R Workshop - Quantitative Science Course Union - UBCO

W. J. Braun

November 19, 2018

Contents

Preface

This short book uses R and several examples to illustrate common statistical situations or vignettes. After a short mention of R and RStudio, the book launches into brief discussions of correlation, contingency tables, t-tests, ANOVA, simple and multiple regression. The assumption of independence between observations is key to the success of these methods, and the assumption of normality also gives many of these methods their mathematical accuracy.

Classical nonparametric methods will also be described as well. These methods are often over-rated, or more accurately, the methods they replace may be under-rated in terms of their robustness to model misspecification. The nonparametrics are not a panacea and really only alleviate misspecification of normality, which is often not a critical assumption.

A vastly more important nonparametric method is bootstrap resampling (and this has recently been recognized through the awarding of the International Statistics Prize in 2018 to Bradly Efron, for his seminal 1979 paper in the area). In order to understand bootstrapping, and as an aid in understanding much of statistical modelling if one wants to avoid the mathematics, simulation is a very useful tool. Notes on the use of simulation tools find their place in the pages that follow.

1

An Overview of R

R is based on the computer language S, developed by John Chambers and others at Bell Laboratories in 1976. Robert Gentleman and Ross Ihaka developed an implementation, and named it R. Gentleman and Ihaka made it open source in 1995, and hundreds of people around the world have contributed to its development.

Although it may be hard for students with little mathematical or computing background to believe, R and RStudio are actually quite friendly tools, but becoming acquainted with them requires a bit of effort. A few hours of playing with R code is all that is really required to achieve modest expertise. Perhaps the most important thing to remember is that there is nothing wrong with making errors when using a programming language like R. You learn from your mistakes, and there is no harm done. Experimentation is the key to learning R, just as it has been the key to science for the past 400 years.

1.1 Downloading and Installing R and RStudio

R can be downloaded for free from `http://cloud.r-project.org`. A *binary version* is usually simplest to use and can be installed in Windows and Mac fairly easily. A binary version is available for Windows Vista or above from the web page `http://cloud.r-project.org/bin/windows/base`. The "setup program" is usually a file with a name like `R-3.4.4-win.exe`. Clicking on this file will start an almost automatic installation of the R system. Clicking "Next" several times is often all that is necessary in order to complete the installation. An R icon will appear on your computer's desktop upon completion.

RStudio is also recommended. You can download the "Open Source Edition" of "RStudio Desktop" from `http://www.rstudio.com/`, and follow the instructions to install it on your computer.

1.2 Key Features of R

1.2.1 Packages

You can do many things with base R, but one of the true strengths of R is the availability of add-on packages. There are literally thousands of *packages*, i.e. `graphics`, `ggplot2`, and `MPV`. A package contains functions and data which extend the abilities of R. Every installation of R contains the base packages (e.g. `base`, `stats`, `graphics`) which are automatically loaded when you start R.

To load an additional package, say, called *DAAG*, type

```
library(DAAG)
```

If you get a warning that the package is can't be found, then the package doesn't exist on your computer, but it can likely be installed. Try

```
install.packages("DAAG")
```

In RStudio, use the `Packages` menu.

1.2.2 Calculations in R

The > sign tells you that R is ready for you to type in a command. For example, you can do arithmetic of any type, including multiplication:

```
1111*1111

## [1] 1234321
```

By hitting the "Enter" key, you are asking R to execute this calculation. Often, you will type in commands such as this into a script window or "pane", as in RStudio, for later execution, through hitting "ctrl-R" or another related keystroke sequence.

You can control the number of digits in the output with the `options()` function. This is useful when reporting final results such as means and standard deviations, since including excessive numbers of digits can give a misleading impression of the accuracy in your results. Compare

```
583/31

## [1] 18.80645
```

with

```
options(digits=3)
583/31

## [1] 18.8
```

1.2.3 Data frames

Most data sets are stored in R as data frames. These are like matrices, but with the columns having their own names.

An example of a data frame that is built into R, or at least into the *datasets* package is `women`, and it can be printed to the screen by typing its name:

```
women

##    height weight
## 1      58    115
## 2      59    117
## 3      60    120
## 4      61    123
## 5      62    126
## 6      63    129
## 7      64    132
## 8      65    135
## 9      66    139
## 10     67    142
## 11     68    146
## 12     69    150
## 13     70    154
## 14     71    159
## 15     72    164
```

You can obtain information about this data frame by typing `help(women)`. This will be the only time in these pages that a data frame will be displayed in full. It is unwise to inspect data frames in this way, as it is far better to use graphical means to display large amounts of data. The `summary()` function provides numerical information which is also a better way of determining the main features of a data frame before doing further exploration of it:

```
summary(women)

##      height          weight
##  Min.   :58.0   Min.   :115
##  1st Qu.:61.5   1st Qu.:124
##  Median :65.0   Median :135
##  Mean   :65.0   Mean   :137
##  3rd Qu.:68.5   3rd Qu.:148
##  Max.   :72.0   Max.   :164
```

Columns can be of different types from each other. An example is the built-in `chickwts` data frame:

```
summary(chickwts)

##      weight           feed
##  Min.   :108   casein   :12
##  1st Qu.:204   horsebean:10
##  Median :258   linseed  :12
##  Mean   :261   meatmeal :11
##  3rd Qu.:324   soybean  :14
##  Max.   :423   sunflower:12
```

If you want to see the first few rows of a data frame, you can use the `head()` function:

```
head(chickwts)

##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean
```

The `tail()` function displays the last few rows. The number of rows can be determined using the `nrow()` function:

```
nrow(chickwts)

## [1] 71
```

Similarly, the `ncol()` function counts the number of columns.

1.2.4 Reading data into a data frame

You will usually have a data set that you wish to read into R. If you have prepared it yourself, you could simply type it into a text file, for example called `mydata.txt`, perhaps with a header indicating column names, and where you use blank spaces to separate the data entries. The `read.table()` function will read in the data for you as follows:

```
mydata <- read.table("mydata.txt", header = TRUE)
```

The object `mydata` now contains the data read in from the external file. You could use any name that you wish in place of `mydata`, as long as the first element of its name is an alphabetic character.

If the data entries are separated by commas and there is no header row, as in the file $wx_l3_2006.txt, youwouldtype$:

```
wx1 <- read.table("wx_l3_2006.txt", header=F, sep=",")
```

Often, your data will be in a spreadsheet. If possible, export it as a `.csv` file and use something like the following to read it in.

```
wx2 <- read.table("wx_l3_fwi_2006-2011.csv", header=F, sep=",")
```

If you cannot export to `.csv`, you can leave it as `.xlsx` and use the `read.xslx()` command in the xlsx package.

Most likely, the data file that you have is not very clean in that there could be missing values or blank spaces in awkward locations, and so on. When reading in a file with columns separated by blanks with blank missing values, you can use code such as

```
dataset1 <- read.table("file1.txt", header=TRUE, sep=" ", na.string=" ")
```

This tells R that the blank spaces should be read in as missing values. Observe the contents of `dataset1`:

```
dataset1
```

```
##    x  y   z
## 1  3  4  NA
## 2 51 48  23
## 3 23 33 111
```

Note the appearance of `NA`. This represents a missing value.

Sometimes, external software exports data files that are tab-separated. When reading in a file with columns separated by tabs with blank missing values, you could use code like

```
dataset2 <- read.table("file2.txt", header=TRUE, sep="\t", na.string=" ")
```

Again, observe the result:

```
dataset2
```

```
##    x   y  z
## 1 33 223 NA
## 2 32  88  2
## 3  3  NA NA
```

If you need to skip the first 3 lines of a file to be read in, use the `skip=3` argument.

1.2.5 Extracting information from data frames

To extract the `height` column from the `women` data frame, use the `$` operator:

```
women$height

##  [1] 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
```

If you want only the chicks who were fed horsebean, you can apply the `subset()` function to the `chickwts` data frame:

```
chickHorsebean <- subset(chickwts, feed == "horsebean")
chickHorsebean

##    weight      feed
## 1     179 horsebean
## 2     160 horsebean
## 3     136 horsebean
## 4     227 horsebean
## 5     217 horsebean
## 6     168 horsebean
## 7     108 horsebean
## 8     124 horsebean
## 9     143 horsebean
## 10    140 horsebean
```

You can now calculate the mean and standard deviation, and so on, of these weights:

```
mean(chickHorsebean$weight) # mean

## [1] 160.2

sd(chickHorsebean$weight)    # standard deviation

## [1] 38.626
```

In order to extract the 4th row from the `chickHorsebean` data frame, type

```
chickHorsebean[4, ]

##   weight      feed
## 4    227 horsebean
```

To extract the element in the 2nd column of the 7th row of `women`, type

```
women[7, 2]

## [1] 132
```

If we want the elements in the 4th through 7th row of the 2nd column of `women`, we can use

```
women[4:7, 2]

## [1] 123 126 129 132
```

Note the use of the `:` operator:

```
4:7

## [1] 4 5 6 7
```

Another built-in data frame is `airquality`. If we want to compute the mean for each of the first 4 columns of this data frame, we can use the `sapply()` function:

```
sapply(airquality[, 1:4], mean)

##   Ozone Solar.R    Wind    Temp
##      NA      NA  9.9575 77.8824
```

The `sapply()` function applies the same function to all columns of the supplied data frame. Note also the very useful functions in Wickham's (2011) plyr package.

1.2.6 Factors

Factors offer an alternative, often more efficient, way of storing character data. For example, a factor with 6 elements and having the two levels, `control` and `treatment` can be created using:

```
grp <- c("control", "treatment",  "control", "treatment", "treatment", "control")
grp

## [1] "control"   "treatment" "control"   "treatment" "treatment" "control"
```

```
grp <- factor(grp)
grp

## [1] control   treatment control   treatment treatment control
## Levels: control treatment
```

Factors

Consider the built-in data frame `InsectSprays`

```
summary(InsectSprays)

##      count       spray
##  Min.   : 0.0   A:12
##  1st Qu.: 3.0   B:12
##  Median : 7.0   C:12
##  Mean   : 9.5   D:12
##  3rd Qu.:14.2   E:12
##  Max.   :26.0   F:12
```

The second column of this data frame is a factor representing the different types of spray used in the associated experiment. The levels of this factor can be listed using the `levels()` function:

```
levels(InsectSprays$spray)

## [1] "A" "B" "C" "D" "E" "F"
```

Factors are a more efficient way of storing character data when there are repeats among the vector elements. This is because the levels of a factor are internally coded as integers.

To see what the codes are for our factor, we can type

```
as.integer(InsectSprays$spray)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3
## [36] 3 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6
## [71] 6 6
```

The labels for the levels are only stored once each, rather than being repeated. We can change the labels for the factor using the `levels()` function as follows:

```
levels(InsectSprays$spray)[3] <- "Raid"
```

Observe the effect of the change in

```
summary(InsectSprays$spray)
```

```
##    A    B Raid    D    E    F
##   12   12   12   12   12   12
```

The `levels()` function also offers a simple way to collapse categories. Suppose we are interested in comparing the first three levels with the last three levels. We can create a new factor for this purpose as follows:

```
InsectSprays$newFactor <- InsectSprays$spray
levels(InsectSprays$newFactor) <- c("A", "A", "A", "B", "B", "B")
```

Check the result:

```
summary(InsectSprays)
```

```
##      count       spray    newFactor
##  Min.   : 0.0   A   :12   A:36
##  1st Qu.: 3.0   B   :12   B:36
##  Median : 7.0   Raid:12
##  Mean   : 9.5   D   :12
##  3rd Qu.:14.2   E   :12
##  Max.   :26.0   F   :12
```

1.2.7 Histograms and simulated normal data

The `hist()` function can be used to draw histograms, and the `rnorm()` function can be used to simulate draws from a normal distribution[1]

A standard normal random variable has a mean of 0 and a standard deviation of 1. Figure **??** shows the results from simulating 2000 standard normal variates, together with a plot of the normal probability density curve, obtained from the `dnorm()` function. Note that we have used the `curve()` function with the `add` argument to overlay the curve. The `col` parameter controls the colour.

```
Z <- rnorm(2000)
hist(Z, prob = TRUE)
curve(dnorm(x), from = -3, to = 3, add = TRUE, col = "blue")
```

[1] often used as a model for noise

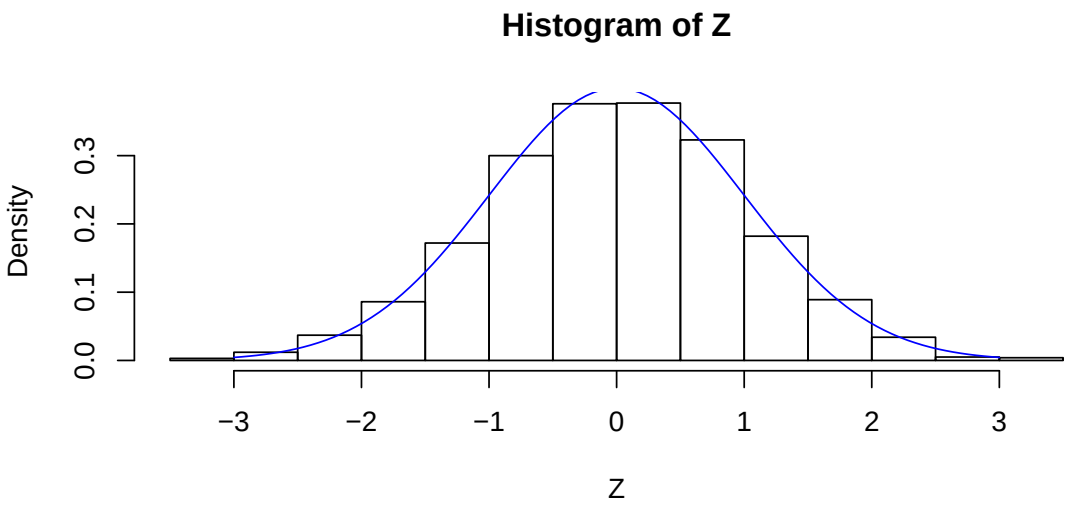

Figure 1.1: Histogram of 2000 standard normal random variates with overlaid density curve (in blue).

1.3 Sources of additional information

John Maindonald has written a comprehensive introduction and overview of R which is a very useful reference for scientists. It can be found at
`https://www.researchgate.net/publication/228702931_The_R_System-An_Introduction_and_Overview`

A handy reference card has been constructed by Jonathan Barron and is available at
`http://www.psych.upenn.edu/~baron/refcard.pdf`

2

An Overview of Statistical Modelling

To a lot of people, the field of statistics can sound frightening, since the underlying theory is based on some sophisticated mathematics which is not easily understood. When simplified to the point where most people can quickly and easily understand, it is then viewed as somewhat boring or unimaginative.

In reality, statistics is an important and powerful discipline which combines elements of art and science. From a certain perspective, it lies at the heart of the scientific method, since when approached properly, it refines the beliefs of an investigator who brings a certain level of knowledge (including possible errors in judgement) about a scientific problem. It does this by allowing the investigator to incorporate new information in the form of data, which may or may not be in numeric form. By appropriate use of probability, the level of uncertainty in the conclusions is measured, either through confidence intervals or p-values, or using a fully probabilistic approach referred to as Bayesian. The latter approach will not be pursued here, although not because it is not important in its own right.

2.1 Types of Data

Before taking measurements or observations on some type of phenomenon, they are unknown. A useful way of coping with this lack of knowledge is based on probability. Probability can allow us to quantify our uncertainty about measurements. For example, before throwing a six-sided die, we know that the number of spots that we will observe follows a specific probability distribution, and we refer to that number as a random variable, which we might refer to as Y, and we can say that the probability that $Y = 4$ is $1/6$, and that the probability that $Y = 7$ or $Y = 1.5$ is 0.

The number of spots on the die, Y, is an example of a count, a type of numeric variable. The number of heads, H, in one toss of a coin is another example of a count, but this time with only two possibilities $H = 0$ or $H = 1$. H is an example of a binary random variable or indicator variable. If you think about it, the numbers 0 or 1 are not actually observed, but rather the head or tail. Therefore, the data is, strictly speaking, not of the form of a numeric variable in this case, but rather a categorical variable, with levels Head and Tail. By using the random variable H, we have converted the categorical variable to numeric by a particular type of coding, but note that the coding was arbitrary, since we could have also defined T to be 0 or 1, depending on the number of tails observed.

Other forms of categorical data are possible as well, such as eye-colour, which might include black, brown, blue, and other. In this case, we would do the numeric coding using three binary variables, B_1, B_2 and B_3, where B_1 is 1 if the eye-colour is black, and 0, otherwise. $B_2 = 1$ for a brown eye and $B_2 = 0$, otherwise. $B_3 = 1$ for a blue eye and $B_3 = 0$, otherwise. All other eye colours are coded automatically as $B_1 = B_2 = B_3 = 0$.

Another important type of data is continuous data. Continuous variables take on measurements that are not necessarily counting numbers, and are expressed as decimals. Temperature, height, weight and time are often thought of as examples of continuous variables. An important distribution for continuous variables is the normal distribution which gives the familiar symmetric bell-shaped curve. Theoretically, normal random variables can take on any kind of value, positive and negative, so there are situations where this is clearly not appropriate. Time-to-event data, such as the time until someone recovers from a disease, or the time until a lightbulb fails, is a data type which is continuous and where the normal distribution is usually not a good approximation.

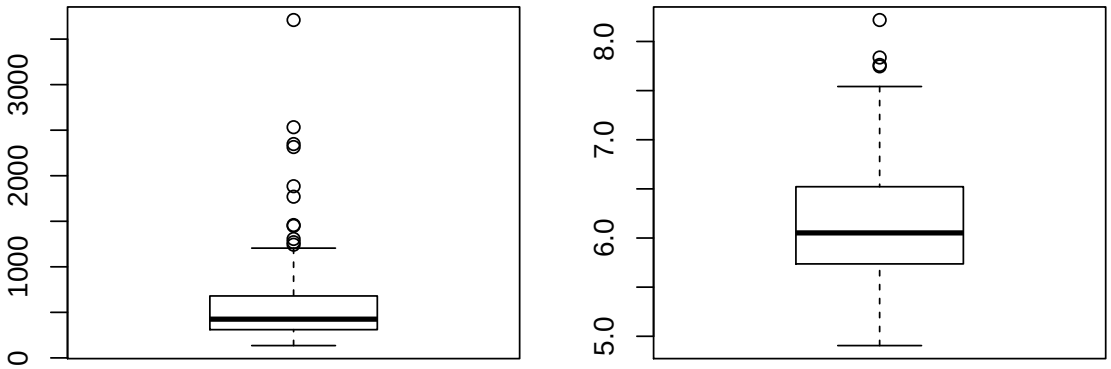

Figure 2.1: Box plot of river lengths, on original and log scales.

Sometimes a transformation, such as a log-transformation or a square root transformation yields a new version of the variable which is better approximated by normality.

2.2 Graphical and Numeric Summaries

An important facet of statistics is univariate analysis, whereby the distribution of a given single random variable is studied, often through the use of summary statistics, such as the mean, median, standard deviation and so on, or through the use of graphics, such as the box plot, histogram or dot chart. A bar chart is the most effective way of conveying categorical data; although pie charts are popular, they have been largely discredited as effective data analysis tools, and should be avoided.

We briefly consider two examples here.

2.2.1 River Lengths - Numeric

The `rivers` data set contains the lengths of 141 important or major North American rivers. A quick numeric summary of these data is obtained through

```
summary(rivers)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     135     310     425     591     680    3710
```

A box plot, as shown in the left panel of Figure fig:box, can be constructed using

```
par(mfrow=c(1,2), mar = c(1, 3, 1, 1))
boxplot(rivers)
boxplot(log(rivers))
```

```
boxplot(rivers)
```

A histogram could be constructed using `hist` in place of `boxplot`. Both types of plot reveal a distribution which is skewed to the right. A normal distribution is not immediately appropriate due to this fact (which is related to the fact that river length cannot be 0). Taking logs and then computing the box plot gives the graph in the right panel of Figure 2.1. The result is much more symmetric; the histogram would be hard to distinguish from a normal distribution.

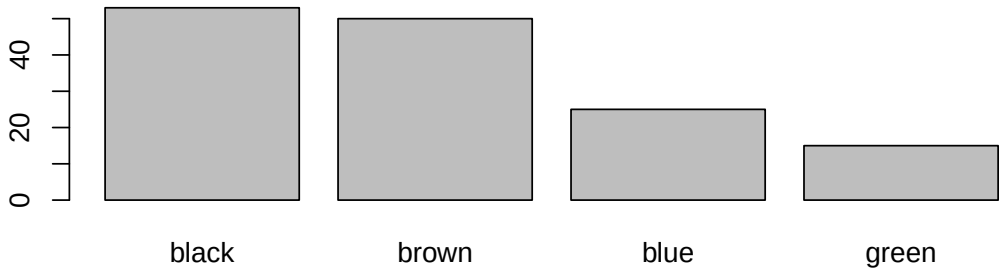

Figure 2.2: Bar chart of brown-haired male eye-colour.

```
boxplot(log(rivers))
```

2.2.2 Eye Colour - Categorical Data

A sample of brown-haired males revealed the following eye colour counts:

black	brown	blue	green
53	50	25	15

The table above provides the best form of numeric summary for this kind of data. Converting to percentages is an equivalent alternative – which hides the total number of data points.

The bar chart is constructed using

```
barplot(c("black" = 53, "brown" = 50, "blue" = 25, "green" = 15))
```

2.3 Classifying Basic Models by Data Type

The heart of statistical modelling lies in determining the relationships between different variables. The goals of statistical analysis are either prediction and explanation. For example, one might want to predict a future value of a random variable, called the response variable, given values of other variables, variously called predictor variables, covariates, or explanatory variables. The latter term is more appropriate when thinking of the modelling problem as one of attempting to explain or understand how the response variable relates or is associated with the other variables.

The following table may be useful in organizing your thoughts as to the best form of analysis for given types of data. It is important to remember that this table does not exhaustive of the kinds of statistical analyses that could be undertaken. The ones listed are the most commonly encountered.

response \ covariates	continuous	categorical	both
continuous	regression, correlation	t-test, ANOVA, Wilcoxon, Sign tests	ANCOVA
categorical	logistic	contingency tables	logistic

Regression refers to both simple and multiple regression (which involves more than 1 covariate). ANOVA refers to the analysis of variance, whereby means of different treatment groups are constrasted, depending on the levels or combination of levels from one or more factors. Factors are essentially another way of referring to categorical variables. Block designs are included in this category, and involve a factor which is not of direct interest to the scientist but which is known or believed to have an affect on the response. By including blocking factors in such analyses, more precision (i.e. less uncertainty) can be gained. ANOVA can also be viewed as a type of regression where the covariates are categorical and are made numeric by the binary variable coding described earlier.

ANCOVA is the analysis of covariance, which can be viewed as regression with both continuous and categorical predictors, or as a way of doing ANOVA (i.e. comparing treatment means), accounting for continuous covariates; this is a way of blocking with continuous covariates.

Logistic regression refers to the modelling of the probability distribution of a binary response variable, and the modern view of statistics sees logistic regression as encompassing contingency table analysis. In other words, contingency table analysis can be accomplished by performing logistic regression with categorical covariates.

Exercises

1. Construct the histogram plot of the data in the `rivers` data set. Describe the shape of the distribution.

2. Construct the histogram of the rivers data on the log scale. Describe the shape now.

3. Why do you think a bar chart is more appropriate than a pie chart for visualizing categorical data?[1]

4. The default colour scheme for most plots in R is gray-scale and not colour. What are the reasons for avoiding colour when visualizing data?[2]

5. The `HairEyeColor` data set in R contains sample information on hair and eye colour for males and females. If you type `HairEyeColor`, you will see two contingency tables of hair colour versus eye colour for the two sexes. You can access the blond female eye colour information directly, by typing

   ```
   HairEyeColor[ ,4, 2]
   ```

   Construct a bar chart for the eye colour of blond females by typing

   ```
   barplot(HairEyeColor[ ,4, 2])
   ```

   What happens when you omit the '2'?

6. Type `help(InsectSprays)` to find information on this data set. Then construct a histogram of the counts of insects in the various experimental plots by using

   ```
   hist(InsectSprays$count)
   ```

   Note the shape of the distribution and re-draw the histogram using the `sqrt` function, that is, by applying a square root transformation to the counts beforehand. How does the distribution shape change.

7. Re-do the previous exercise with box plots. Then try

[1] Discerning differences in areas and angles is more difficult than discerning differences in heights.

[2] Reproducing plots on hard copy often uses gray-scale, and more importantly, a surprisingly large proportion of the human population is colour-blind.

```
boxplot(count ~ spray, data = InsectSprays)
```

and repeat using the square root transformation of the counts. What is the effect of the square root transformation here?[3]

8. In the previous question, what type of data analysis is recommended?[4]

9. Type `help(airquality)` to find information on the `airquality` data set. Then construct a histogram of `airquality$Ozone`. Repeat using a log-transformation. Which is closer to normality?

10. If you were to model Ozone level as it relates to Wind, what analysis technique is recommended?[5] What if you take temperature into account?[6] What if you add in the `Month` variable?[7]

[3] The variability in the different distributions is better approximated by a constant after applying the square root transformation.
[4] ANOVA
[5] Simple regression.
[6] Multiple regression.
[7] There are choices here, but ANCOVA is a simple option.

3

T-tests

The goal of these tests, and the related confidence intervals, is to provide information about the mean of a single population, or about the difference in means of two population. The critical assumption underlying the t-test is that the measurements are independent of each other. In other words, if you know the value of one or more of the observations, you cannot predict another observation or group of observations with improved certainty.

We will use simulation to demonstrate the techniques.

3.1 One sample

We suppose that we have a random sample of measurements from a population with unknown mean μ and variance σ^2. Without telling you, I will simulate such 8 such measurements, storing them in an object called `X`, and we will use a test to determine if the true mean is 0 or not:

```
## [1] 0.89216 2.57622 0.92357 2.59863 2.90734 1.53666 3.42392 0.24498
```

We can calculate the mean and standard deviation for this sample using the `mean()` and `sd` functions:

```
mean(X)

## [1] 1.8879

sd(X)

## [1] 1.1415
```

Clearly, the sample mean is not 0, but the true mean could still be 0, and this result could just be the result of random sampling error. The t-test helps us answer this question:

```
t.test(X, conf.level = .995)

##
##  One Sample t-test
##
## data:  X
## t = 4.68, df = 7, p-value = 0.0023
## alternative hypothesis: true mean is not equal to 0
## 99.5 percent confidence interval:
##  0.26174 3.51413
## sample estimates:
## mean of x
##    1.8879
```

The small p-value indicates strong evidence against the hypothesis that the true mean is 0. In fact, this assertion is correct, since the code used to generate the random sample is as follows:

```
X <- rnorm(8, mean = 1.5) # true mean is 1.5
```

Note that we have used a 99.5% confidence interval to estimate the mean. This differs from the usual 95% that you might have been told to use. If you are conducting multiple tests, you should use caution, and a higher confidence level is a first step, but see the next section for more information.

3.1.1 ASA Statement on p-values

The American Statistical Association (ASA) is taking steps to halt the widespread abuse of p-values in science. In 2016, the ASA released a statement which provides important guidance. Information can be obtained in the article by R. Wasserstein at
`https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf`

Central to Wasserstein's document is the following information:

> The statement's six principles, many of which address misconceptions and misuse of the p-value, are the following:
>
> 1. P-values can indicate how incompatible the data are with a specified statistical model.
> 2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
> 3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
> 4. Proper inference requires full reporting and transparency.
> 5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
> 6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

3.1.2 An example with a skewed population

This time, we simulate a sample of 30 independent observations from an exponential population (i.e. non-normal) with mean 1:

```
Y <- rexp(30)
```

Even though this data set comes from a non-normal population, let's see what happens when we apply the t-test of the hypothesis that the mean is 0.

```
t.test(Y)
```

```
##
##  One Sample t-test
##
## data:  Y
## t = 5.05, df = 29, p-value = 2.2e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.69033 1.63019
## sample estimates:
## mean of x
##    1.1603
```

The p-value is very small, leading us to conclude that the true mean is not 0 (and it isn't).

If we were to test the hypothesis that the true mean is 1 (the truth), we would do the following:

```
t.test(Y, mu = 1)

##
##  One Sample t-test
##
## data:  Y
## t = 0.697, df = 29, p-value = 0.49
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  0.69033 1.63019
## sample estimates:
## mean of x
##    1.1603
```

In this case, the p-value is very large, and the interpretation would be that we do not have enough evidence to reject the null hypothesis.

This example shows that with enough data points, even a substantially skewed population is not a serious enough violation of the assumptions behind the t-test to warrant using an alternative testing method.

3.1.3 *A smaller skewed sample*

This time, we simulate a sample of 5 independent observations from an exponential population (i.e. non-normal) with mean 1:

```
Y <- rexp(5)
```

Let's see what happens when we apply the t-test of the hypothesis that the mean is 0.

```
t.test(Y)

##
##  One Sample t-test
##
## data:  Y
## t = 2.6, df = 4, p-value = 0.06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.071014  2.209623
## sample estimates:
## mean of x
##    1.0693
```

The p-value is not very small, indicating that we don't have enough evidence to reject the null hypothesis.

If we were to test the hypothesis that the true mean is 1 (the truth), we would do the following:

```
t.test(Y, mu = 1)

##
##  One Sample t-test
##
## data:  Y
```

```
## t = 0.169, df = 4, p-value = 0.87
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  -0.071014  2.209623
## sample estimates:
## mean of x
##    1.0693
```

In this case, the p-value is very large, and the interpretation would be that we do not have enough evidence to reject the null hypothesis.

This example shows that with enough data points, even a substantially skewed population is not a serious enough violation of the assumptions behind the t-test to warrant using an alternative testing method.

What if we have more data?

Returning to the test that the mean is 0, let's suppose we have an additional data point:

```
Y <- c(Y, rexp(1))
```

Again, let's see what happens when we apply the t-test of the hypothesis that the mean is 0.

```
t.test(Y)

##
##  One Sample t-test
##
## data:  Y
## t = 2.91, df = 5, p-value = 0.033
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.11632 1.87919
## sample estimates:
## mean of x
##   0.99776
```

The p-value is smaller, indicating that we have some evidence to reject the null hypothesis, but we might want to obtain an even larger sample, in order to be more certain.

3.2 Two independent samples

If we have two random samples that are independent of each other, we can still use a t-test to compare the means of the populations from which they were sampled.

Let's start with `X` and `Y` which were simulated earlier. They come from different populations, and the first has a true mean of 1.5 and the second, a true mean of 1. The t-test will be correct if it gives us a small p-value, indicating strong evidence against the null hypothesis that the true means are the same:

```
t.test(X, Y)

##
##  Welch Two Sample t-test
##
## data:  X and Y
## t = 1.68, df = 12, p-value = 0.12
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.26369  2.04405
## sample estimates:
## mean of x mean of y
##   1.88794   0.99776
```

We actually know, in this case, that the variances are equal, so we could use

```
t.test(X, Y, var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  X and Y
## t = 1.61, df = 12, p-value = 0.13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3179  2.0983
## sample estimates:
## mean of x mean of y
##   1.88794   0.99776
```

Note that the results are not all that different. The conclusions are the same: we don't have enough data to reject the null hypothesis that the means are different. This is not surprising, since the sample sizes are pretty small (8 and 6).

3.3 Two samples - matched pairs

If there is a one-to-one correspondence between measurements in one of the samples with measurements in the other sample, then the appropriate way to compare the means is by taking the differences, and running a one-sample test on the differences. This can be done with the `paired` option in the `t.test()` function.

Let's suppose L is a set of left foot lengths (in cm) for a sample of 15 adult males, and R contains the corresponding right foot lengths. We would be interested in any systematic difference in the lengths of the feet. A simulation model for the case where there is no difference could be the following:

```
L <- rnorm(15, mean =28, sd = 1)
R <- L + rnorm(15, mean = 0, sd = .03)
```

Here we have assumed that the left feet are normally distributed with a mean of 28 cm and a standard deviation of 1 cm. The right feet are not equal to the left feet, but on average there is no difference. The standard deviation of the difference is small.

Let's see what the t-test says:

```
t.test(L, R, paired=TRUE)

##
##  Paired t-test
##
## data:  L and R
## t = -0.417, df = 14, p-value = 0.68
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.016705  0.011262
```

```
## sample estimates:
## mean of the differences
##              -0.0027213
```

The p-value is large indicating that there is no evidence of a difference, in line with the truth.

Let's now simulate from a population where the right feet tend to be slightly larger than the left feet, on average:

```
L <- rnorm(15, mean =28, sd = 1)
R <- L + rnorm(15, mean = 0.02, sd = .03)
```

Does the t-test find the difference?

```
t.test(L, R, paired=TRUE)

##
##  Paired t-test
##
## data:  L and R
## t = -2.74, df = 14, p-value = 0.016
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0402706 -0.0049088
## sample estimates:
## mean of the differences
##                -0.02259
```

The p-value is pretty small, indicating that we have fairly strong evidence that there is a difference between the left and right lengths in this sample.

3.4 Classical nonparametric tests

These tests remain valid for data where the normality assumption clearly does not hold. They have nothing to offer if the more important independence assumption fails. They can all be carried out using the `wilcox.test()` function.

3.4.1 Sign test

The sign test can be used to test for various properties of a population, based on a given random sample.

If we suspected that the left and right feet lengths from the earlier section were non-normal, we could use the sign test to determine whether there is evidence that the right feet are longer than the left feet, by checking the sign of the difference between the right and left feet:

```
sign(R - L)

##  [1] -1  1  1  1  1  1  1  1  1  1  1  1  1 -1  1
```

There seem to be a lot of positives and not so many negatives, so we can compute a p-value for the test against the null hypothesis that there is no difference between left and right foot length by counting the number of positives:

```
Npos <- sum(sign(R-L)>0)
```

and comparing with what we might see in a binomial distribution with $p = .5$, and $n = 15$ trials:

```
binom.test(Npos, 15)

##
##  Exact binomial test
##
## data:  Npos and 15
## number of successes = 13, number of trials = 15, p-value = 0.0074
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.59540 0.98342
## sample estimates:
## probability of success
##                0.86667
```

The p-value is very small, indicating that we have evidence against the null hypothesis. This is in agreement with what we saw in the matched-pairs result.

3.4.2 *Wilcoxon sign-rank test*

The Wilcoxon sign-rank test can be carried out on the simulated foot length data as follows:

```
wilcox.test(L, R,  paired = TRUE)

##
##  Wilcoxon signed rank test
##
## data:  L and R
## V = 17, p-value = 0.012
## alternative hypothesis: true location shift is not equal to 0
```

Again, we have a small p-value, indicating evidence against the null hypothesis that the feet are the same length.

3.4.3 *Mann-Whitney U test*

Recall that `X` is normally distributed and `Y` is exponentially distributed. We can use the Mann-Whitney U test to test for this kind of difference.

```
wilcox.test(X, Y)

##
##  Wilcoxon rank sum test
##
## data:  X and Y
## W = 35, p-value = 0.18
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is large, so the test failed to find the difference.
Let's try another example, this time with different means:

```
X <- rnorm(10)  # mean is 0
Y <- rexp(5) # mean is 1
```

The samples are fairly small, but the test result below confirms that there is a difference between the two populations:

```
wilcox.test(X,Y)

##
##  Wilcoxon rank sum test
##
## data:  X and Y
## W = 7, p-value = 0.028
## alternative hypothesis: true location shift is not equal to 0
```

4

Simple Regression

The yield (y, in kg/plot) was measured for various salinity concentrations (x, measured in units of electrical conductivity). 18 measurements were recorded in a file called *tomato.txt* whose contents appear below:

1.60	59.50
1.60	53.30
1.60	56.80
1.60	63.10
1.60	58.70
3.80	55.20
3.80	59.10
3.80	52.80
3.80	54.50
6.00	51.70
6.00	48.80
6.00	53.90
6.00	49.00
10.20	44.60
10.20	48.50
10.20	41.00
10.20	47.30
10.20	46.10

The first column contains the salinity concentration levels, and the second column contains the yield measurements.

We read these data into R using the `read.table()` function (or using a menu option in RStudio):

```
tom <- read.table("tomato.txt", header=FALSE)
```

Since there is no header, we should apply some sensible names to the data frame:

```
names(tom) <- c("salinity", "yield")
```

We next construct a scatterplot of the data to look for patterns and outliers as in Figure 4.1.

```
plot(yield ~ salinity, data = tom)
```

We have plotted `yield` against `salinity` since we take it that the response variable is yield and the explanatory variable or predictor variable is salinity. We base this on the fact that the yield was measured at various salinity concentrations that appear to have been set by the experimenter.

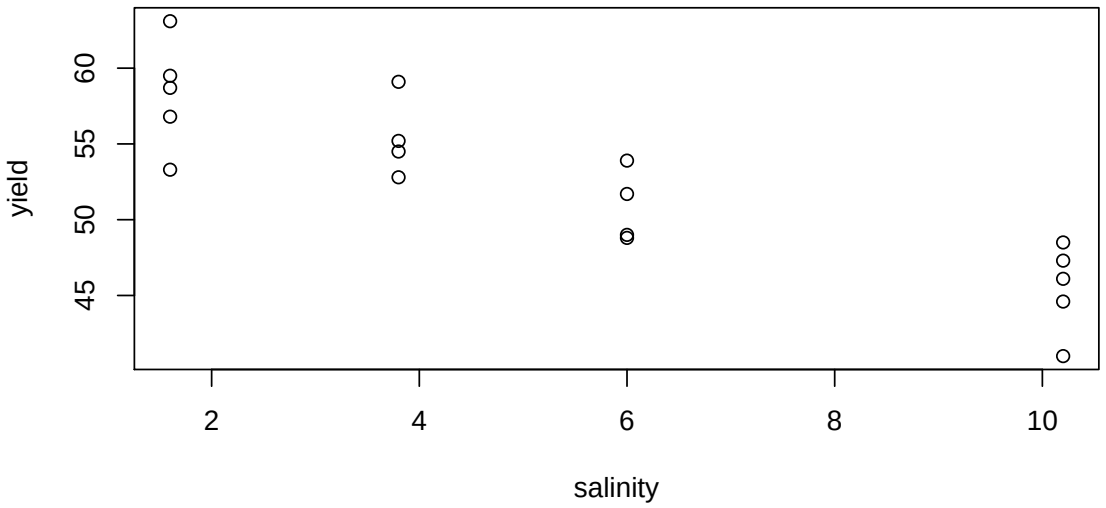

Figure 4.1: Scatterplot of tomato electrical conductivity data.

The scatterplot gives an indication of a clear downward trend as salinity increases. The trend is also vaguely linear. The suggestion in the graph is that yield could be predicted by salinity. We can investigate this with the `lm()` function:

```
tom.lm <- lm(yield ~ salinity, data = tom)
```

Note that the `model formula` used here, i.e. `yield ~ salinity`, is identical to that used in the `plot()` function. This is often the case: if you can figure out a good way of plotting the data, this often suggests the form of analysis.

We can explore the output from the fitted model using the `summary()` function:

```
summary(tom.lm)

##
## Call:
## lm(formula = yield ~ salinity, data = tom)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.956 -1.967  0.173  1.825  4.844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    60.67       1.28   47.37  < 2e-16
## salinity       -1.51       0.20   -7.53  1.2e-06
##
## Residual standard error: 2.83 on 16 degrees of freedom
## Multiple R-squared:  0.78,Adjusted R-squared:  0.766
## F-statistic: 56.7 on 1 and 16 DF,  p-value: 1.21e-06
```

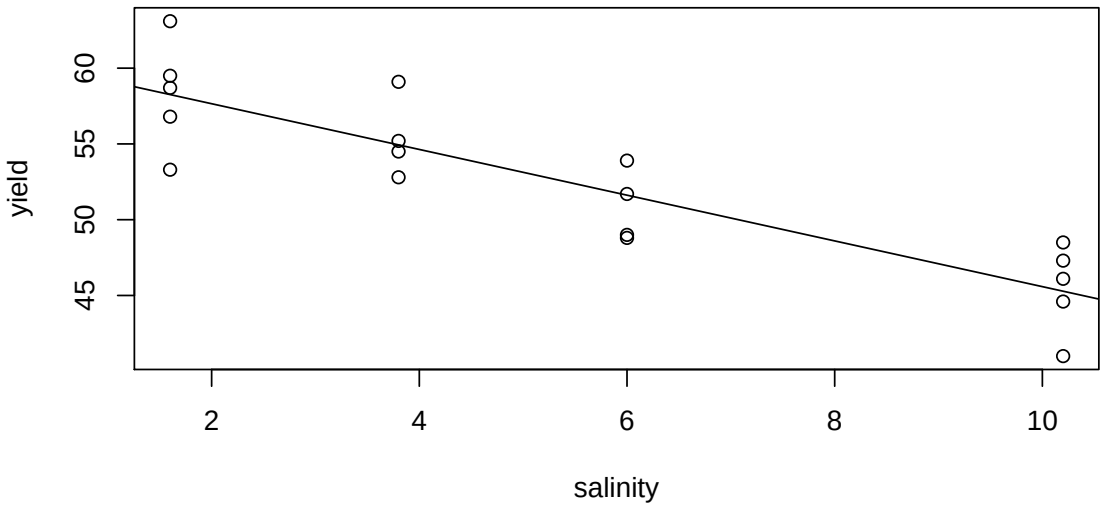

Figure 4.2: Scatterplot of tomato electrical conductivity data with overlaid fitted line.

From the output, we can see the estimates of the intercept and slope of the line. Note that the slope estimate, -1.5088 is negative, corresponding to the negative relation between salinity and yield The intercept is 60.67. In fact, we can overlay the scatterplot of the data with the fitted line using the `abline()` function and the output from the `lm()` function:

```
abline(tom.lm)
```

The F-statistic and corresponding p-value ($1.212e-06$, a very very small number) suggest strongly that the slope is nonzero, so the trend in the data is real, provided certain assumptions are satisfied:

1. the relation between salinity and yield is (at least approximately) linear.

2. the measurements are independent of each other, meaning that knowledge of one measurement does not give you information about any other measurement, beyond what you would be able to predict from the line itself.

3. variability in the yield measurements is the same for all salinity levels.

The first and third assumptions are fairly easy to check, and Figure 4.2 helps to do this. A plot of residuals (differences between what the line would predict and the yield measurements) is a clearer way of checking.

The second assumption is difficult to check. It is connected intimately to the manner in which the data have been collected. We will see later that there are some kinds of dependence that can be assessed, but those methods are not applicable here.

From the output, we also see the R^2 and adjusted R^2 values. Although these quantities are often quoted in the scientific literature and used to justify or validate models, they are actually not very useful, and have little to say about whether a model is valid or not. The proper interpretation of R^2 is as the proportion of variation in the response explained by the model. The coefficient of determination coincides with the square of the Pearson correlation coefficient:

```
cor(tom$salinity, tom$yield)

## [1] -0.88304
```

By itself, this value tells us that the two variables are negatively correlated, meaning that if one were to plot one of the variables against the other, we would see points scattered about a line with negative slope. In fact, we did just that in Figure 4.2, suggesting that there is limited, if any, information to be gleaned from calculating a correlation coefficient, when the power of a regression analysis is at our disposal. If one really wants to compute a Spearman rank correlation, appropriate for data where a linear trend is not in evidence, one can use

```
cor(tom$salinity, tom$yield, method = "spearman")

## [1] -0.89396
```

This also is of limited additional use.

Exercises

1. Consider the data on the model car that was released from various points on a ramp and the distance traveled was measured. The data frame is called `modelcars`, and it consists of two columns, `distance.traveled` and `starting.point`.

```
library(DAAG) # the package containing the model car data set
mcar.lm <- lm(distance.traveled ~ starting.point, data = modelcars)
summary(mcar.lm)$coefficients

##                Estimate Std. Error t value   Pr(>|t|)
## (Intercept)      8.0833     1.0780  7.4988 2.0657e-05
## starting.point   2.0139     0.1312 15.3493 2.8019e-08
```

   Identify the slope and intercept of the line relating distance traveled to starting point.[1] Is there strong evidence of a nonzero slope to this line?[2] Is the slope positive or negative?[3]

2. Write down the two lines of R code which would produce the graph in Figure 4.3. [4]

```
plot(distance.traveled ~ starting.point,
     data = modelcars)
abline(mcar.lm)
```

3. Analyze the `airquality` data to determine the relationship between ozone and wind, by finding the slope and intercept of the linear model. Then plot the line on a scatterplot of the data.

4. Repeat the above analysis using a square root transformation on the Ozone variable. Is this a better way of modelling the data?[5]

[1] intercept: 8.083; slope: 2.014
[2] Yes, the p-value is $2.802e-08$ which is extremely small.
[3] Positive.
[4] plot(distance.traveled starting.point, data = modelcars); abline(mcar.lm)
[5] Yes, the square root transformation reduces some of the nonlinearity which is apparent when working with raw Ozone data.

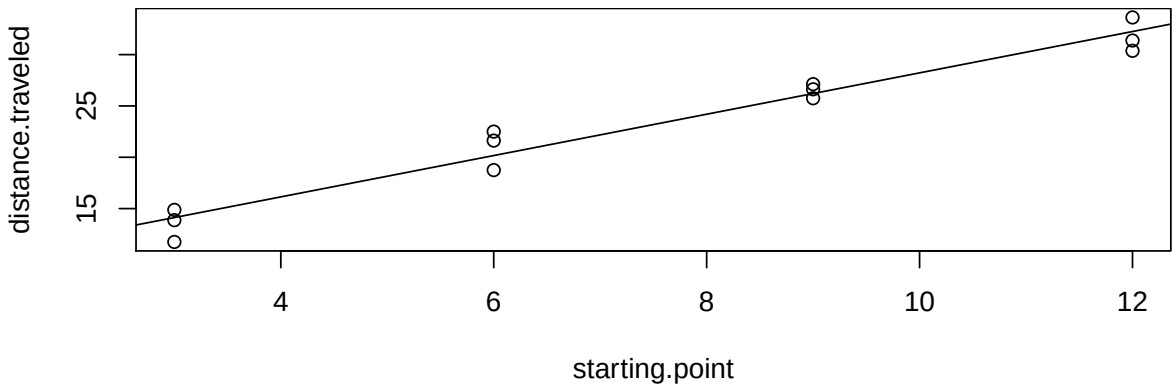

Figure 4.3: Distance travelled by a model car launched from a ramp at various starting points.

5

ANOVA

5.1 One Factor

The `chickwts` data frame contains measurements of the weights of chicks who have been randomly assigned to groups, each of which has been given a different type of feed. It is of interest to know whether the different feed types lead to systematic differences in weight. In other words, does mean weight depend on the type of feed. We refer to feed type as a factor having different levels representing the particular kinds of feed, e.g. linseed, horsebean, and so on.

Side-by-side boxplots, as displayed in Figure 5.1, are a useful way to visualize these data.

```
plot(weight ~ feed, data = chickwts, cex.axis=.75) # feed must be a factor
```

From the graph, it seems that horsebean leads to lower weights than some of the other feed types. It is hard to tell for sure if there is variability between treatments because of the variability within treatments, that is noise due to unmeasured factors.

We can test whether there is a difference in the mean weights statistically with the analysis of variance (ANOVA). A general purpose procedure is as follows:

```
chick.lm <- lm(weight ~ feed, data = chickwts)
anova(chick.lm)

## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value  Pr(>F)
## feed       5 231129   46226    15.4 5.9e-10
## Residuals 65 195556    3009
```

The test statistic compares the variability in the averages with the variability in the noise through an F-statistic. A p-value is computed which gives the strength of evidence against the null hypothesis, that is the hypothesis that there is no difference in the means. A small p-value – and in this case, it is very small – indicates strong evidence against the null hypothesis, in favour of the alternative that there is a difference.

5.2 Two or More Factors

Information on gas mileage for a number of cars is available in `table.b3` of the *MPV* package. Although not from a designed experiment, we will analyze this observational data as if it were. Our objective is to see if mean gas mileage y depends on either or both of carburetor barrels (x_6, viewed as a categorical variable) and type of transmission x_{11}.

```
library(MPV)
b3.lm <- lm(y ~ factor(x6>1)*x11, data = table.b3)
anova(b3.lm)
```

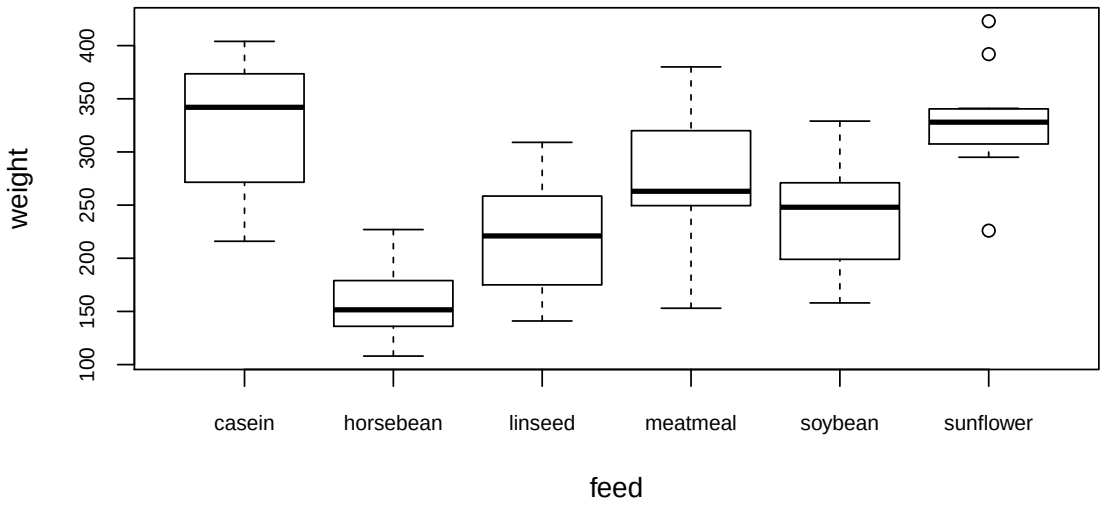

Figure 5.1: Box-plots of chick weight samples for different types of feed.

```
## Analysis of Variance Table
##
## Response: y
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## factor(x6 > 1)       1      0       0    0.02   0.895
## x11                  1    689     689   40.85 6.4e-07
## factor(x6 > 1):x11   1     76      76    4.50   0.043
## Residuals           28    472      17
```

A number of points need to be made. First, `x6` is recorded in the data frame as if it is continuous (and it could be treated as such), so in order to treat it as categorical, we use the `factor()` function. We have also coded the variable to have more than 1 barrel or to have 1 barrel.

The second point is the use of the use of $\star$ which forces the model to include both of the factors as well as interactions between the factors. Interaction effects can play an important role in modelling, since they reflect situations where, for example, changing the level of one factor, might have different effects, depending on the level of the other factor. In this case, increasing number of carburetor barrels might affect gas mileage differently for automatic than for manual transmission. The output can help us determine if this is actually the case.

In fact, the p-value for the interaction effect is just under 5%, so there is moderate evidence of an interacting effect between these two variables. The effect of the number of carburetor barrels on gas mileage could be different for manual and automatic transmissions. If this were part of an actual research study, it is critical that this result would be reproduced in a designed experiment. Publishing a marginally significant result such as we obtained here, based on a small sample coming from an observational study would be irresponsible and reckless. Unfortunately, this analysis mirrors too closely for a lot of what passes as scientific research in the published literature.

5.3 Randomized Block Design

The penicillin data frame in the `BHH2` package contains data coming from a randomized block design. There appear to be two factors, one is the treatment, the other is called `blend`. The `blend` factor is not of direct interest in the study, but has been included in order to reduce the noise in the data which would have otherwise been due to unmeasured factors.

```
library(BHH2) # contains penicillin.data
data(penicillin.data) # loads data
m1<-lm(yield~treat+blend, data=penicillin.data)
anova(m1)

## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## treat      3     70    23.3    1.24  0.339
## blend      4    264    66.0    3.50  0.041
## Residuals 12    226    18.8
```

Because of the way the experiment has been designed, through randomization of subjects to treatment and blocking groups, there is no reason to expect an interaction effect. The p-value for `treat` provides the measure of the strength of evidence against the hypothesis that the mean response does not depend on treatment.

Exercises

1. Consider the data in `PlantGrowth` and conduct an analysis of variance to determine if the mean dried yield weight of the plants under study differs depending on whether the plants were grown under control conditions or under either of two different treatment conditions.

   Visualize the data with side-by-side boxplots.

6

Multiple Regression

The data frame `table.b4` in the *MPV* library contains the following columns:

```
y sale price of the house (in thousands of dollars)
x1 taxes (in thousands of dollars)
x2 number of baths
x3 lot size (in thousands of square feet)
x4 living space (in thousands of square feet)
x5 number of garage stalls
x6 number of rooms
x7 number of bedrooms
x8 age of the home (in years)
x9 number of fireplaces
```

There are 24 observations on these variables in the data frame. A natural question to ask is whether any or all of the given variables or covariates could be used to predict the sale price of a house.

The multiple regression approach is to consider a linear model of the form

$$y = \beta_0 + \sum_{j=1}^{9} \beta_j x_j + \varepsilon.$$

If the β coefficients were known, then we could predict house price y, to within the unknown noise value ε, for a new house in the same area (and era) from which the data were sampled, provided the information on taxes, number of baths, and so on. If, for example, the ε term is modelled as a normal random variable with mean 0 and variance σ^2, we could provide an interval which would contain the true price of the house with a given probability.

The elements of ε are assumed to be uncorrelated random variables with mean 0 and common variance σ^2. The mean or expected value of y for the given values of the covariates is then

$$E[y] = \beta_0 + \sum_{j=1}^{9} \beta_j x_j.$$

(The "E" notation stands for "Expected Value".)

6.1 Fitting the model

The `lm()` function will take care of the coefficient estimation, variance estimation, t and F and p-value calculations in one function call. For example, if we want to relate house price, y to x_1, x_3 and x_6, use

```
house.lm <- lm(y ~ x1 + x3 + x6, data=table.b4)
```

We can view the output from this, using the `summary` function as in

```
summary(house.lm)
```

or to see the coefficient estimates and their statistical properties only, use

```
summary(house.lm)$coefficients
```

If we include all of the covariates, we can use the dot notation in the model formula:

```
house.lm <- lm(y ~ ., data=table.b4)
```

```
summary(house.lm)$coefficients
```

```
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 14.927648    5.91285  2.52461 0.024283
## x1           1.924722    1.02990  1.86884 0.082711
## x2           7.000534    4.30037  1.62789 0.125836
## x3           0.149178    0.49039  0.30420 0.765447
## x4           2.722808    4.35955  0.62456 0.542304
## x5           2.006684    1.37351  1.46099 0.166096
## x6          -0.410124    2.37854 -0.17243 0.865570
## x7          -1.403235    3.39554 -0.41326 0.685678
## x8          -0.037149    0.06672 -0.55679 0.586461
## x9           1.559447    1.93750  0.80488 0.434347
```

The output above lists a number of things. Our focus is principally on the `Estimate` column, since that gives us the estimates of the coefficients β. The intercept is 14.93 and the coefficient of x_1 is 1.92, and so on.

Together with these estimates are estimates of the standard errors. These provide an assessment of the amount of uncertainty is associated with the corresponding coefficient estimate. Clearly, the estimate of β_6 has a relatively large degree of uncertainty associated with it, since the standard error is much larger than the absolute value of the estimate itself.

This highlights an important problem when applying multiple regression techniques: over-fitting. When using a limited amount of data to estimate a large number of parameters in this case, 10, the degree of uncertainty in the estimates rises quickly. It is often better to carefully decide which covariates to include in a model based on other considerations. Use any known science or other information to help make these choices. For example, it might be known that taxes and living space are highly predictive of sale price. In that case, focus on those variables immediately in order to more precisely estimate their coefficients.

```
house.lm <- lm(y ~ x1 + x4, data=table.b4)
summary(house.lm)$coefficients
```

```
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  11.5447    3.17474 3.63643 1.5444e-03
## x1            2.9195    0.57599 5.06869 5.0972e-05
## x4            3.1574    3.29833 0.95729 3.4931e-01
```

Notice how the standard error estimates for the coefficients of x_1 and x_4 are less than before, although the standard error for the living space variable still exceeds the absolute value of the coefficient, so there is considerable uncertainty left there.

6.2 Estimating and predicting

The model can now be used to estimate the expected house price for houses with x_1 taxes and x_4 amount of living space using the formula

$$\widehat{y} = 11.5 + 2.92x_1 + 3.15x_4.$$

This can be accomplished in R using the `predict()` function. For instance, suppose we want to estimate the mean sale price for homes with $2000 taxes and 3000 square feet of living area. Use

```
predict(house.lm, newdata = data.frame(x1 = 2, x4 = 3))

##      1
## 26.856
```

The mean price for such a home is $25856. This estimate highlights an important point: this data set is old and it applies to a particular location. In order to understand the housing market in a particular location and time, it is necessary to use the relevant data.

Note also that the `predict()` function has been used. It can be used both for estimating the mean price of a house or predicting the price of a specific house. Prediction uncertainty is usually much larger than estimation uncertainty; both are incorporated in the `predict()` function. Use `interval = "confidence"` for estimation and `interval = "predict"` for prediction. For example,

```
predict(house.lm, newdata = data.frame(x1 = 2, x4 = 3), interval = "predict")

##      fit    lwr    upr
## 1 26.856 10.233 43.479
```

This says that with 95% probability, the house we are looking at with taxes of $2000 and 3000 square feet of living area is priced between $10232 and $43479.

```
predict(house.lm, newdata = data.frame(x1 = 2, x4 = 3), interval = "confidence")

##      fit   lwr    upr
## 1 26.856 11.42 42.292
```

This says that with 95% confidence, we can say that the mean price of houses with taxes of $2000 and 3000 square feet of living area is between $11420 and $42292.

6.3 Assessing the model

In the past, the term model validation was used when describing the process of deciding whether a model was appropriate or not. This incorrectly conveys the sense that there is a correct model; it is now recognized that all models are incorrect at some level, but some are more useful for certain purposes than others might be. Thus, the term model assessment is now favoured, since it conveys a sense of checking the appropriateness of a given model as opposed to checking its validity.

Although there are a number of statistics that are often (ab)-used to do this assessment, a graphical approach is usually the best way to understand whether there are substantive problems with a given model. In particular, graphs of residuals are the best way to decide if a model is failing severely. A residual is the difference between the observed response value and the value predicted by the model. As such, residuals are effectively predictors of the noise term ε in the model.

Since the noise term is assumed to have mean 0, constant variance and to have no internal correlations, we can often simply look at a graph of the residuals plotted against observation number, fitted value, or a predictor variable to look for patterns. Clear patterns are a sign of severe model failure. Figure 6.1 displays the residuals plotted against the fitted values, with a smooth fitted overlaid red curve which can guide the eye to any systematic patterns. In this case, the curve is not substantially different from a flat line, which would be ideal. There does not seem to be a clear pattern in the residuals in this case.

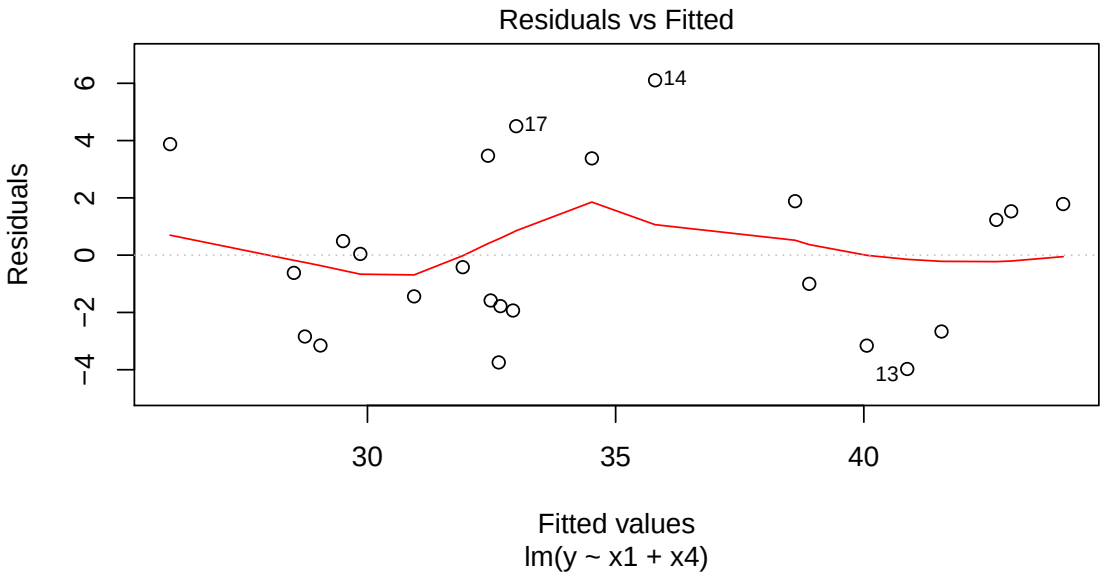

Figure 6.1: Plot of residuals for house price model against fitted values.

```
plot(house.lm, which = 1)
```

Another important plot, such as in Figure 6.2, concerns the influence of individual data points on the model fit. Large values of Cook's distance are suggestive of difficulties which should be remedied. For assistance with such problems, it is probably best to consult your local statisticians for help. The values seen in the current case are not worrisome.

```
plot(house.lm, which = 4)
```

6.4 Significance of regression

The F-test for significance of regression gives us a way of deciding whether any of the regression coefficients should be nonzero. In other words, the coefficients (apart from the intercept) are assumed to be 0's under the null hypothesis and there must be at least one nonzero coefficient if the alternative hypothesis is true. An F distribution is used to conduct the test. The p-value based on the test gives us the strength of evidence supporting the case that at least one regression coefficient is nonzero, where, as usual, small values indicate more evidence than large values would.

We can test whether the coefficients of x_1 and x_4 are both 0 in a variety of ways, including by looking at all output from `summary(house.lm)`. In order to see more clearly what is happening, you can use the `anova()` function to decide between the null model (one with only an intercept) and the model we have already fit:

```
house0.lm <- lm(y ~ 1, data = table.b4) # intercept only model
anova(house.lm, house0.lm) # test significance of regression

## Analysis of Variance Table
##
```

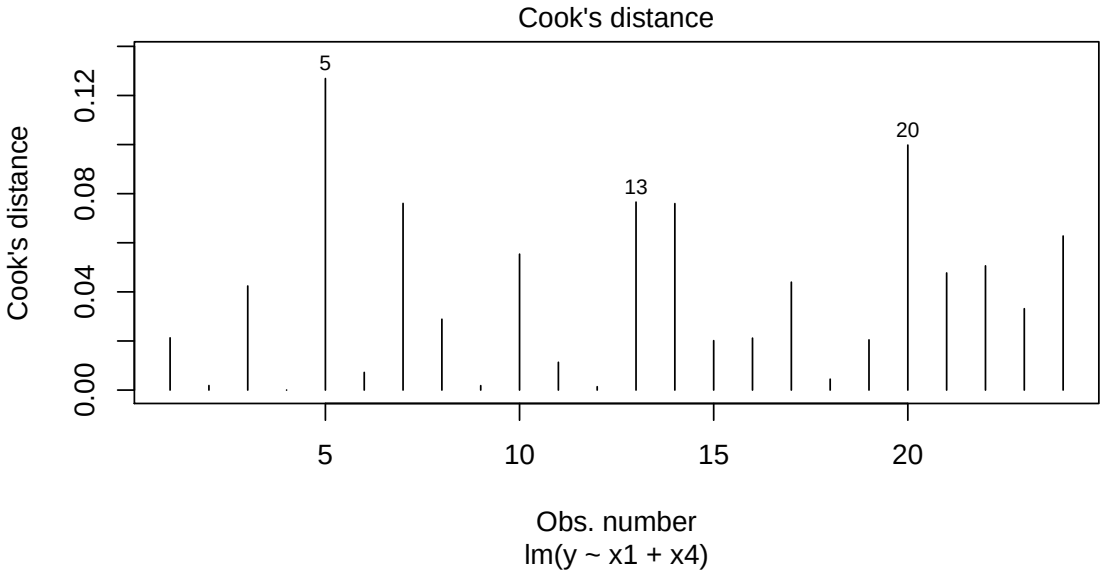

Figure 6.2: Influence diagnostic plot for house price models.

```
## Model 1: y ~ x1 + x4
## Model 2: y ~ 1
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     21 185
## 2     23 829 -2      -644 36.6 1.4e-07
```

The p-value is small indicating that at least one of the coefficients is nonzero. The nice thing about this approach is that we can use it to decide if more variables should be added to the model. For example, let's see if there are any other variables to add, after adding x_1 and x_2:

```
houseAll.lm <- lm(y ~ ., data = table.b4) # fit with all variables
anova(houseAll.lm, house.lm) # compare All- and two-variable models

## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
## Model 2: y ~ x1 + x4
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     14 122
## 2     21 185 -7     -63.1 1.04   0.45
```

This analysis tells us that once we have taken taxes and living area into account, there is no point in adding additional variables into the model. There is no evidence to suggest that the additional coefficients are nonzero.

Exercises

1. Consider the gas mileage data in `table.b3` of the *MPV* package.

(a) Fit a multiple regression model to estimate mean gas mileage y for cars with x_7 number of transmission speeds and having weight x_{10}.

(b) Assess the model using the residual plot.

(c) Use the model to estimate mean gas mileage for cars having weight 5000 pounds and 4 transmission speeds. Use a 95% confidence interval.

(d) Use the F-test for signifiance of regression to decide if any of the coefficients for your fitted model are nonzero.

(e) Use another F-test to decide if variables x_1, x_2, x_4 or x_5 should be added into the model you have already developed.

7

ANCOVA

The analysis of covariance (ANCOVA) allows us to model continuous responses as linear functions of continuous and categorical covariates. In this way, it can be viewed as a relatively straightforward extension of multiple regression. It can also be viewed as an extension of ANOVA whereby there is a blocking factor which is continuously measured. For a categorical covariate with two levels, there would be two lines in the regression: parallel if there is no interaction effect; two different slopes if there is an interaction effect.

The `ToothGrowth` data frame in R concerns the length of odontoblasts, cells connected with the growth of teeth, in a sample of 60 guinea pigs. One of three dose levels of vitamin C were supplied to the guinea pigs in one of two forms: `supp = VC` refers to ascorbic acid and `supp = OJ` refers to orange juice. Figure 7.1 displays the data.

```
library(lattice)
xyplot(len ~ sqrt(dose)|supp, data = ToothGrowth)
```

The figure shows that there are possibly two different lines relating length to vitamin C dose; it is possible that there is a treatment effect.

We use `lm()` to check this, first by allowing for two intercepts and two slopes:

```
TG.lm <- lm(len  ~ sqrt(dose)*supp, data = ToothGrowth)
summary(TG.lm)

##
## Call:
## lm(formula = len ~ sqrt(dose) * supp, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.987 -2.677 -0.172  2.738  7.413
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.48       2.63    0.94    0.350
## sqrt(dose)           17.48       2.43    7.18  1.7e-09
## suppVC              -12.02       3.72   -3.24    0.002
## sqrt(dose):suppVC     8.00       3.44    2.33    0.024
##
## Residual standard error: 3.87 on 56 degrees of freedom
## Multiple R-squared:  0.758,Adjusted R-squared:  0.745
## F-statistic: 58.3 on 3 and 56 DF,  p-value: <2e-16
```

Looking at the interaction between the square root of dose and `supp`, we see a fairly small p-value which is highly suggestive of different slopes. There is no reason to consider the model without different slopes (which

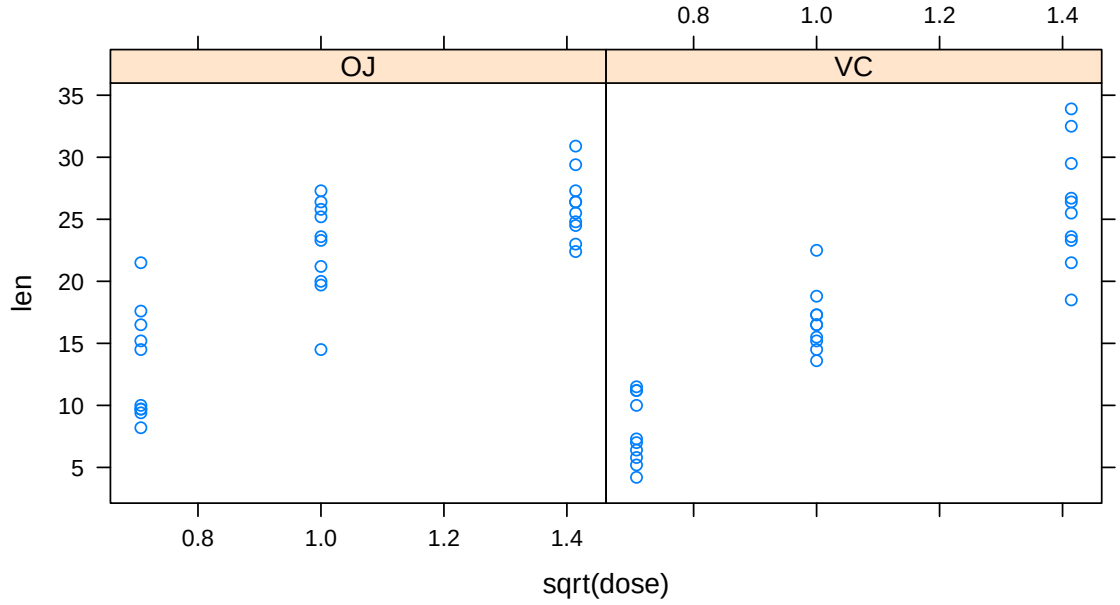

Figure 7.1: Tooth growth data: length of tooth versus square root of vitamin C dose, for each of the two treatment methods.

would have been obtained by replacing the $\star$ with $+$). The value 8 indicates that for `VC`, the slope is 8 units higher than for `OJ`: $17.48 + 8 = 25.48$. The intercept for `VC` is 12.02 units lower: $2.478 - 12.024 = -9.546$.

We can graphically summarize the data with a simple scatterplot with the lines overlaid:

```
plot(len ~ sqrt(dose), pch = as.numeric(supp), data = ToothGrowth)
abline(2.478, 17.479) # OJ line
abline(2.478-12.024, 17.479 + 8, lty=2) # VC line, dashed
```

Exercises

1. Fit the model with two parallel lines to the tooth growth data. What are the intercepts for the VC and OJ lines? What is the slope? Plot the data with the two lines overlaid.

2. Consider the data in `airquality` which relate to Ozone levels in New York. Construct a model which relates Ozone level to temperature and wind, taking Month into account, as a factor. Is there evidence that Month should be included in the model? Is there an interaction between Month and wind? between Month and temperature?

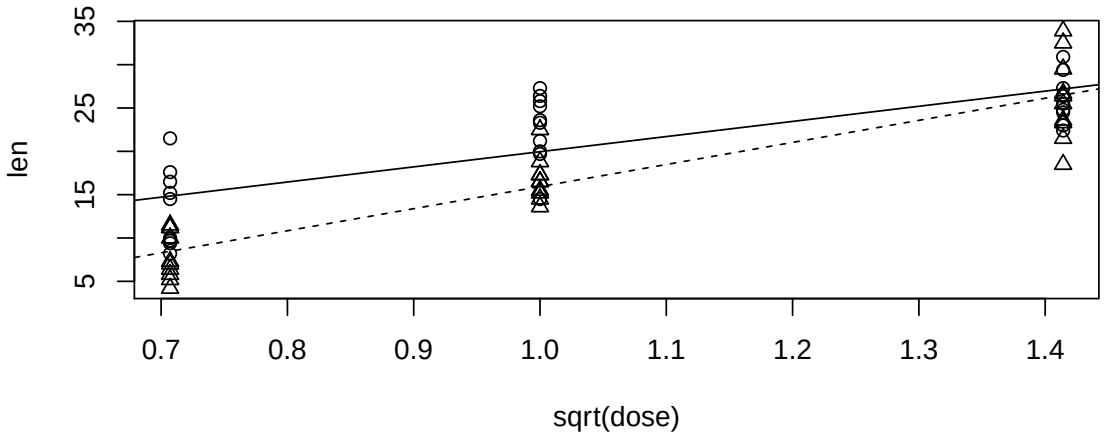

Figure 7.2: Tooth growth data: length of tooth versus square root of vitamin C dose, for each of the two treatment methods. The solid line corresponds to the orange juice treatment, and the dashed line corresponds to the ascorbic acid treatment

8

Logistic Regression

8.1 Modelling binary responses

The data in `p13.1` in the *MPV* package describes successes and failures of surface-to-air missiles as they relate to target speed. The data are plotted in Figure 8.1, with successes on the vertical axis being represented by a '1' and failures being represented by a '0'.

Such binary data are not nearly normally distributed, so it the efficacy of least-squares becomes very questionable here. In this section, we indicate what could and should not be done with least-squares for such data.

```
library(MPV)
plot(p13.1, xlab = "target speed", ylab = "success/failure")
```

The first observation to make is that fitting a straight line to such data makes no sense, since the plotted points do not at all scatter about such a line. Furthermore, if such a line were to be fit to the data, it would necessarily take values outside the interval $[0, 1]$ on subsets of the domain; interpretation of such values would be difficult. In fact, the preferred interpretation of output arising from the fitting of models to such data is that of probability. That is, useful models can provide answers to questions such as, "What is the probability of success at a given target speed?" Since probabilities must lie within the interval $[0, 1]$, we must consider models based on nonlinear functions.

There are many functions which have values in $[0, 1]$. For example, the absolute value of the sine function is a candidate. Such a function might be appropriate if there were oscillatory or periodic behaviour to be modelled, but often, the desired model behaviour is monotonic (either increasing or decreasing). For the current example, we might reasonably believe that the probability of success decreases as target speed increases.

Perhaps the most popular function for this purpose is the logistic function

$$p(x) = \frac{e^x}{e^x + 1}.$$

The function is sketched in Figure 8.2.

```
curve(exp(x)/(1 + exp(x)), from = -3, to = 3, ylab="p(x)")
```

A bit of algebra allows us to express x in terms of p, yielding the logit function:

$$\ell(p) = \log\left(\frac{p}{1-p}\right).$$

While p is restricted to take values between 0 and 1, the logit function can take any possible value, so relating the logit function to a straight line or other linear combination is a possibility. For example,

$$\ell(p(x)) = \beta_0 + \beta_1 x$$

which means that we can express the probability of an event in terms of a covariate x, using a linear function, but the probability is related to the linear function through the logit. This kind of model where a linear function

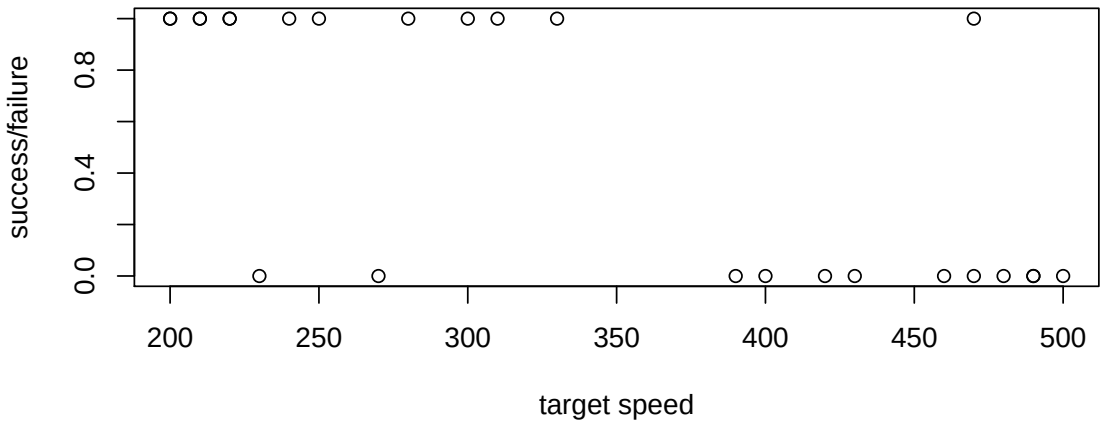

Figure 8.1: Surface-to-air missile successes (1) and failures (0) as they relate to target speed (in knots).

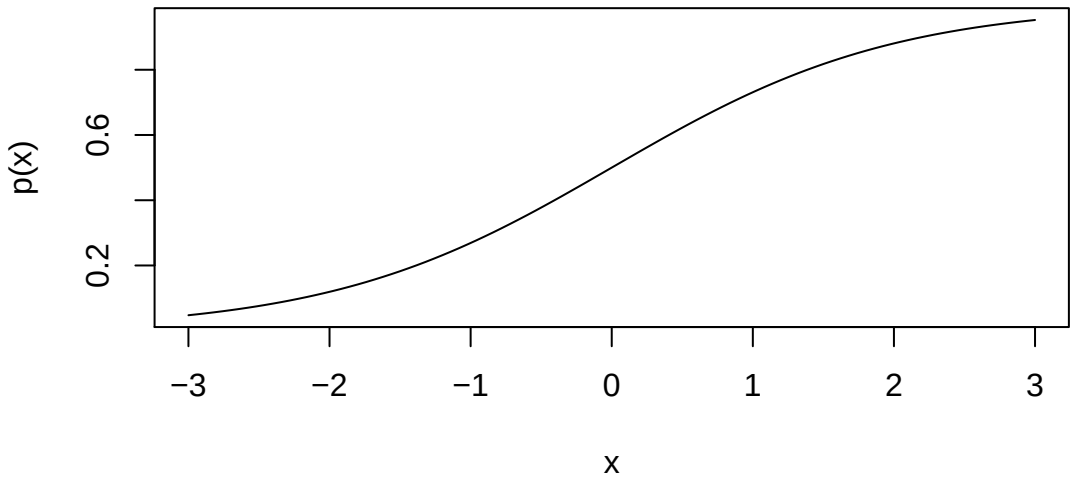

Figure 8.2: The logistic function.

of the covariate(s) is related to a function of the expected response is called a generalized linear model. The logit is an example of a link function, since it links the expected response, in this case the probability $p(x)$ to the linear function of the covariate(s). Other link functions that are popular are the probit, and the complementary log-log. The probit is the inverse of the normal probability distribution function. All of these alternatives are available for use in the `glm()` function through the `binomial()` family function.

To fit the logistic regression model to the missile success data, try

```
p13.glm <- glm(y ~ x, data = p13.1, family = binomial)
summary(p13.glm)

##
## Call:
## glm(formula = y ~ x, family = binomial, data = p13.1)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.062  -0.487   0.392   0.548   2.168
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.07088    2.10900    2.88   0.0040
## x           -0.01770    0.00608   -2.91   0.0036
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.617  on 24  degrees of freedom
## Residual deviance: 20.364  on 23  degrees of freedom
## AIC: 24.36
##
## Number of Fisher Scoring iterations: 4
```

Note that we did not specify the link function; the default choice with the binomial family is the logit.

The Coefficient part of the output tells us that the logit of the probability of success as a linear function of target speed has intercept 6.07 and slope -.0177. Standard error estimates for these parameter estimates are supplied and indicate, in particular, that the slope is clearly negative.

The line itself is not as interesting as the estimated logistic curve which is plotted in Figure 8.3 together with the original data. The curve can now be used to read off specific probabilities of success at the various speeds. Note that in order to obtain the curve, we have used the `predict()` function with `type = "response"`; without specifying `type`, the default is to use the predictions on the linear scale.

```
plot(p13.1, xlab = "target speed", ylab = "success/failure")
newspeeds <- 200:500 # speeds at which we can predict using the fitted model
lines(newspeeds, predict(p13.glm, newdata=data.frame(x = newspeeds),
     type = "response"))
```

Other features of the `glm()` output should be discussed. The dispersion parameter has been taken to be 1. We are assuming that there is no clustering in the data which would have possibly led to overdispersion: the case where the variance exceeds what would be expected under a binomial model. If there is a belief that clustering is occurring (not likely in this example), the `quasibinomial` family should be used instead.

The null deviance refers to a quantity that is calculated for a model that does not include the covariate, in this case `speed`. You can view it and the residual deviance as a generalization of the notion of sum of squares. The residual deviance is calculated for the model which includes the covariate and is considerably smaller than the null deviance, suggesting that the covariate is making a difference to the model fitting. This

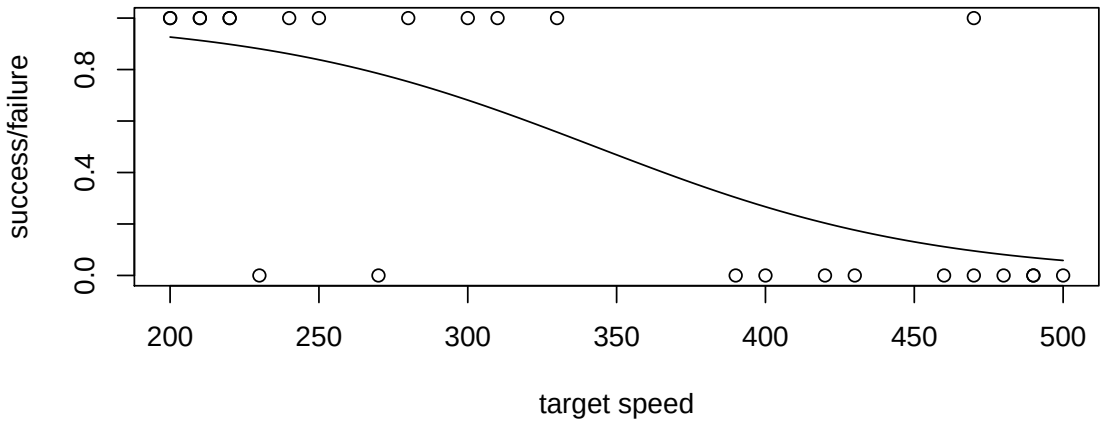

Figure 8.3: Surface-to-air missile successes (1) and failures (0) as they relate to target speed (in knots) with overlaid logistic curve.

agrees with the small p-value, but the comparison with the degrees of freedom, 23, is additionally useful. Under the assumption that the model is correct, the expected value of the residual deviance should be the number of degrees of freedom. Here it is a bit below the degrees of freedom, but not too far off. This is suggestive of a well-fitting model.

8.2 Presence-Absence Data

The `frogs` data in the *DAAG* library contains data on the presence or absence of Southern Corroboree frogs at a number of locations in the Snowy Mountains. Presence is coded as 1 and absence is coded as 0. A number of other covariates are recorded, including `distance` to nearest extant population, `NoOfPools` - the number of potential breeding pools, `meanmin` - the mean minimum Spring temperature, and `meanmax` - the mean maximum Spring temperature. Other variables are also listed, but we will focus on these in order to model the probability of detecting the presence of a frog at a given location:

```
library(DAAG)  # package containing the frogs data set
frogs.glm <- glm(pres.abs ~ log(distance) +
     log(NoOfPools) + meanmin + meanmax, family = binomial, data = frogs)
```

```
summary(frogs.glm)
```

```
##
## Call:
## glm(formula = pres.abs ~ log(distance) + log(NoOfPools) + meanmin +
##     meanmax, family = binomial, data = frogs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.975  -0.722  -0.278   0.797   2.574
##
```

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      18.527      5.267    3.52  0.00044
## log(distance)    -0.755      0.226   -3.34  0.00084
## log(NoOfPools)    0.571      0.215    2.65  0.00800
## meanmin           5.379      1.193    4.51  6.5e-06
## meanmax          -2.382      0.623   -3.82  0.00013
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 279.99  on 211  degrees of freedom
## Residual deviance: 197.66  on 207  degrees of freedom
## AIC: 207.7
##
## Number of Fisher Scoring iterations: 5
```

The fitted model is

$$\widehat{\text{logit}(p)} = 18.5 - .755\log(d) + 0.57\log(N) + 5.38\text{min} - 2.38\text{max}$$

where d is distance, N is number of pools and min and max refer to the mean minimum and mean maximum temperature variables.

The residual deviance is 197.7 on 207 degrees of freedom which is a reasonable value. Thus, the model appears to be an adequate summary of the data.

8.3 Contingency Tables

Binary responses are actually coded categorical variables with 2 levels, and when the covariates are also categorical variables, one can use contingency table analysis. In fact, contingency tables can handle categorical responses with more than 2 levels.

The basic idea of contingency table analysis is to compare the observed counts in the cells of a table with what might be expected if the response and covariates were independent.

An example of a table is counts of individual males by eye color and hair color:

```
 HairEyeColor[,,1]

##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
```

For example, 32 males in the sample had Brown eyes and Black hair. We can use R to compute the expected counts under the assumption eye color and hair color are not associated:

```
HE.chisq <- chisq.test(HairEyeColor[,,1])

## Warning in chisq.test(HairEyeColor[, , 1]): Chi-squared approximation may
be incorrect

HE.chisq$expected
```

```
##        Eye
## Hair     Brown   Blue   Hazel   Green
##   Black 19.670 20.272  9.4337  6.6237
##   Brown 50.229 51.767 24.0896 16.9140
##   Red   11.943 12.308  5.7276  4.0215
##   Blond 16.158 16.652  7.7491  5.4409
```

Thus, under the assumption of independence, instead of 32 Brown-eyed Black-haired males in a sample of this size, we would expect 19.67 and so on. The discrepancy between 32 and 19.67 and all other discrepancies are aggregated into a statistic which is compared with a chisquare distribution in order to obtain a p-value which quantifies the evidence against the hypothesis of no-association.

For this problem, there is a warning suggesting that some of the cells are too small and that the test may not be accurate, so a simulation method can be employed to make the test result more accurate:

```
HE.chisq <- chisq.test(HairEyeColor[,,1], simulate.p.value = TRUE)
HE.chisq

##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  HairEyeColor[, , 1]
## X-squared = 41.3, df = NA, p-value = 5e-04
```

The p-value is very small indicating that there is evidence of an association between hair and eye color for males.

Exercises

1. Estimate the logit of the probability of missile success at a speed of 400 knots. Calculate the probability of missile success. (For this, you can either use the logistic formula, or the `predict()` function in R, using the correct `type`.)

2. The `p13.2` data frame in the *MPV* package has 20 observations on home ownership as it relates to family income. Fit a logistic regression model to the data and use the output to

   (a) identify the logit of the probability of home ownership as a linear function of family income.

   (b) determine if the logistic model is reasonable.

   (c) estimate the probability that a family with an income of $40000 owns their home.

3. The data in `HairEyeColor[,,2]` concern hair and eye color for a sample of females. Conduct a test to see if hair and eye color are associated.

Bibliography

[1] W. John Braun and Duncan Murdoch (2016). *A First Course in Statistical Programming with R* Second Edition. Cambridge University Press.

[2] John H. Maindonald and W. John Braun (2006). *Data Analysis and Graphics.* Third Edition. Cambridge University Press.

[3] John H. Maindonald and W. John Braun (2015). DAAG: Data Analysis and Graphics Data and Functions. R package version 1.22. https://CRAN.R-project.org/package=DAAG

[4] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[5] Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

Index