ANOVA and Linear Regression Using R

DR LEONARD KITI & ALEX MUTUKU

ANOVA

Comparison of means form several groups: The basic ANOVA situation

When the exposure variable has more than two categories, we often wish to compare the mean outcomes from each of the groups defined by these categories.

Example, - Wish to examine how the systolic blood pressure measurements collected as part of the routine hospital visits vary with age groups (3 categories) or sex or ethnic groups or socioeconomic classes.

ANOVA

Main Question: Do the means of the quantitative variables depend on which group the individual is in?

If categorical variable has only 2 values: 2-sample t-test

ANOVA allows for 3 or more groups

ANOVA

Exploratory Data analysis

Graphical investigation:

1. Side-by-side box plots
2. Multiple histograms

What ANOVA does

H_0: The means of all the groups are equal.

H_a: Not all the means are equal

Doesn't say how or which ones differ. Follow up with "multiple comparisons"

Whether the differences between the groups are significant depends on;-

1. The difference in the means
2. The standard deviations of each group
3. The sample sizes

Assumptions of ANOVA

Each group is approximately normal

- Check this by looking at histograms and/or normal quantile plots, or use assumptions
- Standard deviations of each group are approximately equal

ANOVA

Description of the data
A sample of individual diagnosed with a particular disease and categorized as either 1 –" low diseased" 2 – "middle level diseased" and 3 – "highly diseased". 4 drugs known to Control the systolic blood pressure were administered to the sampled individuals and their Systolic blood pressure recorded after 24 hours.

Question of interest

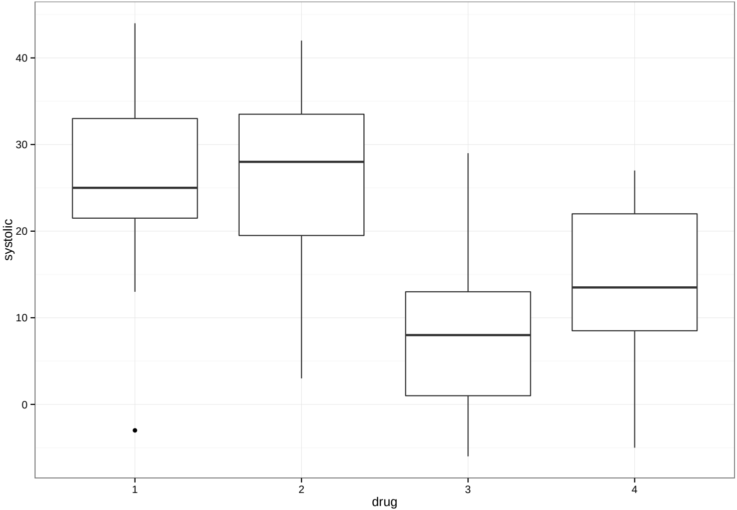
1. Compare the means of systolic blood pressure by the drug administered, is there a significant difference in the drugs administered – One way ANOVA
2. Compare the means of the systolic blood pressure by the drug administered given the disease status of the individuals. – Two way ANOVA

```r
data<-read.csv("data/systolic.csv", header=T)
data$drug<-as.factor(data$drug) #
data$disease<-as.factor(data$disease) #
```

```r
summm(data$systolic, by=data$drug ,graph = F)
```

```
## For data$drug = 1
##  obs. mean   median  s.d.   min.   max.
##  15   26.07  25      11.677 -3     44
##
## For data$drug = 2
##  obs. mean   median  s.d.   min.   max.
##  15   25.53  28      11.618 3      42
##
## For data$drug = 3
##  obs. mean   median  s.d.   min.   max.
##  12   8.75   8       10.019 -6     29
##
## For data$drug = 4
##  obs. mean   median  s.d.   min.   max.
##  16   13.5   13.5    9.324  -5     27
```

```
ggplot(data, aes( drug, systolic )) + geom_boxplot() + theme_b
```



How ANOVA works

ANOVA measures two sources of variation in the data and compares their relative sizes

1. Variation BETWEEN groups for each data value look at the difference between its group mean and the overall mean

$$(\bar{x}_i - \bar{x})^2$$

2. Variation WITHIN groups for each data value we look at the difference between that value and the mean of its group

$$(x_{ij} - \bar{x}_i)^2$$

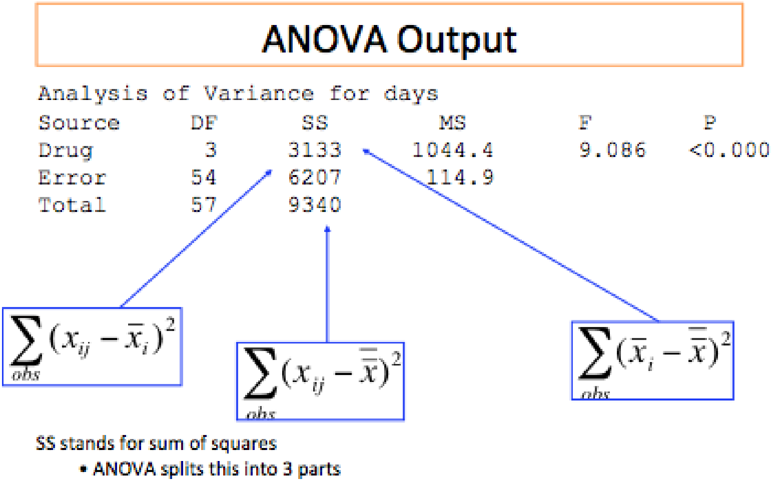
The ANOVA F-statistic is a ratio of the Between Group Variaton divided by the Within Group Variation:

$$F = \frac{Between}{Within} = \frac{MSG}{MSE}$$

A large F is evidence against H_0, since it indicates that there is more difference between groups than within groups

To get the P-value, we compare to $F(k-1, n-k)$ -distribution $k-1$ degrees of freedom in numerator (# groups -1) $n-k$ degrees of freedom in denominator

```
one.way<-aov(data$systolic~data$drug)
```



ANOVA Output

```
Analysis of Variance for days
Source        DF       SS          MS         F         P
Drug           3     3133      1044.4     9.086    <0.000
Error         54     6207       114.9
Total         57     9340
```

$$\sum_{obs}(x_{ij}-\bar{x}_i)^2$$

$$\sum_{obs}(x_{ij}-\bar{\bar{x}})^2$$

$$\sum_{obs}(\bar{x}_i-\bar{\bar{x}})^2$$

SS stands for sum of squares

- ANOVA splits this into 3 parts

Normality and Variance Homogeneity test

Test of homogeneity of variances

```
bartlett.test(data$systolic,data$drug)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data$systolic and data$drug
## Bartlett's K-squared = 1.0063, df = 3, p-value = 0.7997
```

Normality test

```
shapiro.test(resid(one.way))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(one.way)
```

LINEAR REGRESSION

Linear Regression with One Predictor variable

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that a response or outcome variable can be predicted from the other, or others.

Examples of regression method;-

1. The length of hospital stay of a surgical patient can be predicted by utilizing the relationship between the time in the hospital and the severity of the operation.
2. Sales of a product can be predicted by utilizing relationship between sales and amount of advertising expenditures.

Linear regression involves modelling a continous outcome variable with one or more explanatory variables.

With all data analysis the first step is always to explore the data. In this case, scatter plots are very useful in determining whether or not the relationships between the variables are linear.

Assumptions of Linear Regression

1. Variables are normally distributed
2. Variables are related linearly
3. Errors are identically and independently distributed (normally) with mean zero and equal variance
4. Covariance between predictors and error terms is zero
5. No correlation between the predictors

Data Description

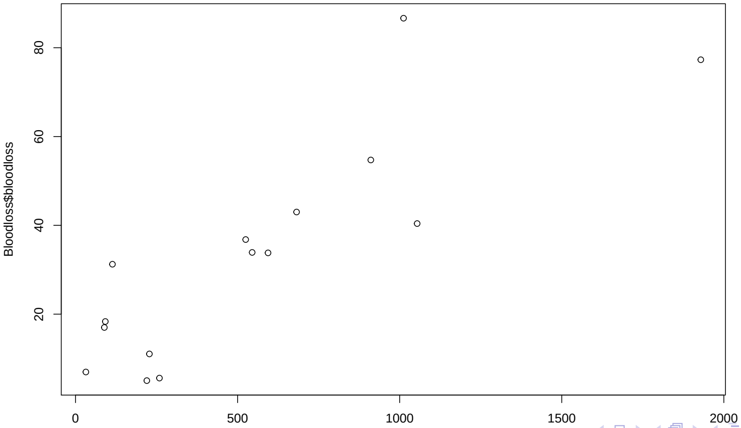
Dataset (Bloodloss.csv) to be used for this section concerns the relationship between Hookworm infection and blood loss from a study conducted in Kibera in 2004.

The dataset has 3 variables (id, worm,bloodloss) and 15 records.

The objective of this analysis is to examine the relationship between these variables

Graphical representation - Scatter plots

```
Bloodloss<-read.csv("data/Bloodloss.csv", header=T)
plot(Bloodloss$worm, Bloodloss$bloodloss)
```



Fitting a linear Regression

```
model1<-lm(Bloodloss$bloodloss ~ Bloodloss$worm)
```

```
summary (model1)
```

```
##
## Call:
## lm(formula = Bloodloss$bloodloss ~ Bloodloss$worm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.846 -10.812   0.750   4.356  34.390
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.847328   5.308569   2.043   0.0618 .
## Bloodloss$worm  0.040922   0.007147   5.725 6.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
## Residual standard error: 13.74 on 13 degrees of freedom
```

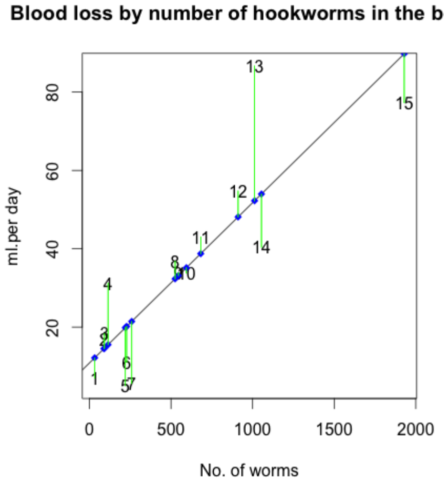
The first section shows the formula that was “called”. The second section gives the distribution of residuals. The third section gives coefficients of the intercept and the effect of 'worm'on blood loss.

Interpreting the coeeficients

The intercept is 10.8 meaning that when there are no worms, the blood loss is estimated to be 10.8 ml per day. However, it's not significantly different from zero as the P value is 0.0618.

The coefficient of 'worm' is 0.04 indicating that each worm will cause an average of 0.04 ml of blood loss per day, which is highly significant from zero.The multiple R-squared value of 0.716 indicates that 71.6% of the variation in the data is explained by the model.

From the analysis, it is clear that blood loss is associated with number of hookworms. On average, each worm may cause 0.04 ml of blood loss. The remaining uncertainity of blood loss, apart from hookworm, is explained by random variation or other factors that were not measured.



Fitted and observed points

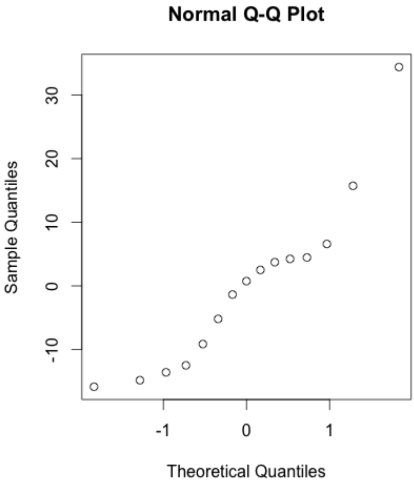
Diagnostics checks

Checking for normality of residuals If you plot the histogram of the residuals it will be difficult to conclude normality due to the small sample size.

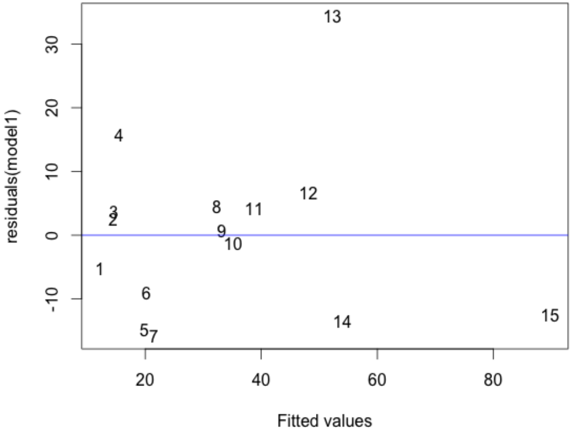
Better to plot the residuals against the expected normal score or to use the Shapiro-Wilk test

```r
shapiro.test(residuals(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model1)
## W = 0.8978, p-value = 0.0882
```


Normal Q-Q Plot
Sample Quantiles
-10
0
10
20
30
-1
0
1
Theoretical Quantiles

Checking on independence There is no obvious pattern. The residuals are quite independent of the expected values


residuals(model1)
Fitted values