Biometry 721/821 Workbook

Biometrical Applications and Data Analysis in R

Department of Genetics

Faculty of Agrisciences

Stellenbosch University

May 2024

Using this workbook

By working through this workbook, you will learn how to perform many different types of statistical analyses using *R*. *R* is a language and environment for statistical computing and graphics (*R-project.org*). In this workbook the title *R* will mostly refer to *RStudio*. *RStudio* is a user-interface for the *R* language.

Prerequisites

As this book predominantly explains the practical side of the biometrical tools, it is assumed that you have already passed at least a set of introductory statistics courses such as Biometry 212 and Biometry 242.

Review relevant introductory statistics notes before working through each chapter. The notes of the Biometry 212 and Biometry 242 modules should be sufficient for background material.
Additionally, the following book is *required* reading for the module:

- ➢ Zar JH. 2014. *Biostatistical Analysis*, Fifth Edition. Pearson Education Limited, UK.
    - o A free online version is available on the SUN library web service.
- ➢ Clewer AG and Scarisbrick DH. 2001. *Practical Statistics and Experimental Design for Plant and Crop Science*. West Sussex, UK: Wiley.

The software that will be used throughout this workbook is *R* (latest version of *R* during the writing of this document is version 4.3.3).

Topics

The following topics will be covered in this workbook:

1. Data handling, descriptive statistics
2. t-tests for: one population, two independent populations, two dependent populations.
3. Analysis of variance: Completely Randomised design, Randomised Block design, Latin Square design; with multiple comparisons
4. Analysis of variance: Factorial experiments, Split plot designs,; with multiple comparisons
5. Non-parametric tests: Wilcoxon rank-sum test, Wilcoxon sign rank-sum test, Kruskal-Wallis test, Friedman test
6. Simple linear regression; correlation

7. Polynomial regression
8. Multiple linear regression; with stepwise and all-subsets selection procedures
9. Analysis of covariance
10. Analysis of variance: repeated measures; with multiple comparisons
11. Analysis of categorical data: contingency table; χ^2 test of association; Logistic regression

Author: Dr T Pepler
Revised by: Mr CS van der Westhuizen, Miss C Bester
Unit for Biometry
Genetics Department
Stellenbosch University

May 2024

Chapter 1 Data handling and descriptive statistics

1. Introduction

In this chapter, you will learn how to import *txt* files into *R*, and how to use *R* to calculate various descriptive statistics, both numerical and graphical.
In the process, some elementary concepts will be revised.

Background reading material

- Biometry 212: note sets 2, 3 and 4
- Clewer and Scarisbrick: chapters 2 and 3

2. Online course page (SUNLearn)

Throughout the semester, the data sets necessary to complete the exercises and tests will be made available on the Biometry 721/821 online course page at:

http://learn.sun.ac.za/

Make sure you are able to access this page and its contents. It is important not to work directly with the data files on the course page. Instead, make your own copies of these copied files. The files can be found in the `Data` section.

3. Importing data into *R*

Make sure you are able to import data from *txt* (and Excel) into *R*. This is something you will need to do very often during the semester.
First, create your own copy of the file `chap1data.xlsx` available on the online course page, of which the data in the `Data1` sheet originally are from Simonoff (2003). Inspect the contents of the file. Then save it as a comma-delimited CSV file. We will use this data set for illustrational purposes in this chapter.

We will use three methods to import data into *R*:

Method 1 – importing a CSV file or other txt files (`read.csv` function)

The `read.csv` function reads data from a comma separated values (csv) file into *R*. An example of the function is displayed below. Please note that *R* uses a double backslash when stating the directory. Please note that read.csv is used when the seperators were commas. If the seperators were decimals, read.csv2 shouldbe used.

```
chap1data1 <-
read.csv(“C:\\Analysis\\chap1data1.csv”,sep=”,”,header=T)
```

For a detailed explanation of the `read.csv` function use the help function `?read.csv`.

Method 2 – copying data directly from Excel using the clipboard

The `read.table` function has a specific argument that allows one to copy / paste data from Excel directly into *R*. This method is very quick and convenient, but it has the disadvantage of only importing small data sets. Copy the data in Excel which you desire to import into *R* so that the section is highlighted. Then submit the following code in *R*.

```
chap1data1 <- read.table(file=“clipboard”,header=T)
```

For a detailed explanation of the `read.table` function use the help function `?read.table`.

Method 3 – importing data with *RStudio*

1. In *Rstudio*, in the environment console click on `Import Dataset`, and then click on `From Text File…`
2. Navigate towards and select your own copy of the CSV file and click `Open`.
3. Enter a name for the data set.
4. Select `Yes` if your data has headings in the first row.
5. If there is any missing data in the file, use `na.strings` to specify how these are indicated. (It is therefore a good idea to use the same missing data indicator consistently throughout the file.)
6. Select `strings as factors`.
7. Click `Import`.

After importing a data set into *R* the name of the data set should display in the environment console. To view the data set, click on the name of the data set. The data set should then appear in the editor.

4. Exporting data from *R*

We can use the `write.table` function to export data from *R* to a txt file. Below is an example of where an object, `results1` is exported to a CSV file.

```
write.table(results1,”C:\\Analysis\\Output\\results.csv”,
sep=”,”,quote=F,row.names=F)
```

For a detailed explanation of the `write.table` function use the help function `?write.table`.

5. Descriptive statistics (C&S chapters 2 and 3)

Make sure you can distinguish between the following *types of data*. Can you give examples of each?

- Qualitative (Categorical)
- Quantitative
- Discrete
- Continuous

Make sure you can distinguish between the following scales of data. Can you give examples of each?
For qualitative (categorical) data:

- Nominal
- Ordinal

For quantitative data:

- Interval
- Ratio

Make sure that you are able to calculate the following descriptive statistics using *R*, and understand how they are to be interpreted. Practise using appropriate columns of the imported `chap1data1.csv` data set. For example, calculate descriptive statistics for the variable `Age` (the 4th column).

Measures of *location*:	Description	*R* function
Minimum	smallest value	`min(chap1data1[,4])`
Maximum	largest value	`max(chap1data1[,4])`
Arithmetic mean	$\bar{x} = \frac{\sum x_i}{n}$	`mean(chap1data1[,4])`
Median	“middle” value	`median(chap1data1[,4])`
Mode	value that occurs with the highest frequency	modeest::mfv(chap1data1[,4])
First quartile (Q_1)	where 25% of the values are smaller	`quantile(chap1data1[,4],0.25)`
Third quartile (Q_3)	where 75% of the values are smaller	`quantile(chap1data1[,4],0.75)`

Measure of *variability* (dispersion)	Description	*R* function
Range	difference between maximum and minimum (very sensitive to outliers)	`range(chap1data1[,4])`
Interquartile range (IQR)	$Q_3 - Q_1$ (robust to outliers)	`quantile(chap1data1[,4],0.75) - quantile(chap1data1[,4],0.25)`
Variance	$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$	`var(chap1data1[,4])`
Standard deviation	$s = \sqrt{s^2}$	`sd(chap1data1[,4])`
Coefficient of variation	$CV = \frac{s}{\bar{x}} \times 100\%$	`sd(chap1data1[,4])/mean(chap1data1[,4])`
Standard error of the mean	$s_{\bar{x}} = \frac{s}{\sqrt{n}}$	`sd(chap1data1[,4])/sqrt(length(chap1data1[,4]))`

Other data summary measures	Description	*R* function
Sum	$\sum x_i$	`sum(chap1data1[,4])`
Count	n	`length(chap1data1[,4])`

Some of the descriptive statistics mentioned above can be obtained by using the following methods:

Method 1 – the `summary` function

We can use the `summary` function to calculate the descriptive statistics like the minimum, maximum, 1st and 3rd quartiles, median and the mean. To calculate these statistics submit the following code into the *R* console. Calculate the summary statistics for variable `Age` which is the 4th variable in the chap1data1 data set.

```
summary(chap1data$’Age’) or

summary(chap1data1[,4]) or summary(chap1data1[,”Age”])
```

Take note of the different methods of specifying what column you want to use shown above. You can use any of the methods. They will not all be show each time in the code below.

We can also calculate these summary statistics for all the variables in a single line of code by using the `sapply` function.

```
sapply(chap1data1,summary)
```

Method 2 – the `numSummary` function in the `RcmdrMisc` package

The `numSummary` function from the `RcmdrMisc` package calculates provides various numeric summaries and statistics such as the mean, standard deviation, standard error of the mean, IQR, coefficient of variation, skewness, kurtosis and quantiles. This function also provides calculations by groups.

```
numSummary(chap1data1[,"Age"])
```

Make sure you are able to construct the following types of graphs, and that you know for which types of data they are appropriate:

- Bar chart (for categorical or discrete numerical data)
- Line plot (similar to a bar chart, used to emphasize discreteness of the variable)
- Histogram (for continuous data)
- Pie chart (for categorical data measured on a nominal scale)
- Scatter plot (to inspect the relationship between two variables)
- Box plot (to inspect the distribution of a numerical variable measured on a continuous scale)

We will use two plotting systems in *R* for this course, 1) *base* plotting and 2) the *ggplot* system.

Construct the abovementioned list of graphs with *R base* plotting.

For a categorical variable, we need to het the frequencies of each category as well. The table function can be used for this, as shown below.

Bar chart – *Shoes*

```
barplot(table(chap1data1[,"Shoes"]),main="Type of
Shoes",ylab="Frequency",xlab="Type of Shoes",col=”green”)
```

Line plot – *Type.of.running*

```
barplot(table(chap1data1[,"Type.of.running"]),main="Type of
running",ylab="Frequency",xlab="Type of running",width=1,space=100)
```

Histogram – *Age*

```
hist(chap1data1[,"Age"],main="Histogram for Age",xlab="Age")
```

Pie chart – *Type.of.running*

```
pie(table(chap1data1[,"Type.of.running"]),main="Pie chart for types
of running")
```

Scatterplot – *Age* and *Miles.per.week*

```
plot(x=chap1data1[,"Age"],y=chap1data1[,"Miles.per.week"], main="Age
vs Miles per week",ylab="Miles per week",xlab="Age")
```

Box plot – *Miles.per.week*

```
boxplot(chap1data1[,"Miles.per.week"],main="Miles per week
Boxplot",ylab="Miles per week",xlab="")
```

Box plot by groups – *Miles.per.week* by *Shoes*

```
boxplot(Miles.per.week~Shoes,data=chap1data1,main="Miles per week by
Shoes Boxplot",ylab="Miles per week",xlab="Shoes")
```

To use the *ggplot* plotting system you need to install the package first and load it into *R*.

Install and load *ggplot*

1. In the miscellaneous window, click on *Packages*.
2. Click on Install, then type "ggplot2" beneath Packages (separate multiple with space or comma):
3. Make sure *Install dependencies* is selected, then click on *Install*.
4. After installation, load the package into *R*: `package(ggplot2)`

Construct the same graphs (except the line plot) with *ggplot* plotting system.

Bar plot – *Shoes*

```
barplot <- ggplot(data=chap1data1,aes(x=Shoes)) +
geom_bar(fill="gray",color="black",width=.5)
```

```
barplot + ggtitle("Barplot of Types of Shoes") + labs(x="Types of
Shoes",y="Frequency") +
theme(plot.title=element_text(size=20),axis.title=element_text(size=
14))
```

Bar plot by groups – *Shoes* by *Male*

```
barplot2 <- ggplot(data=chap1data1,aes(x=Shoes,y=Miles.per.week))
```

```
barplot2 + geom_bar(aes(fill=Male),
position=position_dodge(),stat="identity")  +
scale_fill_manual(values=c("#CC6666","#66CC99"))
```

Histogram – *Age*

```
histo <- ggplot(chap1data1,aes(x=Age))
histo + geom_histogram(binwidth=6)
```

Pie chart – *Shoes*

```
pie    <-    ggplot(data=chap1data1,aes(x=factor(1),fill=Shoes))    +
geom_bar(width = 1)
pie + coord_polar("y")
```

Scatter plot – *Runs.per.week* vs *Age*

```
ggplot(data=chap1data1,aes(y=Runs.per.week,x=Age)) +
geom_point(shape=3,color="darkblue")
```

Scatter plot – Runs.per.week vs Age by Distance

```
scatter <-
ggplot(data=chap1data1,aes(y=Runs.per.week,x=Age,color=Distance))
scatter + geom_point(size=2) +
scale_color_manual(values=c("#CC6666","#66CC99"))
```

Scatter plot – *Runs.per.week* vs *Age* by *Distance* (with trend lines)

```
scatter2 <-
ggplot(data=chap1data1,aes(y=Runs.per.week,x=Age,color=Distance))
scatter2 + geom_point(size=2) +
scale_color_manual(values=c("#CC6666","#66CC99")) +
geom_smooth(method=lm,se=F)
```

Box plot – Miles.per.week by Distance

```
ggplot(chap1data1,aes(factor(Distance),Miles.per.week)) +
geom_boxplot()
```

6. Exercises

Exercise 1

For the data in C&S Example 2.1 (p. 11) (`chap1exer1.csv`), calculate the following statistics regarding the sample plant height. See if you can import the data with the following function:
matrix(c(14.8,15.2,17.4,11.6,12.5),ncol=1)

- Mean
- Median
- Variance
- Standard deviation
- Standard error of the mean
- Coefficient of variation

Exercise 2

See if you can import the data below with the following function:

data.frame(Yield=c(8.1,8.7,9.2,7.8,8.4,9.4))

- report the mean, median and mode.
- report the variance and standard deviation
- produce a line plot

Exercise 3

The data in C&S Example 3.2 (p. 18) (`chap1exer3.csv`), presents the yields (g) of small equal-sized plots of barley. Determine the following:

- mean
- mimimum and maximum
- standard deviation
- range
- first and third quartiles
- interquartile range

and also

- produce a histogram
- produce a box plot.

Chapter 2 - t-tests

1. Introduction

In this chapter, you will learn how to use the t-test to:

- compare the mean of a *single* population to a fixed value
- compare means of two *independent* populations with one another
- compare means of two *dependent* populations with one another

We will also touch upon the F-test for equal variance, and various tests of normality. Make sure you understand each of the following concepts:

- Population
- Sample
- Hypotheses:
- Null hypothesis
- Alternative hypothesis
- Two-sided alternative hypothesis
- Left one-sided alternative hypothesis
- Right one-sided alternative hypothesis
- Test statistic
- Significance level
- Critical value
- p-value
- Rejection of the null hypothesis
- Non-rejection of the null hypothesis
- Confidence interval

Note that we *always* make our conclusion in terms of the null hypothesis, i.e.

- Reject the null hypothesis, or
- Do not reject the null hypothesis.

Do not refer to "acceptance" of any hypothesis – a hypothesis can never be proven true; the only conclusion that can be made is that the hypothesis is either plausible or implausible based on the statistical evidence.

Background reading material

- Biometry 212: note set 16
- Biometry 242: note sets 1 and 3
- Clewer and Scarisbrick: chapters 5 and 6

2. One sample *t*-test (C&S 5.2, p. 39)

Import into *R* the `chap2data1.csv` file. The data represent 15-day comb masses of male chicks that each received a certain sex hormone. Use *R* to test the following hypotheses:

- That the true population mean comb mass is equal to 120g vs. the (two-sided) alternative that the true population mean is not equal to 120g.
- That the 15-day comb masses of chicks given sex hormone A are on average equal to 90g vs. the (one-sided right) alternative that they are greater than 90g.
- That the average 15-day comb mass for chicks on this hormone is equal to 100g vs. the (one-sided left) alternative that they are less than 100g.

One sample *t*-test:

```
t.test(x, y = NULL,
        alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

The *t*-test is based upon two assumptions. Therefore, before you can believe the results of a *t*-test, you have to make sure that the assumptions are tenable for the particular data set. The one-sample *t*-test assumes:

- Normally-distributed data
- Independent observations

Proper *randomisation* in the sampling procedure should ensure independence of the observations.

To test the normality assumption we may use the Shapiro-Wilk test. In addition, histograms and quantile-quantile (Q-Q) plots can be drawn as rough indications of the level of normality, or the lack of it.

Shapiro-Wilk test

```
shapiro.test(x)
```

Where *x* is the single variable which you want to test.

QQ plot

```
qqPlot(chap2data1[,"Comb.mass"],main="QQ plot",ylab="Comb
mass",xlab="Quantiles")
```

Investigate the assumption of normality for the *Comb mass* data.

Remember, tests and graphs become more accurate the more data we have, all else being equal.

If the data are not normal, the data can be transformed, or a *non-parametric* test can be performed instead of a *t*-test. This will be discussed in detail in a later chapter.

3. Two independent samples *t*-test (C&S 6.2, p. 51)

Import into *R* the file `chap2data2.csv`.

The data represent 15-day comb masses of eleven male chicks that each received sex hormone A, and 15-day comb masses of eleven other male chicks that each received sex hormone B. The chicks were randomly assigned to receive one of the two sex hormones.

When the true *means* of two populations are to be compared using a *t*-test, it is first necessary to determine whether we can assume equality of the population *variances* or not. Depending on whether this assumption of *homoscedasticity* can be made, we will use either a *pooled t-test* or a t-test for *unequal variances* to compare the true means. To test for homoscedasticity of two populations we make use of the *F*-test (C&S 6.5, p. 58).

Test the following hypotheses about the population mean comb mass using *R*:

- That the mean comb mass for hormone A chicks is the same as that for hormone B chicks.
- That the mean comb mass for hormone A chicks is 20g greater than that for hormone B chicks.
- That the mean comb mass for hormone B chicks is 12g less than that for hormone A chicks vs. the alternative that the mean mass difference is *greater* than 12g.

F-test for homoscedasticity

```
with(chap2data2, tapply(Comb.mass, Hormone,  var, na.rm=TRUE))
var.test(Comb.mass~Hormone,alternative="two.sided",data=chap2data2)
```

t-test for two independent samples

```
t.test(x, y,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
   conf.level = 0.95)
```

The assumptions underlying this *t*-test are the same as those of the one-sample *t*-test: normally-distributed data and independent observations. In fact, the data from both samples *separately* must be normally distributed. These assumptions can be tested in the same way as for the one-sample case. Do so using the imported two-hormone data.

Checking normality assumption (two or more groups simultaneously):

```
shapiro.test(chap2data2$Comb_mass[chap2data2$Hormone == "A"])
shapiro.test(chap2data2$Comb_mass[chap2data2$Hormone == "B"])
```

Test the normality of the hormone treatment groups separately for the comb mass data. As before, if the data are not normally distributed, a *non-parametric* test must be performed instead of a *t*-test.

If the data consist of paired observations, consider a *t*-test for *dependent* samples, explained in the next section.

4. Two dependent samples *t*-test (C&S 5.8, p. 46)

Often, the data from two samples are not unrelated but *paired*. For example, ten people are about to embark on a diet: they are weighed *before* and again *after*. The *same* people are weighed twice.
This is done either by

1 telling the software to perform a *paired t-test*, or
2 manually calculating a column of differences in the paired values and then performing a one-sample *t*-test on this new column.

Import the *before* and *after* masses from the file `chap2data3.csv`. Note the format of the data. Now test the following hypotheses:

- That the diet has no effect (or is ineffective).

- That on average the diet leads to weight-loss of 10kg.
- That on average the diet leads to weight-loss of 12kg vs. the alternative that the average weight-loss is less than 12kg.

Paired samples *t*-test:

```
t.test(x, y,
        alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = TRUE, var.equal = FALSE, conf.level = 0.95)
```

This test is based on the assumption that the *paired differences* are normally distributed and independent of one another. The assumption of normality can be tested in the same way as in the one sample case, but based on a column of differences in the paired values. Use *R* to check whether the paired differences of the weight loss data came from a normally distributed population.

Once again, if the assumption of normality is not valid, the differences can be transformed, or a *non-parametric* test can be performed instead of a *t*-test.

5. Power analysis

Consider the following definitions:

Type I error – rejecting the null hypothesis when it is true (false positive)

Type II error – not rejecting the null hypothesis when it is false (false negative)

Power – the probability to not commit a type II error.

A power analysis investigates and uses the relationship between power and sample size to determine the optimal sample size for your experiment. This relationship is investigated under various circumstances of natural variation, level of significance and effect sizes (smallest detectable differences).

2.5.1 Learning by example

Consider again the data in the file `chap2data3.csv`. Let’s investigate the relationship between power and sample size for the comb.mass variable given that the estimate of the natural variation is the pooled variance, the level of significance is equal to 5% and that we want to detect a difference between the two hormones as small as 35g. Calculate the power for sample sizes of 5, 7, 9, 11, 13 and 15. Visualize the results with a scatter plot. Investigate this relationship for changes in the smallest detectable difference (delta) values. Use 20g, 35g and 50g in the analysis. What is the effect of changing the sample size and the smallest detectable difference on power?

```
#estimate the pooled variance
hA <- chap2data2[chap2data2$Hormone=="A",][,1]
hB <- chap2data2[chap2data2$Hormone=="B",][,1]
SSA <- sum(hA^2)-sum(hA)^2/length(hA)
SSB <- sum(hB^2)-sum(hB)^2/length(hB)
sp2 <- (SSA+SSB)/(length(hA)+length(hB)-2)
#OR
sp2 <- anova(lm(Comb.mass~Hormone,data=chap2data2))[2,3]
#Example, for n=11
power.t.test(power=NULL,n=11,sd=sp2^.5,delta=35)
#Extract only the value for power
power.t.test(power=NULL,n=11,sd=sp2^.5,delta=35)$power
#Tabulate all results
#set delta
delta <- 35
pwrTable <- data.frame(sample_size=c(5,7,9,11,13,15),

pwr=c(power.t.test(power=NULL,n=5,sd=sp2^.5,delta= delta)$power,

power.t.test(power=NULL,n=7,sd=sp2^.5,delta= delta)$power,

power.t.test(power=NULL,n=9,sd=sp2^.5,delta= delta)$power,

power.t.test(power=NULL,n=11,sd=sp2^.5,delta= delta)$power,

power.t.test(power=NULL,n=13,sd=sp2^.5,delta= delta)$power,

power.t.test(power=NULL,n=15,sd=sp2^.5,delta= delta)$power))
#plot results
plot(pwrTable)
```

6. Exercises

Exercise 1

With the data from C&S Example 5.1 (p. 41) (`chap2exer1.csv`), test if the mean linseed yield is equal to 2.0. Use a 5% significance level and clearly state your conclusion.

Exercise 2

Do C&S Exercise 5.1 (p. 45) (`chap2exer2.csv)`. Test if the population mean is different from 8. Carry out the appropriate test, starting (and testing) any assumptions you make.

Exercise 3

With the data from C&S Table 5.2 (p. 48) (`chap2exer3.csv`)
Perform a paired *t*-test to see if the mean yield from the two wheat varieties differ. Use a 5% significance level. Calculate the sample differences between the two varieties (per farm), and perform a test to see if the mean difference is equal to zero.

Exercise 4

Use *R* to do C&S Exercise 6.2 (p. 61) (`chap2exer4.csv`). That is, test if Varieties A and B are significantly different. Data are the yields (in t/ha).

Chapter 3

1. Introduction

In this chapter, you will learn how to perform an Analysis of variance (ANOVA) for three different experimental designs:

- the Completely Randomised (CR) design,
- the Randomised Block (RB) design, and
- the Latin Square (LS) design.

Although ANOVA refers to the analysis of variance, it in fact entails a comparison of means. However, the level of variation around the means (i.e. the *within-group* variance) is used during the comparison.

The following terminology is important:

- **response** (BMT 242 NS 6 section 1): what is measured, also often called the *yield*
- **treatment** (BMT 242 NS 6 section 1): the treatment protocol/substance each of the experimental units are subjected to
- **experimental design** (BMT 242 NS 6 section 1): how the experimental units are assigned to the different treatments
- **experimental unit** (BMT 242 NS 5 section 2): the smallest bit of experimental material which can receive a *single* treatment
- **replications** (C&S 9.9, p. 128; BMT 242 NS 10): the number of experimental units to which each of the treatments are applied
- **pseudo-replication** (BMT 242 NS 10): treating duplicate measurements taken on single experimental units as if they are replications
- **repeated measure** (C&S 16.7, p. 252; BMT 242 NS 10): measuring the same response on the same experimental units repeatedly over time

As you work through the chapter, make sure that you

- understand and are able to recognise each of the experimental designs;
- know how to set up a field plan for each of the designs;
- know in each case what the primary question is that ANOVA seeks to answer;
- know how to properly interpret the ANOVA output and, if necessary, how to follow it up with multiple comparison tests and contrast-based hypothesis tests;
- are aware of the underlying assumptions, and know how to test them.

Background reading material

Biometry 242: note sets 6, 7, 8, 9 and 10

Clewer and Scarisbrick: chapters 9, 10, 11 and 13

2. Completely Randomised design (C&S chapter 9, p. 102)

In a Completely Randomised design, the n experimental units are completely randomly assigned to $g \geq 2$ treatments. If each treatment is applied the same number of times ($n_1 = n_2 = \cdots = n_g = \frac{N}{g}$), the design is said to be *balanced*.

Investigate how to use *Excel* to lay out the field plan for a Completely Randomised design with $n = 24$ experimental units and $g = 4$ treatments (C&S 9.2, p. 103).

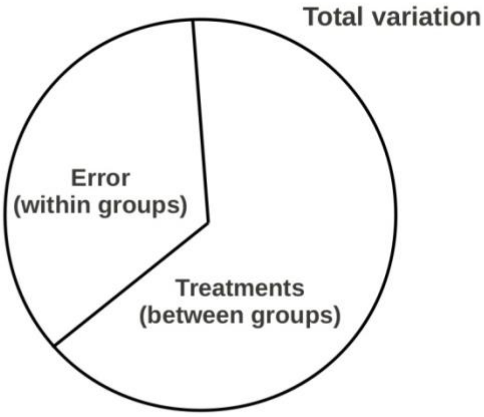

Figure 3.1: Sources of variation in a Completely Randomised experimental design.

An ANOVA of the Completely Randomised design allows us to determine statistically whether all g treatments have an equivalent effect, on average, or not.

The responses are modelled as follows (C&S 9.4, p. 108):

$$Y = \mu_i + \epsilon \tag{3.1}$$

For $i = 1, \ldots, g$, with

- Y: the response measured on the experimental plot;
- μ_i: the mean response of the i^{th} treatment;
- ϵ: the unexplained "error"

Under this formulation, the hypothesis of main interest (i.e. whether the population treatment means differ significantly) is:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$$

$$H_1 : \mu_1 ; \mu_2 ; \ldots ; \mu_g \textit{ are not all equal}$$

Alternatively, but equivalently, the model in Equation 3.1 can be rewritten as:

$$Y = \mu + \alpha_i + \epsilon \qquad (3.2)$$

- Y: the response measured on the experimental plot;
- μ: the overall mean;
- α_i: the mean effect of the i^{th} treatment (its deviation from the overall mean);
- ϵ: the unexplained “error".

Under this formulation, the hypothesis of main interest (i.e. whether the population treatment means differ significantly) is:

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_g = 0$

$H_1: \alpha_1, \alpha_2, \ldots \alpha_g$ are not all equal to 0.

The preceding hypotheses can be tested by interpreting the appropriate F-ratio and corresponding p-value from an ANOVA table obtained from *R* (C&S 9.5.3, p. 114).

Perform an one-way ANOVA

```
<model name> <- summary(aov(<response variable>~<treatment
variable>,data=<data set name>))
```

Import into *R* the file `chap3data1.csv`.

The data indicate the amount of fat absorbed whilst doughnuts are prepared using one of four types of fat. Note the “flattened” format of the data.

Perform an ANOVA on the data. Clearly state the hypothesis that gets tested, and interpret the results of the hypothesis test. Ignore other aspects at this stage. Note that *R* imports the *Oil type* variable as a numeric type variable by default. This variable should first be converted to a factor type variable before the ANOVA can be performed.

The results from an ANOVA are valid only under the following assumptions (C&S 14.1, p.213)

- the errors (i.e. residuals from the linear model in Equation 3.1) are normally distributed;
- the errors have equal variance for the different treatments (homoscedasticity);
- the errors are independent of one another.

Normality: To test for *normality*, each treatment's observations can be tested separately as in previous assignments using the Shapiro-Wilk test. Perform this test for normality as described in Chapter 2.

Alternatively, the residuals from the fitted linear model can be tested for normality. This is the preferred way, as the normality assumption concerns the normality of the model residuals. All of the residuals together also constitutes as larger sample on which to perform the normality test, leading to greater power for the test. (i.e. deviations from normality will be detected more easily).

Shapiro- Wilk test for normality on the residuals

```
<data set name>$residuals <- <model name>$residuals

shapiro.test(<data set name>$residuals)
```

We can also make use of the diagnostic plots to investigate the normality of the data.

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

plot(<model name>)
```

Inspect the *Residuals vs Fitted* plot for deviations from homoscedasticity. Inspect the *Normal Q-Q* plot for deviations from normality.

Homoscedasticity: To test for *homoscedasticity*, Bartlett's or Levene's tests can be used. Bartlett's test is sensitive to deviations from normality.

Bartlett's test for homoscedasticity

```
bartlett.test(<response variable>~ <treatment variable>,
data=chap3data1)
```

Levene's test for homoscedasticity

```
leveneTest(<response variable>~ <treatment variable>, data=<data set
   name>, center="median")
```

Independence. The test for *independence* is especially important when the observations are recorded in time-sequence. It can be tested graphically by plotting the residuals (observed errors) in the order in which the observations were recorded. No obvious pattern should be discernible.

Check whether the first two assumptions (normality and homoscedasticity) of the fitted linear model for the doughnut data are valid.

When the null hypothesis of equal treatment effects is *rejected*, we only know that all the treatments do not have the same effect. However, we do not know exactly which treatments differ from which others. To determine this, we have to perform additional *multiple comparison tests*.

Important: the results from multiple comparison tests *should only be interpreted if the null hypothesis for the treatments has been rejected.*

There is a more general way to perform ANOVA, using the `lm()` and `Anova()` functions in *R*. It is the only way to perform ANOVA for experiments more complicated than the Completely Randomised design.

General way to perform ANOVA

```
<model name> <- lm(<response variable> ~ <treatment variable>,
   data=<data set name>)

summary(<model name>)

Anova(<model name>, type="II")
```

The multiple comparison procedures of *Tukey* is designed to keep the overall significance level α^* fixed at a certain level, say 0.05, for *all possible contrasts simultaneously*. It is therefore a good method to use when there are a large number of comparisons to be made – in such cases the risk of a Type I error (i.e. incorrectly rejecting the null hypothesis) with methods not compensating for multiple testing (such as Fisher's LSD procedure) becomes very high.

For more information, see Biometry 242, note sets 6 and 7, or Clewer and Scarisbrick chapter 9.

3. Randomised Block design (C&S chapter 10, p. 132)

In a Randomised Block design, the n experimental units are *randomly assigned* within each of b blocks to $g \geq 2$ treatments. Compared to a Completely Randomised design, the randomisation is more restricted.

The advantage of the design is that more accurate comparisons can be made between the g treatments, all else being equal (C&S 10.1.2, p. 133).

Investigate how to use *Excel* to lay out the field plan for a Randomised Block design with $n = 24$ experimental units, $g = 4$ treatments and $b =$ 6 blocks (C&S 10.1.1, p. 133).

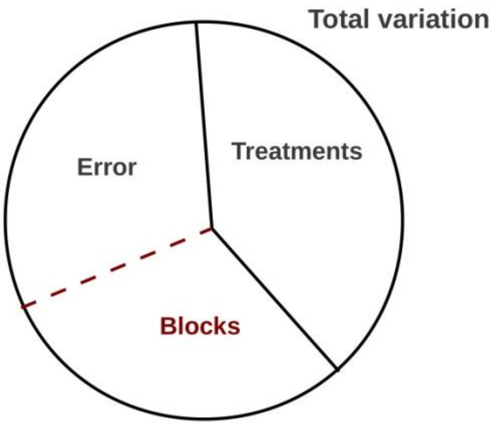

Figure 3.2: Sources of variation in a Randomised Block experimental design.

An ANOVA of the Randomised Block design allows us to determine statistically whether all g treatments have an equivalent average effect or not, whilst taking into account potential differences amongst the blocks. The responses are modelled as follows (C&S 10.6, p. 138)

$$Y = \mu + \alpha_i + \beta_j + \epsilon$$

for $i = 1, \dots, g$ and $j = 1, \dots, b$, with

- Y: the response measured on the experimental plot;
- μ: the overall mean;
- α_i: the mean effect of the i^{th} treatment;
- β_j: the mean effect of the j^{th} block;
- ϵ: the unexplained error variation.

The hypothesis of main interest is the same as before:

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_g = 0$

$H_1: \alpha_1, \alpha_2, \dots \alpha_g$ are not all equal to 0.

The preceding hypothesis can be tested by looking up the appropriate F-test p-value from an ANOVA table obtained from *R*.

Import the file `chap3data2.csv` into *R*.

> *Five different seed treatments (Check, Arason, Spergon, Semesan and Fermate) were each applied to a number of soybean seeds. The seeds were planted on experimental plots in five blocks (the blocks were formed based on elevation). The response of interest is the emergence rate of the soybeans.*

Perform and ANOVA for the Randomised Block design.

ANOVA for Randomised Block design

```
<model name> <- lm(<response variable> ~ <blocking variable> +
   <treatment variable>, data=<data set name>)
summary(<model name>)
Anova(<model name>, type="II")
```

To investigate the assumptions follow the same procedure as discussed earlier in the chapter.

For more information, see Biometry 242 note set 8, or Clewer and Scarisbrick chapter 10.

4. Latin Square design (C&S chapter 11, p. 149)

The Latin Square design is an efficient way of double blocking: $n = g^2$ experimental units are *randomly* assigned so that each treatment appears exactly once within each of g row blocks and exactly once within each of g column blocks. The randomisation used in a Latin Square design is more restrictive than that used in a regular Randomised Block design (C&S 11.1, p. 149).

The advantage of the design is its efficiency (a relatively small number of observations is required). A disadvantage, however, is that the number of treatments, the number of row blocks, and the number of column blocks must all be equal (C&S 11.1.1, p. 150).

Investigate how to use *Excel* to lay out the field plan for a Latin Square design starting from the *standard design* for $g = 4$ treatments (C&S 11.2, p. 151).

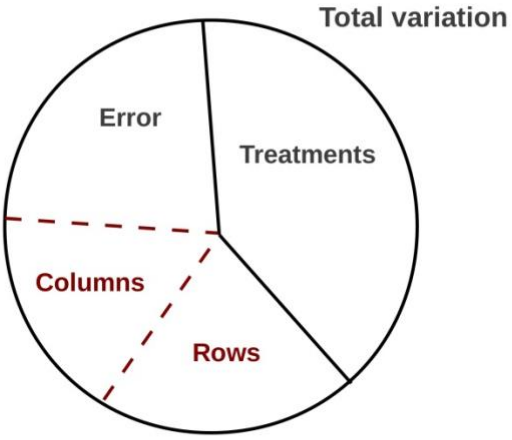

Figure 3.3: Sources of variation in a Latin Square experimental design.

An ANOVA of the Latin Square design allows us to determine statistically whether all g treatments have an equivalent average effect or not, whilst taking into account potential differences amongst the two sets of blocks. The responses are modelled as follows (C&S 11.4, p. 155):

$$Y = \mu + r_j + c_k + \alpha_i + \epsilon$$

for $i = 1, \dots, g, j = 1, \dots, g$ and $k = 1, \dots, g,$ with

- Y: the response measured on the experimental plot;
- μ: the overall mean;
- r_j: the mean effect of the j^{th} row block;
- c_k: the mean effect of the k^{th} column block;
- α_i: the mean effect of the i^{th} treatment;
- ϵ: the unexplained error variation.

The hypothesis of main interest is the same as before:

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_g = 0$

$H_1: \alpha_1, \alpha_2, \dots \alpha_g$ are not all equal to 0.

The preceding hypothesis can be tested by looking up the appropriate F-test p-value from an ANOVA table obtained from *R*.

Import the file `chap3data3.csv` into *R*.

> *Four types of wheat were each planted on one plot in each of four different rows. Four treatments (A, B, C and D) were applied to the plots in a Latin Square design. The wheat yield per plot was measured at harvest time.*

Perform an ANOVA for the Latin Square design. Except for the model specification, all other details are the same as for the Randomised Block design.

ANOVA for Latin Square design

```
<model name> <- lm(<response variable> ~ <row variable> + <column
   variable> + <treatment variable>, data=<data set name>)

summary(<model name>)

Anova(<model name>, type="II")
```

Test the hypothesis of interest, and perform multiple comparisons and contrast-based hypothesis tests if appropriate.

For more, see Biometry 242 note set 9, or Clewer & Scarisbrick chapter 11.

5. Exercises

Exercise 1

With the data from C&S Table 9.4 (p. 107) (`chap3exer1.csv`), perform an ANOVA to test if the yields from the four wheat varieties differ. If so, follow up with multiple comparison tests to find which of the varieties differ from one another. Clearly state your conclusions and also check the assumptions of the linear model.

Exercise 2

Do C&S Exercise 10.1 (p. 147) (`chap3exer2.csv`). An experiment was carried out to compare the effects of three cultivation systems (A, B and C) with traditional seedbed preparation (D) on the total dry matter yield of kale. The trial was laid out in five randomised blocks. Plot yields in t/ha are provided in the csv file. Follow the same instructions as with Exercise 1.

Exercise 3

Do C&S Exercise 11.1 (p 157) (`chap3exer3`.csv). An experiment was designed to test the effects of five levels of phosphate fertiliser (P1, P2, P3, P4 and P5) on the yield of potato tuber dry matter. Yields (kg DM/plot) is presented in the csv file. Follow the same instructions as with Exercise 1.

Chapter 4 - ANOVA for factorial designs

1. Introduction

In this chapter, you will learn how to perform an ANOVA for *factorial designs*, including *split plot designs*.

Make sure to distinguish between the *experimental* design (such as the Completely Randomised, Randomised Block or Latin Square designs) and the *treatment* design (i.e. a *simple* treatment design as was used in the previous chapters, or a *factorial* treatment design as will be discussed in this chapter).

The following terminology is important (C&S 12.3, p. 163):

- main effect
- interaction

As you work through the chapter, make sure that you

- understand and are able to recognise a factorial design;
- know how to set up a field plan for a factorial design;
- know what the questions are that an ANOVA for a factorial design seeks to answer;
- know how to properly interpret the ANOVA output and, if necessary, how to follow it up with multiple comparison tests and contrast-based hypothesis tests;
- are aware of the assumptions of the linear models, and know how to check the validity thereof.

Background reading material

- Biometry 242: note sets 11 and 12
- Clewer and Scarisbrick: chapters 12, 13 and 16

2. Two-way factorial designs (C&S 12.5, p. 166)

In a *two-way* factorial design, the n experimental units are randomly assigned to *combinations* of treatments from two factors: say, a treatments from Factor A, and b treatments from Factor B. If all treatment combinations are applied the same number of times, the design is said to be *balanced*.

Usually the experimental units are completely randomly assigned to the different combinations of treatments. However, it is possible to incorporate blocks.

Investigate how to use *Excel* to lay out the field plan for a two-way factorial design with $n = 24$ experimental units and $a = 4$ and $b = 3$ treatments from two factors, A and B, respectively. Randomise completely.

Consider the data contained in the file `chap4data1.csv`:

> *As part of an investigation of toxic agents, 48 rats were assigned to receive 3 poisons (I, II, III) and 4 treatments (i, ii, iii, iv). The response was the reciprocal of the survival time in tens of*

hours (in other words, 1 divided by the survival time in tens of hours). Therefore the smaller the response, the longer the time the rat lived.

In an ANOVA for a two-way factorial design, the following hypotheses are tested:

1. that there is *interaction* between the treatments of Factor A and the treatments of Factor B;
2. in the absence of interaction, that the treatments of Factor A differ from one another;
3. in the absence of interaction, that the treatments of Factor B differ from one another.

Note the numbering above: The hypothesis of interaction is *always* tested first. *Only if there is no significant interaction* can the other two hypothesis tests be performed, in any order.

The model can formally be specified as follows (C&S 12.5.6, p. 172):

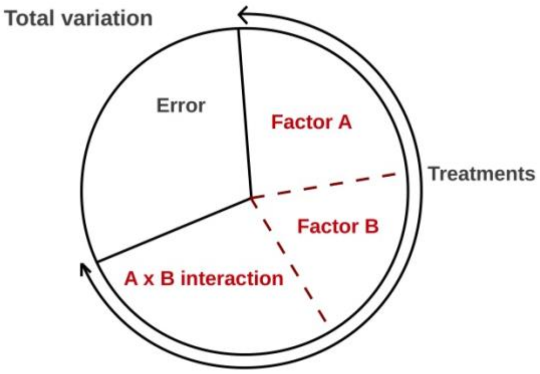

Figure 4.1: Sources of variation in a Completely Randomised factorial design.

$$Y = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon \qquad (4.1)$$

for $i = 1, \ldots, a$ and $j = 1, \ldots, b$, with

- Y: the response measured on the experimental plot;
- μ: the overall mean;
- α_i: the main effect of the i^{th}level of Factor A;
- β_j: the main effect of the j^{th}level of Factor B;
- $(\alpha\beta)_{ij}$:the interaction between the i^{th}level of Factor A and the j^{th}level of Factor B;
- ε: the unexplained error.

In this notation, the three hypothesis tests are:

1. Test for interaction:

$H_0: (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{ab} = 0$

H_1: At least one $(\alpha\beta)_{ij}$ is not equal to 0.

2. Test for Factor A main effects:

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$

$H_1: \alpha_1, \alpha_2, \ldots \alpha_a$ are not all equal to 0. (at least one α_i is not equal to 0)

3. Test for Factor B main effects:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_b = 0$

$H_1: \beta_1, \beta_2, \ldots \beta_b$ are not all equal to 0. (at least one β_j is not equal to 0)

The preceding hypotheses can be tested by looking up the appropriate F-test p-values from an ANOVA table, in the appropriate order.

The ANOVA table can be obtained from R using the following code:

ANOVA for factorial design

```
<model name> <- lm(<response variable> ~ <treatment variable 1> +
   <treatment variable 2> + < treatment variable 1>*<treatment
   variable 2>, data=<data set name>)
summary(<model name>)
Anova(<model name>, type="II")
```

Test the appropriate hypotheses and interpret the results.

Assumptions (C&S 14.1, p. 213): The ANOVA results are only valid if the usual assumptions are tenable: normality, homoscedasticity and independence. Check the diagnostic plots for the fitted linear model to determine whether the normality and homoscedasticity assumptions are justified.

If the interaction is not statistically significant but one (or both) of the factors is significant, perform post-hoc multiple comparison tests for the factor(s).

But what if there is significant interaction? Then we cannot interpret the main effects in isolation of each other. In such a case, if $a < b$, perform separate one-way ANOVAs of Factor B for each of the levels of Factor A. Otherwise, if $a > b$, perform *separate* one-way ANOVAs of Factor A for each of the levels of Factor B. If $a = b$, choose either of these two routes. When we interpret these separate one-way ANOVAs, we need to clearly state in each case for which specific treatment (of the other factor) the results are true.

If the number of treatment combinations is not too large, another option is to perform a one-way ANOVA of the *treatment combinations*, thus considering the levels of Factor A and the levels of Factor B together. In this case there will be $a \times b$ treatment combinations to compare. If the null hypothesis for the treatment combinations is rejected in the ANOVA, follow up with post-hoc multiple comparison tests to determine which treatment combination(s) can be used to obtain optimal yields.

Consider the data contained in the file `chap4data.csv`:

> *In all, four levels of protein and four kinds of fibre are fed to cows for a certain period. Each combination of protein and fibre is fed to three cows only, randomly selected from 48. Each cow's mass gain in kilograms is recorded over the period.*

Analyse the data fully using *R*.

3. Three-way factorial designs (C&S 12.7.1, p. 177)

Three-way factorials are similar to two-way factorials, except that there are three factors, say A, B and C, with a, b and c treatments, respectively.

The experimental units are assigned to the three-way combinations of the treatments.

Formally,

$$Y = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon \qquad (4.2)$$

Notice the *three-way* interaction term $(\alpha\beta\gamma)_{ijk}$, as well as the three two-way interaction terms: $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ and $(\beta\gamma)_{jk}$. An effect can be interpreted only if all the higher-order interactions that involve it are *not* significant. Otherwise the analysis proceeds analogously to the two-way factorial case.

For practise, consider the data in the files `chap4data3.csv` and `chap4data4.csv`:

> **`Data3`**: *Thirty-six adults (18 males and 18 females) participated in a study to compare blood pressure meters (sphygmomanometers) produced by three manufacturers. Persons of each gender were randomly assigned to six groups of three each. Three groups of each gender had systolic blood pressure measurements taken at entry to the experiment; the other three groups had measurements taken after resting for 10 minutes*.

Tip: When specifying the linear model in *R*, using the form

```
<response variable> <- <treatment variable 1>*<treatment variable
2>*<treatment variable 3>
```

will include all three main effects, the three two-way interactions as well as the three-way interaction.

> **`Data4`**: *A further example. Focus on the interpretation of the ANOVA table here; ignore the testing of assumptions.*

4. Higher-order factorial designs (C&S 12.7.2, p. 179)

With more factors, it is possible to have designs with higher-order interactions. Again, an effect can be interpreted only if all the higher-order interactions that involve it are not significant. Higher-order interactions in themselves are very difficult to interpret, and are often ignored.

5. Split plot designs (C&S chapter 16)

It is sometimes necessary to perform a factorial experiment where the levels of one of the factors have to be applied to large plots, whereas the other factor can be applied to smaller units.

Example:

> *Two methods of soil preparation (Factor A) and two levels of a fertilizer (Factor B) have to be compared. The preparation of soil can only be done on a complete strip, whereas different quantities of fertilizer can be applied to smaller plots. The two levels of Factor A will each be allocated randomly to a number of strips, where after each strip will be divided into two and the two levels of Factor B will be applied randomly to each half of a strip.*

We refer to the larger plots on which A was applied as the *main plots* and to the smaller plots within the main plots on which B was applied as the *sub plots*.

One can also use a split plot experimental design when a limited number of plots on which to perform the experiment is available, and one of the factors is more important than the other. The less important factor will then be applied to the main plots, and the more important factor to the sub plots. The reason for this is that the sub plot treatments is repeated a larger number of times than the main plot treatments and the more important factor and interaction is thus measured with *greater precision* than the less important factor.

Disadvantages of the design:

- one of the factors forfeits precision of measurement
- the experiment requires more effort to lay out
- the analysis of the data and interpretation of the results are more difficult than for other treatment designs

The most important difference between the analysis of the data for a split plot design and designs where one does not have to deal with treatments within treatments, is that *two error terms must be calculated*, one for the main plot treatments and one for the sub plot treatments and the interaction.

Important: in hypothesis testing, the *main plot error* must be used to calculate the F-statistic for the main plot factor, and the *sub plot error* must be used to calculate the F-statistics for the sub plot factor and interaction.

4.5.1 Split plot design: Randomised Block design for a two-factor experiment

Consider a Randomised Block design for a two-factor experiment, which is the most common split plot design.

Letting Factor A be the main plot factor, and Factor B the sub plot factor, the experiment is set up as follows: within each block, the levels of Factor A are randomly assigned to each of the main plots. Within each main plot, the levels of Factor B are then randomly assigned to each of the sub plots. The linear model for this design looks very similar to that of a plain randomised block two-way factorial design, with just an additional error term:

$$Y = \mu + \delta_k + \alpha_i + \varepsilon^a + \beta_j + (\alpha\beta)_{ij} + \varepsilon^b$$

- Y: the response measured on the experimental (sub)plot;
- μ: the overall mean;
- δ_k: the effect of the k^{th} block;
- α_i: the main effect of the i^{th}level of Factor A (main plot factor);
- ε^a: the unexplained *main plot* error;
- β_j: the main effect of the j^{th} level of Factor B (sub plot factor);
- $(\alpha\beta)_{ij}$: the interaction between the i^{th} level of Factor A and the j^{th} level of Factor B;
- ε^b: the unexplained *sub plot* error.

Consider the data contained in the file `chap4data5.csv`:

> *An experiment was performed on the fluorometric determination of the riboflavin content of dread cabbage leaves. The two factors were M, the sample size (0.25g, 1g) and PP, the effect of permanganate-peroxide on the determination. The experimental design was a random blocks design that was repeated on three successive days. The two levels of Factor PP were applied to three main plots each (one per day), and then each main plot was divided into two sub plots to which the two levels of Factor M were applied.*

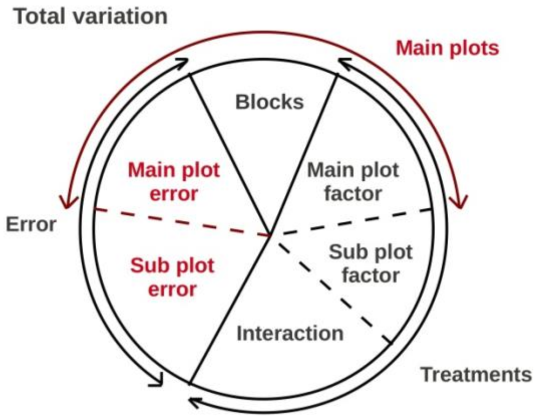

Figure 4.2: Sources of variation in a Randomised Block split plot design.

ANOVA for RB split plot design

```
<model name> <- aov(<response variable>~<block variable> + <main plot
   factor>*<subplot factor> + Error(<block variable>/<main plot
   factor>/<subplot factor>),data=<data set name>)
summary(<model name>)
```

To perform multiple comparison tests after the ANOVA analysis, following the following instructions:

1 Load the scripts `splitplotHSD.R` (available on the SUNLearn page). (This has been included in the markdown file as well.

2 Submit the script in *R*.

3 Use the `splitplotHSD()` function the following parameters:

- `y = <data set name[,"<response variable>"]>`
- `trt = <data set name[,"<factor>"]>` (i.e. the factor for which you want to perform the multiple comparisons)
- `DFerror = <DF for the error>` (either the main plot or subplot error, depending on which factor you are performing the multiple comparisons for)
- `MSerror = <mean squared error>` (either the main plot or subplot error, depending on which factor you are performing the multiple comparisons for)

Thus for the example data (`chap4data5.csv`), you should use the following lines of code in the *R* script:

```
> chap4data5.model1 <- aov(Conc~Block + PP*Size +
   Error(Block/PP/Size),data=chap4data5)
> summary(chap4data5.model1)
```

This should give the following output:

```
> summary(chap4data5.model1)

Error: Block
      Df Sum Sq Mean Sq
Block  2  3.762   1.881

Error: Block:PP
          Df Sum Sq Mean Sq F value  Pr(>F)   
PP         1  716.1   716.1   144.2 0.00687 **
Residuals  2    9.9     5.0                   
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Block:PP:Size
          Df Sum Sq Mean Sq F value Pr(>F)
Size       1  36.40   36.40   3.720  0.126
PP:Size    1  13.02   13.02   1.331  0.313
Residuals  4  39.14    9.79
```

To perform post-hoc multiple comparison tests for factor *PP*, load and submit the `splitplotHSD.R` script in *R* and then use the following code:

```
$groups
      trt    means M
1 Without 39.81667 a
2 With    24.36667 b
```

It is clear that the *Without* level of factor *PP* gives the greatest mean yield, and differs significantly from level *With*.

4.5.2 Split plot design: Completely Randomised design for a two-factor experiment

For a Completely Randomised split plot design, the ANOVA analysis in *R* is slightly different.

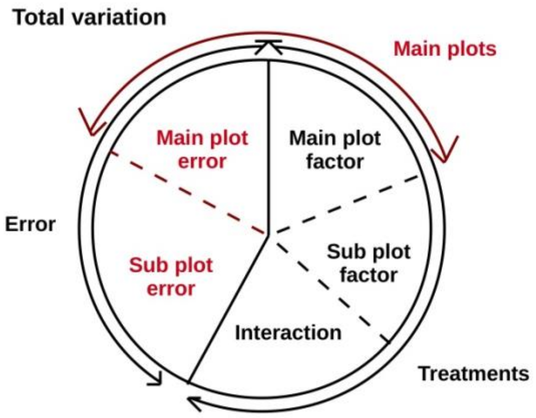

Figure 4.3: Sources of variation in a Completely Randomised split plot design.

The linear model looks as follows:

$$Y = \mu + \alpha_i + \varepsilon^a + \beta_j + (\alpha\beta)_{ij} + \varepsilon^b$$

- Y: the response measured on the experimental (sub)plot;
- μ: the overall mean;
- α_i: the main effect of the i^{th} level of Factor A (main plot factor);
- ε^a: the unexplained *main plot* error;
- β_j: the main effect of the j^{th} level of Factor B (sub plot factor);
- $(\alpha\beta)_{ij}$: the interaction between the i^{th} level of Factor A and the j^{th} level of Factor B;
- ε^b: the unexplained *sub plot* error.

Consider the data contained in the file `chap4data6.csv` (C&S 16.5, p. 250):

> *A factorial experiment was conducted to investigate the effects of five nutrient levels and three temperatures (20, 25, 30 degrees Celsius) on the fresh weight yields (g per pot) of potted plants. Six cabinets were used for the experiment, with two cabinets being randomly allocated to each of the three temperature levels. Within each cabinet, each nutrient level was assigned randomly to one of five potted plants (randomly placed). There are thus six main plots (cabinets), and five subplots (potted plants) within each main plot.*

ANOVA for CR split plot design

```
<model name> <- aov(<response variable>~<main plot factor>*<subplot
   factor> + Error(<main plot identifier>/<main plot factor>/<subplot
   factor>),data=<data set name>)
summary(<model name>)
```

To perform multiple comparison tests after the ANOVA analysis, following the following instructions:

2. Load the scripts `splitplotHSD.R` (available on the SUNLearn page) (This has been included in the markdown file as well.
3. Submit the script in *R*.
4. Use the `splitplotHSD()` function the following parameters:

- `y = <data set name[,”<response variable>”]>`
- `trt = <data set name[,”<factor>”]>` (i.e. the factor for which you want to perform the multiple comparisons)
- `DFerror = <DF for the error>` (either the main plot or subplot error, depending on which factor you are performing the multiple comparisons for)
- `MSerror = <mean squared error>` (either the main plot or subplot error, depending on which factor you are performing the multiple comparisons for)

Thus for the example data (`chap4data6.csv`), you should use the following lines of code in the *R* script:

```
> chap4data6.model1 <- aov(Yield~Temp*Nutrient.level +
Error(Cabinet/Temp/Nutrient.level),data=chap4data6)
```

```
> summary(chap4data6.model1)
```

This should give the following output:

```
> summary(chap4data6.model1)

Error: Cabinet
          Df Sum Sq Mean Sq F value Pr(>F)  
Temp       2  75.82   37.91   12.87 0.0337 *
Residuals  3   8.84    2.95                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Cabinet:Temp:Nutrient.level
                    Df Sum Sq Mean Sq F value  Pr(>F)    
Nutrient.level       4 1241.9  310.48 393.510 1.3e-12 ***
Temp:Nutrient.level  8   20.4    2.56   3.239   0.033 *  
Residuals           12    9.5    0.79                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant interaction between the *Temperature* and *Nutrient level*, and the main effects of these two factors can therefore not be interpreted in isolation of each other. One solution to this is to perform one-way ANOVA's for *Nutrient Level* at each of the levels of the *Temperature* variable.

In the *R* output, make sure to use the correct *F*-statistic and *p*-value for the main plot factor (not those given in the ANOVA table!).

6. Exercises

Exercise 1

Do the ANOVA analysis for C&S Exercise 12.1 (p. 180) (`chap4exer1.csv`) and state your conclusions about the main effects and interaction. An experiment was carried out to investigate the effect of row spacing on the yield of three varieties (V1, V2, V3) of dwarf French beans. The spacings were: S1=20cm, S2=40cm, S3=60cm, D40=80cm. Factor A had three levels (variety), and factor B had four levels. Treatment combinations were arranged in randomised blocks.

Exercise 2

Do the ANOVA analysis for C&S Exercise 16.1 (p. 253) (`chap4exer2.csv`) and state your conclusions about the main effects and interaction. An experiment was laid out to compare the effects of three concentrations of a chemical seed dressing with a control on the yield of oats. The experiment was laid out in five randomised blocks. Yields in kg per subplot are shown in the csv file.

Chapter 5 - Non-parametric statistics

1. Introduction

In this chapter, we will discuss some *non-parametric tests*. These tests are typically performed when dealing with very small sample sizes or when the assumption of normality is severely violated, and the classical techniques therefore cannot be employed.

In short, when we have a very small sample or one for which the normality assumption is not valid:

- we use a *Wilcoxon rank-sum test* instead of a t*-test* for *two independent* samples
- we use a *Wilcoxon matched-pair signed-rank test* instead of a t-test for *two dependent* samples
- we use a *Kruskal-Wallis test* instead of a completely randomised (one-way) ANOVA
- we use *Friedman's test* instead of a randomised block design ANOVA

As you work through the chapter, make sure that you:

- know when a particular non-parametric test is appropriate;
- appreciate that even a non-parametric test may be inappropriate; it may have its own assumptions.

In particular, all the non-parametric techniques described here make the assumption that *the distributions of the groups being compared has the same general shape.* The Wilcoxon rank-sum test (corresponding to a t-test for two independent samples) has the additional requirement of *homoscedasticity*.

Background reading material

- Biometry 242: note set 13
- Clewer and Scarisbrick: chapter 19

2. Wilcoxon rank-sum test (C&S 19.5, p. 229)

The Wilcoxon rank-sum test is also known as the *Mann-Whitney U test.*

The Wilcoxon rank-sum test is the non-parametric equivalent of the t-test for two independent samples.

The test is appropriate to use when the data is not normally distributed, but *the two treatment groups come from distributions with the same general shape and a common variance (homoscedastic).*

We can test for normality using Shapiro-Wilk's test, and for homoscedasticity using Bartlett's or Levene's test.

The hypothesis of interest is slightly different to the parametric case:

H_0: The *locations* of the two populations are the same

H_1: The *locations* of the two populations differ

We should remember that when the sample size is small, the tests for normality and homoscedasticity themselves might have very low power to detect deviations from these assumptions.

As a rule of thumb, when there are 10 or fewer observations in either group ($n_1 < 10$ or $n_2 < 10$), it is good practice to perform a Wilcoxon rank-sum test.

Apart from small samples, another situation in which it is obvious for the non-parametric test to be used is when the data are *ordinal* instead of numerical.

For example, a response variable may consist of ratings of say, 1 to 5. Such ratings are in fact ordered categories, not regular numbers with which we can perform normal arithmetic. The data are "ordinal".

Consider the data in the file `chap5data1.csv`.

> *A new treatment for headaches is being tested. The new treatment is given to 15 patients, while a second lot of 15 patients is given aspirin. Each patient is asked to indicate the effectiveness of the medication (*$1 =$ *ineffective to* $5 =$ *effective).*

Use *R* to perform a test to determine whether there is a significant difference between the two medications. First make sure that the assumption of homoscedasticity is tenable.

Wilcoxon rank-sum test

```
with(<data set name>,tapply(<response variable>, <factor variable>,
   median, na.rm=TRUE))
wilcox.test(<response variable>~ <factor variable>
   ,alternative="two.sided",data=<data set name>)
```

Note that both *approximate* and *exact* p-values can be calculated. However, for larger data sets the calculation of exact p-values may take a long time and should be avoided – in such cases the approximate p-values should be sufficient.

Interpret the p-values (test results) clearly.

3. Wilcoxon matched-pair signed-rank test (C&S 19.4, p. 297)

The Wilcoxon matched-pair signed-rank test is the non-parametric equivalent of the t-test for two dependent samples (paired t-test).

The test is appropriate to use when the differences are not normally distributed, but the assumption is made that *the differences still come from a symmetrical distribution.*

Consider the data in the file `chap5data2.csv`.

> *The 'no. 2 leaf' on each of n=8 tobacco plants were treated as follows: After dividing each leaf in two half-leaves, one half-leaf was rubbed with a rag that was soaked in Virus preparation A, and the other half-leaf was rubbed with a rag that was soaked in Virus preparation B. It was decided in a random manner which half-leaf would receive treatment A and which half-leaf treatment B. The response of interest was the number of virus lesions (small dark rings) that were caused.*

Although the eight leaves were chosen in a random manner and the two treatments were applied in a random manner to each half of each leaf the two members of each pair are not independent so that the differences between the two treatments cannot be calculated as the difference between two means. Furthermore the sample size is very small and the data is not normally distributed. A non-parametric test will therefore be more appropriate than the parametric t-test for dependent samples.

Import the data in *R*. Draw a histogram of the differences to check that the distribution thereof is more or less symmetrical. Perform a Wilcoxon matched-pairs test to determine whether the number of lesions caused by the two viruses is the same.

Wilcoxon matched-pairs test

```
with(<data set name>,median(<variable 1>-<variable 2>, na.rm=TRUE))

with(<data set name>,wilcox.test(<variable 1>, <variable 2>,
   alternative='two.sided',

  paired=TRUE))
```

As with the Wilcoxon rank-sum test, you can also perform an *exact* test, but the computational time may be much greater for larger data sets.

4. Kruskal-Wallis test (C&S 19.6, p. 302)

The Kruskal-Wallis test is the non-parametric equivalent of the completely randomised (one-way) ANOVA.

The test is suitable for use with small data sets obtained from completely randomised experiments with *2 or more treatments*, when the data is not normally distributed. However, it still assumes that *the data in the treatment groups come from distributions that have the same general shape.*

We can test for normality using Shapiro-Wilk's test, as in previous chapters.

As with the Wilcoxon rank-sum test, the Kruskal-Wallis test can also be used to analyse ordinal data (such as ratings of 1 to 5).

The hypothesis of interest is:

H_0: The *locations* of the g populations are the same

H_1: The *locations* of all g populations are not the same

Consider the data in the file `chap5data3.csv`.

> *In an experiment with four insecticides (A, B, C and D) that were each repeated five times in a completely random manner on cabbages, caterpillars were counted on the plants two weeks later.*

Because counts typically have a *skewed* distribution (rather than symmetric as expected under the normal distribution) it is sensible to perform a Kruskal-Wallis test on these data.

Perform a Kruskal-Wallis test using *R*, to determine whether there are significant differences between the insecticides in terms of their effect on caterpillar numbers.

Kruskal-Wallis test

```
with(<data set name>, tapply(<response variable>,<factor
   variable>,median,na.rm=TRUE))
kruskal.test(<response variable>~<factor variable>, data=<data set
   name>)
```

If there seem to be significant differences between the treatment groups, the Kruskal-Wallis test can be followed up with pairwise Wilcoxon rank-sum tests, to determine which of the treatment groups differ significantly from each other. Remember that *the family-wise Type I error rate will become inflated* in this case, so it will be necessary to adjust the significance level for the tests (to compare the p-values against) manually.

We can also use the following tests to perform multiple comparisons which will automatically take into account the inflation of *family-wise Type I error rate*.

Nemenyi test

```
posthoc.kruskal.nemenyi.test(<response variable> ~ <factor
   variable>,dist="Tukey",data=<data set name>)
```

Dunn test

```
posthoc.kruskal.dunn.test((<response variable> ~ <factor
   variable>,p.adjust.method="bonferroni",data=<data set name>)
```

5. Friedman’s test (C&S 19.7, p. 304)

The Friedman test is the non-parametric equivalent of the randomised block design.

The test is appropriate when the ANOVA residuals are not normally distributed.

Consider the data in the file:

> *Ten wine tasters were required to evaluate each of four types of semi-sweet wines by arranging them in order o fpreference.*

If each of the ten tasters are regarded as a block and the preferences are ranked, it means that the Friedman test can be applied directly to the data. Make note that the format of the input data file is different from the usual.

Use *R* to perform a Friedman test to determine whether there are differences between the ranks allocated to the four wines.

Friedman test

```
friedman.test(na.omit(with(<data set name>, cbind(<treatment 1>,
   <treatment 2>,…))))
```

6. Exercises

Exercise 1

See C&S Example 19.4 (p. 300) (`chap5exer1.csv`). Perform a non-parametric test on the data to determine if the two wheat varieties differ significantly. State your conclusion clearly.

Exercise 2

See C&S Example 19.5 (p. 303) (`chap5exer2.csv`). Using the data, perform a non-parametric test to determine if the yields of the four wheat varieties are equal. State you conclusion clearly.

Exercise 3

See C&S Example 19.6 (p. 305) (`chap5exer3.csv`). Perform a non-parametric test on the data to determine if the four wheat varieties that were compared using a Random Block design differed.

Exercise 4

See C&S Example 19.3 (p. 298) (`chap5exer4.csv`). Perform a non-parametric test on the data to determine if there is a difference between the two fungal strains.

Chapter 6 - Simple linear regression

1. Introduction

In this chapter, we discuss *simple linear regression*. In simple linear regression, the relationship between a single *dependent* numerical variable and a single *independent* numerical variable is modelled as a straight line. As you work through the chapter, make sure to note:

- what the technique does;
- what diagnostic checks can be performed;
- what the underlying assumptions are;
- how the model can be used to make predictions;
- how the data can be transformed to better conform to the assumptions.

Background reading material

- Biometry 212: note set 7
- Clewer and Scarisbrick: Chapter 7 and 14

2. The basics (C&S 7.1-7.5, p. 63)

In simple linear regression, we have a single dependent (or *response)* variable Y which we would like to explain in terms of a single independent (or *explanatory*) variable X. The way Y is explained in terms of X is particularly easy: The best-fitting straight line through the scatter plot of Y against X is found. The mathematical model is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where

- Y is the value of the dependent variable;
- β_0 is the *intercept* of the best-fitting line;
- β_1 is the *slope* of the best-fitting line;
- X is the value of the independent variable;
- ϵ is the part of Y which is not explained by the regression line (i.e. the *residual* or *error*).

For the model to be usable, the slope (β_1) and the intercept (β_0) of the mathematical model must be estimated from the data. But on what basis should the estimates be found? Of all lines, the “best” one is said to be the one which has the smallest sum of squared errors, or the *least squares*. The estimates (b_0 and b_1) are therefore chosen to be exactly those numbers which minimise:

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

This is done automatically in *R*; we simply get the estimates from the output. The formula of the fitted model is:

$$\hat{Y} = b_0 + b_1 X.$$

Estimating β_0 and β_1 is merely a mathematical operation; it does not involve any statistics. However, in order to test *hypotheses* regarding β_1 (or β_0), we do require statistics.

The hypothesis of main interest is:

H_0: $\beta_1 = 0$ (No straight-line relationship between X and Y)

H_1: $\beta_1 \neq 0$ (Straight-line relationship between X and Y)

Alternatively, if $b_1 > 0$, we may be interested in the *right one-sided* alternative hypothesis:

H_1: $\beta_1 > 0$ (*Positive* straight-line relationship between X and Y)

Or, if $b_1 < 0$, we may be interested in the *left one-sided* alternative hypothesis:

H_1: $\beta_1 < 0$ (*Negative* straight-line relationship between X and Y)

Consider the data in the file `chap6data1.csv`:

> *The body weights (in kilograms) and resting metabolic rates (in kcal/24hr) of 44 women were measured.*

Use *R* to draw a scatter plot of the data. Put `BodyWeight` on the horizontal axis, and `MetabolicRate` on the vertical axis. Determine if a straight line fits the data well.

Use *R* to fit a simple linear regression of metabolic rate onto body weight. Take note of the order in which the variables are mentioned in the preceding phrase: Regression of dependent variable (Y) *onto* independent/explanatory variable (X).

Simple linear regression

```
<model name> <- lm(<response variable>~<independent
   variable>,data=<data set name>)

summary(<model name>)

Anova(<model name>)

confint(<model name>,level=0.95)
```

Give the equation of the fitted regression line. Where appropriate, clearly *interpret* the parameter estimates (the slope and intercept). Determine whether *greater* weight is associated with *greater* metabolism.

3. Diagnostic checks (C&S, p. 72)

Although the line that is fitted will always be the best possible straight line, it may still offer only a poor account of the relationship between X and Y. X and Y may not have a straight-line relationship, for example, or there may not be a significant relationship between the two variables at all.

The *coefficient of determination*, or R^2, is a measure of how well the fitted line explains Y in terms of X. Suppose that $R^2 = 0.5$. We interpret this by saying that $0.5 \times 100 = 50\%$ of the variation in the dependent variable (Y) is accounted for by the regression model on the independent variable (X). The remaining variation is unaccounted for by X; it may be due to other factors.

Determine and interpret the coefficient of determination of the metabolism data.

After the least squares line is found, some points may lie far above or below this best line. Such points (which are sufficiently vertically distanced from the line) are called *outliers*.

Other points may be atypical in a different way. Intuitively speaking, observations with *high leverage* lie horizontally (i.e. in the X-space) far from the other points. They may have an exaggerated effect on the choice of best line; these observations punch above their weight.

Neither outliers nor high leverage observations should be removed without inspection from a data set. First check whether a simple error has not been made in the recording or preparation of the data. If so, correct the data; otherwise confer with an expert. Remember that the purpose is to model reality. If such outlying or high leverage observations represent a legitimate part of reality, they should be included.

Various measures have been developed to more easily and accurately identify outliers and/or high leverage observations. To perform such diagnostic checks in *R*, do the following:

Diagnostic measures (for regression model):

```
influence.measures(<model name>)

matrix(rstudent(<model name>),ncol=1)
```

In the descriptions of the diagnostic measures which follow, p is the number of model parameters (i.e. intercept + number of independent variables), while N is the total number of observations. For the metabolism data, $p = 2$ (the intercept and the slope) and $N = 44$.

Use the following code to add the diagnostic measures and the fitted values to the original data set.

```
within(<data set name>, {
   fitted(<model name>)
   residuals(<model name>)
   rstudent(<model name>)
   hatvalues(<model name>)
   cooks.distance(<model name>)
   obsNumber <- 1:nrow(<data set name>)
})
```

The following diagnostic measures will be included amongst the output:

- **Student residual:** Standardised residual, $\frac{residual}{s_{residuals}}$, reflecting the degree to which the corresponding observation deviates from its predicted value.
- **Rstudent:** Reflects the deviation of an observed Y_i from its prediction and should not exceed 2 in absolute value. In other words, it should fall between -2 and 2.
- **HatDiagH:** Measures the distance from the X-value(s) of an observation to the center of the data ($\bar{X}$) and should not exceed $\frac{2p}{N}$. Used to identify potentially influential observations with high *leverage*. For the metabolism data, $\frac{2p}{N} = 0.091$.
- **CovRatio:** Indicates the effect of an observation on the accuracy of estimation and should not exceed $1 \pm \frac{3p}{N}$. For the metabolism data, $1 - \frac{3p}{N} = 0.864$ and $1 + \frac{3p}{N} = 1.136$.
- **Dffits:** Diagnoses the influence of the observations on the predicted values and should not exceed $2\sqrt{\frac{p}{N}}$ in absolute value. In other words, it should fall between $-2\sqrt{\frac{p}{N}}$ and $2\sqrt{\frac{p}{N}}$. For the metabolism data, $2\sqrt{\frac{p}{N}} = 0.426$.
- **Dfbetas:** Diagnoses the influence of the observations on the different regression parameters. Therefore, each parameter in the model (here the intercept and the slope) has a separate Dfbetas value for each observation. These values should not exceed $\frac{2}{\sqrt{N}}$ in absolute value. In other words, they should fall between $-\frac{2}{\sqrt{N}}$ and $\frac{2}{\sqrt{N}}$. For the metabolism data, $\frac{2}{\sqrt{N}} = 0.302$.

Use *R* to perform a diagnostics check of the regression analysis performed earlier on the metabolism data. Do not remove any observations.

Use the following function to calculate the cut-off values in *R*. Submit the code into the *R* console.

Diagnostic measures cut-off values

```
influence.cutoffs <- function(model){
  p <- length(model$coefficients)
  n <- length(model$residuals)
    DFBETAS <- 2/(n^0.5)
  DFFITS <- 2*(p/n)^0.5
  COVRATIO.lwr <- 1-3*p/n
  COVRATIO.upr <- 1+3*p/n
  cook.D <- 4/n
  HATDIAG <- 2*p/n

   list(DFBETAS=DFBETAS,DFFITS=DFFITS,COVRATIO.lwr=COVRATIO.lwr,COVRATIO.upr=COVRATIO.upr,cook.D=cook.D,HATDIAG=HATDIAG)
}
```

The following graphs can also be plotted to investigate for influential values. The `car` package needs to be installed for these plots to work.

Influential observation plots

```
avPlots(<model name>)
leveragePlots(<model name>)
influencePlot(<model name>)
```

4. Assumptions (C&S 7.8, p. 75; 14.1, p. 213)

The assumptions are made about the *regression errors*, ϵ_i, for $i = 1, \dots, n$. These assumptions are tested using the *residuals*. The residuals are assumed to be:

- normally distributed;
- come from the same single distribution, i.e. have the same variability across the range of fitted values (homoscedasticity assumption);
- independently distributed.

Often a detailed diagnostics check will be prompted by an initial investigation of the assumptions.

Checking assumptions (linear regression)

```
par(mfrow=c(2,2))
plot(<model name>)

shapiro.test(<model name>$residuals)
```

The Residuals vs Fitted graph can be used to determine whether the homoscedasticity assumption is tenable. The Shapiro- Wilk test can be used to test for normality.

The assumption of independence of the observations is explicitly tested *only if the times at which the observations were made are known.* In other words, for *time series data.* If this is the case, a *Durbin-Watson test* can be performed.

Checking independence assumption (linear regression):

```
dwtest(<response variable> ~ <independent variable>,
   alternative="two.sided", data=<data set name>)
```

A graph of the residuals (on the y-axis) vs. the times of the observations (on the x-axis) can also be drawn. There should be no obviously discernible pattern.

Besides the ordinary linear regression, there are also various techniques for *repeated-measures* data. These techniques will be discussed in a subsequent chapter; it is important to *not* analyse such data using ordinary linear regression as described here.

Use *R* to test the assumptions underlying the regression analysis performed earlier on the metabolism data.

6.4.1 Assumptions hold: predictions (C&S 7.8.6, p. 80)

Once the regression model has been fitted and if the model assumptions seem valid, it can be used to make predictions by substituting the value of a new *X* into the regression formula.

What is the difference between *interpolation* and *extrapolation*? Why is it important to distinguish between the two? Why may extrapolation be inappropriate/dangerous in some situations?

What is the difference between a *confidence interval* and a *prediction interval* for a new prediction? Make very sure you know how to interpret each of these intervals. For the intervals to be valid, the assumptions underlying the regression analysis must hold.

To obtain the confidence intervals and prediction intervals for *new (predicted) observations*, a new data set (containing the values of the independent variables for the new observations) must be created first. Use the following steps to obtain the interval estimates for new observations in *R:*

Confidence and prediction interval estimates (new data)

```
<new data set name> <- data.frame(<independent variable>=c(<value to
   be predicted>))

predict(<model name>,newdata=<new data set
   name>,interval="confidence")

predict(<model name>,newdata=<new data set
   name>,interval="prediction")
```

Use the following code to plot the confidence and/or prediction intervals on the scatter plot by using the `ggplot` system.

```
<new data set name 2> <- cbind(<data set name>,predict(<model
   name>,interval="prediction"))

ggplot(<new data set name 2>,aes(x=<independent
   variable>,y=<response variable>)) + geom_point() +
   geom_line(aes(y = fit),colour = "blue", size = 1) +
   stat_smooth(method=lm) + geom_ribbon(aes(ymin = lwr, ymax = upr),
   fill = "blue", alpha = 0.2)
```

- Use *R* to predict the metabolic rate of a woman who weighs 70kg, using the fitted linear regression model.
- Use *R* to determine and *interpret* a 95% *confidence* interval when $X = 70$ based on the earlier analysis.
- Use *R* to determine and *interpret* a 95% *prediction* interval when $X = 70$ based on the earlier analysis.

6.4.2 Assumptions fail: transformation (C&S 14.2, p. 219)

If the assumptions underlying the regression of Y onto X are seriously violated, one possible solution is to *transform* the X and/or Y variables into new X^* and Y^* variables, for example, which can be used to perform a regression analysis for which the assumptions hopefully hold. Some possible transformations of Y:

- For positive and exponentially increasing values, perhaps $Y^* = \log_e Y$ ($Y^* = lnY$).
- For non-negative and exponentially increasing values, perhaps $Y^* = \log_e(Y+1)$ ($Y^* = \ln{(Y+1)}$).
- For counts, perhaps $Y^* = \sqrt{Y+0.5}$.
- For proportions, perhaps $Y^* = \sin^{-1}(\sqrt{Y})$.
- Another alternative: The reciprocal, $Y^* = \frac{1}{Y}$.

After a prediction of Y^* is made using X^*, the prediction still needs to be *back-transformed* appropriately so that the prediction is given in terms of the original Y values, not those of Y^*.

Consider the data in the file `chap6data2.csv`:

> *A professor is interested in investigating the association between the marks obtained (Y, out of 40) and the time spent studying for a test by students (X, in minutes).*

Investigate the effect of the following transformations on the assumptions of homoscedasticity and normality when compared to the original Y:

- $Y^* = \log_e Y$
- $Y^* = \frac{1}{Y}$

Use the best transformation to predict the mark of a student who studied for 52 minutes.

5. Exercises

Exercise 1

With the data from C&S Example 7.1, p. 64 (`chap6exer1.csv`): Perform a simple linear regression of Area (Y) on Stem length (X) and give the equation of the best-fitting regression line, as well as the coefficient of determination (R^2). Interpret the R^2 value. Also test the assumptions underlying this regression model, and clearly state your conclusions with regard to the validity of the fitted model.

Exercise 2

With the data from C&S Example 7.2, p. 69 (`chap6exer2.csv`): Perform a simple linear regression of wheat yield (Y) on nitrogen fertilizer (X) and give the equation of the best-fitting regression line. Predict the wheat yield for a nitrogen fertilizer level of $X = 80$kg nitrogen per hectare (also include the confidence and prediction intervals in your answer, and clearly interpret).

Exercise 3

Do C&S Exercise 7.2 (p. 84). The data is in the file `chap6exer3.csv`. A study was undertaken to find out if tree diameter measurements 1.5m above ground level can be used to predict heights for a certain species. The measurements of 12 trees are shown in the csv file. Complete a thorough regression analysis (significance, R-squared, assumptions).

Chapter 7 - Multiple linear regression

1. Introduction

In this chapter, we discuss both *polynomial regression* and *multiple linear regression*. In polynomial regression, the relationship between a single numerical response variable and a *single* numerical explanatory variable is modelled as a *polynomial* curve, either quadratic, cubic, quartic, etc. In multiple linear regression, the relationship between a single numerical response variable and *multiple* numerical explanatory variables is modelled as a *hyperplane*. As you work through the chapter, make sure to note:

- what each technique does;
- how each technique can be used to make predictions.

Diagnostic checks and tests of assumptions are performed in the same way as for simple linear regression.

Background reading material

- Biometry 212: note set 7 and 8
- Biometry 242: note set 14
- Clewer and Scarisbrick: Chapter 8

2. Polynomial regression (C&S 8.2 – 8.3, p. 87)

In polynomial regression, we have a single response variable Y which we would like to explain in terms of a single explanatory variable X. Instead of explaining Y in terms of X as a straight line (as is done in simple linear regression), we make use of a polynomial curve. For example, we can explain Y in terms of a *quadratic* polynomial in X (a "parabola"). In this case the mathematical model is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

where

- Y is the value of the response variable;
- X is the value of the explanatory variable;
- $\beta_0, \beta_1, \beta_2$ are the coefficients of the polynomial;
- ϵ is that part of Y which is not explained by the polynomial (i.e. the error).

The parameters β_0, β_1 and β_2 must be estimated from the data. The parameter estimates b_0, b_1 and b_2 are estimated by R using the method of ordinary least squares (OLS). The formula of the fitted model is:

$$\hat{Y} = b_0 + b_1 X + b_2 X^2.$$

Hypotheses concerning the parameters can be tested. Perhaps the most interesting hypothesis is whether the coefficient of the quadratic term is equal to zero (in the population):

H_0: $\beta_2 = 0$ (No quadratic relationship between X and Y)

H_1: $\beta_2 \neq 0$ (Quadratic relationship between X and Y)

If the null hypothesis is not rejected, a simple linear regression should be fitted instead, with the simple linear regression model used for further analyses.

How about *cubic* or *quartic* polynomial regression? How can we recognize these by their number of *turning points*? How does this in general relate to their *order*?

Consider the data in the file `chap7data1.csv`:

> *The manufacturer of an additive to petrol claims that his product improves the fuel consumption of cars. In a series of experiments, different quantities of the additive are poured into the petrol tank of a test car (recorded in cm^3 per litre). The resultant fuel consumption is also recorded, in litres per 100km.*

Use *Excel* to fit *trend lines* of orders 1 to 4 through the data points.

Include the model formulae and R^2 values on the graphs. Comment on the changes in goodness-of-fit.

Important: Beware of fitting too many parameters (called *overfitting*)! As a rule of thumb, for a well estimated regression model you need about eight observations per parameter to be estimated.

Draw a scatter plot of the data in *R* and include a trend line on the plot. Use *R* to fit all polynomial regression models, from linear to quadratic. Perform hypothesis tests on each newly introduced parameter as the models are expanded. Comment on the results.

Polynomial regression

```
<model name> <- lm(<response variable> ~ <independent variable> +
   I(<independent variable>^2), data=<data set name>)
summary(<model name>)
```

To assess the assumptions and diagnostic measures use the same procedures that were introduced in Chapter 6.

For a straight-line model there are $p = 2$ parameters (including the intercept), for a quadratic model there are $p = 3$ parameters, and so on. The number of parameters, p, affects the acceptable ranges of the different diagnostic measures discussed in Chapter 6. Calculate the diagnostic measures for the cubic model of fuel consumption, and comment on the results.

The underlying *assumptions* (normality, homoscedasticity and independence of errors) can be tested in the same way as for simple linear regression. Investigate the validity of the assumptions underlying the cubic model of fuel consumption.

If the assumptions hold, *predictions* can be made in the same way as for simple linear regression. Based on the cubic model for fuel consumption, predict the fuel consumption when 5.5 cm^3 of the additive is added per litre of petrol. Interpret both the 95% prediction interval and the 95% confidence interval for this prediction. **Note**: *Extrapolation* can be even more dangerous in the case of polynomial regression than in the case of simple linear regression.

3. Multiple linear regression (C&S 8.5, p. 100)

In multiple linear regression, we have a dependent variable Y which we would like to explain in terms of the *multiple* independent variables $X_1, X_2, \dots, X_q$ simultaneously.

The mathematical model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + \epsilon$$

where

- Y is the value of the response variable;
- X_j is the value of the j^{th} explanatory variable;
- β_0 is the intercept, and β_j is the partial slope of the j^{th} explanatory variable;
- ϵ is that part of Y which is not explained by the regression (i.e. the error).

The parameters, $\beta_0, \beta_1, \beta_2, \dots \beta_q$, must be estimated from the data. The parameter estimates, $b_0, b_1, b_2, \dots, b_q$, are estimated by R using the method of least squares. The formula of the fitted model is:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_q X_q$$

Hypotheses concerning the parameters can be tested. For example, for a particular $j = 1, \dots, q$:

H_0: $\beta_j = 0$ (X_j does not “contribute” to the explanation of Y)

H_1: $\beta_j \neq 0$ (X_j does “contribute” to the explanation of Y)

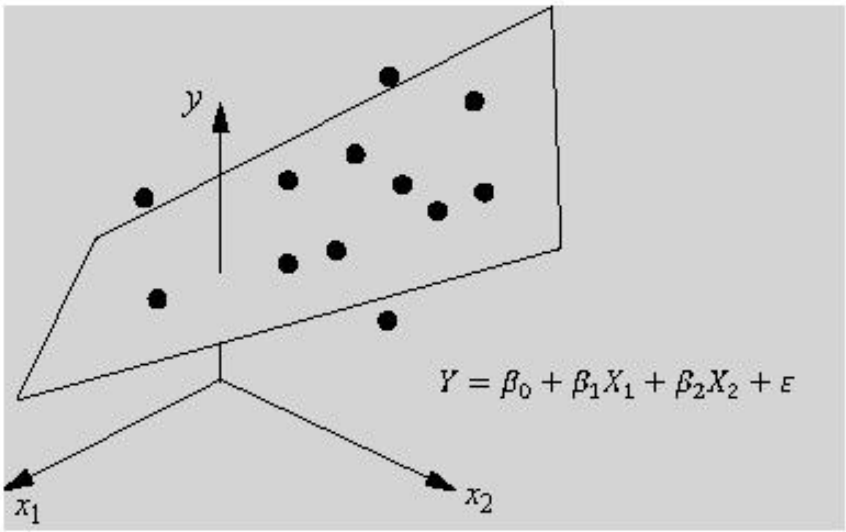

Figure 7.1: A multiple regression model (with two explanatory variables) represented as a hyperplane in three-dimensional space. The dots surrounding the hyperplane represent the observed data.

Consider the data in the file `chap7data2.csv`:

> *It is known that mammals toxified by certain medicines, insecticide and carcinogens can be detoxified by certain enzymes. A study was subsequently performed on chickens: The response variable is the percentage detoxification of the poison Malathion; the explanatory variables are the activities of five enzymes.*

Use *R* to fit a multiple linear regression model of the percentage detoxification on the five enzymes. It is usually prudent to draw scatter plots of the response variable against each of the explanatory variables, to determine the nature of the relationship (if any) between them. In *R*, an easy way to visualise relationships between a (small) number of variables, is to use a *scatter plot matrix*.

Scatter plot matrix

```
plot(chap7data2)
```

or

```
scatterplotMatrix(~<variable 1>+<variable 2>+…,
 reg.line=lm,smooth=FALSE, spread=FALSE, span=0.5,ellipse=FALSE,
 levels=c(.5,.9),id.n=0,diagonal='density',data=<data set name>)
```

Multiple linear regression

```
<model name> <- lm(<response variable>~ . , data=<data set name>)
```

How do we choose which explanatory variables to keep in the model, and which to exclude? We rely on the *principle of parsimony*: We would like our final model to be as simple as possible, but not too simple. Various approaches can be followed in order to obtain a parsimonious model.

Variable selection (multiple linear regression)

```
stepwise(<full model name>, direction='backward/forward',
   criterion='AIC')
```

See `?stepwise` for other options that are available for the `direction` argument. There are four options in total, *backward*, *forward*, *backward/forward* and *forward/backward*.

```
regsubsets(<response variable >~<independent variable
   1>+<independent variable 2> + …,data=<data set
   name>,nbest=1,nvmax=6)
```

One can also fit all possible subsets and then choose the best model, based on a certain criteria. Load the script `allSubsets.R` into the *R* console. Execute the following code.

Constructing all the possible subsets with listed performance measures

```
allSubsets(<data set name>,y.name="<response variable
   name>",perf.measure="adj.r.squared")
```

Unlike R^2, AIC, BIC, Adjusted R^2 and Mallows' C_p penalise for the number of explanatory variables in the model. Mallow's C_p is given by:

$$C_p = \frac{SSE_p}{s_\varepsilon^2} - (n - 2p),$$

where s_ε^2 is the mean squared error from the model with the largest number of explanatory variables. For a good-fitting model, $C_p \approx p$, where $p = q + 1$ (i.e. the number of parameters, including the intercept, in the model under consideration).

Use R to find a parsimonious multiple linear regression model of the percentage detoxification. Investigate using the six variable selection methods mentioned above.

Diagnostic measures can be calculated in the same way as before. Perform a diagnostics check of the best fitting model of detoxification according to Mallows' C_p criterion.

The underlying *assumptions* (normality, homoscedasticity and independence of errors) can be tested in the same way as before:

- **Normality:** test the normality of the model residuals (all together).
- **Homoscedasticity:** check the plot of the residuals vs. predicted Y-values for any obvious patterns.
- **Independence:** only applicable for time series data (measurements taken serially over time).

Investigate the validity of the assumptions of the best fitting model of detoxification selected by using Mallows' C_p criterion.

If the assumptions hold, *predictions* can be made exactly as before. Using the best fitting model of detoxification selected by using the *AIC* criterion, predict the percentage detoxification when the five explanatory variables have values 270, 290, 140, 320 and 95, respectively. Interpret both the 95% prediction interval and the 95% confidence interval for this prediction.

As a further example, consider the data in the file `chap7data3.csv`:

> *X_1 is the amount of tricalcium aluminate, X_2 is the amount of tricalcium silicate, X_3 is the amount of tetracalcium alomino ferrite, X_4 is the amount of dicalcium silicate, and Y is the heat evolved in calories per gram of cement.*

Find a parsimonious multiple linear regression model of Y on the four explanatory variables. Use Mallows' C_p criterion.

4. Exercises

Exercise 1

With the data from C&S Example 8.1, p. 90 (`chap7exer.csv`): Fit a quadratic regression model of wheat yield (Y) on nitrogen fertilizer (X) and give the equation of the regression line, as well as the coefficient of determination (R^2). Also check that the assumptions of this polynomial regression model are justified. Does the quadratic model fit the data better than the simple linear regression (straight-line) model?

Exercise 2

The file `chap7exer2.csv` contains sales data for a number of pharmacies. Fit a multiple linear regression model of Volume (i.e. the sales volume) on the other variables (excluding the pharmacy number), and find the best fitting model according to Mallows' C_p criterion. For this final model:

- Test the model assumptions
- Interpret the regression coefficients (parameter estimates)
- Interpret the coefficient of determination (R^2 value)
- Predict the sales volume for a new pharmacy with
  * `Floor space` $= 4500$
  * `Prescription sales` $= 25$
  * `Parking` $= 42$
  * `Shopping centre` $= 1$
  * `Income` $= 13$

Chapter 8 Analysis of covariance

1. Introduction

In this chapter, we study analysis of covariance (ANCOVA, or also referred to as ANOCOVA). As you work through the chapter, make sure to note:

- the links with both linear regression and analysis of variance (ANOVA);
- how the inclusion of a covariate increases the precision of the analysis
- and the underlying assumptions of the ANCOVA method.

Background reading material

- Clewer and Scarisbrick: Chapter 17

2. The basics (C&S 17.3, p. 260)

Suppose that five sheep with a mean age of 55 weeks received a basic ration (Feed A), and has a mean mass of $\bar{Y}_A = 141kg$. Further suppose a second group of sheep with a mean age of 65 weeks received a vitamin B12 supplement together with the basic ration (Feed B), and the mean mass of this group is $\bar{Y}_B = 170kg$. The mass difference between the two groups is:

$$\bar{Y}_B - \bar{Y}_A = 29kg$$

which is the effect of

(Age Contribution) + (B12 contribution).

However, the effect of age and B12 cannot be distinguished from each other. We say that the effect of B12 is *confounded* with the effect of age in the observed difference, $\bar{Y}_B - \bar{Y}_A$. In order to obtain an unconfounded estimate of the effect of B12, we must make use of Y values that had been adjusted for the differences in age. If we were to know what the relationship between age and mass is, we could, for example, predict what the mass of each group should be, given mass $= 60$ weeks, so that the difference between the two values would reflect the treatment difference only.

In the analysis of covariance, a set of treatments are compared with each other *after* an adjustment has been made for differences that exist between experimental units that can be attributed to a different measurable factor, rather than to the treatments. In the case of the example above we would include the age of the sheep in the analysis as a so-called *covariate*.

The assumption is that there is a linear relationship between the covariate (call it X) and the response variable (call it Y), and that this relationship is used to do the adjustment. This analysis thus combines the theory of the general linear model (and ANOVA) and linear regression.

Although an analysis of covariance can be executed with any design, it is often used as an alternative to the randomised block design.

3. Analysis of covariance: completely randomised design (C&S 17.4, p. 260)

Suppose a completely randomised design in which a second variable, other that the variable in which we are interested, is measured. For example, y_{ij} may be the yield of a crop observed in an experiment where three treatments are compared. A model that supposes differences between treatment means for these data is as follows:

$$Y = \mu_i + \varepsilon \qquad (8.1)$$

Suppose that an additional set of observations, namely the yield of weeds on the plots, X, is also available. The amount of weeds may have an effect on yield in the sense that they compete with the crop for sunlight, water, nutrients and space:

Group (i)	1		2		3	
Variables	X	Y	X	Y	X	Y
Observations	0.4 0.3 0.2 0.5	5 2 3 1	0.1 0.8	4 7	0.9 0.6 0.7	6 9 8
Totals	1.4	11	0.9	11	2.2	23

If we wish to investigate a linear relationship between X and Y in the first group, we can specify the linear model as:

$$Y = \mu_1 + \beta_1 \mathbf{X} + \varepsilon \qquad (8.2)$$

where β_1 is the slope in group 1. Similarly for the other two groups. If one is satisfied that the slope (or trend) with regard to the weeds is the same for the three groups, it leads to the model:

$$Y = \mu_i + \beta \mathbf{X} + \varepsilon \qquad (8.3)$$

where β is the *common* slope (i.e. the slope is equal for all the groups). Not that a model with different treatment means for each of the groups supposes that the *intercepts* of the regression lines differ from each other. As before, we want to fit a series of models and consider the goodness of fit of the models, as one can then make use of convenient ANOVA techniques. A sequence of models can be fitted to test for different intercepts (i.e. differences between the treatment means) and different slopes (i.e. differences between the groups with regard to the effect of the covariate on the response):

- **Common intercept / mean, no covariate effect:** Treatments do not differ, covariate does not have an effect on the response.
$$Y = \mu + \varepsilon \qquad (8.4)$$
- **Common intercept, common slope:** Treatments do not differ, but the covariate has an effect on the response.
$$Y = \mu + \beta \mathbf{X} + \varepsilon \qquad (8.5)$$
- **Different intercepts, common slope:** Treatments differ, but the effect of the covariate on the response is the same for all treatment groups (parallel lines).
$$Y = \mu_i + \beta \mathbf{X} + \varepsilon \qquad (8.6)$$

- **Common intercept, different slopes:** Treatments do not differ, but the effect of the covariate on the response differs between the groups (different trends).

$$Y = \mu + \beta_i \mathbf{X} + \varepsilon \tag{8.7}$$

- **Different intercepts, different slopes:** Treatments differ, and the effect of the covariate on the response also differs between the groups.

$$Y = \mu_i + \beta_i \mathbf{X} + \varepsilon \tag{8.8}$$

Analysis of covariance can thus also be used to test for *differences in trend* in different groups. The linear model for ANCOVA (i.e. the different intercepts, common slope model) as discussed in this chapter is based on the following assumptions:

- **Normality:** The residuals from the model should be normally distributed.
- **Homoscedasticity:** The residuals should have the same variability across the range of fitted values.
- **Common slope:** The slopes of the covariate regression lines are equal for all the groups (only applicable if a "common slope" model is fitted).

8.3.2 Learning by example

Consider the data in the file `chap8data.csv`.

> *In a completely randomised design, three treatments (A, B and C) were each applied to ten patients for the treatment of leprosy. Scores for leprosy bacilli before (X) and after (Y) treatment were recorded.*

Use *R* to perform an ANOCOVA in the following way:

Analysis of covariance (ANCOVA):

What is your conclusion about the effectiveness of the three leprosy treatments?

Common slope assumption for the *leprosy* data (`chap8data1.csv`)

```
chap8data1.model2 <- lm(Y ~ X + Treatment +X*Treatment,
   data=chap8data1)
summary(chap8data1.model2)
```

Do the normality, homoscedasticity and common slope assumptions seem justified for the leprosy data?

Next, consider the data in the file `chap8data2.csv`:

> *Three treatments (A, B and C) were allocated completely at random to each of five lambs. The mass increase per day (Y) of each lamb was then measured over a specific time period. The initial ages (X) were also recorded.*

Perform an ANCOVA to determine whether the effects of the three treatments on the mass increase of the lambs differ, using the initial age as a covariate in the model. Did the initial age of the lambs have an effect on their mass increase? Do the three treatments differ with regard to the mass increase? Are the assumptions of the linear model for the ANCOVA justified?

To do multiple comparison tests after the ANCOVA (to determine which of the treatment means differ significantly from each other), use the following steps:

Multiple comparison tests (after ANCOVA)

Get the *lsmeans-script.R* file (you need to do this only once in your *R* session):

Download the *lsmeans-script.R* file from the SUNLearn page.
Open the file as a script in *R*.
Submit the contents of the *lsmeans-script.R* file in *R*.

To perform the multiple comparisons tests:

Submit the script in the *R* console window.

```
>LSmultcomp(<model name>,which=”<treatment variable>”)
```

4. Covariance analysis of other experimental designs

An analysis of covariance can be executed with any design, as for example with the randomised block design or the Latin square design.

Investigate how to use *R* to perform an ANCOVA for the randomised block design experiment in the file in `chap8data3.csv`:

> *Y is the corn yield for six different corn varieties (labelled A – F) planted in four blocks. The stand (variable X) is included as a covariate in the analysis.*

Do the corn varieties differ significantly from each other? Which variety(-ies) can be used to maximise the corn yield? Does the stand (covariate) have an effect on the corn yield?

5. Exercises

Exercise 1

With the data from C&S Table 17.1, p. 257 (`chap8exer1.csv`):

- Perform an ANOCOVA to determine whether the leaf area (Y) differs for the two fertilizers. Give the resulting `Type III SS` table.
- Clearly state your conclusions with regard to the treatments and the covariate.

Exercise 2

With the data from C&S Example 17.2, p. 261 (`chap8exer2.csv`):

- Perform an ANCOVA to determine whether the present wheat yield (Y) differs for the three treatments. Give the resulting `Type III SS` table.
- Carry out multiple comparison tests to see which of the treatments differ from each other.
- Check whether the model assumptions (normality, homoscedasticity, common slope) are justified.

Chapter 9 - Analysis of variance for repeated measures

1. Introduction

In this chapter, we study the ANOVA for experiments with *repeated measures*. As you work through the chapter, make sure to note:

- That repeated measurements (per subject) are dependent, and this dependence should be taken into account when constructing linear models on such data.
- The difference between *fixed* effects (controlled by the experimenter) and *random* effects (usually related to the subject).
- How dependence in repeated measurements can be modelled using various covariance structures.
- How the goodness-of-fit of various models can be compared using the AIC (remember: a *smaller* AIC value indicates a better fit).

Background reading material

- Clewer and Scarisbrick: p. 252 – 253
- https://ourcodingclub.github.io/tutorials/mixed-models/#what

2. Learning by example

Consider the data in the file `chap9data1.csv`:

> *The data show the masses of 10 calves at four different time points (months 3, 4, 5 and 6), with the calves having been put on one of two different feeds.*

The question is whether or not the weights of the calves differ significantly, depending on the feed.

Unlike the data analysed in previous chapters, we have to take special care here because the weight of a calf in a particular month is likely to be correlated to some extent to its weight in the previous month(s). Therefore, at least some of the observations are *dependent* on some other observations. In the techniques studied earlier, independence was one of the underlying assumptions.

It is important to recognise the role that each variable plays in this example:

- *Weight* is the numerical response (dependent) variable.
- *Food*, *Calf* and *Month* are all explanatory variables. We will model them as categorical (classification) variables.
- *Food* is the treatment variable, manipulated by the experimenter. We will thus model it as a *fixed effect*.
- *Calf* represents the subjects (the experimental units). The weight measurements were made on the calves. As the experimenter had little control over the characteristics of the specific calves on which the experiment was performed, and the calves in the sample are a small randomly chosen part of a much larger population, we will model *Calf* as a *random effect*.
- The repeated measures occur over the variable *Month*. *Month* therefore represents the within-subject effect. As the experimenter had control over which months to make the measurements, we will model *Month* as a *fixed effect*.

Import the data into *R*. Remember to code the explanatory variables as factors. Use the following steps to perform the ANOVA for the repeated measures design:

ANOVA for repeated measures for Example 1

```
chap9data1.model1 <- lme(Weight~Food + Month +
   Food:Month,random=~1|Calf,data=chap9data1)

summary(chap9data1.model1)

Anova(chap9data1.model1)
```

In the output, take note of the estimated correlation structure as well as the AIC value. The *Akaike Information Criterion* (AIC) can be used to compare various models to one another.

With the dependence between observations now having been taken into account, hypotheses concerning the fixed effects (`Food`, `Month` and `Food*Month`, for instance) can be tested in exactly the same way as described in the previous chapters concerning ANOVA. If significant differences are found in these effects, post-hoc multiple comparison tests can also be performed as before.

Multiple comparison tests (repeated measures design)

```
LSmultcomp(chap9data1.model1,which="Month")
```

Various graphs can be drawn to further elucidate.

Investigate the calf data fully, also taking into account the alternative correlation structures below.

3. Exercises

Exercise 1

The `chap9exer1.csv` file contains a sample of 30 participants whose interest in voting was measured at three different ages (10, 15 and 20 years). The interest values are represented on a scale that ranges from 1 to 5 and indicate how interested each participant was in voting at each given age. (Source: http://www.r-bloggers.com/r-tutorial-series-one-way-repeated-measures-anova/;downloaded on 31 January 2014)

Perform an ANOVA to determine whether the voting interest of the participants changed over times.

Exercise 2

The `chap9exer2.csv` file contains part of the data for a study of oral condition of cancer patients conducted at the Mid-Michigan Medical Centre. The oral conditions of the patients were measured and recorded at the initial stage, at the end of the second week, at the end of the fourth week, and at the end of the sixth week. The variables age, initial weight and initial cancer stage of the patients were also recorded. Patients were divided into two groups at random: One group received a placebo and the other group received aloe juice treatment. (Source: Mid-Michigan Centre, Midland, Michigan, 1999: A study of oral condition of cancer patients.)

The variables in the data set are:

- `ID:` Patient ID
- `Treat:` Treatment group, either 0 = placebo, or 1 = aloe juice
- `Time:` Time point when the oral condition was measured
- `Condition:` Oral condition of the patient

Perform an ANOVA to determine whether there was a difference in oral condition between the (aloe juice) treatment and control groups. Also add a $Treat \times Time$ interaction term to the linear model.

Chapter 10 Categorical data analysis

1. Introduction

In this chapter, we will discuss the analysis of count data, specifically the use of *contingency tables* and *logistic regression*. As you work through the chapter, make sure to note:

- The properties of binary data (i.e. variables that can only take one of two values, eg. 0 or 1) and why it cannot be analysed like numerical data.
- The use of contingency tables to summarise count data and test for association between two (or more) variables.
- That the logistic regression model is a generalised linear model (as is simple linear and multiple linear regression), but with a different *link function.*

Background reading material

- Biometry 212: note set 13
- Clewer and Scarisbrick: Chapter 18
- Quinn and Keough: Extract from Chapter 13, p. 360 – 371 (available on SUNLearn)

2. Contingency tables

Up to this point we have concerned ourselves with data measured on a numerical scale (interval or ratio). Such data can be represented on a numbered line, and therefore has arithmetic meaning.

However, we often encounter situations where the levels of the variable are identified by names or ranks only (nominal or ordinal), and one may be interested in the number of observations that occur at each level of the variable. We refer to such data as *categorical data* or *counts*. Categorical data are displayed in the form of *contingency tables* (so called because they often originate from surveys). Such a table served as a summary of one or more variables in the form of a one or more directional frequency distribution. Consider the case where the students in a class are classified on the grounds of gender and whether they are colour blind, or not. Such variables that can assume one of two possible values are known as *binomial* variables.

Gender	Sight		Total
	Colour blind	Not colour blind	
Male	45	126	171
Female	30	212	242
Total	75	338	413

The above table is knows as a 2×2 (number of rows $\times$ number of columns in general $r \times c$) contingency table, in which each member of the group is classified according to two factors or attributes with two classes each. For a 2×2 contingency table such as the example above, we can calculate the *odds* of, for example, a male being colour blind, as:

$$\frac{f_{11}}{f_{22}} = \frac{45}{126} = 0.357,$$

which means that, if we know that a person is male, the probability of him being colour blind is 64% (i.e. $(1 - 0.357) \times 100$) less than the probability that he is not colour blind. We can also calculate the odds of a female person being colour blind as:

$$\frac{f_{21}}{f_{22}} = \frac{30}{212} = 0.142.$$

The *odds ratio* for colour blindness with regard to gender is defined as:

$$\frac{f_{11}}{f_{12}} \div \frac{f_{21}}{f_{22}} = \frac{0.357}{0.142} = 2.52,$$

which can be interpreted as follows: The odds of a male being colour blind is about 2.5 times the odds of a female being colour blind.

By using the marginal totals and grand total of a contingency table, we can calculate the *expected frequency* (under the hypothesis that the two factors in the table are independent of each other) of the cell in the i^{th} and j^{th} column as:

$$e_{ij} = \frac{f_{i\bullet} \times f_{\bullet j}}{f_{\bullet\bullet}}, \tag{10.1}$$

Where $f_{i\bullet}, f_{\bullet j}$ and $f_{\bullet\bullet}$refers to the observed i^{th} row total, observed j^{th} column total and observed grand total, respectively.

Calculate the expected frequencies for the *Gender vs. Sight* table.

We can calculate the following chi-squared test statistic (equation 10.2):

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}}$$

where

- f_{ij} is the observed cell frequency;
- e_{ij} is the expected cell frequency;
- r is the number of rows in the contingency table;
- c is the number of columns in the contingency table;

to test the hypothesis:

H_0: There is no association between the row and column factors (*independent factors*).

H_1: There is an association between the row and column factors (*dependent factors*).

Under the null hypothesis, the test statistic in (10.2) is χ^2 distributed with $(r-1) \times (c-1)$ degrees of freedom. However, the chi-square test of association is only valid for sufficiently large samples. If any of the expected cell frequencies are smaller than five, *Fisher's exact test* should rather be used.

10.2.1 Learn by example

Consider the data in the file `chap10data1.csv`:

> *In a completely randomised experiment, a woman tasted eight cups of tea: four of which the milk had been added first and four of which the tea had been added first. She knew there were four cups of each type and had to taste/guess which four had the milk added first. The order of presenting the cups to her was randomised.*

To test whether the woman could really taste which cups had the milk added first (i.e. that there is an association between the order of the milk being added and her tasting/guessing the order correctly), use *R* to analyse the data as follows:

Chi-square test of association (for *chap10data1.csv*)

```
chap10.data1.table1 <- xtabs(~Poured+Guess, data=chap10data1)

chap10.data1.test1 <- chisq.test(chap10.data1.table1,
 correct=FALSE)
```

Interpret the estimated odds ratio for the tea pouring data. What is your conclusion about the woman's tasting ability?

For another exercise, consider the data in the file `chap10data2.csv`:

> *The number of attacks by Great White sharks on South Africans have been recorded by province and the type of activity the victim was involved in at the time of the attack, for the period up to the year 2000.*

Test the hypothesis that there is no association between the type of activity and province, with regards to the shark attack victims.

3. Logistic regression (Quinn & Keough 13.2, p. 360 – 371)

One very important application of linear models in biology is to model response variables that are binary (eg. presence/absence, alive/dead). The explanatory variables can be either continuous and/or categorical. Because of the binary nature of the response variable, we need to use a technique called *logistic regression*, which is a type of *generalised linear model (GLM)*. This type of model estimates $\pi(x)$, the probability that the binary response (Y) equals one for a given value of the explanatory variable (X).

The logistic regression model is nonlinear, with a sigmoidal shape. The change in the probability that Y equals one for a given change in X is greatest for values of X near the middle of the range, rather than for values at the extremes.

The error terms from the logistic model have a *binomial distribution* – they are not normally distributed as for simple linear and multiple linear regression. Therefore ordinary least squares (OLS) estimation of the model is *not* appropriate, and we need to use *maximum likelihood (ML)* estimation.

10.3.1 Simple logistic regression (Quinn & Keough 13.2.1, p. 360)

Suppose the purpose is to model the probability that a binary variable Y (taking the values 0 or 1) will be equal to one, using a single explanatory/independent variable X. The *simple logistic regression* model is:

$$\pi(x) = \frac{e^{\beta_0+\beta_1\mathbf{X}+\varepsilon}}{1+e^{\beta_0+\beta_1\mathbf{X}+\varepsilon}}, \qquad (10.3)$$

where

- $\pi(x)$ is the probability that Y equals one for a given value of X;
- e is the constant $e = 2.718\ldots$;
- β_0 is the *intercept* (constant);
- β_1 is the *slope*, which measures the rate of change in $\pi(x)$ for a given change in X;
- ε is the unexplained error variation.

To fit the logistic model, $\pi(x)$ is transformed with the *logit* link function (i.e. the natural logarithm of the odds),

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \qquad (10.4)$$

so that the right-hand side of the logistic model closely resembles the familiar linear model,

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1\mathbf{X} + \varepsilon. \qquad (10.5)$$

The logit link function does two important things:

1 It ensures that $\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$ ranges from $-\infty$ to $+\infty$ whereas $\pi(x)$ is constrained to the $[0,1]$ interval.
2 It models the binomial distribution of the errors.

The interpretation of the β_1 coefficient is that, for each unit increase in the value of X, the odds of Y being equal to one changes with a factor of e^{β_1}.

To estimate the probability of Y being equal to one (i.e. $\pi(x)$) for a given value of X, we need to estimate the values of β_0 and β_1 and substitute the given X-value into equation (10.3).

10.3.2 Learn by example

Consider the data in the file `chap10data3.csv`:

> *The birth weights (in gram) were recorded for 2950 babies born in a number of South African hospitals during the year 2009. It was also recorded whether each of the babies died (Yes/No) before their discharge from the hospital.*

Use the following steps in *R* to fit a simple logistic regression model to predict the probability of a baby dying before discharge from the hospital:

Simple logistic regression

```
<model name> <- glm(<response variable> ~ <independent variable>,
   family=binomial(), data=<data set name>)
summary(<model name>)
```

ROC curve

Submit the code in the script file *ROC-curve.R* from the *SUNLearn* page. Submit the code in the *R* console. Then run the code:

```
ROC.curve(<model name>,response=<data set name>$<response variable>)
```

Does birth weight have an effect on a baby's probability of dying? Interpret the coefficient of birth weight. How do you interpret the ROC curve?

The *receiver operating characteristic (ROC)* curve provides a way to assess the goodness of the model fit. The *area under the curve (AUC)* is calculated, and can obtain a maximum value of 1. The greater the AUC, the better the model fits the data. The AUC can thus be used to compare different models fitted to the same data set.

Another model fit statistic is the AIC value, which is also given in the *R* output. In the same way as for other linear models, a smaller AIC value indicates a better fit.

10.3.3 Multiple logistic regression (Quinn & Keough 13.2.2, p. 365)

Logistic regression can easily be extended to situations with multiple explanatory variables. The general *multiple logistic regression* model for q explanatory variables is,

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + \varepsilon. \tag{10.6}$$

The interpretation of the $\beta_j, j = 1, \ldots, q$ coefficients is that, for each one unit increase in the value of X_j (while holding the remaining explanatory variables constant), the odds of Y being equal to one changes with a factor of e^{β_j}.

Categorical variables can also be incorporated in the logistic regression model by converting them to dummy variables. *R* does this automatically for any variables specified as `factor type` variables.

To estimate the probability of Y being equal to one (i.e. $\pi(x)$) for given values of the X variables, we need to estimate the values of all $q + 1$ of the β_j coefficients and substitute the given X_j-values into the following equation:

$$\widehat{\pi(x)} = \frac{e^{b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_q X_q}}{1 + e^{b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_q X_q}} \tag{10.7}$$

10.3.4 Learn by example

Consider the data in the file `chap10data4.csv`:

> *The delivery mode (Caesarean or Vaginal delivery), birth weights (in gram), Apgar score at 1 minute after birth (on a scale of 0 – 10), and temperature (in degrees Celsius) were recorded for 2950 babies born in a number of South African hospitals during the year 2009. It was also recorded whether each of the babies died (Yes/No) before their discharge from the hospital.*

Use the following steps in *R* to fit a multiple logistic regression model to predict the probability of a baby dying before discharge from the hospital:

Multiple logistic regression:

```
chap10data4.model1 <- glm(Died ~ ., family=binomial(),
   data=chap10data4)
summary(chap10data4.model1)
```

As in the case of multiple linear regression (Chapter 7), stepwise selection procedures can be used to find a good fitting multiple logistic regression model. To perform stepwise selection on a (full) fitted multiple logistic regression model, follow the steps outline in Chapter 7.

Which of the variables have an effect on a baby's probability of dying?

4. Exercises

Exercise 1

With the data from C&S Example 18.12, p. 284 (`chap10exer1.csv`): Test whether seed colour and petal colour in the field beans are associated. State the hypothesis being tested, and give the test statistic and *p*-value from the *R* output. Clearly state your conclusion.

Exercise 2

With the data from C&S Example 18.14, p. 288 (`chap10exer2.csv`): Test the null hypothesis that the proportion of cuttings which root is the same for both growth hormones, versus the alternative that they are not. Give the test statistic and *p*-value from the *R* output, and clearly state your conclusion.

Exercise 3

The file `chap10exer3.csv` contains the following variables recorded for 2842 babies born in South African hospitals during the year 2009:

- **Maternal race**: Black, White, Asian, Other
- **Delivery mode**: Caesarean, Vaginal
- **Gender**: Male, Female
- **Birth weight** in gram
- **Apgar score** at one minute (out of 10)
- **Temperature** of the baby within one hour from birth (in degrees Celsius)
- Whether the baby stayed in the hospital for more than 21 days (**Long Stay**: Yes, No)

Use stepwise selection (using *forward/backward*) to find a good fitting logistic regression model to predict the probability of a baby staying for more than 21 days in the hospital.

Interpret the effects on the odds ratio for all the variables in the final model. Also obtain the ROC curve for the final model, and calculate the area under the ROC curve (AUC value).

Bibliography

Clewer AG and Scarisbrick DH. 2001. *Practical statistics and experimental design for plant and crop science.* J.Wiley. ISBN 9780471899082.

Quinn GP and Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge University Press. ISBN 9780521009768.

Simonoff JS. 2003. *Analyzing categorical data.* Springer texts in statistics. Springer. ISBN 9780387007496.

Zar JH. 2014. *Biostatistical Analysis*, Fifth Edition. Pearson Education Limited, UK. ISBN 9781292024042