups (18-39/ ≥40).

We are interested in determining the answers to the following questions: Does the amount of drug have an effect on the HDL level? Does the age of the patient have an effect on the HDL level? Is there an interaction between age and amount of drug? We begin by reading in the data and making some exploratory plots of the data.

```
> dat = read.table("C:/Documents and Settings/Administrator/Desktop/
                                        W2024/Cholesterol.txt",header=TRUE)
> dat
      Drug      Age    Value
1   Placebo  young     4
2   Placebo  young     3
3   Placebo  young    -1
…..
17    10mg   old     8
18    10mg   old     7
```

To make side-by-side boxplots:

```
> par(mfrow=c(1,2))
> plot(Value ~ Drug + Age, data=dat)
```



Judging by the boxplots there appears to be a difference in HDL level for the different drug levels. However, the difference is less pronounced between the two age groups.

An interaction plot displays the levels of one factor on the x-axis and the mean response for each treatment on the y-axis. In addition, it shows a separate line connecting the means corresponding to each level of the second factor. When no interaction is present the lines should be roughly parallel.

These types of plots can be used to determine whether an interaction term should be included in our ANOVA model.

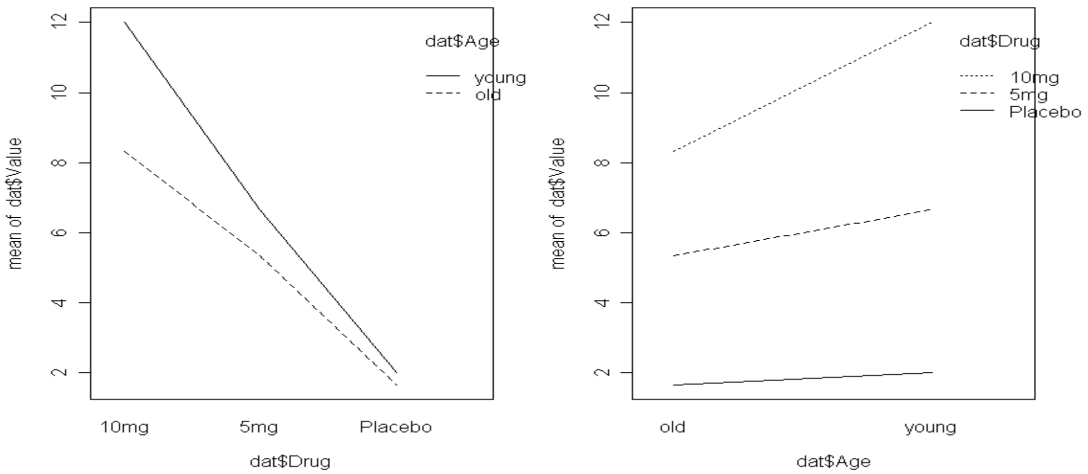
Interaction plots can be made in R using the command

```
>  interaction.plot(factorA, factorB, Response)
```

If we switch the order of factor A and B we alter which variable is plotted on the x-axis.

Ex. Create an interaction plot for the HDL data.

```
> interaction.plot(dat$Drug, dat$Age, dat$Value)
> interaction.plot(dat$Age, dat$Drug, dat$Value)
```



The interaction plots look roughly parallel, but to confirm we fit a two-way ANOVA with an interaction term:

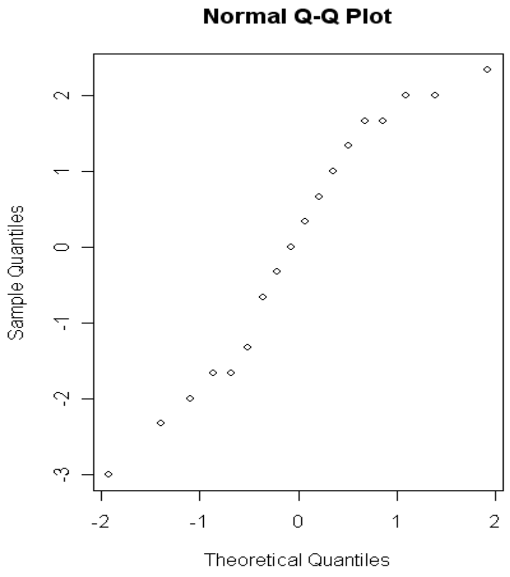
```
> results = lm(Value ~ Drug + Age + Drug*Age, data=dat)
> anova(results)
```

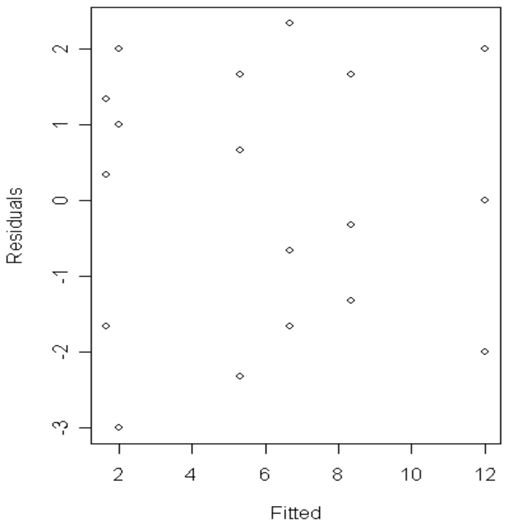
```
Analysis of Variance Table
Response: Value
            Df   Sum Sq   Mean Sq   F value     Pr(>F)
Drug         2   208.333   104.167   25.6849   4.611e-05 ***
Age          1    14.222    14.222    3.5068    0.08568 .
Drug:Age     2     8.778     4.389    1.0822    0.36974
Residuals 12      48.667     4.056
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Studying the output of the ANOVA table we see that there is no evidence of a significant interaction effect (*F*=1.08, *p*=0.370). We therefore cannot conclude that there is an interaction between age and the amount of drug taken. The test for the main effect of treatment (*F*=25.68, *p*<0.0001) shows a significant drug effect on the HDL level. Finally, the test for the main effect of age (*F*=3.51, *p*=0.0857) tells us there is not enough evidence to conclude that there is a significant age effect.

After fitting an ANOVA model it is important to always check the relevant model assumptions. This includes making QQ-plots and residual plots.

```
> qqnorm(results$res)
> plot(results$fitted,results$res,xlab="Fitted",ylab="Residuals")
```





Neither plot indicates a significant violation of the normality assumption.

C. Analysis of Covariance (ANCOVA)

Analysis of covariance (ANCOVA) combines features of both ANOVA and regression. It augments the ANOVA model with one or more additional quantitative variables, called covariates, which are related to the response variable. The covariates are included to reduce the variance in the error terms and provide more precise measurement of the treatment effects. ANCOVA is used to test the main and interaction effects of the factors, while controlling for the effects of the covariate.

Ex. A company studied the effects of three different types of promotions on the sales of a specific brand of crackers:

- Treatment 1 – The crackers were on their regular shelf, but free samples were given in the store,
- Treatment 2 - The crackers were on their regular shelf, but were given additional shelf space.
- Treatment 3 - The crackers were given special display shelves at the end of the aisle in addition to their regular shelf space.

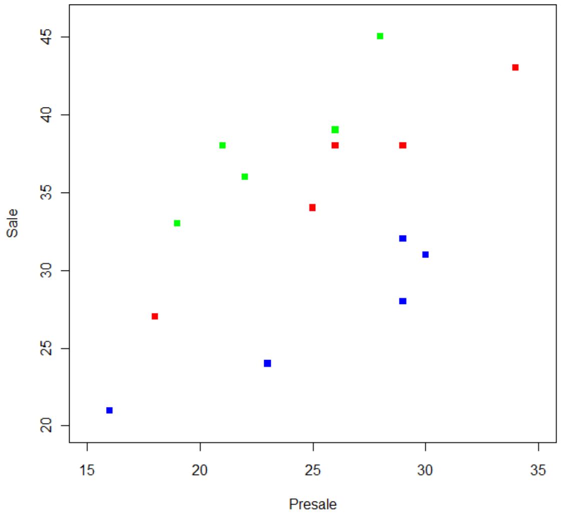
The company selected 15 stores to participate in the study. Each store was randomly assigned one of the 3 promotion types, with 5 stores assigned to each promotion. Data was collected on the number of boxes of crackers sold during the promotion period, y, as well as the number sold during the proceeding time period, denoted x.

The following commands read the data:

```
> dat = read.table("C:/Users/Documents/W2024/Cracker.txt",header=TRUE)
> dat
  Treatment Sale Presale
1         T1   38      21
2         T2   43      34
3         T3   24      23
4         T1   39      26
….
14         T2   34      25
15         T3   28      29
```

Prior to performing ANCOVA it is sensible to make a scatter plot of the response variable against the covariate, using separate symbols for each level of the factor(s). This allows one to verify the assumptions that there is a linear relationship between the covariate and the response variable, and that all treatment regression lines have the same slope.

```
> plot(Presale[Treatment == 'T1'], Sale[Treatment == 'T1'], xlab='Presale',
                    ylab='Sale', xlim=c(15,35), ylim=c(20,46), pch=15, col='green')
> points(Presale[Treatment == 'T2'], Sale[Treatment == 'T2'], pch=15, col='red')
> points(Presale[Treatment == 'T3'], Sale[Treatment == 'T3'], pch=15, col='blue')
```



As it appears that both the linearity and equal slopes assumptions required for ANCOVA are valid we are able to proceed with the analysis. The following code performs a one-way ANCOVA model, controlling for the sales in the proceeding time period:

```
> results = lm(Sale ~ Presale + Treatment)
> anova(results)
Analysis of Variance Table
Response: Sale
            Df   Sum Sq   Mean Sq   F value     Pr(>F)
Presale      1   190.68   190.678    54.379   1.405e-05 ***
Treatment    2   417.15   208.575    59.483   1.264e-06 ***
Residuals   11    38.57     3.506
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The output tells us that the three cracker promotions differ in effectiveness (F = 59.48, p-value < 0.0001). One can now continue by using multiple comparison techniques to determine how they differ. Note that we also need to check the residuals to determine whether the other model assumptions hold.