Further Statistical Analysis using R

Mark Dunning, Matt Eldridge and Sarah Vowler *

Last Document revision: October 19, 2015

Contents

*Acknowledgements: Sarah Dawson

1 Course Introduction

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.". R.A. Fisher, 1938

The goals of statistical methods could be summarised as follows:

- drawing conclusions about a population by analysing data on just a sample;
- evaluating the uncertainty in these conclusions; and,
- designing the sampling approach so that valid and accurate conclusions can be made from the data collected.

The statistical approach used is dependent on the data type. In this document we will describe methods for comparing multiple groups (**ANOVA** (analysis of variance), and **non-parametric** alternatives), and **Linear Regression**. We will assume you are already familiar with methods to perform one-sample or two-sample tests, as described in our Introductory Statistics course. After describing the assumptions of each test, we will give a worked example in R.

1.1 Exploratory Analysis

Before conducting a formal analysis of our data, it is always a good idea to run some exploratory checks of the data:

- To check that the data has been read in or entered correctly;
- To identify any outlying values and if there is reason to question their validity, exclude them or investigate them further;
- To see the distribution of the observations and whether the planned analyses are appropriate.

It's always a good idea to calculate some summary statistics for your data, such as the mean and standard deviation, or the median and inter-quartile range if your data is skewed. You should also consider whether there may be outliers in your data (but do not remove them from the analysis without good reason) or whether there may be missing data. Summary statistics were covered in detail in our Introductory Statistics course.

1.2 Statistical Tests - basic setup

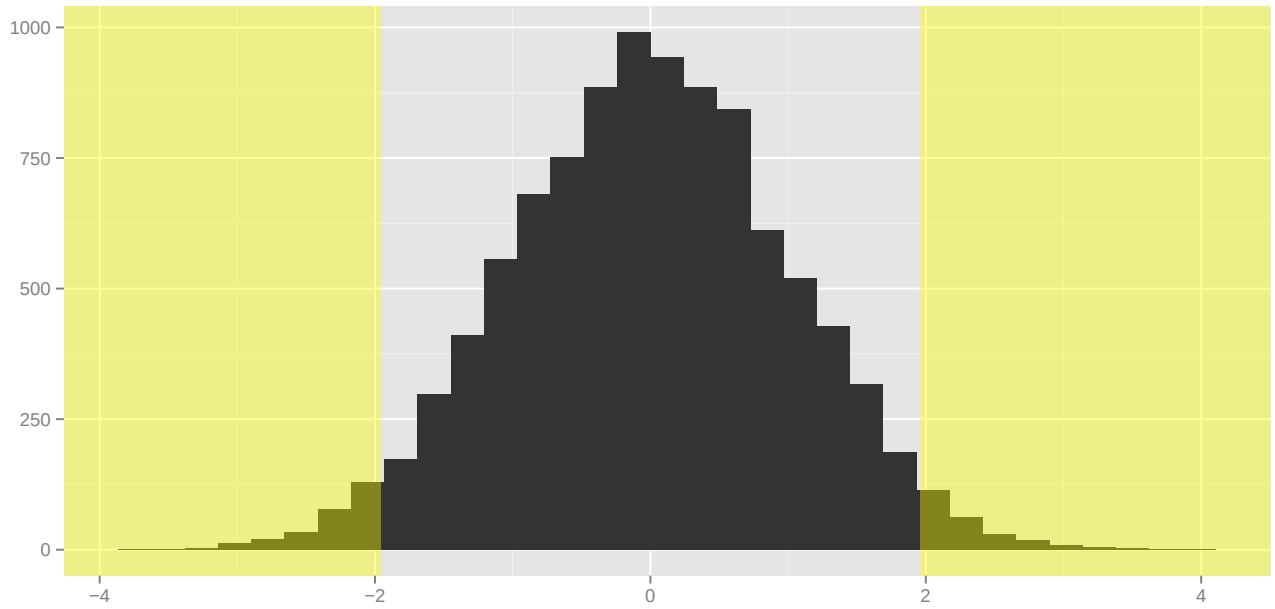
There are four key steps in every statistical test:

- 1. Formulate a **null hypothesis**, H_0. This is the working hypothesis that we wish to disprove.
- 2. Under the assumption that the **null hypothesis** is true, calculate a **test statistic** from the data.
- 3. Determine whether the **test statistic** is more extreme than we would expect under the **null hypothesis**, i.e. look at the p-**value**.
- 4. Reject or do not reject the **null hypothesis**.

As the name suggests, the null hypothesis typically corresponds to a **null** effect.

For example, there is **no difference** in the measurements in group 1 compared with group 2. A small p-value indicates that the probability of observing such a test statistic as small under the assumption that the null hypothesis is true. If the p-value is below a pre-specified **significance level**, then this is a **significant result** and, we would conclude, there is evidence to reject the null hypothesis.

The **significance level** is most commonly set at 5% and may also be thought of as the **false positive rate**. That is, there is a 5% chance that the null hypothesis is true for data-sets with test statistics corresponding to p-values of less than 0.05 i.e. we may wrongly reject the null hypothesis when the null hypothesis is true (false positive).



Equally, we may make **false negative** conclusions from statistical tests. In other words, we may not reject the null hypothesis when the null hypothesis is, in fact, not true. However, statisticians tend not to talk about false negative rates, preferring instead to refer to **power**, which is 1-false negative rate.

The **power** of a statistical test will depend on:

- The **significance level** - a 5% test of significance will have a greater chance of rejecting the null than a 1% test because the strength of evidence required for rejection is less.
- The **sample size** - the larger the sample size, the more accurate our estimates (e.g. of the mean) therefore, we can differentiate between the null and alternative hypotheses more clearly.
- The **size of the difference or effect** we wish to detect - bigger differences (i.e. alternative hypotheses) are easier to detect than smaller differences.
- The **variability**, or standard deviation, of the observations - the more variable our observations, the less accurate our estimates therefore, it is more difficult to differentiate between the null and alternative hypotheses.

	Null hypothesis does not hold	**Null hypothesis holds**
Reject null hypothesis	Correct *True Positive*	Wrong *False positive*
Do no reject null hypothesis	Wrong *False negative*	Correct *True negative*

Table 1: Error definitions

2 R Introduction

To install R visit www.r-project.org. In the 'Getting Started' box half-way down the page follow the 'download R' link. On the next page choose the appropriate operating system for your computer from the three 'Download R for...' options.

Following this link will start the installation of R. If you get a security warning select 'Run'. Follow the directions in the install wizard to install R. We haven chosen to run R through the popular Rstudio interface, which you will also need to install. The version of R used to write this manual is 3.2.1. Please bear in mind that the version number you download may be different as new versions are released yearly.

This manual, and the accompanying practical will assume some familiarity with the R statistical language. In particular, you should be familiar with the following concepts:

- Using the RStudio program
- Setting your working directory
- Creating variables and basic object types; in particular vectors and data frames
- Using built-in R functions
- Using R to get help on functions
- Subset operations for vectors and data frames using the [] notation
- Reading tabular data into R
- Basic plots; scatter plots, boxplot and histogram

Several Online videos are available that cover this materials. For example

- http://shop.oreilly.com/product/0636920034834.do
- http://blog.revolutionanalytics.com/2012/12/coursera-videos.html
- http://bitesizebio.com/webinar/20600/beginners-introduction-to-r-statistical-software

2.1 R packages Used

In order to run the examples in this manual, and the practical, you will need to execute the following command in R to install the required packages. Or if you prefer, packages can be installed via the **Packages** tab in RStudio.

```
install.packages(c("tidyr","beeswarm","RColorBrewer","clinfun","RVAideMemoire"))
```

You will not need to type this command more than once, so long as you keep the same version of R. Once these packages have been installed, you will need to load each of libraries into R when you require them. e.g.

```
library(tidyr)
```

Alternatively, packages can be loaded from the "Packages" tab in the lower-right panel of RStudio.

2.2 Formatting data for Statistical Testing in R

Before conducting a statistical test in R we often need to manipulate our data into a particular format. If you are familiar with working with numerical data in Excel (or similar), then you probably represent your data in tabular format. See Table 2 for an example taken from the seminal 'Tidy Data' paper by Hadley Wickham [1].

	Name	treatmenta	treatmentb
1	John Smith	-	18.00
2	Jane Doe	4	1.00
3	Mary Johnson	6	7.00

Table 2: An imaginary dataset in human-readable format.

Table 2 is easy for humans to comprehend; we can easily scan the table and identify that John Smith has a measurement of 18 for 'treatmenta' and no observation for 'treatmentb'. However, it is not very efficient for computers to deal with data in this format. For the majority of statistical testing that is performed in R, we think of the observations in our datasets as being explained by a series of factors, or explanatory variables. For example, in our imaginary dataset we have a measurement for each patient that belongs to either treatment group a, or treatment group b. Thus, the factor in this case would be treatment. Table 3 shows the same dataset, but in a *'tidy'* format with *variables* as columns and each row forming a different *observation*. Sometimes this is also referred to as *long* data, as opposed to the *wide* format of the original table.

Several options exist in R to transform our data in this manner; such as the `stack` function from base R, to the *reshape* and *reshape2* packages. However, we will describe the *tidyr* package, which is the evolution of *reshape* and *reshape2* and is nicely integrated with other advanced manipulation tools such as *dplyr* and *ggplot2*.

We will now illustrate the how to transform our imaginary dataset into a suitable format for statistical analysis using *tidyr*. The goal of such an operation is to take the `treatmenta` and `treatmentb` columns

	name	treatment	n
1	Jane Doe	a	4.00
2	Jane Doe	b	1.00
3	John Smith	a	
4	John Smith	b	18.00
5	Mary Johnson	a	6.00
6	Mary Johnson	b	7.00

Table 3: The same imaginary dataset, but in tidy format.

and create two separate columns; one which indicates whether a particular observation was made for treatment a or treatmentb, and the corresponding values. In the language of *tidyr*, we say the the treatment variable is a *key*.

```
messy <- data.frame(Name = c("John Smith","Jane Doe","Mary Johnson"),
                    treatmenta = c(NA,4,6),treatmentb=c(18,1,7))
```

The function to be used is called `gather` [1]. The function is sometimes able to take a data frame as input and 'guess' what the tidy form of that dataset should look like. Finer control is possible, such as specifying column names in the output. In this example, we specify that we want a column in the output called `'treatment'` for our new variable (the key), and `'n'` for the corresponding values. The final argument is for specifying which columns in the input data are to be decomposed.

```
library(tidyr)
tidy <- gather(messy,treatment,n,treatmenta,treatmentb)
tidy
```

```
##           Name  treatment  n
## 1   John Smith treatmenta NA
## 2     Jane Doe treatmenta  4
## 3 Mary Johnson treatmenta  6
## 4   John Smith treatmentb 18
## 5     Jane Doe treatmentb  1
## 6 Mary Johnson treatmentb  7
```

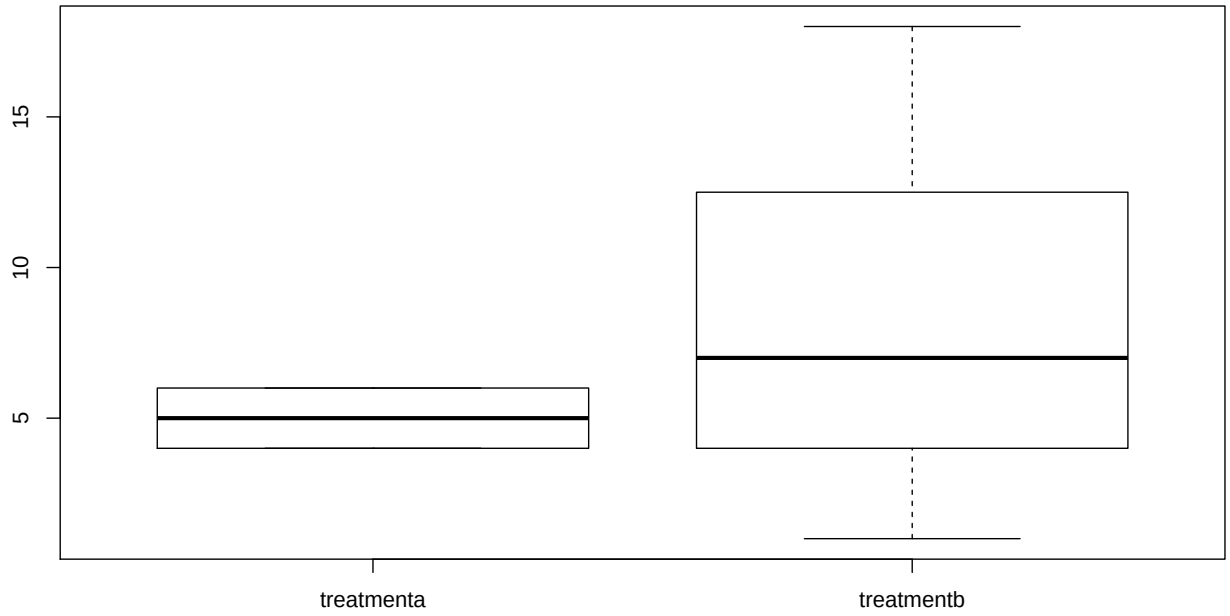
Note that we can use shortcuts to select columns, such as selecting all columns in the range `treatmenta` to `treatmentb`, or all columns except `Name`. You should see that these give the same result.

```
gather(messy,treatment,n,treatmenta:treatmentb)
gather(messy,treatment,n,-Name)
```

With our data in this format, we can now specifiy various formulas using the tilde ($\sim$) syntax. You may have already seen the following syntax which will plot a boxplot with explanatory variable `treatment` on the x-axis, and observations `n` on the y-axis.

[1] Don't forget that you can get help on this function using `?gather`

```
boxplot(tidy$n~tidy$treatment)
```



As we often have the variable and values that we wish to plot in the same data frame, there are a couple of shortcuts to avoid using the $ syntax (which is a bit in-elegant). Firstly, functions such as `boxplot` and other statistical testing functions (e.g. `t.test`) have a `data` argument. This allows the name of a data frame to be specified and the variables can be referred to by name.

```
boxplot(n~treatment,data=tidy)
```

The `with` function also performs the same task;

```
with(tidy, boxplot(n~treatment))
```

3 Comparing Multiple Groups

ANOVA stands for *an*alysis *o*f *v*ariance. There are three main types of ANOVA: one-way, two-way and repeated measures. In this course our main focus will be on the one-way ANOVA.

3.1 One-way ANOVA

The two-sample t-test is useful when we have just two groups of continuous data to compare. When we want to compare more than two groups, a one-way ANOVA can be used to simultaneously compare all groups, rather than carrying out several individual two-sample t-tests. The main advantage of doing this is that it reduces the number of tests being carried out, meaning that the type I error rate (the probability of seeing a significant result just by chance) does not become inflated.

A one-way ANOVA compares group means by partitioning the variation in the data into **between group** variance and **within group variance** (see Table 4).

Source of Variation	**Degrees of Freedom**	**Sum of Squares**	**Mean Squares**	$F_{k-1,N-k}$
Between groups	$k-1$	$BSS = \sum_k^{i=1} n_i y_i^2 - \frac{T^2}{N}$	$S_B^2 = \frac{BSS}{k-1}$	
Within groups	$N-k$	$WSS = S - \sum_{i=1}^{k} n_i y_i^2$	$S_W^2 = \frac{WSS}{N-k}$	
Total	$N-1$	TSS = BSS + WSS		$\frac{S_B^2}{S_W^2}$

Table 4: One-way ANOVA table

The between group variance is divided by the within group variance to give an F statistic. This tells us the ratio of between group variation to within group variation. A large F-value implies that there are significant differences between groups and conversely a small F-value implies there are not significant differences between groups. Giving this a little bit of thought, the idea becomes more intuitive:

- If the variance within groups is small, but between groups the variance is very large, we can infer that there are likely to be differences between groups. Following this theory, the F statistic is calculated by dividing the between group variation by the within group variation. So in this scenario, we divide a large number by a comparatively small number and this leaves us with a large(ish) number for our F statistic, corresponding to a small p-value.
- If the variance within groups is large, but between groups the variance is also large, it is more difficult to know whether the groups truly differ in their mean value. In this scenario, the F statistic is calculated by dividing a large number by another large number - depending on the relative size of these two large numbers we may be left with a large or small number for our F statistic. The same is true when the variance within and between groups are both small.
- Finally, if the variance within group is large, but between groups the variance is very small, we can infer that there are unlikely to be differences between groups. In this scenario, the F statistic is calculated by dividing a small number by a large number and this will result in a small(ish) number for our F statistic, corresponding to a large p-value.

Luckily, you won't need to do the calculations from Table 4 by hand as R will do these all for you, but it's good to have an appreciation of what the test is actually doing in the background.

Obviously the outcome of the one-way ANOVA depends on the data available. If there is high variation in the data, a much larger sample size will be needed to detect a difference between groups. Likewise, if we are interested in detecting very small differences between groups a larger sample size will also be required.

3.1.1 ANOVA assumptions

There are several assumptions behind the one-way ANOVA:

- Random selection of observations
- Normally distributed response (assessed for each group separately)
- Approximately equal variance across the groups
- Independent observations

Like most statistical tests, ANOVA assumes that there is a random selection of observations from the population of interest. If this is not the case then any results obtained from the test on the sample of observations may not generalise to the population as a whole.

The main assumption of ANOVA is that the distribution of the response variable should be normally distributed for each group being compared. This can be assessed prior to fitting the ANOVA by constructing a histogram of the response variable for each group being compared. This can also be assessed after fitting the ANOVA by constructing a normal probability plot of the residuals (sometimes called a QQ-plot).

Another important assumption is that there is approximately equal variance across the groups being compared. This assumption is important because of the way the F-test in the ANOVA uses the pooled variance across groups. If one group has a much larger variance than another group, the results of the F-test may not be valid. The equal variance assumption can be assessed using either Bartlett's or Levene's test (REMEMBER! this adds to the multiple testing problem), or visually using a histogram plotted separately for each group.

The final assumption of the one-way ANOVA is the independence of observations. There is no easy way of assessing independence, so a lot of people overlook this assumption. However, a little thought about where the data comes from and how it was collected can give us a good indication of whether the observations are independent or not. Things like taking observations from related individuals or having multiple measurements per subject will cause the independence assumption to be invalid. You should ask the question "is there any reason why any of the measurements are more likely to be similar than any others?" If the answer is no then it is safe to assume independence of observations.

When the F-test provides a significant result, it tells us that there is at least one difference in the groups. However, it does not tell us group is different. We may be interested in making comparisons between pairs of groups to identify where the difference lies and estimate the size of the difference (the effect size). This can be done by using unpaired two-sample t-tests. If we wish to make multiple comparisons we must be careful to adjust for multiple testing. R provides several different types of multiple-testing adjustment, each suiting different types of comparisons. These are discussed in more detail in the section 3.1.2.

3.1.2 Choosing the correct post-hoc test

Tukey	Compare all pairs of columns
Bonferroni	Compare all pairs of columns OR compare selected pairs of columns
Dunnett	Compare all columns vs. control column
Trend test	Test for linear trend in means across columns

Table 5: Multiple-testing adjustment methods

3.1.3 One-way ANOVA Example

The protein expression level was measured in 5 cell types from a single cell line. We want to know whether there are any differences in the expression level between the five different cell types. The raw data are given in Table 6. These data come from the Babraham Bioinformatics course Statistical Analysis using GraphPad Prism.

	A	B	C	D	E
1	0.40	0.26	0.24	1.04	0.74
2	1.50	0.47	0.25	2.78	0.99
3	0.98	0.42	1.01	0.82	1.26
4	0.33	0.64	0.77	1.65	1.50
5	0.75	0.32	0.47	0.49	0.30
6	1.48	0.65	0.47	0.97	0.34
7	1.18	0.43	0.46	1.39	0.77
8	0.33	0.67	0.65	3.24	1.94
9	1.42	0.43	0.41	1.12	2.62
10	2.09	0.70	0.81	2.82	1.42
11	1.37	0.79	1.20	1.27	0.73
12	1.23	0.89	1.08	1.60	2.09
13			0.34	1.98	1.52
14			1.98	9.32	1.67
15			1.39	2.31	3.40
16			1.12	4.19	2.16
17			3.14	1.73	2.31
18			2.78	5.16	1.32

Table 6: Protein Expression data

Our **null hypothesis** is that the mean value is the same in each of the five groups.

Our **alternative hypothesis** is that the mean value is different in one or more of the five groups.

These data can be read using the `read.csv` function in R, which will create a *data frame* representation.

```
proteinData <- read.csv("protein-expression.csv")
```

At this point, it is a good idea to inspect the data to make sure they have been imported correctly. Sometimes R will read data without complaint, but create an object that you can't actually use for

analysis. If you are using RStudio, the command `View(proteinData)` will bring-up a display of the dataset. Otherwise the following commands will tell you about the dimensions of the data, first few lines and numerical summary of each column.

```
head(proteinData)
```

```
##      A    B    C    D    E
## 1 0.40 0.26 0.24 1.04 0.74
## 2 1.50 0.47 0.25 2.78 0.99
## 3 0.98 0.42 1.01 0.82 1.26
## 4 0.33 0.64 0.77 1.65 1.50
## 5 0.75 0.32 0.47 0.49 0.30
## 6 1.48 0.65 0.47 0.97 0.34
```

```
dim(proteinData)
```

```
## [1] 18  5
```

```
summary(proteinData)
```

```
##        A                B                C                D                E
##  Min.   :0.3300   Min.   :0.2600   Min.   :0.2400   Min.   :0.490   Min.   :0.300
##  1st Qu.:0.6625   1st Qu.:0.4275   1st Qu.:0.4625   1st Qu.:1.157   1st Qu.:0.825
##  Median :1.2050   Median :0.5550   Median :0.7900   Median :1.690   Median :1.460
##  Mean   :1.0883   Mean   :0.5558   Mean   :1.0317   Mean   :2.438   Mean   :1.504
##  3rd Qu.:1.4350   3rd Qu.:0.6775   3rd Qu.:1.1800   3rd Qu.:2.810   3rd Qu.:2.053
##  Max.   :2.0900   Max.   :0.8900   Max.   :3.1400   Max.   :9.320   Max.   :3.400
##  NA's   :6        NA's   :6
```

3.1.4 Checking the model assumptions

A boxplot of these data can be created using the `boxplot` function, which allows us to compare the median and inter-quartile range (IQR) of each cell type. Optionally, we can use the *beeswarm* package to overlay individual points on the plot. See Figure 1

```
library(beeswarm)
library(RColorBrewer)
boxplot(proteinData,xlab="Cell Type",ylab="Protein Expression",main="Protein Expression")
beeswarm(proteinData, add=TRUE,pch=16,col=brewer.pal(5,"Set1"),method="swarm")
```

comment: The RColorBrewer package is also used to define a colour palette for the dataset.

Figure 1 shows us that the median value varies between the five groups. We can also see that the data is skewed for cell types A and D, as the bar in the middle, which shows the median, is not equally between the two outer bars, which show the lower and upper quartiles. In addition, there are extreme values (indicated by the dots above the boxplot) present for cell types C and D. We can also see that protein expression levels in some groups are much more variable than in others; the protein expression levels in group B are very consistent, but for groups C, D and E they are much more varied.

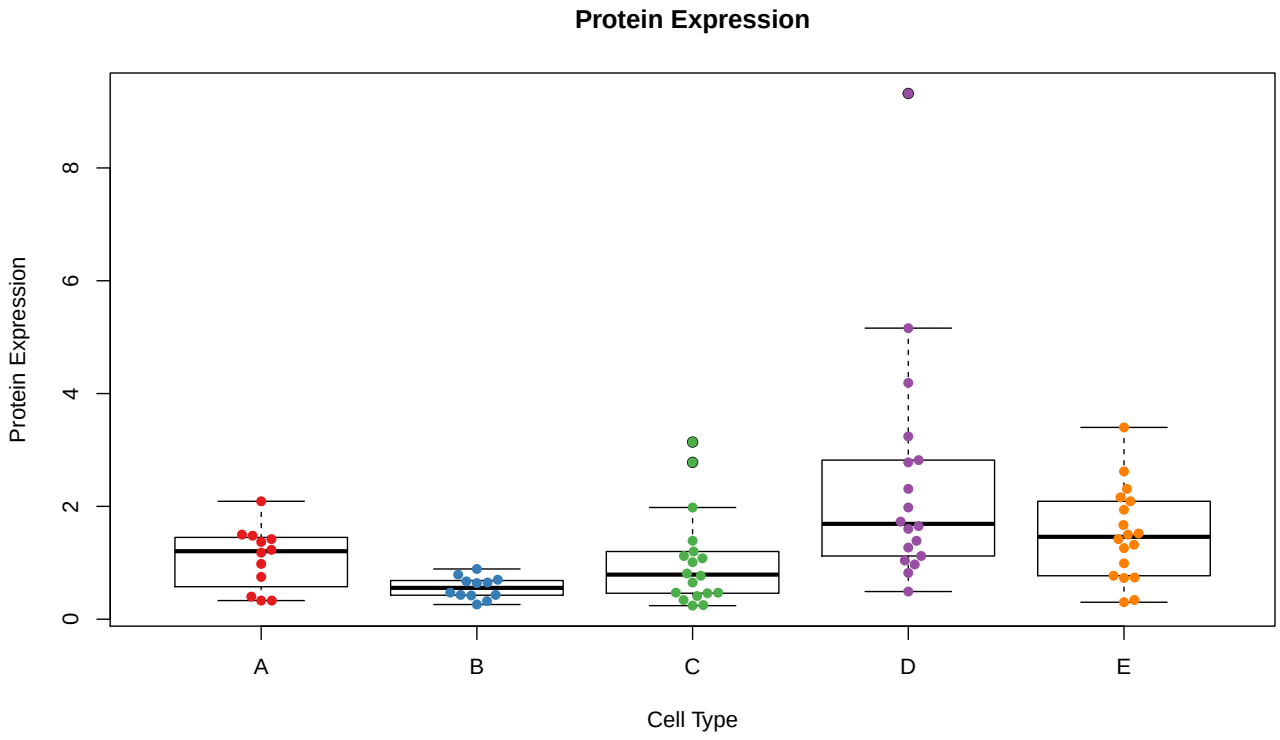


Figure 1: Boxplot of the protein expression levels for five cell types

We can make a similar assessment of the data using histograms (plotted separately for each group in our dataset). See Figure 2. In particular we can use these histograms to make an assessment of normality and constant variance which are two of the assumptions behind ANOVA (Section 3.1.1). Note the difference in the range of the x-axis on the different plots.

```
par(mfrow=c(2,3))
cols <- brewer.pal(5,"Set1")
hist(proteinData$A,xlab="A",col=cols[1],main="")
hist(proteinData$B,xlab="B",col=cols[2],main="")
hist(proteinData$C,xlab="C",col=cols[3],main="")
hist(proteinData$D,xlab="D",col=cols[4],main="")
hist(proteinData$E,xlab="E",col=cols[5],main="")
```

comment: If you are more-familiar with programming, you could write this chunk of code with a for loop (or similar)

In its current format, the data are unsuitable for analysis using one-way ANOVA as both the normality assumption and the constant variance assumption are violated. We can sometimes overcome this issue by transforming the data, and in this case we can use a log-transformation to normalise the data (Note: you do not need to do this step if the assumptions of normality and constant variance hold). This can be carried out within R using the `log` function.

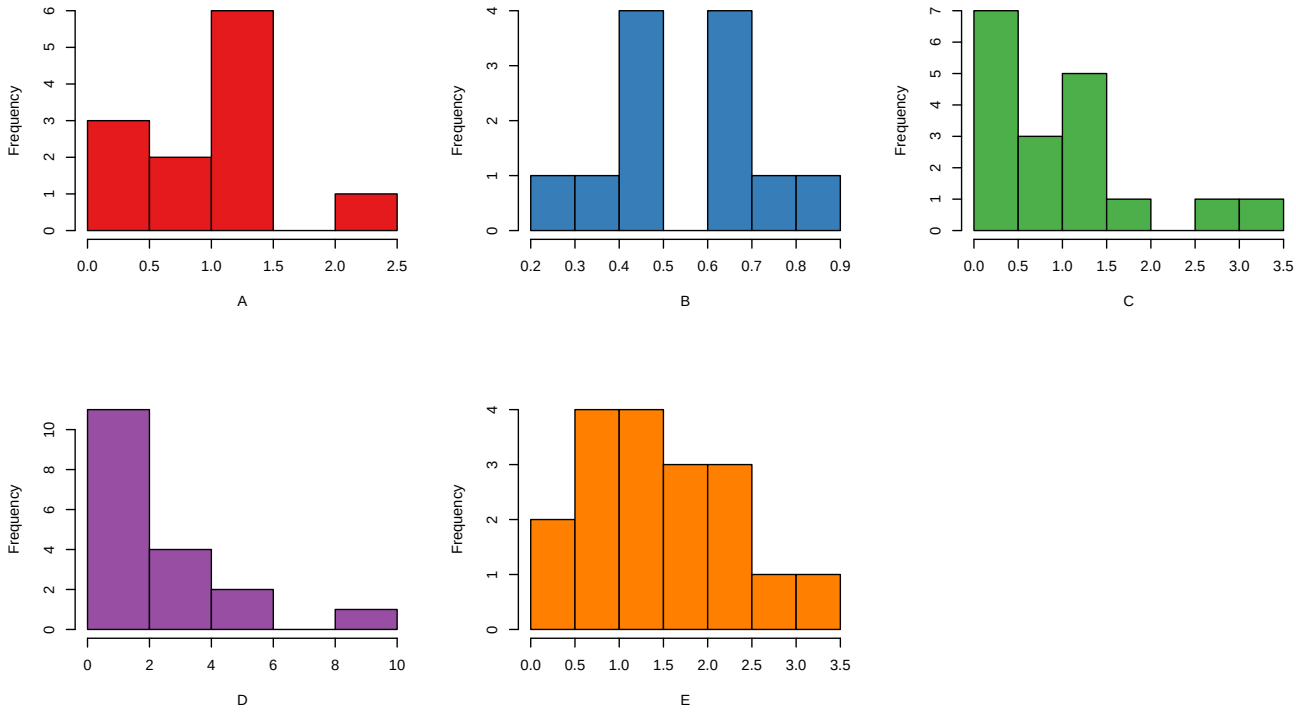


Figure 2: Histogram of the protein expression levels for five cell types

```
proteinData.ln <- log(proteinData)
head(proteinData.ln)
```

```
##              A          B            C           D           E
## 1 -0.91629073 -1.3470736 -1.427116356  0.03922071 -0.30110509
## 2  0.40546511 -0.7550226 -1.386294361  1.02245093 -0.01005034
## 3 -0.02020271 -0.8675006  0.009950331 -0.19845094  0.23111172
## 4 -1.10866262 -0.4462871 -0.261364764  0.50077529  0.40546511
## 5 -0.28768207 -1.1394343 -0.755022584 -0.71334989 -1.20397280
## 6  0.39204209 -0.4307829 -0.755022584 -0.03045921 -1.07880966
```

comment: If you are not sure what the `log` is doing, don't forget that you can bring-up the help page: `?log`

Hopefully the transformation will result in data that are normally distributed enough to meet the normality assumption. To check this, create another set of histograms in the same way you did on the untransformed data (See Figure 3)

```
par(mfrow=c(2,3))
hist(proteinData.ln$A,xlab="A",col=cols[1],main="")
hist(proteinData.ln$B,xlab="B",col=cols[2],main="")
hist(proteinData.ln$C,xlab="C",col=cols[3],main="")
```

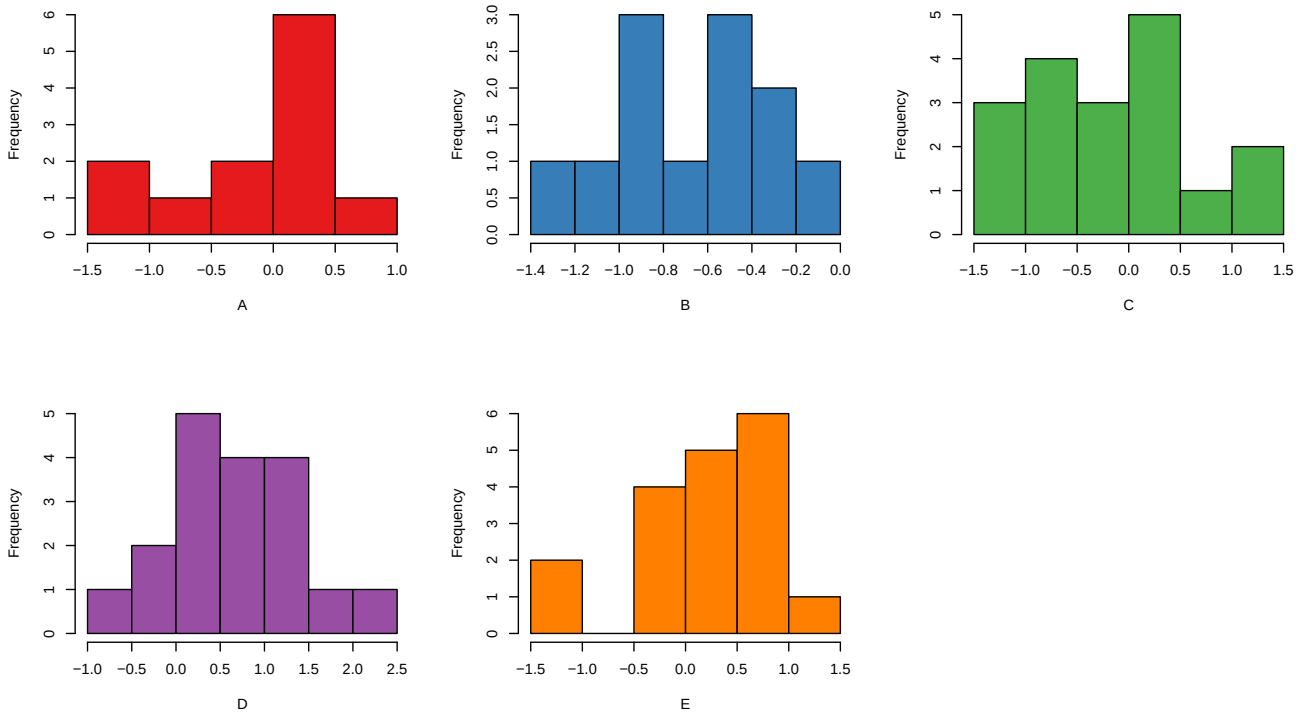


Figure 3: Histogram of the natural-log-transformed protein expression levels for five cell types

```
hist(proteinData.ln$D,xlab="D",col=cols[4],main="")
hist(proteinData.ln$E,xlab="E",col=cols[5],main="")
```

3.1.5 Fitting the model

Figure 3 now shows that most of the cell types are approximately normally distributed, though cell type A is still a little questionable. We'll proceed for now, as the one-way ANOVA is quite robust to small deviations from normality, but we must bear this in mind when interpreting out results. We can also see from Figure 3 that the variance in each of the five groups is now much more similar. They do NOT need to be perfectly the same, and in this case they are similar enough for the assumption of equal variance to be reasonable. Now that we are happy that the assumptions are reasonable, we can perform the one-way ANOVA.

When perfoming linear regression or ANOVA in R, it is more convenient to transform our data into **long** format, which can be done using the `gather` function from the package *tidyr*. For comprehensive descriptions of data manipulation and tidy data, see Hadley Wickham's paper [1] or video [2] on the subject.

Once we have transformed our data, the `aov` function fits the analysis of variance model to our data. Diagnostic plots of the model fit can be visualised using the `plot` function (Figure 4). Of the most

interest is the QQ-plot, which allows us to assess whether the distribution of the response variable is sufficiently normally distributed for each group being compared.

```
library(tidyr)
anovaData <- gather(proteinData.ln)
head(anovaData)
```

```
##   key       value
## 1   A -0.91629073
## 2   A  0.40546511
## 3   A -0.02020271
## 4   A -1.10866262
## 5   A -0.28768207
## 6   A  0.39204209
```

```
mod <- aov(value~ key, data=anovaData)
mod
```

```
## Call:
##    aov(formula = value ~ key, data = anovaData)
##
## Terms:
##                       key Residuals
## Sum of Squares  14.26860  32.05574
## Deg. of Freedom        4        73
##
## Residual standard error: 0.6626611
## Estimated effects may be unbalanced
## 12 observations deleted due to missingness
```

```
par(mfrow=c(2,2))
plot(mod)
```

The summary function allows us to assess the significance of the model:-

```
summary(aov(mod))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## key          4  14.27   3.567   8.123 1.78e-05 ***
## Residuals   73  32.06   0.439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 12 observations deleted due to missingness
```

One of the assumptions of the one-way ANOVA, which we have already explored with our scatter plot, is that the variances in each of the five groups are approximately equal. We can formally test this assumption when carrying out a one-way ANOVA, using a Bartlett's test (although its use is cautioned due to multiple testing issues). The results provide a p-value indicating whether equal variance can be assumed, with a value less than 0.05 suggesting that equal variance should not be assumed. Note that

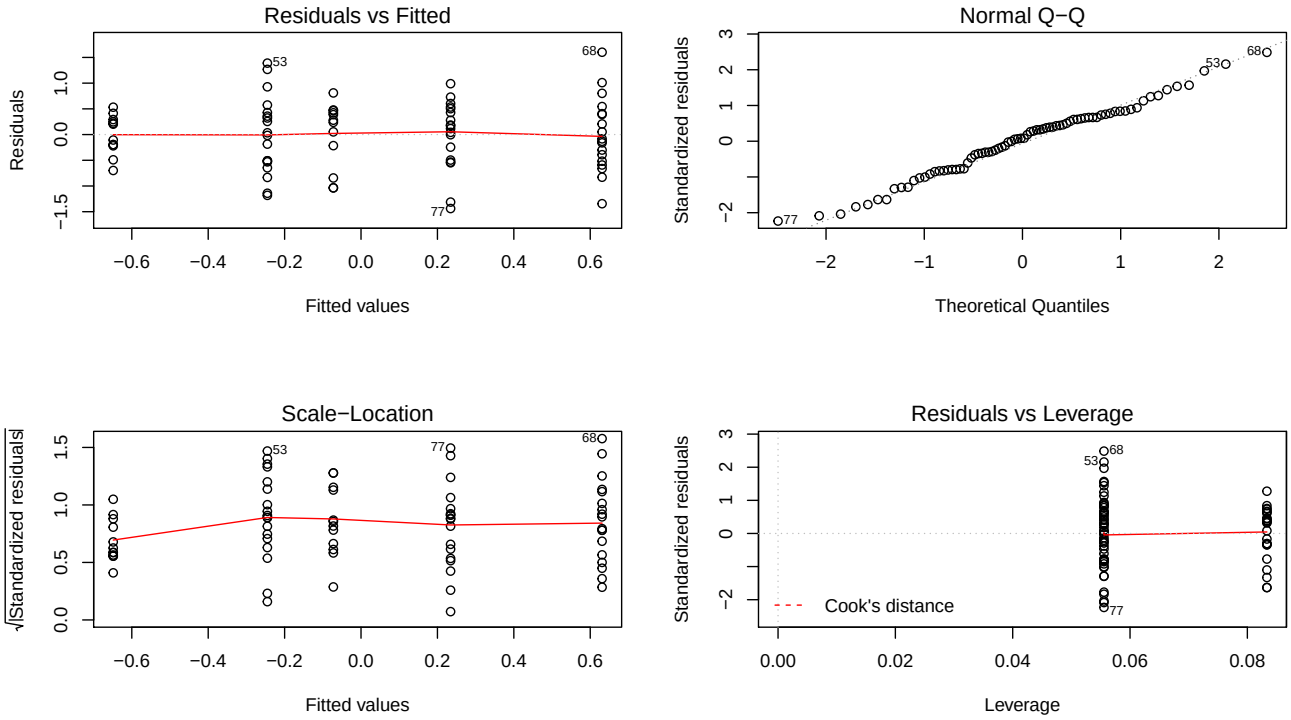


Figure 4: Visualising the results of the ANOVA model

if log-tranformed data are being used for ANOVA, the Bartlett's test should also be performed on the logged data.

```
bt <- bartlett.test(value~key,data=anovaData)
bt

##
##  Bartlett test of homogeneity of variances
##
## data:  value by key
## Bartlett's K-squared = 5.8261, df = 4, p-value = 0.2125
```

In this example, the p-value was 0.213, so there is no evidence to suggest that the variances in each of the five groups aren't approximately equal to each other. If the Bartlett's test gives a significant p-value, we cannot assume equal variances across the groups and a non-parametric test, such as the Kruskal-Wallis test, should be used instead of a one-way ANOVA.

The equal variance assumption (on the log-tranformed data) is reasonable in this example, so we can go ahead and use the one-way ANOVA to analyse the data. The results of the one-way ANOVA provide a p-value of < 0.0001 which is statistically significant. This suggests that there is evidence of a difference in the mean log-transformed (or geometric mean of the original) protein expression levels between two or more of the five cell types. As the result of the one-way ANOVA was significant, we may be interested in

making further comparisons between pairs of groups, and we can do this with the post-hoc test results. Note: if the one-way ANOVA result had not been significant we would usually stop here and not look at the post-hoc test results.

```
post.tests<- TukeyHSD(mod)
post.tests
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = value ~ key, data = anovaData)
##
## $key
##           diff        lwr       upr     p adj
## B-A -0.5762181 -1.3329262 0.1804901 0.2187264
## C-A -0.1726875 -0.8634644 0.5180894 0.9560187
## D-A  0.7034259  0.0126490 1.3942027 0.0438762
## E-A  0.3068774 -0.3838995 0.9976543 0.7265567
## C-B  0.4035305 -0.2872463 1.0943074 0.4809387
## D-B  1.2796439  0.5888671 1.9704208 0.0000183
## E-B  0.8830955  0.1923186 1.5738723 0.0054767
## D-C  0.8761134  0.2582638 1.4939630 0.0015431
## E-C  0.4795649 -0.1382847 1.0974145 0.2023355
## E-D -0.3965485 -1.0143981 0.2213012 0.3841989
```

The post-hoc tests are actually just unpaired t-tests, but the results reported are adjusted for multiple testing. In this example, we want to compare all pairs of groups, but sometimes there may be specific groups that you wish to compare. You should plan your comparisons before starting your analysis and it is better, at least from a statistical viewpoint, to perform the least number of comparisons that will sufficiently answer your question(s). As we want to perform more than one t-test on this data (in fact we are performing 10 pairwise comparisons!), we must be careful to adjust for multiple testing. Here we see that there is a significant difference between cell types A v D (p value: 0.0438762), a very significant difference between cell types B v E (p value: 0.0054767) and C v D (p value: 0.0015431), and an extremely significant difference between cell types B v D (p value: 1.8302674×10^{-5}).

Just because a result is statistically significant does not mean that it is biologically or clinically important. You can refer to the mean difference column to judge whether a difference of the size seen in the data is likely to be of biological or clinical importance, but remember that these values are now on the natural log (ln) scale! You can transform back to the original scale by taking the exponential of the mean difference: in this example, the mean difference between B and D is 1.28 on the natural log scale. On the original scale this translates to $e^{1.28} = 0.562$, this is now the difference in geometric rather than the usual arithmetic means.

3.2 The Kruskal-Wallis test

This tests if k independent samples are drawn from the same population. As the Kruskal-Wallis test ranks the values, it is more powerful than the Median test (3.4), which only looks at direction of differences. The Kruskal-Wallis test is derived from the one-way ANOVA (3.1), but uses ranks rather than actual observations. It is also the extension of the Mann-Whitney U test to greater than two groups. However, it will provide the same result as the Mann-Whitney U test if carried out on two groups.

3.2.1 Assumptions

The assumptions of the Kruskal-Wallis test are:

- 1. The data have been collected from a randomly selected set of observations.
- 2. The dependent variable is at least at the ordinal level of measurement, i.e. the data are able to be ranked.
- 3. There are more than two independent groups.
- 4. There is independence of observations within each group and between the groups. There are no repeated measures or multiple response categories.
- 5. The shapes of the distributions of the groups are similar.

3.2.2 Null Hypothesis

If the last assumption holds then the hypotheses are:

H_0: The medians in the k groups are equal.

H_A: There is a difference in medians between the k groups.

If the last assumption does not hold:

H_0: The k groups have the same shape and location.

H_A: The k groups have a different shape and location.

The alternative hypothesis can be directional or non-directional. If a significant result is obtained then post-hoc testing can be used to see where any differences lie.

3.2.3 Method

If the assumptions are met the test can be used in the following way:

- 1. Determine the null and alternative hypothesis and α the level of significance for the test.
- 2. Rank all observations from lowest to highest.
- 3. Calculate the sum of the ranks for each group.
- 4. Calculate the average rank in each group, R_i, and the overall average rank, R .

- 5. Calculate the test statistic H,

$$H = 12\frac{\sum n_i(R_i - R)^2}{N(N+1)} \tag{1}$$

  where n_i = the number of observations in group i, N = the total sample size
- 6. Compare this value with the χ^2 distribution with $k-1$ degrees of freedom. If the statistic is bigger than the critical value in the chi-square table, the result is significant. If the result is significant, then pairwise post-hoc tests can be carried out.

3.2.4 Example

The data are the reduction in weekly headache activity for three treatment groups, expressed as a percentage of the baseline data (example from Altman [3]). Note that we are assuming that the groups have a similar distribution as we are testing for a difference in medians.

Relaxation / response feedback	Relaxation alone	Untreated
62.00	69.00	50.00
74.00	43.00	-120.00
86.00	100.00	100.00
74.00	94.00	-288.00
91.00	100.00	4.00
37.00	98.00	-76.00

The null and alternative hypothesis for our test are as follows:-

H_0: The three samples come from populations with the same median. H_A: At least one sample comes from a population with a different median.

Computing the ranks then gives the following table:

Relaxation/response feedback	Rank	Relaxation alone	Rank	Untreated	Rank
62.00	8.00	69.00	9.00	50.00	7.00
74.00	10.50	43.00	6.00	-120.00	2.00
86.00	12.00	100.00	17.00	100.00	17.00
74.00	10.50	94.00	14.00	-288.00	1.00
91.00	13.00	100.00	17.00	4.00	4.00
37.00	5.00	98.00	15.00	-76.00	3.00
Rank sum	59.00		78.00		34.00
(mean)	9.83		13.00		5.67

We then have all the values that we need in order to compute the test statistic:

$R = \frac{N+1}{2} = \frac{18+1}{2} = 9.5$

$R_1 = 9.833;\ R_2 = 13;\ R_3 = 5.667$

$H = \frac{12(6(9.833-9.5)^2+6(13-9.5)^2+6(5.667-9.5)^2)}{18(18+1)} = 5.69$

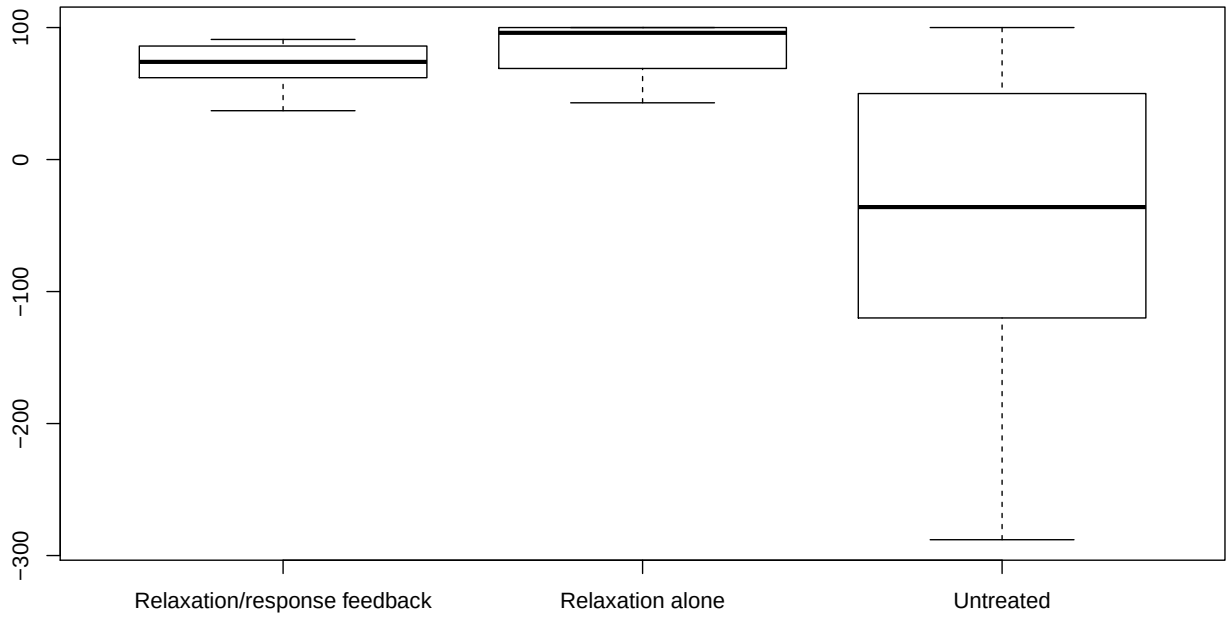
However, $\chi^2_{2,0.05} = 5.99, 5.69 < 5.99$. Therefore there is not sufficient evidence to reject the null hypothesis at the 5% level of significance, $p = 0.06$

3.2.5 Analysis in R

A quick boxplot of the raw data reveal that the assumptions for One-Way ANOVA (3.1.1) are not satisfied and we need to take a non-parametric approach.

```
headache <- matrix(c(62,74,86,74,91,37,
                     69,43,100,94,100,98,
                     50,-120,100,-288,4,-76), ncol=3)

colnames(headache) <- c("Relaxation/response feedback","Relaxation alone","Untreated")
boxplot(headache)
```



To see how to compute the statistic in R, we first have to structure the data in a more convenient format using the `gather` function from the *tidyr* package (2.2). We can then use the `kruskal.test` function, which applies the Kruskal-Wallis test.

```
library(tidyr)
headache <- data.frame(headache)
headache <- gather(headache)
kt <- kruskal.test(value~key,data=headache)
kt
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  value by key
## Kruskal-Wallis chi-squared = 5.7254, df = 2, p-value = 0.05711
```

```
names(kt)
```

```
## [1] "statistic" "parameter" "p.value"   "method"    "data.name"
```

```
kt$statistic
```

```
## Kruskal-Wallis chi-squared
##                    5.72545
```

```
kt$p.value
```

```
## [1] 0.05711293
```

3.2.6 Presentation of Results

The results of the Kruskal-Wallis test could be reported in the following way:

> The results of the Kruskal-Wallis test indicate that there was no evidence of differences in the reduction in weekly headache activity for three treatment groups (2k-w=5.69, p=0.051).

3.2.7 Advantages and Limitations

The Kruskal-Wallis test is very simple and easy to understand and use. It does not require equal sample sizes between groups. It is one of the most popular non-parametric tests. It is also one of the most powerful (i.e. it is very likely to reject the null hypothesis given that it is false) non-parametric tests for continuous data.

3.3 Friedman's test

The Friedman test extends the previously mentioned Wilcoxon signed ranks test to more than two repeated values at more than two time points. Alternatively, to more than two matched blocks where the individuals of each block are randomly assigned to a group.

The test examines the ranks at the different time points or in matched blocks and tests whether the continuous underlying distribution of the variables is the same. It is the non-parametric equivalent of the repeated measures ANOVA.

3.3.1 Null Hypothesis

H_0: There is no difference in median between the groups being tested.

H_A: There is at least one difference in median between the groups.

This is a non-directional alternative hypothesis. If the alternative hypothesis is to be directional a more powerful way of analysing the data would be to carry out planned comparisons, with the appropriate correction for multiple testing. Alternatively, overall tests of significance followed by post-hoc tests can be used.

3.3.2 Assumptions

- 1. The data to be analysed are continuous and at least at the ordinal level of measurement.
- 2. The data from a randomly selected sample are either multiple observations from a single sample across more than two time periods or conditions. Otherwise, the data are blocks of matched subjects in which the subjects from a given block are each randomly assigned to one of the three or more conditions.
- 3. The subjects or blocks of subjects are independent; that is, the results within one block do not have an influence on the results within the other blocks.

3.3.3 Method

- 1. Construct the null and alternative hypotheses.
- 2. Construct a two-way table with N (the number of subjects or matched sets of subjects) rows and k (the number of conditions or data collection periods) columns.
- 3. Rank each person's scores from lowest to highest and sum ranks in each column.
- 4. If the null hypothesis is not true then the sum of the columns will vary from column to column. The Friedman test examines the extent to which these column sums vary from what is expected using the following formula:

$$F_r = \frac{12}{Nk(k+1)} \sum_j R_j^2 - 3N(k+1) \tag{2}$$

  where: R_j = the sum of the ranks for column j N = the number of subjects k = the number of periods or conditions.
- 5. Look F_r up in tables of Friedmans distribution.
- 6. Reject the null hypothesis in favour of the alternative hypothesis if the F_r value is greater than (or equal to) the value in the tables.

Note: If N and k are sufficiently large, then F_r can be compared to a χ^2 distribution on $k-1$ degrees of freedom.

3.3.4 Example

The data for this example is taken from Rubin and Peter's paper [4]. The Friedman test will be used to study whether or not hydralazine would relieve high blood pressure in the lungs.

Person	Before	After48Hours	After6Months
1	22.20	5.40	10.60
2	17.00	6.30	6.20
3	14.10	8.50	9.30
4	17.00	10.70	12.30

Table 7: Total pulmonary resistance before and after hydralazine

Person	Before	Rank	After48Hours	Rank	After6Months	Rank
1.00	22.20	3.00	5.40	1.00	10.60	2.00
2.00	17.00	3.00	6.30	2.00	6.20	1.00
3.00	14.10	3.00	8.50	1.00	9.30	2.00
4.00	17.00	3.00	10.70	1.00	12.30	2.00
		12.00		5.00		7.00

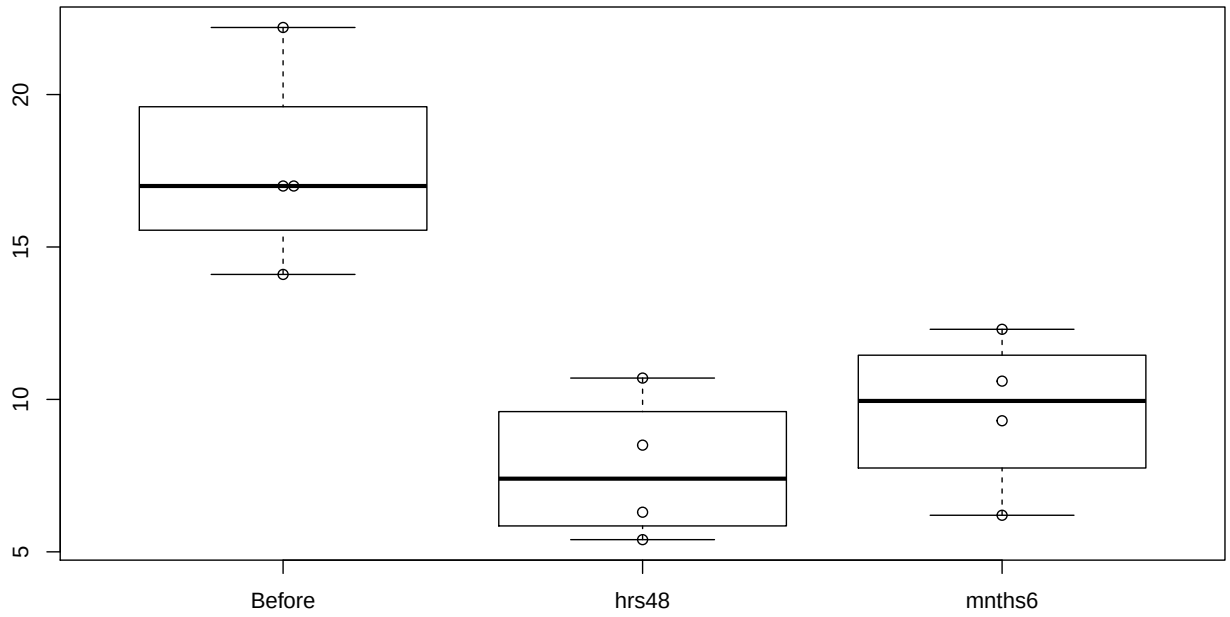
$$F_r = \frac{12}{4\times3(3+1)}[12^2 + [5^2 + 7^2] - [3 \times 4(3+1)] = 6.5$$

As 6.5 is the same as the value in the Friedman's distribution table there is sufficient evidence to reject the null hypothesis and conclude that at least one group is different from the others.

3.3.5 Analysis in R

We create a data frame representation in R, and produce the boxplots:-

```
rubinPeters <- data.frame(Person = 1:4,
                          Before=c(22.2,17,14.1,17),
                          hrs48 = c(5.4,6.3,8.5,10.7),
                          mnths6 = c(10.6,6.2,9.3,12.3))
boxplot(rubinPeters[,2:4])
beeswarm(rubinPeters[,2:4],add=TRUE)
```




```
fr <-  friedman.test(as.matrix(rubinPeters[,-1]))
fr

##
##  Friedman rank sum test
##
## data:  as.matrix(rubinPeters[, -1])
## Friedman chi-squared = 6.5, df = 2, p-value = 0.03877
```

3.3.6 Post-hoc testing

If the Friedman test shows that there is a difference in medians between the groups it is possible to carry out post-hoc testing to see which groups there is actually a difference between. This is done by comparing average ranks in all the pairs or comparing to baseline. The null hypothesis that there is no difference in mean ranks between the pairs, will be rejected if the absolute value of these differences is greater than a specified critical value. If the following condition holds, the null hypothesis will be rejected:

$$|\bar{R}_i - \bar{R}_j| \geq Z_\alpha/[k(k-1)]\sqrt{\frac{k(k+1)}{6N}} \tag{3}$$

where; R_i = the mean rank in period or condition i, R_j = the mean rank in period or condtion j, Z_α = the critical z value for α', $\alpha' = \alpha/[k(k-1)]$, k = the number of periods or conditions, N = the

number of subjects.

In the example above, the average ranks for the three time points are 3, 1.25 and 1.75. Since k= 3 and $\alpha = 0.05$ the critical value of the z-statistic is a z for which α' = 0.05 / 3(2) = 0.0083. Looking this value up in the tables for the Normal distribution gives z = 2.39. The critical value therefore, is:

$$2.39\sqrt{\frac{3(4)}{6(4)}} = 1.68$$

The absolute values of the three comparisons are:

$$|\bar{R}_1 - \bar{R}_2| = |3 - 1.25| = 1.75 > 1.68$$

$$|\bar{R}_1 - \bar{R}_3| = |3 - 1.75| = 1.25 < 1.68$$

$$|\bar{R}_2 - \bar{R}_3| = |1.25 - 1.75| = 0.5 < 1.68$$

The comparison between before treatment and 48 hours after treatment is the only one that is greater than the critical value of 1.68. Therefore, we can conclude that according to the post-hoc approach, hydralazine only relieves high blood pressure in the lungs 48 hours after the treatment. This effect was not maintained 6 months after treatment.

It is also possible to use the Wilcoxon ranked sign tests for post-hoc testing. The procedure is carried out in the same way as described before. However, the Bonferroni correction must be applied to allow for multiple testing. That is the critical value of α becomes $\alpha' = \alpha/k$ where k is the number of tests to be carried out and α is the original significance level. The value of α' is the one looked at in the table or that the output p-value is compared against.

3.3.7 Presentation of the Results

The results of the Friedman test could be reported in the following way:

> The results of the Friedman test indicate that there is a significant difference in median total pulmonary resistance across the three time periods. Therefore, we can conclude that hydralazine alters total pulmonary resistance (p=0.042).
>
> Post-hoc analyses with the adjustment of the two-tailed level to 0.0083 indicated that there were decreases in total pulmonary resistance from before treatment (Md=17.0) to 48 hours after treatment (Md= 7.4). There was no evidence of other differences.

Note: That post-hoc testing was carried out here on a very small sample size as an illustration, in reality post-hoc testing would not be carried out on such a small sample size.

3.3.8 Advantages and Limitations

The Friedman test is very versatile and can be used with randomised block designs and multiple observations of a single sample. It is useful when the dependent variable is not consistent with a normal distribution.

There are some drawbacks however, it is possible for the medians not to change and there still to be significant differences between groups. Although it is often referred to as the Friedman two-way ANOVA by ranks, it is restricted to within group comparisons. It is not possible to test between group comparisons. This is a major disadvantage in clinical research as it is not possible to make experimental-control group comparisons. Each group can be analysed separately and compare their results. However, it is not possible to test a group and time interaction with independent groups.

3.3.9 Summary

Friedman's test, tests the null hypothesis that k related groups come from the same population. For each case, the k observations are ranked from 1 to k; the test statistic is based on these ranks.

After establishing a difference between one of the variables, post-hoc testing can be carried out to decide which of the groups are actually different. An appropriate method for allowing for multiple testing must be carried out.

3.4 Median Test

Ths Median test tests whether two or more independent samples are drawn from populations with the same median using the χ^2 statistic. It can be used when the assumptions of similarity of distributions for the Mann-Whitney U and Kruskal-Wallis (3.2) tests are not met.

3.4.1 Null Hypothesis

H_0: There is no difference in medians amongst the groups being studied.

H_A: There is at least one difference in medians amongst the groups being studied.

3.4.2 Assumptions

The assumptions of the Median test are:

- 1. The dependent variable is at least at the ordinal level of measurement. The data are from two or more groups
- 2. The groups are independent and a subject can only be in one of the groups.
- 3. The assumptions of the χ^2 test apply to the second half of the test (if these are not met then Fisher's exact test should be used.

3.4.3 Method

If these assumptions are met then the test can be carried out in the following way:

- 1. Construct the null and alternative hypotheses and decide on α the level of significance for the test.
- 2. Treat the data as a single sample and calculate the overall median.
- 3. Separate the data into the various groups and classify the observations in each group as either above, below or equal to the overall median. Calculate the number above and below or equal to the median, in each group.
- 4. Arrange these values into a $2 \times c$ contingency table, where the two rows are: or $\leq$ to the overall median. The c columns are the groups.
- 5. Calculate the χ^2 statistic for the table, if the assumptions hold.
- 6. Compare the value of the chi-square statistic with the value in the tables on $(c-1)$ degrees of freedom (where c is the number of groups) at the pre-specified level of α
- 7. Reject the null hypothesis of equal medians if X^2 exceeds the critical value of the χ^2 distributions.
- 8. If the null hypothesis is rejected, it is then possible to do post-hoc testing on the individual groups to see which ones are significantly different. This again will be using the median test, but applied to pairs of groups.

3.4.4 Example

There are three groups with different types of dementia (data from Sanjana Nyatsanza, Fulbourn hospital). Below are the patients scores on a mini mental state examination (MMSE). The median test will be used to see if there is a significant difference between the groups.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Group1	19.00	7.00	17.00	28.00	21.00	6.00	21.00	19.00	27.00	8.00	25.00
Group2	16.00	22.00	30.00	24.00	22.00	23.00	22.00	28.00	29.00	29.00	0.00
Group3	4.00	9.00	30.00	29.00	25.00	22.00	25.00	26.00	27.00	18.00	10.00

1. Stating the null hypothesis and significant level

H_0: There is no difference in medians between the groups.

H_A: There is a difference in medians between the groups.

$\alpha = 0.05$

2. The overall median (i.e. the one in the middle when ranked in order) is 22.

3. Classify the values in each group as above or below 22.

19	7	17	28	21	6	21	19	27	8	25
<=22	<=22	<=22	>22	<=22	<=22	<=22	<=22	>22	<=22	>22
16	22	30	24	22	23	22	28	29	29	0
<=22	<=22	>22	>22	<=22	>22	<=22	>22	>22	>22	<=22
4	9	30	29	25	22	25	26	27	18	10
<=22	<=22	>22	>22	>22	<=22	>22	>22	>22	<=22	<=22

	Group1	Group2	Group3	Total
<=22	8.00	5.00	5.00	18.00
>22	3.00	6.00	6.00	15.00
Total	11.00	11.00	11.00	33.00

5. Find the expected values for each of the cells:

6.00	6.00	6.00
5.00	5.00	5.00

and calculate the χ^2 statistic

$\chi^2 = \sum_{ij} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

$= \frac{(8-6)^2}{6} + \frac{(5-6)^2}{6} + \frac{(5-6)^2}{6} + \frac{(3-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(6-5)^2}{5}$

$= 0.667 + 0.167 + 0.167 + 0.8 + 0.2 + 0.2 = 2.20$

6. The χ^2 statistic for $\alpha = 0.05$ is 5.99

7. $\chi^2 = 2.20 < 5.99$, there is insufficient evidence to reject the null hypothesis that the medians are the same for all three groups, $p = 0.33$

8. As this result is not significant, post-hoc testing could not be carried out as there are no significant differences between the groups.

The Median test is very straightforward and easy to apply and is particularly useful when the exact values of the scores (especially those at the extremes) are unknown. The test only considers two states for the scores, above and below (or equal to) the median and does not take the size of the differences into account. Therefore, the Median test is less powerful than the Mann-Whitney U and Kruskal-Wallis tests.

3.4.5 Analysis in R

The Median test is implemented in the *RVAideMemoire* package (2.1). As usual we need to 'tidy' the input data frame (2.2) in order to define a factor for the test. The `mood.medtest` function implements the test and has the option to compute an exact, or approximate p-value.

```
library(RVAideMemoire)

mmse <- data.frame(Group1=c(19,7,17,28,21,6,21,19,27,8,25),
                   Group2 = c(16,22,30,24,22,23,22,28,29,29,0),
                   Group3 = c(4,9,30,29,25,22,25,26,27,18,10)
)

mmse <- gather(mmse)
```

```
mood.medtest(value~key,data=mmse,exact=FALSE)

##
##  Mood's median test
##
## data:  value by key
## X-squared = 2.2, df = 2, p-value = 0.3329

mood.medtest(value~key,data=mmse)

##
##  Mood's median test
##
## data:  value by key
## p-value = 0.3935
```

3.5 Jonckheere-Terpstra Test

Also known as the *test for ordered alternatives* or a *non-parametric test for trend*. The Jonckheere-Terpstra test is used when the assumption that the independent variable is nominal in the Kruskal-Wallis test (3.2) is violated i.e. the groups have an explicit order. Since it allows the independent variable (the groups) to have an order it is more powerful than the Kruskal-Wallis test (3.2) when the groups are ordered.

3.5.1 Null Hypothesis

H_0: There is no difference in median values between the groups.

H_A: The median values of the groups increase in a specific predetermined sequence.

3.5.2 Assumptions

- 1. The data have been collected from a randomly selected set of observations.
- 2. The data to be analysed are continuous and at least at the ordinal level of measurement.
- 3. The k groups must be ordinal with a predetermined order.
- 4. Under the null hypothesis it is assumed that each sample is from the same population.

3.5.3 Method

- 1. Construct the null and alternative hypotheses and determine the level of significance, α.
- 2. Specify the order of the groups, which need not be equal sized.
- 3. Cast the data into a two-way table with the groups in the pre-specified order, arranged from smallest to largest.
- 4. Within each group order the data from smallest to largest.

- 5. Count the total number of times each value in the first group precedes (is lower than) a value in the subsequent groups this is the precedent count for the group.
- 6. Add 0.5 to each precedent count when a tie (equal value) occurs between groups.
- 7. Find the precedent count for the remaining groups and sum over the groups to give J, the test statistic.
- 8. Compare this value to that in the tables for J. If the statistic is greater than or equal to the critical value in the Jonckheere-Terpstra test tables, the result is significant. If the result is significant, the null hypothesis is rejected in favour of the alternative hypothesis.

3.5.4 Large Sample Size

1. When the sample size is large the distribution of J tends to a normal distribution, with mean:-

$$\mu_j = \frac{N^2 - \sum_{j=1}^{k} n_j^2}{4} \tag{4}$$

and standard deviation

$$\sigma_j = \sqrt{\frac{N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)}{72}} \tag{5}$$

where,

N = total sample size,

n_j = sample size of group j,

k = number of groups,

$\sum$ = sum across groups,

A z statistic can then be calculated as follows:

$$z = \frac{J - \mu_j}{\sigma_j} \tag{6}$$

This z statistic can then be compared to the tables for the normal distribution:

3.5.5 Example

Mcm-2 values were collected in a breast cancer study. The median Mcm-2 value was expected to increase with histological grade (data below). This hypothesis was tested using the Jonckheere-Terpstra test.

Stating assumptions and significance level

1. H_0: The median Mcm-2 value is the same across histological grades.

H_A: There is an increase is median Mcm-2 value as histological grade increases. $\alpha = 0.05$

Grade1	Grade2	Grade3
1.99	4.40	6.94
3.01	9.82	8.04
4.17	10.23	9.82
7.13	11.99	15.75
9.82	11.99	18.30
9.91	13.17	25.01
	13.20	26.40
		28.17

2. The order of the groups is that of increasing histological grade.

5. The precedent counts for each pair of groups is in the table below.

6. 0.5 has been added to each precedent count with a tied value between groups.

Grade1And2	Grade1And3	Grade2And3
7	8.00	8.00
7	8.00	5.50
7	8.00	5.00
6	7.00	5.00
5.5	5.50	5.00
5	5.00	5.00
		5.00
Total:37.5	41.50	38.50

7. $J = 37.5 + 41.5 + 38.5 = 117.5$

8. From the tables ($n_1 = 6, n_2 = 7, n_3 = 8, \alpha = 0.05$) the critical value is 99. As $117.5 > 99$, there is sufficient evidence to reject the null hypothesis and conclude that there is a significant increase in median Mcm-2 value as histological grade increases.

3.5.6 Presentation of results

The results of the Jonckheere-Terpstra test could be reported in the following way:

> The results of the Jonckheere-Terpstra test show that there is a trend for an increase in median Mcm-2 value as histological grade increases (J=117.5, p=0.003).

3.5.7 Analysis in R

Performing the Jonckheere-Terpstra test in R requires the *clinfun* package to be installed (2.1) (we recommend using this version of the test rather than versions in other packages).

```
library(clinfun)
grade <- data.frame(Grade1 = c(1.99,3.01,4.17,7.13,9.82,9.91,NA,NA),
                    Grade2 = c(4.40,9.82,10.23,11.99,11.99,13.17,13.20,NA),
                    Grade3 =c(6.94,8.04,9.82,15.75,18.30,25.01,26.40,28.17))
grade <- gather(grade)
grade$key <- as.numeric(gsub("Grade","",grade$key))
jonckheere.test(grade$value,grade$key)
```

```
##
##  Jonckheere-Terpstra test
##
## data:
## JT = 117.5, p-value = 0.004047
## alternative hypothesis: two.sided
```

3.5.8 Advantages and limitations

The main advantage of the Jonckheere-Terpstra test is that unlike the Kruskal-Wallis test it allows the groups to have an order therefore it is more powerful than the Kruskal-Wallis test if the groups have a pre-specified order. Since the Jonckheere-Terpstra test is a test for trend there is no need for post-hoc tests to see where differences lie after a significant result.

The main limitation of the test is that the groups must have a pre-specified or explicit order. It is not possible to look for an order and then test for a trend. If there is no explicit order then a Kruskal-Wallis test should be used instead.

3.5.9 Summary

The Jonckheere-Terpstra test is a more powerful alternative to the Kruskal-Wallis test when there is an explicit order to the groups. It is a test for trend of increasing medians between the groups. It is a much under-used nonparametric test; often the Kruskal-Wallis test is used where it would have been more appropriate to use the Jonckheere-Terpstra test.

4 Regression

Regression Analysis refers to a set of statistical techniques for modeling the relationship between two or more variables. One of these variables is the **response** variable (or **dependent** variable), while the other variables are known as **explanatory** (or **independent**) variables. Both response and explanatory variables are continuous, i.e. real numbers with decimal places, for example weights, intensities or growth rates.

Regression analysis is widely used for prediction, where the explanatory variables are known as **predictors** and the response variable is the thing that is being predicted.

One way of working out whether regression is the appropriate analysis for your data is to consider the most natural way of plotting the data in order to address the question you are asking. An XY scatter plot would point to regression, whereas analysis of variance (ANOVA) might be more appropriate if a better representation of the data was in the form of a boxplot.

An example of a dataset where regression analysis might be applied is shown in Figure 5. This shows the results of a dose-response experiment where an *E. coli* strain was exposed to various concentrations of the growth inhibitor, lactoferrin. In this example, we are interested in how the growth rate varies with concentration and a regression analysis would be suitable. Regression analysis involves fitting a model to the data that attempts to describe the relationship between the response and explanatory variables, in this case the growth rate and concentration.

There are several types of regression analysis — which you use will depend on the number of explanatory variables and the type of model to be fitted.

- **Simple linear regression** – the simplest and most frequently used, where there is one response variable and one explanatory variable and the relationship can be described through a linear model
- **Multiple linear regression** – fits a linear model using multiple explanatory variables
- **Polynomial regression** – used to test for non-linearity in a relationship
- **Non-linear regression** – to fit a specified non-linear model to the data
- **Non-parametric regression** – used when there is no obvious functional form
- **Logistic regression** –when the response variable is a nominal (or categorical) variable

There are many other techniques that can be used when there are three or more measurement variables, including principal components analysis, hierarchical and non-hierarchical clustering, and multidimensional scaling.

4.1 Linear Regression

Linear regression involves fitting the simplest model of all, a linear model of the form:

$$y = ax + b \tag{7}$$

where y is the response variable and x is a continuous explanatory variable. This should look familiar as the equation for a straight line graph. There are two parameters, a and b. a is the intercept, the

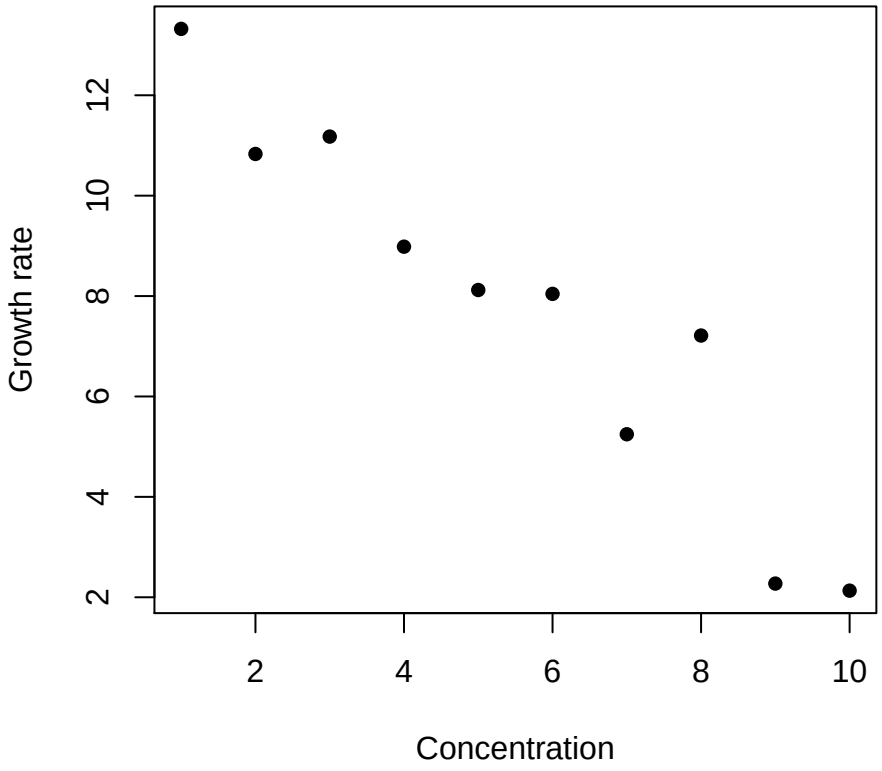


Figure 5: Growth rate of E. coli at various concentrations of lactoferrin.

value of y when $x = 0$, and b is the slope, or gradient, and is equal to the change in y divided by the change in x that brought about the change in y.

4.1.1 The lactoferrin dataset

Let's go back to our example dataset in Figure 5, showing the effect of lactoferrin concentration on the growth rate of *E. coli*. We can read the data into R using the `read.csv` function:

```
data <- read.csv("lactoferrin.csv")
data
```

```
##    conc growth
## 1     1  13.32
## 2     2  10.83
## 3     3  11.18
## 4     4   8.98
## 5     5   8.12
```

```
## 6     6   8.04
## 7     7   5.25
## 8     8   7.21
## 9     9   2.27
## 10   10   2.13
```

We can generate something similar to Figure 5 using the `plot` function:

```
plot(data$conc, data$growth, pch=16, xlab="Concentration", ylab="Growth rate")
```

Figure 5 shows that there is a roughly linear relationship between growth rate and lactoferrin concentration, with the growth rate decreasing with higher concentrations. We could estimate the parameters of a simple linear model by drawing a line through the data by eye and then calculating it's gradient and seeing where the line passes through the y-axis, i.e. the growth rate when the lactoferrin concentration is zero ($x = 0$). The R function `lm` will do this for us using a mathematical technique known as *'least squares'*. Figure 6 shows the resulting line of best fit. Also shown are the **residuals**, the vertical distances between the actual data points and the line of best fit, i.e. between the observed and fitted values.

4.1.2 Fitting the linear model

The aim is to minimize the sum of squares of the residuals (also known as the error sum of squares, **SSE**), i.e. to find the minimum of

$$SSE = \sum_i (y_i - a - bx_i)^2 \tag{8}$$

More formally, the regression model is written as:

$$y_i = a + bx_i + \varepsilon_i \tag{9}$$

where ε_i is the error term and is assumed to be normally distributed:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{10}$$

Now we can write the error sum of squares as:

$$SSE = \sum_i \varepsilon_i^2 = \sum_i (y_i - a - bx_i)^2 \tag{11}$$

Mathematically, we set the derivative of this function with respect to the slope to zero ($dSSE/db = 0$), do the same for the derivative with respect to the intercept ($dSSE/da = 0$), and then solve the resulting simultaneous equations. This is left as an exercise for the more mathematically inclined, or see "Statistics: An Introduction using R" by Michael Crawley [5].

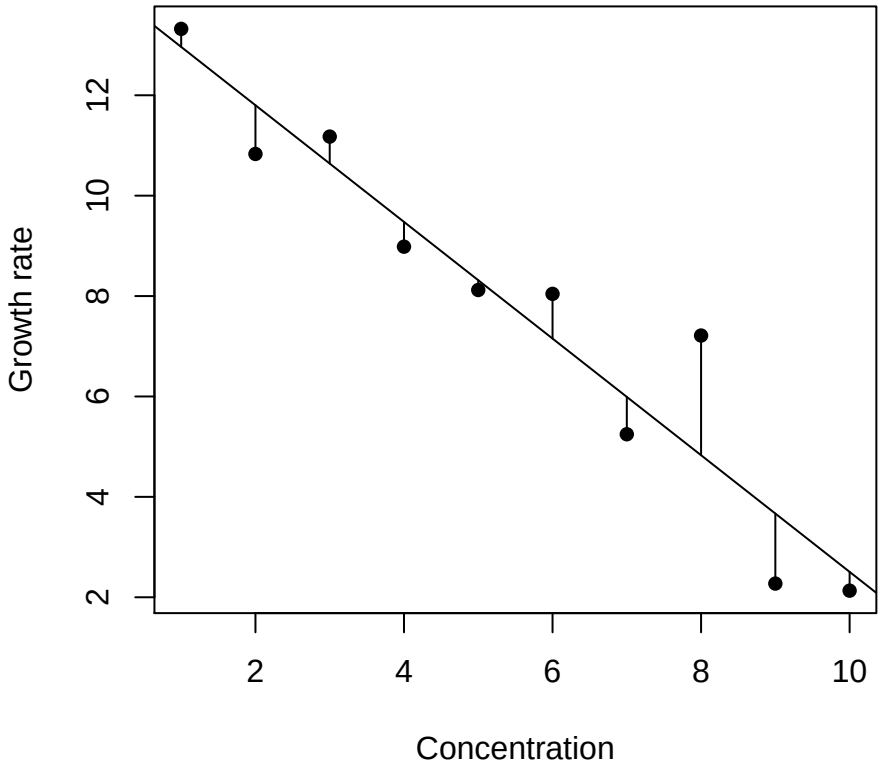


Figure 6: Line of best fit and residuals.

In R, the tilde symbol, '$\sim$', is used in describing a model and can be read as 'is modelled as a function of'. In our example, the model would be written:

$$growth \sim conc \tag{12}$$

The response variable goes on the left of the tilde and the explanatory variable on the right hand side. Our model can be read 'growth rate is modelled as a function of concentration'.

To obtain the parameters of the model, also known as the coefficients, we fit a linear model using the `lm` function:

```
model <- lm(growth ~ conc, data=data)
model

##
## Call:
## lm(formula = growth ~ conc, data = data)
##
```

```
## Coefficients:
## (Intercept)         conc
##       14.12        -1.16
```

We can show the line of best fit on the plot using the `abline` function:

```
plot(data, pch=16, xlab="Concentration", ylab="Growth rate")
abline(model)
```

and the `fitted` function will give us the fitted values for each data point:

```
fitted <- fitted(model)
fitted
```

```
##     1     2     3     4     5     6     7     8     9    10
## 12.96 11.80 10.64  9.48  8.32  7.15  5.99  4.83  3.67  2.51
```

The residuals are the differences between the observed values and the fitted values:

```
residuals <- data$growth - fitted
residuals
```

```
##      1      2      3      4      5      6      7      8      9     10
##  0.359 -0.971  0.537 -0.493 -0.195  0.890 -0.744  2.384 -1.395 -0.374
```

Coefficient, residuals and other properties of the model can be obtained using a number of useful functions or accessed directly through the model object:

```
coefficients(model)
```

```
## (Intercept)        conc
##       14.12       -1.16
```

```
residuals(model)
```

```
##      1      2      3      4      5      6      7      8      9     10
##  0.359 -0.971  0.537 -0.493 -0.195  0.890 -0.744  2.384 -1.395 -0.374
```

```
names(model)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
##  [6] "assign"        "qr"            "df.residual"   "xlevels"       "call"
## [11] "terms"         "model"
```

```
model$coefficients
```

```
## (Intercept)        conc
##       14.12       -1.16
```

```
model$residuals
```

```
##      1      2      3      4      5      6      7      8      9     10
##  0.359 -0.971  0.537 -0.493 -0.195  0.890 -0.744  2.384 -1.395 -0.374
```

A noteable property of linear regression is that the line of best fit passes through the point $(\bar{x}, \bar{y})$ where $\bar{x}$ and $\bar{y}$ are the mean values of x and y respectively.

```
mean(data$conc)
```

```
## [1] 5.5
```

```
mean(data$growth)
```

```
## [1] 7.73
```

```
model$coefficients[1] + model$coefficients[2] * mean(data$conc)
```

```
## (Intercept)
##        7.73
```

4.1.3 Calculating standard errors in the regression parameters

We would like to know how reliable are our estimates for the regression parameters, i.e. the slope and the intercept. For this we consider the total variation in y, represented by the total sum of squares of y:

$$SSY = \sum_i (y_i - \bar{y})^2 \tag{13}$$

The fitted values are represented formally by the symbol $\hat{y}_i$ and the sum of squares of the residuals, SSE, can be written as follows:

$$\hat{y}_i = a + bx_i \tag{14}$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2 \tag{15}$$

Computing the SSE in R is straightforward:

```
SSE <- sum(residuals^2)
SSE
```

```
## [1] 10.8
```

The total variation in y, SSY, can be partitioned into separate components for the variation that is explained by the model, denoted by SSR, and the unexplained variation that is the error sum of squares, SSE, we've already calculated.

$$SSY = SSR + SSE \tag{16}$$

Source	Sum of squares	Degrees of freedom	Mean squares	F ratio
Regression	$SSR = 111.38$	1	111.38	82.84
Error	$SSE = 10.756$	8	$s^2 = 1.345$	
Total	$SSY = 122.136$	9		

Table 8: ANOVA table for the lactoferrin dataset

The variation that is explained by the model is called the regression sum of squares, denoted by SSR, is given by:

$$SSR = \sum_i (\hat{y}_i - \bar{y}_i)^2 \tag{17}$$

Although the `lm` function does all the hard work for you, the following shows how to compute the variances in R from our data:

```
mean_growth <- mean(data$growth)
mean_growth
```

```
## [1] 7.734
```

```
SSY <- sum((data$growth - mean_growth)^2)
SSY
```

```
## [1] 122.1
```

```
SSR <- sum((fitted - mean_growth)^2)
SSR
```

```
## [1] 111.4
```

```
SSE <- SSY - SSR
SSE
```

```
## [1] 10.76
```

These sources of variation are laid out in an analysis of variance table in Table 8. The mean squares column contains values for the variance for each source, calculated as:

$$\text{variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}} \tag{18}$$

The number of degrees of freedom is determined by considering the number of parameters estimated from the data for each sum of squares. For the total sum of squares, $SSY = \sum(y - \bar{y})^2$, there is just one parameter estimated from the data: the mean value, $\bar{y}$. So we have $n - 1$ degrees of freedom, where n is the number of observations (10 in this case). Similarly, in order to calculate the error sum of squares, $SSE = \sum(y - \hat{y})^2 = \sum(y - a - bx)^2$, we need to know the values of two parameters, a and b. These are estimated from the data, so the number of degrees of freedom are $n - 2$.

The number of degrees of freedom for the regression is more difficult to understand but if we consider that the regression degrees of freedom and the error degrees of freedom should add up to the total number of degrees of freedom, then we can see that this value must be 1. In the regression model, the fitted line is specified by two parameters, the slope and the intercept. But the fitted line must go through the mean of the response variable. This fixes the intercept and leaves just one degree of freedom for estimating the slope of the line.

The value of most interest in Table 8 is the error variance, s^2.

$$s^2 = \frac{SSE}{n-2} \tag{19}$$

This is used in computing the standard errors in both the slope and intercept (see [5] for the formulae).

The F ratio is the ratio between the regression variance and the error variance.

```
error_variance <- SSE / (length(data$growth) - 2)
error_variance
```

```
## [1] 1.345
```

```
F_ratio <- SSR / error_variance
F_ratio
```

```
## [1] 82.84
```

This can be used to test for a non-zero slope in the linear regression, where the null hypothesis is that the slope is zero, i.e. there is no dependence of the response variable on the explanatory variable. The ANOVA table and p-value for the test are computed in R by calling the `anova` function:

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: growth
##           Df Sum Sq Mean Sq F value  Pr(>F)
## conc       1  111.4   111.4    82.8 1.7e-05 ***
## Residuals  8   10.8     1.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.1.4 Summarizing the linear model

The `summary` function shows the estimated slope and intercept parameters and their standard errors and is the most useful function for summarizing the results of a regression analysis. In practice you will not calculate the values of SSY, SSE, etc., longhand as above, but instead just call the `summary` function.

```
summary(model)
```

```
##
## Call:
## lm(formula = growth ~ conc, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.395 -0.681 -0.284  0.493  2.384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.125      0.792    17.8  1.0e-07 ***
## conc          -1.162      0.128    -9.1  1.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.16 on 8 degrees of freedom
## Multiple R-squared:  0.912,Adjusted R-squared:  0.901
## F-statistic: 82.8 on 1 and 8 DF,  p-value: 1.71e-05
```

The output includes the standard errors for both the concentration coefficient (slope) and intercept. A t-statistic is given for each coefficient, calculated by dividing the estimated value by the standard error for the coefficient. From this a p-value is also computed.

The coefficients, errors and various statistics can be accessed directly through the summary object returned by the `summary` function or using the `coefficients` function passing the summary object.

```
summary <- summary(model)
names(summary)
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"  "aliased"
##  [6] "sigma"         "df"            "r.squared"     "adj.r.squared" "fstatistic"
## [11] "cov.unscaled"
```

```
summary$fstatistic
```

```
## value numdf dendf
## 82.84  1.00  8.00
```

```
coefficients(summary)
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   14.125     0.7921  17.832 1.002e-07
## conc          -1.162     0.1277  -9.102 1.706e-05
```

```
summary$coefficients
```

```
##             Estimate Std. Error t value  Pr(>|t|)
```

```
## (Intercept)   14.125     0.7921  17.832 1.002e-07
## conc          -1.162     0.1277  -9.102 1.706e-05
```

```
summary$coefficients[1,2]
```

```
## [1] 0.7921
```

Also given is the residual error; this is the square root of the error variance we calculated earlier.

```
summary$sigma
```

```
## [1] 1.16
```

```
sqrt(error_variance)
```

```
## [1] 1.16
```

4.1.5 Confidence intervals for the model parameters

Confidence intervals on the model coefficients are calculated from the t-value and the standard error. Recall that in Student's t-distribution, values of t are the numbers of standard errors to be expected with specified probability and for a given number of degrees of freedom.

$$\text{confidence interval} = t\text{−value} \times \text{standard error} \tag{20}$$

$$CI_{95\%} = t_{(\alpha=0.025, \text{d.f.}=8)} \times \text{s.e.} \tag{21}$$

In R, the `confint` function computes the confidence intervals for the parameters of a model:

```
confint(model, level = 0.95)
```

```
##               2.5 %  97.5 %
## (Intercept) 12.298 15.9516
## conc        -1.456 -0.8675
```

4.1.6 Measuring the degree of fit

The output from the `summary` method includes a value for r^2. This is a measure of the degree of fit and is calculated as the fraction of the total variation in the response variable, y, that is explained by the regression.

$$r^2 = \frac{SSR}{SSY} \tag{22}$$

```
r_squared <- SSR / SSY
r_squared
```

```
## [1] 0.9119
```

This varies from 1 when the regression explains all of the variation in y (all the observed points lie on the line of best fit, $SSR = SSY$ and $SSE = 0$) to 0 when the regression explains none of the variation ($SSE = SSY$, $SSR = 0$). The square root of this quantity, r, is the correlation coefficient (Pearson's product-moment correlation).

```
cor.test(data$growth, data$conc)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$growth and data$conc
## t = -9.1, df = 8, p-value = 2e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9896 -0.8159
## sample estimates:
##    cor
## -0.955
```

```
cor(data$growth, data$conc) ^ 2
```

```
## [1] 0.9119
```

4.1.7 Checking the model assumptions

Aside from the assumption that there is a linear relationship between the explanatory variable and the response variable, linear regression also makes the following assumptions:

- **Constant variance (homoscedasticity)** – the variance of the errors is constant across the range of values for the explanatory variable
- **Independence of errors** – the errors in the response variables are uncorrelated with each other
- **Normality of errors** – the residuals follow a normal distribution

Diagnostic plots for the model can be visualized using the `plot` function (Figure 7). The first graph shows the residuals plotted against the fitted values; ideally these should look random without any structure or pattern in the plot. It would be a problem if there was a clear trend of increasing scatter as the fitted values get larger, i.e. violating the assumption of constant variance. The second plot, the quantile-quantile plot (QQ plot), is also worth looking at as this can highlight problems with non-normality of errors. If the errors are normally distributed this plot should be a straight line. An S-shaped or banana-shaped plot would indicate a need to fit a different model or to transform the data.

```
par(mfrow=c(2,2))
plot(model)
```

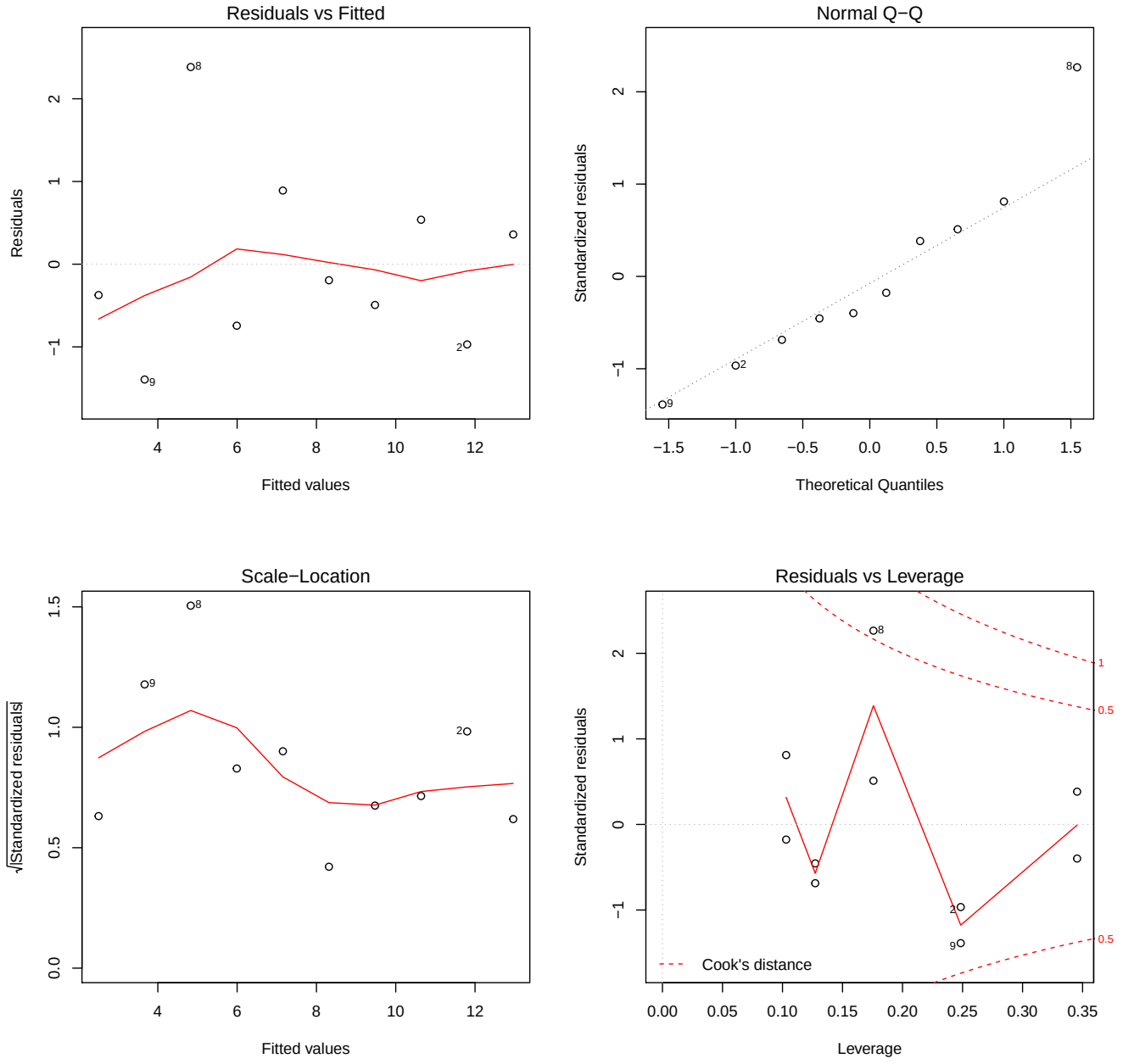


Figure 7: Diagnostic plots for checking the linear model fitted to the lactoferrin data.

Outliers may have a strong influence on the fitted slope and intercept. The residuals vs. leverage plot shows labeled points that represent cases that may need investigating as potentially having undue influence on the regression relationship (cases 2, 8 and 9 in our example dataset).

4.1.8 Using the model for prediction

One use of the regression model is to predict or estimate the value of the response variable for new values of the explanatory variable(s). For our example, we may wish to estimate the growth rate of *E. coli* for some new concentrations of lactoferrin. We can use the `predict` function to do this:

```
concentrations <- data.frame(conc = c(2.5, 6.5, 12.0))
predicted_growth <- predict(model, concentrations)
predicted_growth
```

```
##       1       2       3
## 11.2201  6.5724  0.1819
```

4.2 Beyond Simple Linear Regression

The above introduction to linear regression focuses solely on the simplest case of a linear relationship between a single explanatory variable and the response variable. The relationship often will turn out not to be a straight line in which case we may need to fit a non-linear function or we may be able to convert the relationship into a linear one by transforming the response variable, e.g. by using its logarithm. We may also have several independent, potentially explanatory, variables and may be interested in understanding the functional relationships between these and the dependent variable to see what is causing the variation.

4.2.1 Polynomial regression

One way of assessing whether we have a non-linear relationship is to use polynomial regression. This involves attempting to fit a polynomial function, e.g. a quadratic function of the form:

$$y = a + bx + cx^2 \tag{23}$$

In R the model would be written `y ~x + I(x^2)` where the `I` function is needed to treat the `x^2` term 'as is' (note that the `^` symbol has a different meaning in the R syntax for specifying a model).

Consider the radioactive decay data set shown in Figure 8. The data appear to follow a curve more than a straight line. This is also evident from the residuals diagnostic plot (Figure 9). The residuals for high and low values of x are positive while the residuals for the intermediate values of x are mostly negative.

We can assess the non-linearity in the relationship between x and y by fitting a polynomial regression model. The simplest such model involves adding a quadratic term, x^2.

```
quadratic <- lm(y ~ x + I(x^2))
summary(quadratic)
```

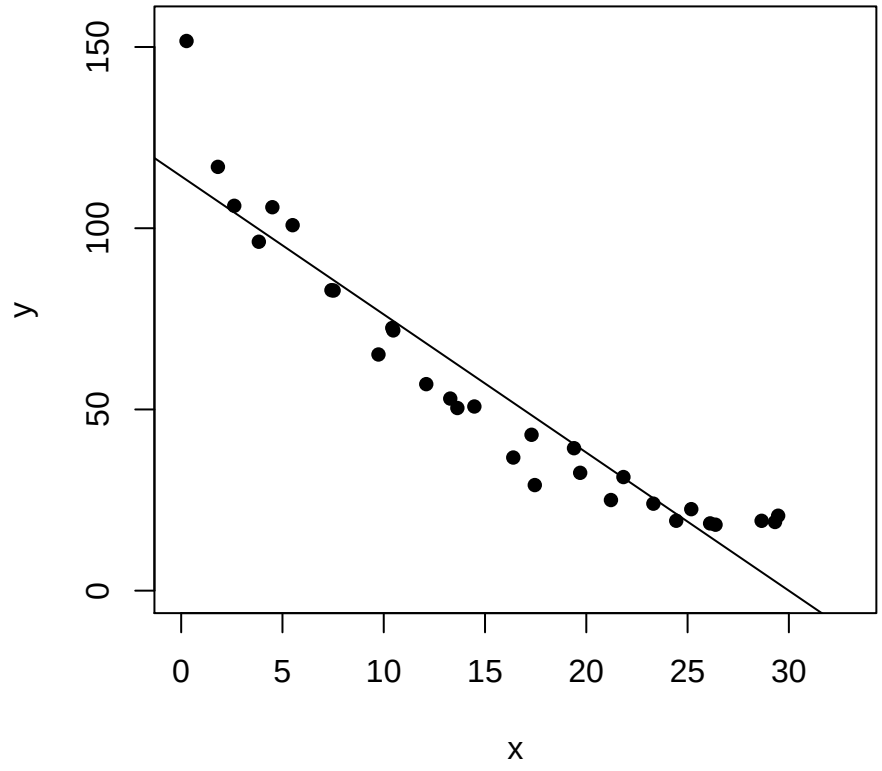


Figure 8: Radioactive decay data set.

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.403  -2.267  -0.429   2.797  16.669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.1447     3.2076   42.76  < 2e-16 ***
## x            -8.2430     0.4803  -17.16  4.8e-16 ***
## I(x^2)        0.1446     0.0152    9.53  4.0e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

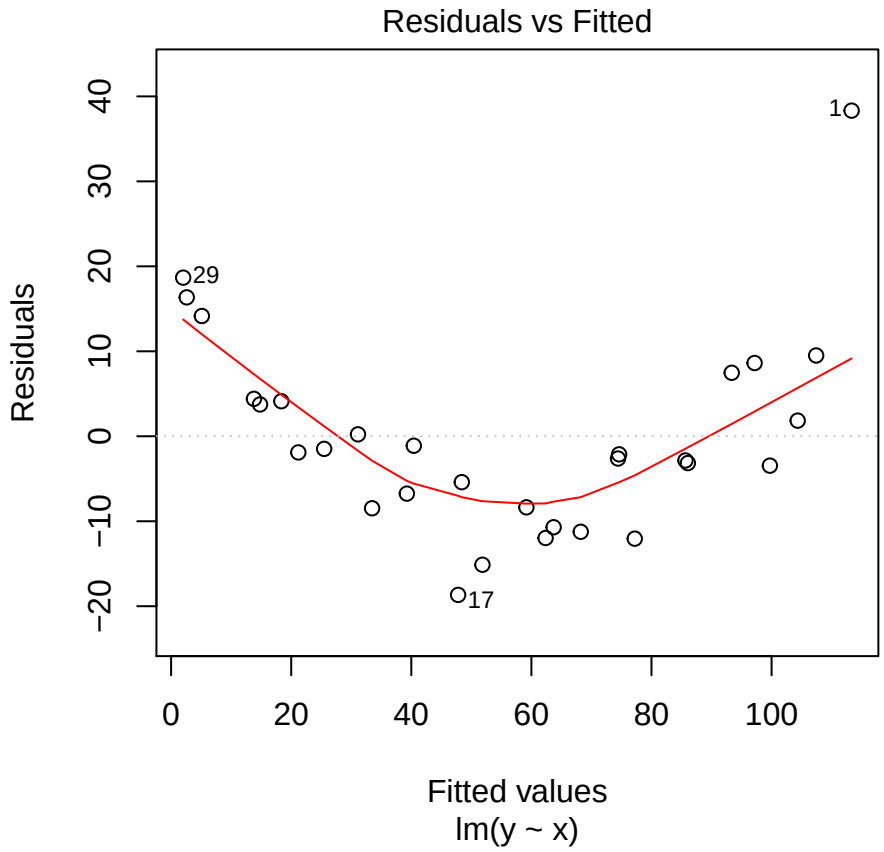


Figure 9: Residuals for the linear model fitted to the radioactive decay data.

```
## Residual standard error: 5.78 on 27 degrees of freedom
## Multiple R-squared:  0.976,Adjusted R-squared:  0.974
## F-statistic:  549 on 2 and 27 DF,  p-value: <2e-16
```

Note that the p-value for the quadratic term is highly significant, providing evidence that the relationship is non-linear. Also note that even though the function we are fitting is non-linear, it is still a linear model. If we replace the x^2 term by z, the relationship would be written $y = a + bx + cz$. Linear models do not necessarily have to involve a straight-line relationship between the response variable and the explanatory variables.

We can also use the `anova` function to compare the linear and quadratic models.

```
linear <- lm(y ~ x)
anova(quadratic, linear)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x + I(x^2)
```

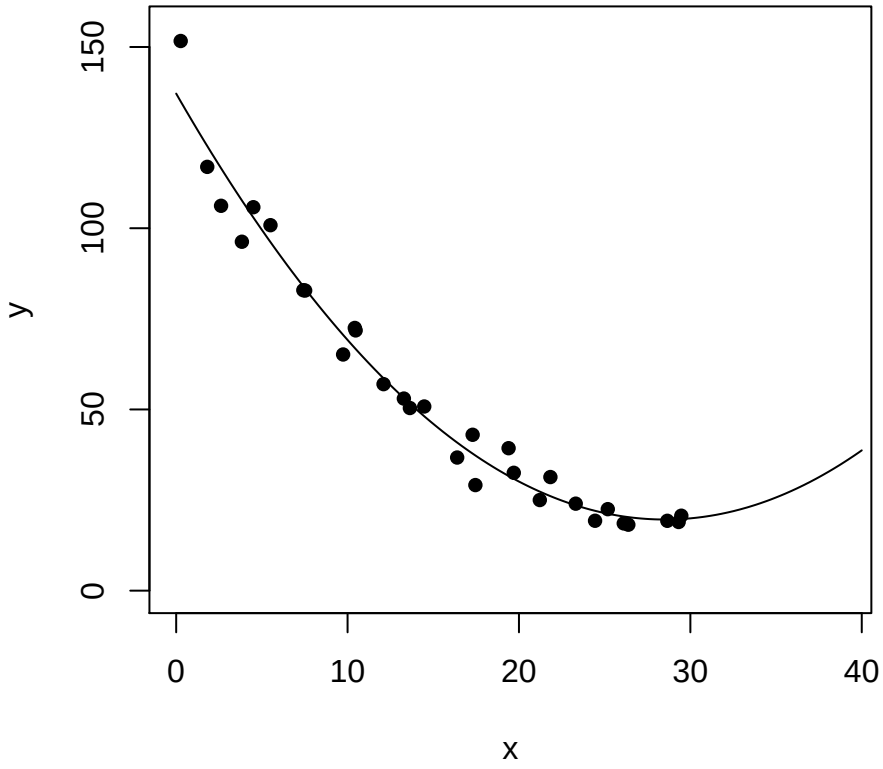


Figure 10: Fitting a quadratic model to the radioactive decay data set.

```
## Model 2: y ~ x
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     27  901
## 2     28 3931 -1     -3030 90.8  4e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the difference in the two models is the same as the significance of the quadratic term in the quadratic model, again indicating the improved fit.

However, the quadratic function is probably not the best model to fit to the decay data. Figure 10 shows the fit from the quadratic function, from which we can see that the fitted response values begin to increase beyond a value of about 30. This is not consistent with the notion that these are decay data.

4.2.2 Linearizing the model by transformation

For our radioactive decay data set, an exponential function of the form $y = ae^{-bx}$ might be more appropriate. This is an example where we can linearize the model by applying a transformation, in this case taking logarithms of both sides of the equation:

$$y = ae^{-bx} \tag{24}$$

$$\ln(y) = \ln(a) - bx \tag{25}$$

If we make $\ln(y)$ the response variable instead of y, we can use `lm` to fit an exponential curve:

```
exponential <- lm(log(y) ~ x)
summary(exponential)
```

```
##
## Call:
## lm(formula = log(y) ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29218 -0.07294  0.00668  0.09593  0.23848
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.93285    0.04391   112.3   <2e-16 ***
## x           -0.07269    0.00247   -29.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.119 on 28 degrees of freedom
## Multiple R-squared:  0.969,Adjusted R-squared:  0.968
## F-statistic:  866 on 1 and 28 DF,  p-value: <2e-16
```

Figure 11 compares the fitted quadratic and exponential models.

4.2.3 Multiple regression

Sometimes we may want to fit a model using more than one explanatory variable. Multiple linear regression involves fitting a linear model of the form:

$$y = a + bx_1 + cx_2 + dx_3 + ... \tag{26}$$

where x_1, x_2, x_3, etc. are our explanatory variables. The R code for fitting this model is:

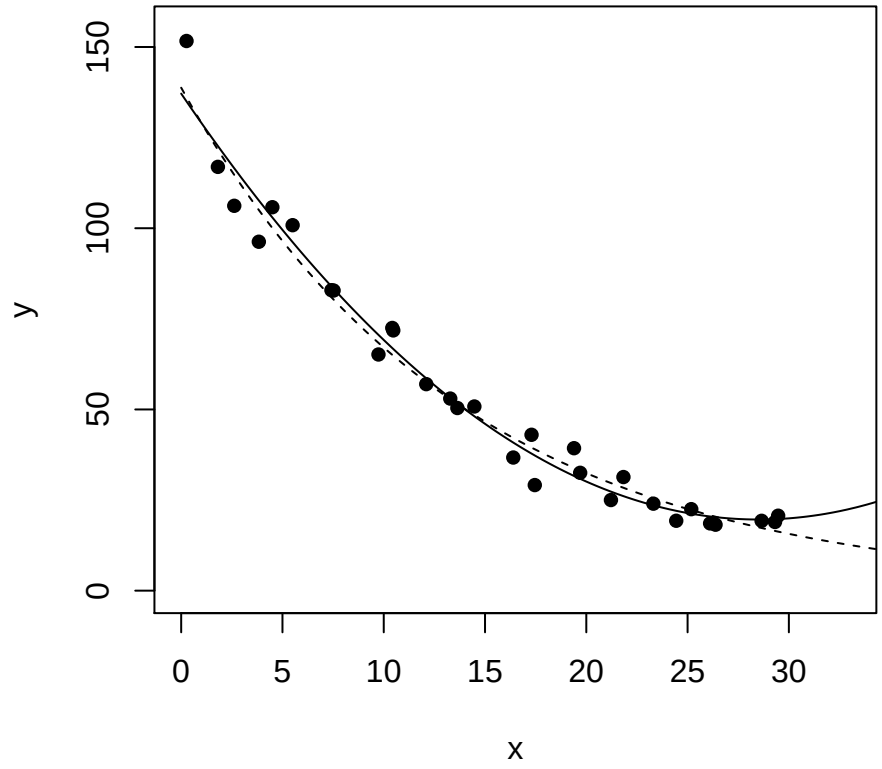


Figure 11: Fitting quadratic (solid line) and exponential (dotted line) models to the radioactive decay data set.

```
model <- lm(y ~ x1 + x2 + x3)
```

The quadratic function we fitted above to the radioactive decay data is an example of multiple linear regression.

Where there are multiple explanatory variables, interaction terms can be added to the model, for example

$$y = a + bx_1 + cx_2 + dx_1x_2 \tag{27}$$

This is specified in R using the ':' symbol:

```
model <- lm(y ~ x1 + x2 + x1:x2)
```

or more succinctly:

```
model <- lm(y ~ x1 * x2)
```

where the '*' symbol indicates inclusion of the explanatory variables and interactions, not multiplication.

Care must be taken not to overfit when creating a multiple regression model for a dataset containing many explanatory variables, particularly if interactions terms are used in constructing the model.

Models can be built up by starting from a simple model with few explanatory variables and adding additional explanatory variables, or we could simplify a very complicated model containing all explanatory variables and interaction terms by iteratively reducing non-signficant terms. These approaches are known as forward selection and backward selection respectively.

Another assumption of multiple regression is that the explanatory variables are not multicollinear. Multicollinearity occurs when two independent variables are highly correlated with each other. If the goal of building the model is prediction, multicollinearity is not that important. However, if the goal is to understand the factors that cause the variation in the response variable, this can lead to some confusion. For example, our model may suggest that an explanatory variable is the leading cause of the variation in the response variable but in reality it is not causative but highly correlated with another explanatory variable that is the actual cause.

4.2.4 Using nominal variables in a multiple regression

Nominal (or categorical) variables can be used in multiple regression. For example, a multiple regression model that attempts to predict blood pressure from explanatory variables such as height, weight, age and the number of hours of exercise per week, might also include sex. In this case, we would create a variable in which every female has a value of 0 and every male a value of 1.

Where there are multiple values of a nominal variable, or multiple categories, some care is required. The usual technique is to create $k-1$ dummy variables, where k is the number of values that the nominal variable can take. For example, if the blood pressure study included socio-economic status as a nominal variable with 5 values, we would create 4 dummy variables representing the first four statuses. For an observation in one of those 4 socio-economic classes, we would set the corresponding dummy variable to 1 and all others to 0. For an observation in the fifth class, we would set all dummy variables to 0.

4.2.5 Non-linear models

In some situations a particular mechanistic model will lend itself to the data. Where this takes the form of a non-linear equation that cannot be linearized by transformation of the response variable or the explanatory variable (or both), the `nls` function can be used in place of `lm`. Initial guesses for each of the parameters will have to be supplied. For example, fitting a non-linear function of the form, $y = a - be^{-cx}$, can be carried out as follows:

```
library(nls)
model <- nls(y ~ a - b * exp(-c * x), start = list(a = 50, b = 100, c = 0.05))
```

This is beyond the scope of what is covered in this introduction to regression analysis; see [5] for more details.

References

[1] Hadley Wickham. Tidy data. *Journal of Statistical Software*, VV. URL: http://vita.had.co.nz/papers/tidy-data.pdf.

[2] Hadley Wickham. Tidy data and tidy tools. URL: https://vimeo.com/33727555.

[3] DG Altman. *Practical Statistics for Medical Research*. Chapman and Hall, 1997.

[4] Rubin LJ and Peter RH. Oral hydralazine therapy for primary pulmonary hypertension. *N. Eng. J. Med*, 203:69–773, 1980.

[5] Michael J. Crawley. *Statistics: An Introduction using R*. Wiley, 2005.