R (BGU course)

Jonathan D. Rosenblatt

2019-03-27

Contents

Chapter 1

Preface

This book accompanies BGU's "R" course, at the department of Industrial Engineering and Management. It has several purposes:

- Help me organize and document the course material.
- Help students during class so that they may focus on listening and not writing.
- Help students after class, so that they may self-study.

At its current state it is experimental. It can thus be expected to change from time to time, and include mistakes. I will be enormously grateful to whoever decides to share with me any mistakes found.

I am enormously grateful to Yihui Xie, who's *bookdown* R package made it possible to easily write a book which has many mathematical formulae, and R output.

I hope the reader will find this text interesting and useful.

For reproducing my results you will want to run `set.seed(1)`.

1.1 Notation Conventions

In this text we use the following conventions: Lower case x may be a vector or a scalar, random of fixed, as implied by the context. Upper case A will stand for matrices. Equality $=$ is an equality, and $:=$ is a definition. Norm functions are denoted with $\|x\|$ for vector norms, and $\|A\|$ for matrix norms. The type of norm is indicated in the subscript; e.g. $\|x\|_2$ for the Euclidean (l_2) norm. Tag, x' is a transpose. The distribution of a random vector is $\sim$.

1.2 Acknowledgements

I have consulted many people during the writing of this text. I would like to thank Yoav Kessler, Lena Novack, Efrat Vilenski, Ron Sarafian, and Liad Shekel in particular, for their valuable inputs.

Chapter 2

Introduction

2.1 What is R?

R was not designed to be a bona-fide programming language. It is an evolution of the S language, developed at Bell labs (later Lucent) as a wrapper for the endless collection of statistical libraries they wrote in Fortran.

As of 2011, half of R's libraries are actually written in C.

2.2 The R Ecosystem

A large part of R's success is due to the ease in which a user, or a firm, can augment it. This led to a large community of users, developers, and protagonists. Some of the most important parts of R's ecosystem include:

- CRAN: a repository for R packages, mirrored worldwide.
- R-help: an immensely active mailing list. Noways being replaced by StackExchange meta-site. Look for the R tags in the StackOverflow and CrossValidated sites.
- Task Views: part of CRAN that collects packages per topic.
- Bioconductor: A CRAN-like repository dedicated to the life sciences.
- Neuroconductor: A CRAN-like repository dedicated to neuroscience, and neuroimaging.
- Books: An insane amount of books written on the language. Some are free, some are not.
- The Israeli-R-user-group: just like the name suggests.
- Commercial R: being open source and lacking support may seem like a problem that would prohibit R from being adopted for commercial applications. This void is filled by several very successful commercial versions such as Microsoft R, with its accompanying CRAN equivalent called MRAN, Tibco's Spotfire, and others.
- RStudio: since its earliest days R came equipped with a minimal text editor. It later received plugins for major integrated development environments (IDEs) such as Eclipse, WinEdit and even VisualStudio. None of these, however, had the impact of the RStudio IDE. Written completely in JavaScript, the RStudio IDE allows the seamless integration of cutting edge web-design technologies, remote access, and other killer features, making it today's most popular IDE for R.
- RStartHere: a curated list of useful packages.

2.3 Bibliographic Notes

For more on the history of R see AT&T's site, John Chamber's talk at UserR!2014, Nick Thieme's recent report in Significance, or Revolution Analytics' blog.

You can also consult the Introduction chapter of the MASS book (Venables and Ripley, 2013).

Chapter 3

R Basics

We now start with the basics of R. If you have any experience at all with R, you can probably skip this section.

First, make sure you work with the RStudio IDE. Some useful pointers for this IDE include:

- Ctrl+Return(Enter) to run lines from editor.
- Alt+Shift+k for RStudio keyboard shortcuts.
- Ctrl+r to browse the command history.
- Alt+Shift+j to navigate between code sections
- tab for auto-completion
- Ctrl+1 to skip to editor.
- Ctrl+2 to skip to console.
- Ctrl+8 to skip to the environment list.
- Code Folding:
  - Alt+l collapse chunk.
  - Alt+Shift+l unfold chunk.
  - Alt+o collapse all.
  - Alt+Shift+o unfold all.
- Alt+"-" for the assignment operator `<-`.

3.0.1 Other IDEs

Currently, I recommend RStudio, but here are some other IDEs:

1. Jupyter Lab: a very promising IDE, originally designed for Python, that also supports R. At the time of writing, it seems that RStudio is more convenient for R, but it is definitely an IDE to follow closely. See Max Woolf's review.
2. Eclipse: If you are a Java programmer, you are probably familiar with Eclipse, which does have an R plugin: StatEt.
3. Emacs: If you are an Emacs fan, you can find an R plugin: ESS.
4. Vim: Vim-R.
5. Visual Studio also supports R. If you need R for commercial purposes, it may be worthwhile trying Microsoft's R, instead of the usual R. See here for installation instructions.
6. Online version (currently alpha): R Studio Cloud.

3.1 File types

The file types you need to know when using R are the following:

- **.R**: An ASCII text file containing R scripts only.
- **.Rmd**: An ASCII text file. If opened in RStudio can be run as an R-Notebook or compiled using knitr, bookdown, etc.

3.2 Simple calculator

R can be used as a simple calculator. Create a new R Notebook (.Rmd file) within RStudio using File-> New -> R Notebook, and run the following commands.

```
10+5
```

```
## [1] 15
```

```
70*81
```

```
## [1] 5670
```

```
2**4
```

```
## [1] 16
```

```
2^4
```

```
## [1] 16
```

```
log(10)
```

```
## [1] 2.302585
```

```
log(16, 2)
```

```
## [1] 4
```

```
log(1000, 10)
```

```
## [1] 3
```

3.3 Probability calculator

R can be used as a probability calculator. You probably wish you knew this when you did your Intro To Probability classes.

The Binomial distribution function:

```
dbinom(x=3, size=10, prob=0.5)  # Compute P(X=3) for X~B(n=10, p=0.5)
```

```
## [1] 0.1171875
```

Notice that arguments do not need to be named explicitly

```
dbinom(3, 10, 0.5)
```

```
## [1] 0.1171875
```

The Binomial cumulative distribution function (CDF):

```
pbinom(q=3, size=10, prob=0.5) # Compute P(X<=3) for X~B(n=10, p=0.5)
```

```
## [1] 0.171875
```

The Binomial quantile function:

```
qbinom(p=0.1718, size=10, prob=0.5) # For X~B(n=10, p=0.5) returns k such that P(X<=k)=0.1718
```

```
## [1] 3
```

Generate random variables:

```
rbinom(n=10, size=10, prob=0.5)
```

```
##  [1] 4 4 5 7 4 7 7 6 6 3
```

R has many built-in distributions. Their names may change, but the prefixes do not:

- **d** prefix for the *distribution* function.
- **p** prefix for the *cummulative distribution* function (CDF).
- **q** prefix for the *quantile* function (i.e., the inverse CDF).
- **r** prefix to generate random samples.

Demonstrating this idea, using the CDF of several popular distributions:

- `pbinom()` for the Binomial CDF.
- `ppois()` for the Poisson CDF.
- `pnorm()` for the Gaussian CDF.
- `pexp()` for the Exponential CDF.

For more information see `?distributions`.

3.4 Getting Help

One of the most important parts of working with a language, is to know where to find help. R has several in-line facilities, besides the various help resources in the R ecosystem.

Get help for a particular function.

```
?dbinom
help(dbinom)
```

If you don't know the name of the function you are looking for, search local help files for a particular string:

```
??binomial
help.search('dbinom')
```

Or load a menu where you can navigate local help in a web-based fashion:

```
help.start()
```

3.5 Variable Asignment

Assignment of some output into an object named "x":

```
x = rbinom(n=10, size=10, prob=0.5) # Works. Bad style.
x <- rbinom(n=10, size=10, prob=0.5)
```

If you are familiar with other programming languages you may prefer the `=` assignment rather than the `<-` assignment. We recommend you make the effort to change your preferences. This is because thinking with `<-` helps to read your code, distinguishes between assignments and function arguments: think of `function(argument=value)` versus `function(argument<-value)`. It also helps understand special assignment operators such as `<<-` and `->`.
Remark. **Style**: We do not discuss style guidelines in this text, but merely remind the reader that good style is extremely important. When you write code, think of other readers, but also think of future self. See Hadley's style guide for more.

To print the contents of an object just type its name

```
x
```

```
##  [1] 7 4 6 3 4 5 2 5 7 4
```

which is an implicit call to

```
print(x)
```

```
##  [1] 7 4 6 3 4 5 2 5 7 4
```

Alternatively, you can assign and print simultaneously using parenthesis.

```
(x <- rbinom(n=10, size=10, prob=0.5))  # Assign and print.
```

```
##  [1] 5 5 5 4 6 6 6 3 6 5
```

Operate on the object

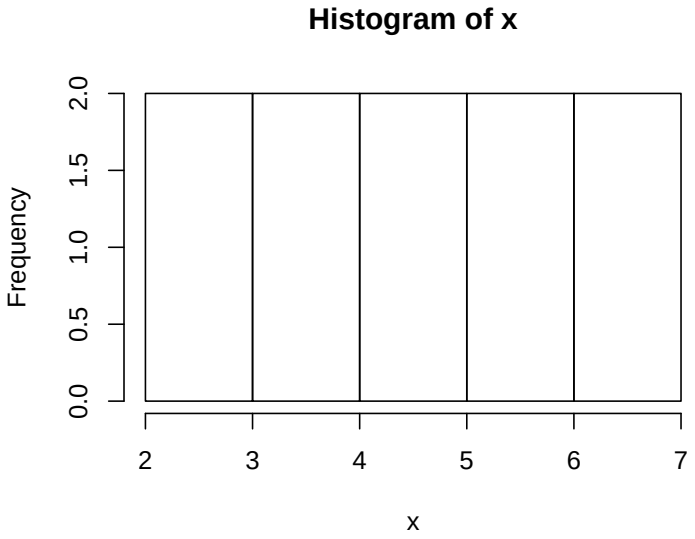
```
mean(x)  # compute mean
```

```
## [1] 5.1
```

```
var(x)  # compute variance
```

```
## [1] 0.9888889
```

```
hist(x) # plot histogram
```



R saves every object you create in RAM[1]. The collection of all such objects is the **workspace** which you can inspect with

```
ls()
```

```
## [1] "x"
```

or with Ctrl+8 in RStudio.

If you lost your object, you can use `ls` with a text pattern to search for it

```
ls(pattern='x')
```

```
## [1] "x"
```

To remove objects from the workspace:

```
rm(x) # remove variable
ls() # verify
```

```
## character(0)
```

You may think that if an object is removed then its memory is freed. This is almost true, and depends on a negotiation mechanism between R and the operating system. R's memory management is discussed in Chapter 15.

[1] S and S-Plus used to save objects on disk. Working from RAM has advantages and disadvantages. More on this in Chapter 15.

3.6 Missing

Unlike typically programming, when working with real life data, you may have **missing** values: measurements that were simply not recorded/stored/etc. *R* has rather sophisticated mechanisms to deal with missing values. It distinguishes between the following types:

1. `NA`: Not Available entries.
2. `NaN`: Not a number.

R tries to defend the analyst, and return an error, or `NA` when the presence of missing values invalidates the calculation:

```
missing.example <- c(10,11,12,NA)
mean(missing.example)
```

```
## [1] NA
```

Most functions will typically have an inner mechanism to deal with these. In the `mean` function, there is an `na.rm` argument, telling *R* how to Remove `NA`s.

```
mean(missing.example, na.rm = TRUE)
```

```
## [1] 11
```

A more general mechanism is removing these manually:

```
clean.example <- na.omit(missing.example)
mean(clean.example)
```

```
## [1] 11
```

3.7 Piping

Because R originates in Unix and Linux environments, it inherits much of its flavor. Piping is an idea taken from the Linux shell which allows to use the output of one expression as the input to another. Piping thus makes code easier to read and write.
Remark. Volleyball fans may be confused with the idea of spiking a ball from the 3-meter line, also called piping. So: (a) These are very different things. (b) If you can pipe, ASA-BGU is looking for you!

Prerequisites:

```
library(magrittr) # load the piping functions
x <- rbinom(n=1000, size=10, prob=0.5) # generate some toy data
```

Examples

```
x %>% var() # Instead of var(x)
x %>% hist()  # Instead of hist(x)
x %>% mean() %>% round(2) %>% add(10)
```

The next example[2] demonstrates the benefits of piping. The next two chunks of code do the same thing. Try parsing them in your mind:

```
# Functional (onion) style
car_data <-
  transform(aggregate(. ~ cyl,
                      data = subset(mtcars, hp > 100),
                      FUN = function(x) round(mean(x, 2))),
            kpl = mpg*0.4251)
```

```
# Piping (magrittr) style
car_data <-
```

[2] Taken from http://cran.r-project.org/web/packages/magrittr/vignettes/magrittr.html

```
mtcars %>%
subset(hp > 100) %>%
aggregate(. ~ cyl, data = ., FUN = . %>% mean %>% round(2)) %>%
transform(kpl = mpg %>% multiply_by(0.4251)) %>%
print
```

Tip: RStudio has a keyboard shortcut for the %>% operator. Try Ctrl+Shift+m.

3.8 Vector Creation and Manipulation

The most basic building block in R is the **vector**. We will now see how to create them, and access their elements (i.e. subsetting). Here are three ways to create the same arbitrary vector:

```
c(10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21) # manually
10:21 # the `:` operator
seq(from=10, to=21, by=1) # the seq() function
```

Let's assign it to the object named "x":

```
x <- c(10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21)
```

Operations usually work element-wise:

```
x+2
```

```
##  [1] 12 13 14 15 16 17 18 19 20 21 22 23
```

```
x*2
```

```
##  [1] 20 22 24 26 28 30 32 34 36 38 40 42
```

```
x^2
```

```
##  [1] 100 121 144 169 196 225 256 289 324 361 400 441
```

```
sqrt(x)
```

```
##  [1] 3.162278 3.316625 3.464102 3.605551 3.741657 3.872983 4.000000
##  [8] 4.123106 4.242641 4.358899 4.472136 4.582576
```

```
log(x)
```

```
##  [1] 2.302585 2.397895 2.484907 2.564949 2.639057 2.708050 2.772589
##  [8] 2.833213 2.890372 2.944439 2.995732 3.044522
```

3.9 Search Paths and Packages

R can be easily extended with packages, which are merely a set of documented functions, which can be loaded or unloaded conveniently. Let's look at the function `read.csv`. We can see its contents by calling it without arguments:

```
read.csv
```

```
## function (file, header = TRUE, sep = ",", quote = "\"", dec = ".",
##     fill = TRUE, comment.char = "", ...)
## read.table(file = file, header = header, sep = sep, quote = quote,
##     dec = dec, fill = fill, comment.char = comment.char, ...)
## <bytecode: 0x3e49070>
## <environment: namespace:utils>
```

Never mind what the function does. Note the `environment: namespace:utils` line at the end. It tells us that this function is part of the **utils** package. We did not need to know this because it is loaded by default. Here are some packages that I have currently loaded:

```
search()
```

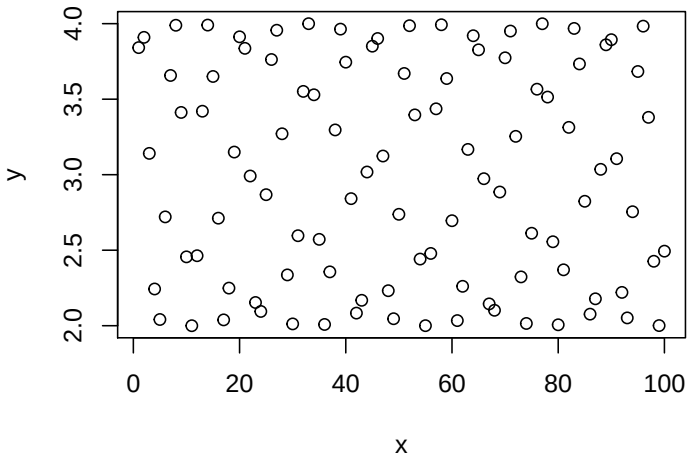
```
##  [1] ".GlobalEnv"           "package:nycflights13" "package:doSNOW"
##  [4] "package:snow"         "package:doParallel"   "package:parallel"
##  [7] "package:iterators"    "package:biganalytics" "package:bigmemory"
## [10] "package:dplyr"        "package:biglm"        "package:DBI"
## [13] "package:MatrixModels" "package:plotly"       "package:kernlab"
## [16] "package:scales"       "package:plyr"         "package:class"
## [19] "package:rpart"        "package:nnet"         "package:e1071"
## [22] "package:glmnet"       "package:foreach"      "package:ellipse"
## [25] "package:nlme"         "package:lattice"      "package:lme4"
## [28] "package:Matrix"       "package:multcomp"     "package:TH.data"
## [31] "package:survival"     "package:mvtnorm"      "package:MASS"
## [34] "package:ggalluvial"   "package:ggplot2"      "package:hexbin"
## [37] "package:data.table"   "package:magrittr"     "tools:rstudio"
## [40] "package:stats"        "package:graphics"     "package:grDevices"
## [43] "package:utils"        "package:datasets"     "package:methods"
## [46] "Autoloads"            "package:base"
```

Other packages can be loaded via the `library` function, or downloaded from the internet using the `install.packages` function before loading with `library`. R's package import mechanism is quite powerful, and is one of the reasons for R's success.

3.10 Simple Plotting

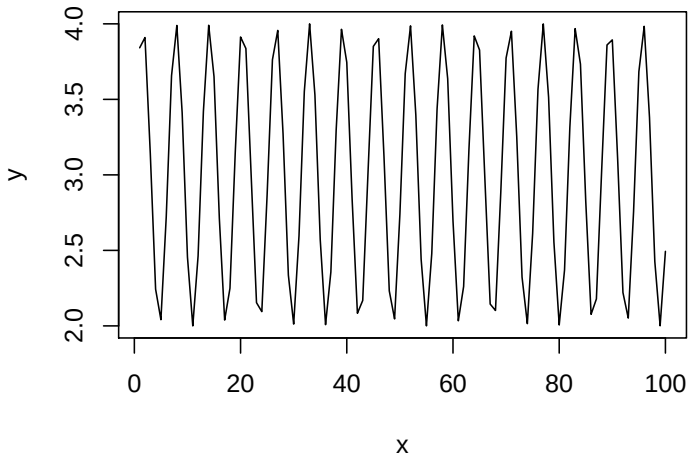
R has many plotting facilities as we will further detail in the Plotting Chapter 12. We start with the simplest facilities, namely, the `plot` function from the **graphics** package, which is loaded by default.

```
x<- 1:100
y<- 3+sin(x)
plot(x = x, y = y) # x,y syntax
```



Given an `x` argument and a `y` argument, `plot` tries to present a scatter plot. We call this the `x,y` syntax. R has another unique syntax to state functional relations. We call `y~x` the "tilde" syntax, which originates in works of Wilkinson and Rogers (1973) and was adopted in the early days of S.

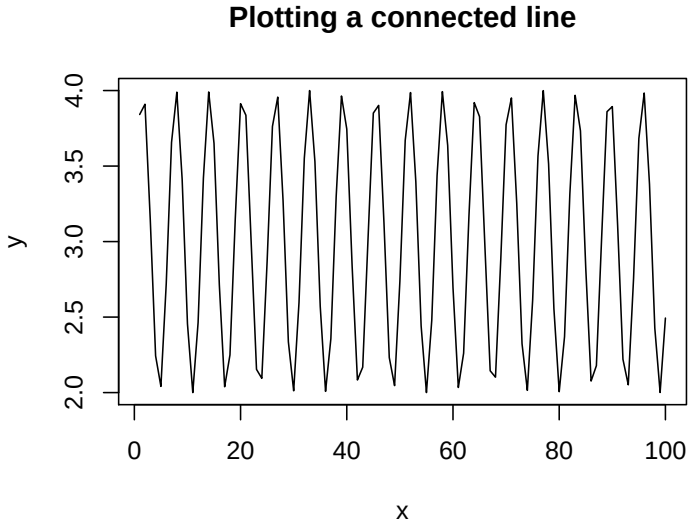
```
plot(y ~ x, type='l') # y~x syntax
```



The syntax `y~x` is read as "y is a function of x". We will prefer the `y~x` syntax over the `x,y` syntax since it is easier to read, and will be very useful when we discuss more complicated models.

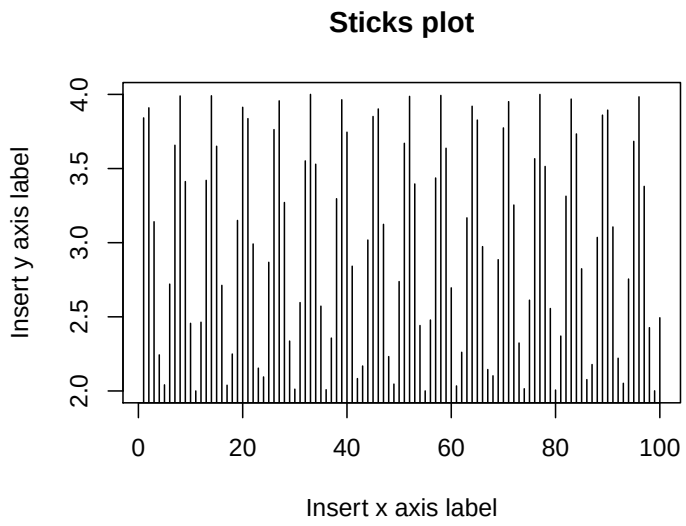
Here are some arguments that control the plot's appearance. We use `type` to control the plot type, `main` to control the main title.

```
plot(y~x, type='l', main='Plotting a connected line')
```



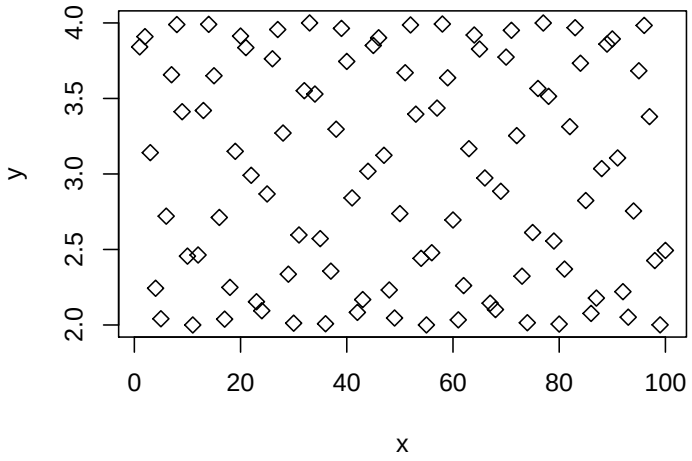
We use `xlab` for the x-axis label, `ylab` for the y-axis.

```
plot(y~x, type='h', main='Sticks plot', xlab='Insert x axis label', ylab='Insert y axis label')
```



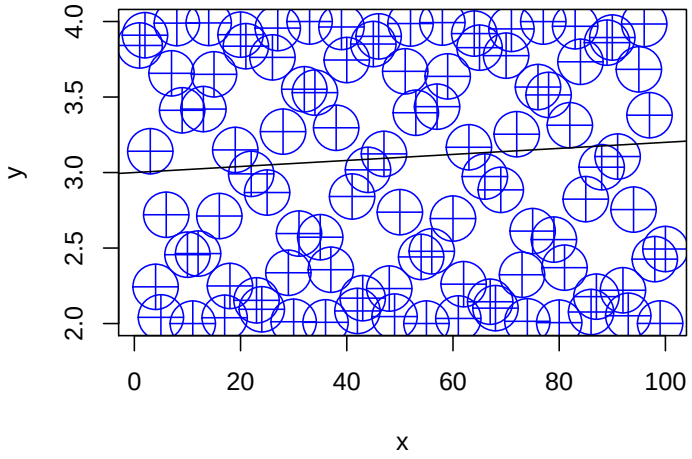
We use `pch` to control the point type (pch is acronym for Plotting CHaracter).

```
plot(y~x, pch=5) # Point type with pcf
```



We use `col` to control the color, `cex` (Character EXpansion) for the point size, and `abline` (y=Bx+A) to add a straight line.

```
plot(y~x, pch=10, type='p', col='blue', cex=4)
abline(3, 0.002)
```



For more plotting options run these

```
example(plot)
example(points)
?plot
help(package='graphics')
```

When your plotting gets serious, go to Chapter 12.

3.11 Object Types

We already saw that the basic building block of R objects is the vector. Vectors can be of the following types:

- **character** Where each element is a string, i.e., a sequence of alphanumeric symbols.
- **numeric** Where each element is a real number in double precision floating point format.
- **integer** Where each element is an integer.
- **logical** Where each element is either TRUE, FALSE, or NA[3]
- **complex** Where each element is a complex number.
- **list** Where each element is an arbitrary R object.
- **factor** Factors are not actually vector objects, but they feel like such. They are used to encode any finite set of values. This will be very useful when fitting linear model because they include information on contrasts, i.e., on the encoding of the factors levels. You should always be alert and recall when you are dealing with a factor or with a character vector. They have different behaviors.

Vectors can be combined into larger objects. A `matrix` can be thought of as the binding of several vectors of the same type. In reality, a matrix is merely a vector with a dimension attribute, that tells R to read it as a matrix and not a vector.

If vectors of different types (but same length) are binded, we get a `data.frame` which is the most fundamental object in R for data analysis. Data frames are brilliant, but a lot has been learned since their invention. They have thus been extended in recent years with the `tbl` class, pronounced [Tibble] (https://cran.r-project.org/web/packages/tibble/vignettes/tibble.html), and the `data.table` class.
The latter is discussed in Chapter 4, and is strongly recommended.

3.12 Data Frames

Creating a simple data frame:

```
x<- 1:10
y<- 3 + sin(x)
frame1 <- data.frame(x=x, sin=y)
```

Let's inspect our data frame:

```
head(frame1)
```

```
##   x      sin
## 1 1 3.841471
## 2 2 3.909297
## 3 3 3.141120
## 4 4 2.243198
## 5 5 2.041076
## 6 6 2.720585
```

Now using the RStudio Excel-like viewer:

```
View(frame1)
```

[3] R uses a **three** valued logic where a missing value (NA) is neither TRUE, nor FALSE.

We highly advise against editing the data this way since there will be no documentation of the changes you made. Always transform your data using scripts, so that everything is documented.

Verifying this is a data frame:

```
class(frame1) # the object is of type data.frame
```

```
## [1] "data.frame"
```

Check the dimension of the data

```
dim(frame1)
```

```
## [1] 10  2
```

Note that checking the dimension of a vector is different than checking the dimension of a data frame.

```
length(x)
```

```
## [1] 10
```

The length of a `data.frame` is merely the number of columns.

```
length(frame1)
```

```
## [1] 2
```

3.13 Exctraction

R provides many ways to subset and extract elements from vectors and other objects. The basics are fairly simple, but not paying attention to the "personality" of each extraction mechanism may cause you a lot of headache.

For starters, extraction is done with the `[` operator. The operator can take vectors of many types.

Extracting element with by integer index:

```
frame1[1, 2]  # exctract the element in the 1st row and 2nd column.
```

```
## [1] 3.841471
```

Extract **column** by index:

```
frame1[,1]
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

Extract column by name:

```
frame1[, 'sin']
```

```
##  [1] 3.841471 3.909297 3.141120 2.243198 2.041076 2.720585 3.656987
##  [8] 3.989358 3.412118 2.455979
```

As a general rule, extraction with `[` will conserve the class of the parent object. There are, however, exceptions. Notice the extraction mechanism and the class of the output in the following examples.

```
class(frame1[, 'sin'])  # extracts a column vector
```

```
## [1] "numeric"
```

```
class(frame1['sin'])  # extracts a data frame
```

```
## [1] "data.frame"
```

```
class(frame1[,1:2])  # extracts a data frame
```

```
## [1] "data.frame"
```

```
class(frame1[2])  # extracts a data frame
```

```
## [1] "data.frame"
```

```
class(frame1[2, ])  # extract a data frame
```

```
## [1] "data.frame"
```

```
class(frame1$sin)  # extracts a column vector
```

```
## [1] "numeric"
```

The `subset()` function does the same

```
subset(frame1, select=sin)
subset(frame1, select=2)
subset(frame1, select= c(2,0))
```

If you want to force the stripping of the class attribute when extracting, try the `[[` mechanism instead of `[`.

```
a <- frame1[1] # [ extraction
b <- frame1[[1]] # [[ extraction
class(a)==class(b) # objects have differing classes
```

```
## [1] FALSE
```

```
a==b # objects are element-wise identical
```

```
##          x
##  [1,] TRUE
##  [2,] TRUE
##  [3,] TRUE
##  [4,] TRUE
##  [5,] TRUE
##  [6,] TRUE
##  [7,] TRUE
##  [8,] TRUE
##  [9,] TRUE
## [10,] TRUE
```

The different types of output classes cause different behaviors. Compare the behavior of `[` on seemingly identical objects.

```
frame1[1][1]
```

```
##    x
## 1  1
## 2  2
## 3  3
## 4  4
## 5  5
## 6  6
## 7  7
## 8  8
## 9  9
## 10 10
```

```
frame1[[1]][1]
```

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

If you want to learn more about subsetting see Hadley's guide.

3.14 Augmentations of the data.frame class

As previously mentioned, the `data.frame` class has been extended in recent years. The best known extensions are the `data.table` and the `tbl`. For beginners, it is important to know R's basics, so we keep focusing on data frames. For more advanced users, I recommend learning the (amazing) `data.table` syntax.

3.15 Data Import and Export

For any practical purpose, you will not be generating your data manually. R comes with many importing and exporting mechanisms which we now present. If, however, you do a lot of data "munging", make sure to see Hadley-verse Chapter 23. If you work with MASSIVE data sets, read the Memory Efficiency Chapter 15.

3.15.1 Import from WEB

The `read.table` function is the main importing workhorse. It can import directly from the web.

```
URL <- 'http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/bone.data'
tirgul1 <- read.table(URL)
```

Always look at the imported result!

```
head(tirgul1)
```

```
##      V1    V2     V3          V4
## 1 idnum   age gender      spnbmd
## 2     1  11.7   male  0.01808067
## 3     1  12.7   male  0.06010929
## 4     1 13.75   male 0.005857545
## 5     2 13.25   male  0.01026393
## 6     2  14.3   male   0.2105263
```

Oh dear. `read.,table` tried to guess the structure of the input, but failed to recognize the header row. Set it manually with `header=TRUE`:

```
tirgul1 <- read.table('data/bone.data', header = TRUE)
head(tirgul1)
```

3.15.2 Import From Clipboard

TODO:datapasta

3.15.3 Export as CSV

Let's write a simple file so that we have something to import

```
head(airquality) #  examine the data to export
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

```
temp.file.name <- tempfile() # get some arbitrary file name
write.csv(x = airquality, file = temp.file.name) # export
```

Now let's import the exported file. Being a .csv file, I can use `read.csv` instead of `read.table`.

```
my.data<- read.csv(file=temp.file.name) # import
head(my.data) # verify import
```

```
##   X Ozone Solar.R Wind Temp Month Day
## 1 1    41     190  7.4   67     5   1
## 2 2    36     118  8.0   72     5   2
## 3 3    12     149 12.6   74     5   3
## 4 4    18     313 11.5   62     5   4
## 5 5    NA      NA 14.3   56     5   5
## 6 6    28      NA 14.9   66     5   6
```

Remark. Windows users may need to use "\" instead of "/".

3.15.4 Export non-CSV files

You can export your R objects in endlessly many ways: If instead of the comma delimiter in .csv you want other column delimiters, look into `?write.table`. If you are exporting only for R users, you can consider exporting as binary objects with `saveRDS`, `feather::write_feather`, or `fst::write.fst`. See (http://www.fstpackage.org/) for a comparison.

3.15.5 Reading From Text Files

Some general notes on importing text files via the `read.table` function. But first, we need to know what is the active directory. Here is how to get and set R's active directory:

```
getwd() #What is the working directory?
setwd() #Setting the working directory in Linux
```

We can now call the `read.table` function to import text files. If you care about your sanity, see `?read.table` before starting imports. Some notable properties of the function:

- `read.table` will try to guess column separators (tab, comma, etc.)
- `read.table` will try to guess if a header row is present.
- `read.table` will convert character vectors to factors unless told not to using the `stringsAsFactors=FALSE` argument.
- The output of `read.table` needs to be explicitly assigned to an object for it to be saved.

3.15.6 Writing Data to Text Files

The function `write.table` is the exporting counterpart of `read.table`.

3.15.7 .XLS(X) files

Strongly recommended to convert to .csv in Excel, and then import as csv. If you still insist see the **xlsx** package.

3.15.8 Massive files

The above importing and exporting mechanisms were not designed for massive files. See the section on the **data.table** package (4), Sparse Representation (14), and Out-of-Ram Algorithms (15) for more on working with massive data files.

3.15.9 Databases

R does not need to read from text files; it can read directly from a database. This is very useful since it allows the filtering, selecting and joining operations to rely on the database's optimized algorithms. Then again, if you will only be analyzing your data with R, you are probably better of by working from a file, without the databases' overhead. See Chapter 15 for more on this matter.

3.16 Functions

One of the most basic building blocks of programming is the ability of writing your own functions. A function in R, like everything else, is an object accessible using its name. We first define a simple function that sums its two arguments

```
my.sum <- function(x,y) {
  return(x+y)
}
my.sum(10,2)
```

```
## [1] 12
```

From this example you may notice that:

- The function `function` tells R to construct a function object.
- Unlike some programming languages, a period (`.`) is allowed as part of an object's name.
- The arguments of the `function`, i.e. `(x,y)`, need to be named but we are not required to specify their class. This makes writing functions very easy, but it is also the source of many bugs, and slowness of R compared to type declaring languages (C, Fortran,Java,...).
- A typical R function does not change objects[4] but rather creates new ones. To save the output of `my.sum` we will need to assign it using the `<-` operator.

Here is a (slightly) more advanced function:

```
my.sum.2 <- function(x, y , absolute=FALSE) {
  if(absolute==TRUE) {
    result <- abs(x+y)
  }
  else{
    result <- x+y
  }
  result
}
my.sum.2(-10,2,TRUE)
```

```
## [1] 8
```

Things to note:

- `if(condition){expression1} else{expression2}` does just what the name suggests.
- The function will output its last evaluated expression. You don't need to use the `return` function explicitly.
- Using `absolute=FALSE` sets the default value of `absolute` to `FALSE`. This is overridden if `absolute` is stated explicitly in the function call.

An important behavior of R is the *scoping rules*. This refers to the way R seeks for variables used in functions. As a rule of thumb, R will first look for variables inside the function and if not found, will search for the variable values in outer environments[5]. Think of the next example.

[4] This is a classical *functional programming* paradigm. If you want an object oriented flavor of R programming, see Hadley's Advanced R book.

[5] More formally, this is called Lexical Scoping.

```
a <- 1
b <- 2
x <- 3
scoping <- function(a,b){
  a+b+x
}
scoping(10,11)
```

```
## [1] 24
```

3.17 Looping

The real power of scripting is when repeated operations are done by iteration. R supports the usual `for`, `while`, and `repated` loops. Here is an embarrassingly simple example

```
for (i in 1:5){
    print(i)
    }
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
```

A slightly more advanced example, is vector multiplication

```
result <- 0
n <- 1e3
x <- 1:n
y <- (1:n)/n
for(i in 1:n){
  result <- result+ x[i]*y[i]
}
```

Remark. **Vector Operations**: You should NEVER write your own vector and matrix products like in the previous example. Only use existing facilities such as `%*%`, `sum()`, etc.
Remark. **Parallel Operations**: If you already know that you will be needing to parallelize your work, get used to working with `foreach` loops in the **foreach** package, rather then regular `for` loops.

3.18 Apply

For applying the same function to a set of elements, there is no need to write an explicit loop. This is such an elementary operation that every programming language will provide some facility to **apply**, or **map** the function to all elements of a set. R provides several facilities to perform this. The most basic of which is `lapply` which applies a function over all elements of a list, and return a list of outputs:

```
the.list <- list(1,'a',mean) # a list of 3 elements from different classes
lapply(X = the.list, FUN = class) # apply the function `class` to each elements
```

```
## [[1]]
## [1] "numeric"
##
## [[2]]
## [1] "character"
##
## [[3]]
## [1] "standardGeneric"
```

```
## attr(,"package")
## [1] "methods"
```

```
sapply(X = the.list, FUN = class) # lapply with cleaned output
```

```
## [1] "numeric"         "character"       "standardGeneric"
```

What is the function you are using requires some arguments? One useful trick is to create your own function that takes only one argument:

```
quantile.25 <- function(x) quantile(x,0.25)
sapply(USArrests, quantile.25)
```

```
##   Murder.25%  Assault.25% UrbanPop.25%     Rape.25%
##        4.075      109.000       54.500       15.075
```

What if you are applying the same function with **two** lists of arguments? Use **mapply**. The following will compute a different quantile to each column in the data:

```
quantiles <- c(0.1, 0.5, 0.3, 0.2)
mapply(quantile, USArrests, quantiles)
```

```
##   Murder.10%  Assault.50% UrbanPop.30%     Rape.20%
##         2.56       159.00        57.70        13.92
```

R provides many variations on `lapply` to facilitate programming. Here is a partial list:

- `sapply`: The same as `lapply` but tries to arrange output in a vector or matrix, and not an unstructured list.
- `vapply`: A safer version of `sapply`, where the output class is pre-specified.
- `apply`: For applying over the rows or columns of matrices.
- `mapply`: For applying functions with more than a single input.
- `tapply`: For splitting vectors and applying functions on subsets.
- `rapply`: A recursive version of `lapply`.
- `eapply`: Like `lapply`, only operates on `environments` instead of lists.
- `Map+Reduce`: For a Common Lisp look and feel of `lapply`.
- `parallel::parLapply`: A parallel version of `lapply` from the package **parallel**.
- `parallel::parLBapply`: A parallel version of `lapply`, with load balancing from the package **parallel**.

3.19 Recursion

The R compiler is really not designed for recursion, and you will rarely need to do so.
See the RCpp Chapter 19 for linking C code, which is better suited for recursion. If you really insist to write recursions in R, make sure to use the `Recall` function, which, as the name suggests, recalls the function in which it is place. Here is a demonstration with the Fibonacci series.

```
fib<-function(n) {
    if (n <= 2) fn<-1
    else fn <- Recall(n - 1) + Recall(n - 2)
    return(fn)
}
fib(5)
```

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

3.20 Strings

Note: this section is courtesy of Ron Sarafian.

Strings may appear as character vectors,files names, paths (directories), graphing elements, and more.

Strings can be concatenated with the super useful `paste` function.

```
a <- "good"
b <- "morning"
is.character(a)
```

```
## [1] TRUE
```

```
paste(a,b)
```

```
## [1] "good morning"
```

```
(c <- paste(a,b, sep = "."))
```

```
## [1] "good.morning"
```

```
paste(a,b,1:3, paste='@@@', collapse = '^^^^')
```

```
## [1] "good morning 1 @@@^^^^good morning 2 @@@^^^^good morning 3 @@@"
```

Things to note:

- `sep` is used to separate strings.
- `collapse` is used to separate results.

The `substr` function extract or replace substrings in a character vector:

```
substr(c, start=2, stop=4)
```

```
## [1] "ood"
```

```
substr(c, start=6, stop=12) <- "evening"
```

The `grep` function is a very powerful tool to search for patterns in text. These patterns are called regular expressions

```
(d <- c(a,b,c))
```

```
## [1] "good"         "morning"      "good.evening"
```

```
grep(pattern = "good",x = d)
```

```
## [1] 1 3
```

```
grep("good",d, value=TRUE, ignore.case=TRUE)
```

```
## [1] "good"         "good.evening"
```

```
grep("([a-zA-Z]+)\\1",d, value=TRUE, perl=TRUE)
```

```
## [1] "good"         "good.evening"
```

Things to note:

- Use `value=TRUE` to return the string itself, instead of its index.
- `([a-zA-Z]+)\\1` is a regular expression to find repeating characters. `perl=TRUE` to activate the Perl "flavored" regular expressions.

Use `gsub` to replace characters in a string object:

```
gsub("o", "q", d) # replace the letter "o" with "q".
```

```
## [1] "gqqd"         "mqrning"      "gqqd.evening"
```

```
gsub("([a-zA-Z]+)\\1", "q", d, perl=TRUE) # replace repeating characters with "q".
```

```
## [1] "gqd"         "morning"     "gqd.evening"
```

The `strsplit` allows to split string vectors to list:

```
(x <- c(a = "thiszis", b = "justzan", c = "example"))
```

```
##         a         b         c
## "thiszis" "justzan" "example"
```

```
strsplit(x, "z") # split x on the letter z
```

```
## $a
## [1] "this" "is"
##
## $b
## [1] "just" "an"
##
## $c
## [1] "example"
```

Some more examples:

```
nchar(x) #  count the nuber of characters in every element of a string vector.
```

```
## a b c
## 7 7 7
```

```
toupper(x) # translate characters in character vectors to upper case
```

```
##         a         b         c
## "THISZIS" "JUSTZAN" "EXAMPLE"
```

```
tolower(toupper(x)) # vice verca
```

```
##         a         b         c
## "thiszis" "justzan" "example"
```

```
letters[1:10] # lower case letters vector
```

```
##  [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"
```

```
LETTERS[1:10] # upper case letters vector
```

```
##  [1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J"
```

```
cat("the sum of", 1, "and", 2, "is", 1+2) #  concatenate and print strings and values
```

```
## the sum of 1 and 2 is 3
```

If you need more than this, look for the stringr package that provides a set of internally consistent tools.

3.21 Dates and Times

Note: This Section is courtesy of Ron Sarafian.

3.21.1 Dates

R provides several packages for dealing with date and date/time data. We start with the `base` package.

R needs to be informed explicitly that an object holds dates. The `as.Date` function convert values to dates. You can pass it a `character`, a `numeric`, or a `POSIXct` (we'll soon explain what it is).

```
start <- "1948-05-14"
class(start)
```

```
## [1] "character"
```

```
start <- as.Date(start)
class(start)
```

```
## [1] "Date"
```

But what if our date is not in the yyyy-mm-dd format? We can tell R what is the character date's format.

```
as.Date("14/5/1948", format="%d/%m/%Y")
```

```
## [1] "1948-05-14"
```

```
as.Date("14may1948", format="%d%b%Y")
```

```
## [1] "1948-05-14"
```

Things to note:

- The format of the date is specified with the `format=` argument. `%d` for day of the month, `/` for separation, `%m` for month, and `%Y` for year in four digits. See `?strptime` for more available formatting.
- If it returns NA, then use the command `Sys.setlocale("LC_TIME","C")`

Many functions are content aware, and adapt their behavior when dealing with dates:

```
(today <- Sys.Date()) # the current date
```

```
## [1] "2019-03-27"
```

```
today + 1 # Add one day
```

```
## [1] "2019-03-28"
```

```
today - start # Diffenrece between dates
```

```
## Time difference of 25884 days
```

```
min(start,today)
```

```
## [1] "1948-05-14"
```

```
month(today)
```

```
## [1] 3
```

3.21.2 Times

Specifying times is similar to dates, only that more formatting parameters are required. The `POSIXct` is the object class for times. It expects strings to be in the format YYYY-MM-DD HH:MM:SS. With `POSIXct` you can also specify the timezone, e.g., `"Asia/Jerusalem"`.

```
time1 <- Sys.time()
class(time1)
```

```
## [1] "POSIXct" "POSIXt"
```

```
time2 <- time1 + 72*60*60 # add 72 hours
time2-time1
```

```
## Time difference of 3 days
```

```
class(time2-time1)
```

```
## [1] "difftime"
```

Things to note:

- Be careful about DST, because `as.POSIXct("2019-03-29 01:30")+3600` will not add 1 hour, but 2 with the result: `[1] "2019-03-29 03:30:00 IDT"`

Compute differences in your unit of choice:

```
difftime(time2,time1, units =  "hour")
```

```
## Time difference of 72 hours
```

```
difftime(time2,time1, units =  "week")
```

```
## Time difference of 0.4285714 weeks
```

Generate sequences:

```
seq(from = time1, to = time2, by = "day")
```

```
## [1] "2019-03-27 22:25:09 IST" "2019-03-28 22:25:09 IST"
## [3] "2019-03-29 23:25:09 IDT" "2019-03-30 23:25:09 IDT"
```

```
seq(time1, by = "month", length.out = 12)
```

```
##  [1] "2019-03-27 22:25:09 IST" "2019-04-27 22:25:09 IDT"
##  [3] "2019-05-27 22:25:09 IDT" "2019-06-27 22:25:09 IDT"
##  [5] "2019-07-27 22:25:09 IDT" "2019-08-27 22:25:09 IDT"
##  [7] "2019-09-27 22:25:09 IDT" "2019-10-27 22:25:09 IST"
##  [9] "2019-11-27 22:25:09 IST" "2019-12-27 22:25:09 IST"
## [11] "2020-01-27 22:25:09 IST" "2020-02-27 22:25:09 IST"
```

3.21.3 lubridate Package

The **lubridate** package replaces many of the **base** package functionality, with a more consistent interface. You only need to specify the order of arguments, not their format:

```
library(lubridate)
ymd("2017/01/31")
```

```
## [1] "2017-01-31"
```

```
mdy("January 31st, 2017")
```

```
## [1] "2017-01-31"
```

```
dmy("31-Jan-2017")
```

```
## [1] "2017-01-31"
```

```
ymd_hms("2000-01-01 00:00:01")
```

```
## [1] "2000-01-01 00:00:01 UTC"
```

```
ymd_hms("20000101000001")
```

```
## [1] "2000-01-01 00:00:01 UTC"
```

Another nice thing in **lubridate**, is that periods can be created with a number of friendly constructor functions that you can combine time objects. E.g.:

```
seconds(1)
```

```
## [1] "1S"
```

```
minutes(c(2,3))
```

```
## [1] "2M 0S" "3M 0S"
```

```
hours(4)
```

```
## [1] "4H 0M 0S"
```

```
days(5)
```

```
## [1] "5d 0H 0M 0S"
```

```
months(c(6,7,8))
```

```
## [1] "6m 0d 0H 0M 0S" "7m 0d 0H 0M 0S" "8m 0d 0H 0M 0S"
```

```
weeks(9)
```

```
## [1] "63d 0H 0M 0S"
```

```
years(10)
```

```
## [1] "10y 0m 0d 0H 0M 0S"
```

```
(t <- ymd_hms("20000101000001"))
```

```
## [1] "2000-01-01 00:00:01 UTC"
```

```
t + seconds(1)
```

```
## [1] "2000-01-01 00:00:02 UTC"
```

```
t + minutes(c(2,3)) + years(10)
```

```
## [1] "2010-01-01 00:02:01 UTC" "2010-01-01 00:03:01 UTC"
```

And you can also extract and assign the time components:

```
t
```

```
## [1] "2000-01-01 00:00:01 UTC"
```

```
second(t)
```

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

```
second(t) <- 26
t
```

```
## [1] "2000-01-01 00:00:26 UTC"
```

Analyzing temporal data is different than actually storing it. If you are interested in time-series analysis, try the **tseries**, **forecast** and **zoo** packages.

3.22 Complex Objects

Say you have a list with many elements, and you want to inspect this list. You can do it using the *Environment* pane in RStudio (Ctrl+8), or using the **str** function:

```
complex.object <- list(7, 'hello', list(a=7,b=8,c=9), FOO=read.csv)
str(complex.object)
```

```
## List of 4
##  $    : num 7
##  $    : chr "hello"
##  $    :List of 3
##   ..$ a: num 7
##   ..$ b: num 8
##   ..$ c: num 9
##  $ FOO:function (file, header = TRUE, sep = ",", quote = "\"", dec = ".",
##      fill = TRUE, comment.char = "", ...)
```

Some (very) advanced users may want a deeper look into object. Try the lobstr package, or the **.Internal(inspect(...))** function described here.

```
x <- c(7,10)
.Internal(inspect(x))
```

```
## @1f076148 14 REALSXP g0c2 [NAM(3)] (len=2, tl=0) 7,10
```

3.23 Vectors and Matrix Products

This section is courtesy of Ron Sarafian.

If you are operating with numeric vectors, or matrices, you may want to compute products. You can easily write your own R loops, but it is much more efficient to use the built-in operations.

Definition 3.1 (Matrix Product)**.** The matrix-product between matrix $n \times m$ matrix A, and $m \times p$ matrix B, is a $n \times p$ matrix C, where:

$$c_{i,j} := \sum_{k=1}^{m} a_{i,k} b_{k,j}$$

Vectors can be seen as single row/column matrices. We can thus use matrix products to define the following:

Definition 3.2 (Dot Product)**.** The dot-product, a.k.a. scalar-product, or inner-product, between row-vectors $x := (x_1, \ldots, x_n)$ and $y := (y_1, \ldots, y_n)$ is defined as the matrix product between the $1 \times n$ matrix x', and the $n \times 1$ matrix y:

$$x'y := \sum_{i} x_i y_i$$

Definition 3.3 (Outer Product)**.** The outer product between row-vectors $x := (x_1, \ldots, x_n)$ and $y := (y_1, \ldots, y_n)$ is defined as the matrix product between the $n \times 1$ matrix x, and the $1 \times n$ matrix y':

$$(xy')_{i,j} := x_i \, y_j$$

Matrix products are computed with the **%*%** operator:

```
x <- rnorm(4)
y <- exp(-x)
t(x) %*% y # Dot product.
```

```
##           [,1]
## [1,] -3.298627
```

```
x %*% y # Dot product.
```

```
##           [,1]
## [1,] -3.298627
```

```
crossprod(x,y) # Dot product.
```

```
##           [,1]
## [1,] -3.298627
```

```
crossprod(t(x),y) # Outer product.
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,] -1.5412664 -0.5513476 -1.7862644 -0.5988587
## [2,]  0.6075926  0.2173503  0.7041748  0.2360800
## [3,] -1.8496379 -0.6616595 -2.1436542 -0.7186764
## [4,]  0.4348046  0.1555399  0.5039206  0.1689432
```

```
crossprod(t(x),t(y)) # Outer product.
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,] -1.5412664 -0.5513476 -1.7862644 -0.5988587
## [2,]  0.6075926  0.2173503  0.7041748  0.2360800
## [3,] -1.8496379 -0.6616595 -2.1436542 -0.7186764
## [4,]  0.4348046  0.1555399  0.5039206  0.1689432
```

```
x %*% t(y) # Outer product
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,] -1.5412664 -0.5513476 -1.7862644 -0.5988587
## [2,]  0.6075926  0.2173503  0.7041748  0.2360800
## [3,] -1.8496379 -0.6616595 -2.1436542 -0.7186764
## [4,]  0.4348046  0.1555399  0.5039206  0.1689432
```

```
x %o% y # Outer product
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,] -1.5412664 -0.5513476 -1.7862644 -0.5988587
## [2,]  0.6075926  0.2173503  0.7041748  0.2360800
## [3,] -1.8496379 -0.6616595 -2.1436542 -0.7186764
## [4,]  0.4348046  0.1555399  0.5039206  0.1689432
```

```
outer(x,y) # Outer product
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,] -1.5412664 -0.5513476 -1.7862644 -0.5988587
## [2,]  0.6075926  0.2173503  0.7041748  0.2360800
## [3,] -1.8496379 -0.6616595 -2.1436542 -0.7186764
## [4,]  0.4348046  0.1555399  0.5039206  0.1689432
```

Things to note:

- The definition of the matrix product has to do with the view of a matrix as a linear operator, and not only a table with numbers. Pick up any linear algebra book to understand why it is defined this way.
- Vectors are matrices. The dot product, is a matrix product where $m = 1$.
- `*` is an element-wise product, whereas `%*%` is a dot product.
- While not specifying whether the vectors are horizontal or vertical, R treats the operation as $(1 \times n) * (n \times 1)$.
- `t()` is the vector/ matrix transpose.

Now for matrix multiplication:

```
(x <- rep(1,5))
```

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

```
(A <- matrix(data = rep(1:5,5), nrow = 5, ncol = 5, byrow = TRUE)) #
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    2    3    4    5
## [2,]    1    2    3    4    5
## [3,]    1    2    3    4    5
## [4,]    1    2    3    4    5
## [5,]    1    2    3    4    5
```

```
x %*% A # (1X5) * (5X5) => (1X5)
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    5   10   15   20   25
```

```
A %*% x # (5X5) * (5X1) => (1X5)
```

```
##      [,1]
```

```
## [1,]   15
## [2,]   15
## [3,]   15
## [4,]   15
## [5,]   15
```

```
0.5 * A
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  0.5    1  1.5    2  2.5
## [2,]  0.5    1  1.5    2  2.5
## [3,]  0.5    1  1.5    2  2.5
## [4,]  0.5    1  1.5    2  2.5
## [5,]  0.5    1  1.5    2  2.5
```

```
A %*% t(A) # Gram matrix
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]   55   55   55   55   55
## [2,]   55   55   55   55   55
## [3,]   55   55   55   55   55
## [4,]   55   55   55   55   55
## [5,]   55   55   55   55   55
```

```
t(x) %*% A %*% x # Quadratic form
```

```
##      [,1]
## [1,]   75
```

Can I write these functions myself? Yes! But a pure-R implementation will be much slower than %*%:

```
my.crossprod <- function(x,y){
  result <- 0
  for(i in 1:length(x)) result <- result + x[i]*y[i]
  result
}
x <- rnorm(1e8)
y <- rnorm(1e8)
system.time(a1 <- my.crossprod(x,y))
```

```
##    user  system elapsed
##   6.445   0.007   6.452
```

```
system.time(a2 <- sum(x*y))
```

```
##    user  system elapsed
##    0.22    0.14    0.36
```

```
system.time(a3 <- c(x%*%y))
```

```
##    user  system elapsed
##   0.349   0.000   0.349
```

```
all.equal(a1,a2)
```

```
## [1] TRUE
```

```
all.equal(a1,a3)
```

```
## [1] TRUE
```

```
all.equal(a2,a3)
```

```
## [1] TRUE
```

3.24 Bibliographic Notes

There are endlessly many introductory texts on R. For a list of free resources see CrossValidated. I personally recommend the official introduction Venables et al. (2004), available online, or anything else Bill Venables writes.

For Importing and Exporting see (https://cran.r-project.org/doc/manuals/r-release/R-data.html). For working with databases see (https://rforanalytics.wordpress.com/useful-links-for-r/odbc-databases-for-r/). For a little intro on time-series objects in R see Cristoph Sax's blog. For working with strings see Gaston Sanchez's book. For advanced R programming see Wickham (2014), available online, or anything else Hadley Wickham writes. For a curated list of recommended packages see here.

3.25 Practice Yourself

1. Load the package **MASS**. That was easy. Now load **ggplot2**, after looking into `install.pacakges()`.
2. Save the numbers 1 to 1,000,000 (`1e6`) into an object named `object`.
3. Write a function that computes the mean of its input. Write a version that uses `sum()`, and another that uses a `for` loop and the summation `+`. Try checking which is faster using `system.time`. Is the difference considerable? Ask me about it in class.
4. Write a function that returns `TRUE` if a number is divisible by 13, `FALSE` if not, and a nice warning to the user if the input is not an integer number.
5. Apply the previous function to all the numbers in `object`. Try using a `for` loop, but also a mapping/apply function.
6. Make a matrix of random numbers using `A <- matrix(rnorm(40), ncol=10, nrow=4)`. Compute the mean of each column. Do it using your own loop and then do the same with `lapply` or `apply`.
7. Make a data frame (`dataA`) with three columns, and 100 rows. The first column with 100 numbers generated from the $\mathcal{N}(10, 1)$ distribution, second column with samples from $Poiss(\lambda = 4)$. The third column contains only `1`.
   Make another data frame (`dataB`) with three columns and 100 rows. Now with $\mathcal{N}(10, 0.5^2)$, $Poiss(\lambda = 4)$ and `2`. Combine the two data frames into an object named `dataAB` with `rbind`. Make a scatter plot of `dataAB` where the x-axes is the first column, the y-axes is the second and define the shape of the points to be the third column.
8. In a sample generated of 1,000 observations from the $\mathcal{N}(10, 1)$ distribution:
   1. What is the proportion of samples smaller than 12.4 ?
   2. What is the 0.23 percentile of the sample?
9. Nothing like cleaning a dataset, to practice your R basics. Have a look at RACHAEL TATMAN collected several datasets which BADLY need some cleansing.

You can also self practice with DataCamp's Intoroduction to R course, or go directly to exercising with R-exercises.

Chapter 4

data.table

`data.table` is an excellent extension of the `data.frame` class. If used as a `data.frame` it will look and feel like a data frame. If, however, it is used with it's unique capabilities, it will prove faster and easier to manipulate. This is because `data.frame`s, like most of R objects, make a copy of themselves when modified. This is known as passing by value, and it is done to ensure that object are not corrupted if an operation fails (if your computer shuts down before the operation is completed, for instance). Making copies of large objects is clearly time and memory consuming. A `data.table` can make changes in place. This is known as passing by reference, which is considerably faster than passing by value.

Let's start with importing some freely available car sales data from Kaggle.

```
library(data.table)
library(magrittr)
auto <- fread('data/autos.csv')
```

```
## Warning in fread("data/autos.csv"): Found and resolved improper
## quoting out-of-sample. First healed line 5263: <<2016-03-29
## 16:46:46,"_SPARDOSE"______Polo_1_4___6N1___60PS___5Tuerer____FESTPREIS,privat,Angebot,
## 500,control,limousine,1999,manuell,60,polo,150000,12,benzin,volkswagen,ja,
## 2016-03-25 00:00:00,0,59581,2016-03-30 11:46:58>>. If the fields are
## not quoted (e.g. field separator does not appear within any field), try
## quote="" to avoid this warning.
```

```
View(auto)
```

```
dim(auto) #  Rows and columns
```

```
## [1] 371824     20
```

```
names(auto) # Variable names
```

```
##  [1] "dateCrawled"         "name"                "seller"
##  [4] "offerType"           "price"               "abtest"
##  [7] "vehicleType"         "yearOfRegistration"  "gearbox"
## [10] "powerPS"             "model"               "kilometer"
## [13] "monthOfRegistration" "fuelType"            "brand"
## [16] "notRepairedDamage"   "dateCreated"         "nrOfPictures"
## [19] "postalCode"          "lastSeen"
```

```
class(auto) # Object class
```

```
## [1] "data.table" "data.frame"
```

```
file.info('data/autos.csv') # File info on disk
```

```
##                    size isdir mode               mtime               ctime
## data/autos.csv 68439217 FALSE  644 2019-02-24 23:52:04 2019-02-24 23:52:04
```

```
##                             atime  uid  gid   uname  grname
## data/autos.csv 2019-03-27 22:15:55 1000 1000 johnros johnros
```

```
gdata::humanReadable(68439217)
```

```
## [1] "65.3 MiB"
```

```
object.size(auto) %>% print(units = 'auto') # File size in memory
```

```
## 103.3 Mb
```

Things to note:

- The import has been done with `fread` instead of `read.csv`. This is more efficient, and directly creates a `data.table` object.
- The import is very fast.
- The data after import is slightly larger than when stored on disk (in this case).

Let's start with verifying that it behaves like a `data.frame` when expected.

```
auto[,2] %>% head
```

```
##                                                 name
## 1:                                        Golf_3_1.6
## 2:                                  A5_Sportback_2.7_Tdi
## 3:                         Jeep_Grand_Cherokee_"Overland"
## 4:                                  GOLF_4_1_4__3T\xdcRER
## 5:                         Skoda_Fabia_1.4_TDI_PD_Classic
## 6: BMW_316i___e36_Limousine___Bastlerfahrzeug__Export
```

```
auto[[2]] %>% head
```

```
## [1] "Golf_3_1.6"
## [2] "A5_Sportback_2.7_Tdi"
## [3] "Jeep_Grand_Cherokee_\"Overland\""
## [4] "GOLF_4_1_4__3T\xdcRER"
## [5] "Skoda_Fabia_1.4_TDI_PD_Classic"
## [6] "BMW_316i___e36_Limousine___Bastlerfahrzeug__Export"
```

```
auto[1,2] %>% head
```

```
##          name
## 1: Golf_3_1.6
```

But notice the difference between `data.frame` and `data.table` when subsetting multiple rows. Uhh!

```
auto[1:3] %>% dim # data.table will exctract *rows*
```

```
## [1]  3 20
```

```
as.data.frame(auto)[1:3] %>% dim # data.frame will exctract *columns*
```

```
## [1] 371824      3
```

Just use columns (`,`) and be explicit regarding the dimension you are extracting…

Now let's do some `data.table` specific operations. The general syntax has the form `DT[i,j,by]`. SQL users may think of `i` as `WHERE`, `j` as `SELECT`, and `by` as `GROUP BY`. We don't need to name the arguments explicitly. Also, the `Tab` key will typically help you to fill in column names.

```
auto[,vehicleType,] %>% table # Exctract column and tabulate
```

```
## .
##                andere        bus     cabrio      coupe kleinwagen
##      37899       3362      30220      22914      19026      80098
##      kombi  limousine        suv
##      67626      95963      14716
```

```
auto[vehicleType=='coupe',,] %>% dim # Exctract rows
```

```
## [1] 19026    20
```

```
auto[,gearbox:model,] %>% head # exctract column range
```

```
##      gearbox powerPS model
## 1:   manuell       0  golf
## 2:   manuell     190
## 3: automatik     163 grand
## 4:   manuell      75  golf
## 5:   manuell      69 fabia
## 6:   manuell     102   3er
```

```
auto[,gearbox,] %>% table
```

```
## .
##           automatik   manuell
##     20223     77169    274432
```

```
auto[vehicleType=='coupe' & gearbox=='automatik',,] %>% dim # intersect conditions
```

```
## [1] 6008   20
```

```
auto[,table(vehicleType),] # uhh? why would this even work?!?
```

```
## vehicleType
##                andere        bus     cabrio      coupe kleinwagen
##      37899       3362      30220      22914      19026      80098
##      kombi  limousine        suv
##      67626      95963      14716
```

```
auto[, mean(price), by=vehicleType] # average price by car group
```

```
## Warning in gmean(price): The sum of an integer column for a group was more
## than type 'integer' can hold so the result has been coerced to 'numeric'
## automatically for convenience.
```

```
##    vehicleType         V1
## 1:              20124.688
## 2:       coupe  25951.506
## 3:         suv  13252.392
## 4:  kleinwagen   5691.167
## 5:   limousine  11111.107
## 6:      cabrio  15072.998
## 7:         bus  10300.686
## 8:       kombi   7739.518
## 9:      andere 676327.100
```

The `.N` operator is very useful if you need to count the length of the result. Notice where I use it:

```
auto[.N-1,,] # will exctract the *last* row
```

```
##            dateCrawled                   name seller offerType price
## 1: 2016-03-20 19:41:08 VW_Golf_Kombi_1_9l_TDI privat   Angebot  3400
##    abtest vehicleType yearOfRegistration gearbox powerPS model kilometer
## 1:   test       kombi               2002 manuell     100  golf    150000
##    monthOfRegistration fuelType      brand notRepairedDamage
## 1:                   6   diesel volkswagen
##            dateCreated nrOfPictures postalCode            lastSeen
## 1: 2016-03-20 00:00:00            0      40764 2016-03-24 12:45:21
```

```
auto[,.N] # will count rows
```

```
## [1] 371824
```

```
auto[,.N, vehicleType] # will count rows by type
```

```
##    vehicleType     N
## 1:             37899
## 2:       coupe 19026
## 3:         suv 14716
## 4:  kleinwagen 80098
## 5:   limousine 95963
## 6:      cabrio 22914
## 7:         bus 30220
## 8:       kombi 67626
## 9:      andere  3362
```

You may concatenate results into a vector:

```
auto[,c(mean(price), mean(powerPS)),]
```

```
## [1] 17286.2996   115.5414
```

This `c()` syntax no longer behaves well if splitting:

```
auto[,c(mean(price), mean(powerPS)), by=vehicleType]
```

```
##     vehicleType           V1
##  1:             20124.68801
##  2:                71.23249
##  3:       coupe  25951.50589
##  4:       coupe    172.97614
##  5:         suv  13252.39182
##  6:         suv    166.01903
##  7:  kleinwagen   5691.16738
##  8:  kleinwagen     68.75733
##  9:   limousine  11111.10661
## 10:   limousine    132.26936
## 11:      cabrio  15072.99782
## 12:      cabrio    145.17684
## 13:         bus  10300.68561
## 14:         bus    113.58137
## 15:       kombi   7739.51760
## 16:       kombi    136.40654
## 17:      andere 676327.09964
## 18:      andere    102.11154
```

Use a `list()` instead of `c()`, within `data.table` commands:

```
auto[,list(mean(price), mean(powerPS)), by=vehicleType]
```

```
## Warning in gmean(price): The sum of an integer column for a group was more
## than type 'integer' can hold so the result has been coerced to 'numeric'
## automatically for convenience.
```

```
##    vehicleType         V1        V2
## 1:             20124.688  71.23249
## 2:       coupe  25951.506 172.97614
## 3:         suv  13252.392 166.01903
## 4:  kleinwagen   5691.167  68.75733
## 5:   limousine  11111.107 132.26936
## 6:      cabrio  15072.998 145.17684
```

```
## 7:         bus  10300.686 113.58137
## 8:       kombi   7739.518 136.40654
## 9:      andere 676327.100 102.11154
```

You can add names to your new variables:

```
auto[,list(Price=mean(price), Power=mean(powerPS)), by=vehicleType]
```

```
## Warning in gmean(price): The sum of an integer column for a group was more
## than type 'integer' can hold so the result has been coerced to 'numeric'
## automatically for convenience.
```

```
##    vehicleType      Price     Power
## 1:              20124.688  71.23249
## 2:       coupe  25951.506 172.97614
## 3:         suv  13252.392 166.01903
## 4:  kleinwagen   5691.167  68.75733
## 5:   limousine  11111.107 132.26936
## 6:      cabrio  15072.998 145.17684
## 7:         bus  10300.686 113.58137
## 8:       kombi   7739.518 136.40654
## 9:      andere 676327.100 102.11154
```

You can use `.()` to replace the longer `list()` command:

```
auto[,.(Price=mean(price), Power=mean(powerPS)), by=vehicleType]
```

```
## Warning in gmean(price): The sum of an integer column for a group was more
## than type 'integer' can hold so the result has been coerced to 'numeric'
## automatically for convenience.
```

```
##    vehicleType      Price     Power
## 1:              20124.688  71.23249
## 2:       coupe  25951.506 172.97614
## 3:         suv  13252.392 166.01903
## 4:  kleinwagen   5691.167  68.75733
## 5:   limousine  11111.107 132.26936
## 6:      cabrio  15072.998 145.17684
## 7:         bus  10300.686 113.58137
## 8:       kombi   7739.518 136.40654
## 9:      andere 676327.100 102.11154
```

And split by multiple variables:

```
auto[,.(Price=mean(price), Power=mean(powerPS)), by=.(vehicleType,fuelType)] %>% head
```

```
## Warning in gmean(price): The sum of an integer column for a group was more
## than type 'integer' can hold so the result has been coerced to 'numeric'
## automatically for convenience.
```

```
##    vehicleType fuelType     Price     Power
## 1:              benzin 11820.443  70.14477
## 2:       coupe   diesel 51170.248 179.48704
## 3:         suv   diesel 15549.369 168.16115
## 4:  kleinwagen   benzin  5786.514  68.74309
## 5:  kleinwagen   diesel  4295.550  76.83666
## 6:   limousine   benzin  6974.360 127.87025
```

Compute with variables created on the fly:

```
auto[,sum(price<1e4),] # Count prices lower than 10,000
```

```
## [1] 310497
```

```
auto[,mean(price<1e4),] # Proportion of prices lower than 10,000
```

```
## [1] 0.8350644
```

```
auto[,.(Power=mean(powerPS)), by=.(PriceRange=price>1e4)]
```

```
##    PriceRange    Power
## 1:      FALSE 101.8838
## 2:       TRUE 185.9029
```

You may sort along one or more columns

```
auto[order(-price), price,] %>% head # Order along price. Descending
```

```
## [1] 2147483647   99999999   99999999   99999999   99999999   99999999
```

```
auto[order(price, -lastSeen), price,] %>% head# Order along price and last seen . Ascending and descending.
```

```
## [1] 0 0 0 0 0 0
```

You may apply a function to ALL columns using a Subset of the Data using `.SD`

```
count.uniques <- function(x) length(unique(x))
auto[,lapply(.SD, count.uniques), vehicleType]
```

```
##    vehicleType dateCrawled  name seller offerType price abtest
## 1:                  36714 32891      1         2  1378      2
## 2:       coupe      18745 13182      1         2  1994      2
## 3:         suv      14549  9707      1         1  1667      2
## 4:  kleinwagen      75591 49302      2         2  1927      2
## 5:   limousine      89352 58581      2         1  2986      2
## 6:      cabrio      22497 13411      1         1  2014      2
## 7:         bus      29559 19651      1         2  1784      2
## 8:       kombi      64415 41976      2         1  2529      2
## 9:      andere       3352  3185      1         1   562      2
##    yearOfRegistration gearbox powerPS model kilometer monthOfRegistration
## 1:                101       3     374   244        13                  13
## 2:                 75       3     414   117        13                  13
## 3:                 73       3     342   122        13                  13
## 4:                 75       3     317   163        13                  13
## 5:                 83       3     506   210        13                  13
## 6:                 88       3     363    95        13                  13
## 7:                 65       3     251   106        13                  13
## 8:                 64       3     393   177        13                  13
## 9:                 81       3     230   162        13                  13
##    fuelType brand notRepairedDamage dateCreated nrOfPictures postalCode
## 1:        8    40                3          65            1       6304
## 2:        8    35                3          51            1       5159
## 3:        8    37                3          61            1       4932
## 4:        8    38                3          68            1       7343
## 5:        8    39                3          82            1       7513
## 6:        7    38                3          70            1       5524
## 7:        8    33                3          63            1       6112
## 8:        8    38                3          75            1       7337
## 9:        8    38                3          41            1       2220
##    lastSeen
## 1:    32813
## 2:    16568
## 3:    13367
## 4:    59354
## 5:    65813
```

```
## 6:    19125
## 7:    26094
## 8:    50668
## 9:     3294
```

Things to note:

- `.SD` is the data subset after splitting along the `by` argument.
- Recall that `lapply` applies the same function to all elements of a list. In this example, to all columns of `.SD`.

If you want to apply a function only to a subset of columns, use the `.SDcols` argument

```
auto[,lapply(.SD, count.uniques), by=vehicleType, .SDcols=price:gearbox]
```

```
##    vehicleType price abtest vehicleType yearOfRegistration gearbox
## 1:              1378      2           1                101       3
## 2:       coupe  1994      2           1                 75       3
## 3:         suv  1667      2           1                 73       3
## 4:  kleinwagen  1927      2           1                 75       3
## 5:   limousine  2986      2           1                 83       3
## 6:      cabrio  2014      2           1                 88       3
## 7:         bus  1784      2           1                 65       3
## 8:       kombi  2529      2           1                 64       3
## 9:      andere   562      2           1                 81       3
```

4.1 Make your own variables

It is very easy to compute new variables

```
auto[,log(price/powerPS),] %>% head # This makes no sense
```

```
## [1]      Inf 4.567632 4.096387 2.995732 3.954583 1.852000
```

And if you want to store the result in a new variable, use the `:=` operator

```
auto[,newVar:=log(price/powerPS),]
```

Or create multiple variables at once. The syntax `c("A","B"):=.(expression1,expression2)`is read "save the **list** of results from expression1 and expression2 using the **vector** of names A, and B".

```
auto[,c('newVar','newVar2'):=.(log(price/powerPS),price^2/powerPS),]
```

4.2 Join

data.table can be used for joining. A *join* is the operation of aligning two (or more) data frames/tables along some index. The index can be a single variable, or a combination thereof.

Here is a simple example of aligning age and gender from two different data tables:

```
DT1 <- data.table(Names=c("Alice","Bob"), Age=c(29,31))
DT2 <- data.table(Names=c("Alice","Bob","Carl"), Gender=c("F","M","M"))
setkey(DT1, Names)
setkey(DT2, Names)
DT1[DT2,,]
```

```
##    Names Age Gender
## 1: Alice  29      F
## 2:   Bob  31      M
## 3:  Carl  NA      M
```

```
DT2[DT1,,]
```

```
##    Names Gender Age
## 1: Alice      F  29
## 2:   Bob      M  31
```

Things to note:

- A join with `data.tables` is performed by indexing one `data.table` with another. Which is the outer and which is the inner will affect the result.
- The indexing variable needs to be set using the `setkey` function.

There are several types of joins:

- **Inner join**: Returns the rows along the intersection of keys, i.e., rows that appear in **all** data sets.
- **Outer join**: Returns the rows along the union of keys, i.e., rows that appear in **any** of the data sets.
- **Left join**: Returns the rows along the index of the "left" data set.
- **Right join**: Returns the rows along the index of the "right" data set.

Assuming `DT1` is the "left" data set, we see that `DT1[DT2,,]` is a right join, and `DT2[DT1,,]` is a left join. For an inner join use the `nomath=0` argument:

```
DT1[DT2,,,nomatch=0]
```

```
##    Names Age Gender
## 1: Alice  29      F
## 2:   Bob  31      M
```

```
DT2[DT1,,,nomatch=0]
```

```
##    Names Gender Age
## 1: Alice      F  29
## 2:   Bob      M  31
```

4.3 Reshaping data

Data sets (i.e. frames or tables) may arrive in a "wide" form or a "long" form. The difference is best illustrated with an example. The `ChickWeight` data encodes the weight of various chicks. It is "long" in that a variable encodes the time of measurement, making the data, well, simply long:

```
ChickWeight %>%  head
```

```
## Grouped Data: weight ~ Time | Chick
##   weight Time Chick Diet
## 1     42    0     1    1
## 2     51    2     1    1
## 3     59    4     1    1
## 4     64    6     1    1
## 5     76    8     1    1
## 6     93   10     1    1
```

The `mtcars` data encodes 10 characteristics of 32 types of automobiles. It is "wide" since the various characteristics are encoded in different variables, making the data, well, simply wide.

```
mtcars %>% head
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

```
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Most of *R*'s functions, with exceptions, will prefer data in the long format. There are thus various facilities to convert from one format to another. We will focus on the `melt` and `dcast` functions to convert from one format to another.

4.3.1 Wide to long

`melt` will convert from wide to long.

```
dimnames(mtcars)
```

```
## [[1]]
##  [1] "Mazda RX4"           "Mazda RX4 Wag"       "Datsun 710"
##  [4] "Hornet 4 Drive"      "Hornet Sportabout"   "Valiant"
##  [7] "Duster 360"          "Merc 240D"           "Merc 230"
## [10] "Merc 280"            "Merc 280C"           "Merc 450SE"
## [13] "Merc 450SL"          "Merc 450SLC"         "Cadillac Fleetwood"
## [16] "Lincoln Continental" "Chrysler Imperial"   "Fiat 128"
## [19] "Honda Civic"         "Toyota Corolla"      "Toyota Corona"
## [22] "Dodge Challenger"    "AMC Javelin"         "Camaro Z28"
## [25] "Pontiac Firebird"    "Fiat X1-9"           "Porsche 914-2"
## [28] "Lotus Europa"        "Ford Pantera L"      "Ferrari Dino"
## [31] "Maserati Bora"       "Volvo 142E"
##
## [[2]]
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
mtcars$type <- rownames(mtcars)
melt(mtcars, id.vars=c("type")) %>% head
```

```
##                type variable value
## 1         Mazda RX4      mpg  21.0
## 2     Mazda RX4 Wag      mpg  21.0
## 3        Datsun 710      mpg  22.8
## 4    Hornet 4 Drive      mpg  21.4
## 5 Hornet Sportabout      mpg  18.7
## 6           Valiant      mpg  18.1
```

Things to note:

- The car type was originally encoded in the rows' names, and not as a variable. We thus created an explicit variable with the cars' type using the `rownames` function.
- The `id.vars` of the `melt` function names the variables that will be used as identifiers. All other variables are assumed to be measurements. These can have been specified using their index instead of their name.
- If not all variables are measurements, we could have names measurement variables explicitly using the `measure.vars` argument of the `melt` function. These can have been specified using their index instead of their name.
- By default, the molten columns are automatically named `variable` and `value`.

We can replace the automatic namings using `variable.name` and `value.name`:

```
melt(mtcars, id.vars=c("type"), variable.name="Charachteristic", value.name="Measurement") %>% head
```

```
##                type Charachteristic Measurement
## 1         Mazda RX4             mpg        21.0
## 2     Mazda RX4 Wag             mpg        21.0
## 3        Datsun 710             mpg        22.8
## 4    Hornet 4 Drive             mpg        21.4
## 5 Hornet Sportabout             mpg        18.7
## 6           Valiant             mpg        18.1
```

4.3.2 Long to wide

`dcast` will convert from long to wide:

```
dcast(ChickWeight, Chick~Time, value.var="weight")
```

```
##    Chick  0  2  4  6   8  10  12  14  16  18  20  21
## 1     18 39 35 NA NA  NA  NA  NA  NA  NA  NA  NA  NA
## 2     16 41 45 49 51  57  51  54  NA  NA  NA  NA  NA
## 3     15 41 49 56 64  68  68  67  68  NA  NA  NA  NA
## 4     13 41 48 53 60  65  67  71  70  71  81  91  96
## 5      9 42 51 59 68  85  96  90  92  93 100 100  98
## 6     20 41 47 54 58  65  73  77  89  98 107 115 117
## 7     10 41 44 52 63  74  81  89  96 101 112 120 124
## 8      8 42 50 61 71  84  93 110 116 126 134 125  NA
## 9     17 42 51 61 72  83  89  98 103 113 123 133 142
## 10    19 43 48 55 62  65  71  82  88 106 120 144 157
## 11     4 42 49 56 67  74  87 102 108 136 154 160 157
## 12     6 41 49 59 74  97 124 141 148 155 160 160 157
## 13    11 43 51 63 84 112 139 168 177 182 184 181 175
## 14     3 43 39 55 67  84  99 115 138 163 187 198 202
## 15     1 42 51 59 64  76  93 106 125 149 171 199 205
## 16    12 41 49 56 62  72  88 119 135 162 185 195 205
## 17     2 40 49 58 72  84 103 122 138 162 187 209 215
## 18     5 41 42 48 60  79 106 141 164 197 199 220 223
## 19    14 41 49 62 79 101 128 164 192 227 248 259 266
## 20     7 41 49 57 71  89 112 146 174 218 250 288 305
## 21    24 42 52 58 74  66  68  70  71  72  72  76  74
## 22    30 42 48 59 72  85  98 115 122 143 151 157 150
## 23    22 41 55 64 77  90  95 108 111 131 148 164 167
## 24    23 43 52 61 73  90 103 127 135 145 163 170 175
## 25    27 39 46 58 73  87 100 115 123 144 163 185 192
## 26    28 39 46 58 73  92 114 145 156 184 207 212 233
## 27    26 42 48 57 74  93 114 136 147 169 205 236 251
## 28    25 40 49 62 78 102 124 146 164 197 231 259 265
## 29    29 39 48 59 74  87 106 134 150 187 230 279 309
## 30    21 40 50 62 86 125 163 217 240 275 307 318 331
## 31    33 39 50 63 77  96 111 137 144 151 146 156 147
## 32    37 41 48 56 68  80  83 103 112 135 157 169 178
## 33    36 39 48 61 76  98 116 145 166 198 227 225 220
## 34    31 42 53 62 73  85 102 123 138 170 204 235 256
## 35    39 42 50 61 78  89 109 130 146 170 214 250 272
## 36    38 41 49 61 74  98 109 128 154 192 232 280 290
## 37    32 41 49 65 82 107 129 159 179 221 263 291 305
## 38    40 41 55 66 79 101 120 154 182 215 262 295 321
## 39    34 41 49 63 85 107 134 164 186 235 294 327 341
## 40    35 41 53 64 87 123 158 201 238 287 332 361 373
## 41    44 42 51 65 86 103 118 127 138 145 146  NA  NA
## 42    45 41 50 61 78  98 117 135 141 147 174 197 196
## 43    43 42 55 69 96 131 157 184 188 197 198 199 200
## 44    41 42 51 66 85 103 124 155 153 175 184 199 204
## 45    47 41 53 66 79 100 123 148 157 168 185 210 205
## 46    49 40 53 64 85 108 128 152 166 184 203 233 237
## 47    46 40 52 62 82 101 120 144 156 173 210 231 238
## 48    50 41 54 67 84 105 122 155 175 205 234 264 264
## 49    42 42 49 63 84 103 126 160 174 204 234 269 281
## 50    48 39 50 62 80 104 125 154 170 222 261 303 322
```

Things to note:

- `dcast` uses a formula interface (`~`) to specify the row identifier and the variables. The LHS is the row identifier, and the RHS for the variables to be created.
- The measurement of each LHS at each RHS, is specified using the `value.var` argument.

4.4 Bibliographic Notes

`data.table` has excellent online documentation. See here. See here for **joining**. See here for more on **reshaping**. See here for a comparison of the `data.frame` way, versus the `data.table` way. For some advanced tips and tricks see Andrew Brooks' blog.

4.5 Practice Yourself

1. Create a matrix of ones with `1e5` rows and `1e2` columns. Create a `data.table` using this matrix.
    1. Replace the first column of each, with the sequence $1, 2, 3, \ldots$.
    2. Create a column which is the sum of all columns, and a $\mathcal{N}(0, 1)$ random variable.
2. Use the cars dataset used in this chapter from kaggle Kaggle.
    1. Import the data using the function `fread`. What is the class of your object?
    2. Use `system.time()` to measure the time to sort along "seller". Do the same after converting the data to `data.frame`. Are data tables faster?

Also, see DataCamp's Data Manipulation in R with data.table, by Matt Dowle, the author of *data.table* for more self practice.

Chapter 5

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a term coined by John W. Tukey in his seminal book (Tukey, 1977). It is also (arguably) known as *Visual Analytics*, or *Descriptive Statistics*. It is the practice of inspecting, and exploring your data, before stating hypotheses, fitting predictors, and other more ambitious inferential goals. It typically includes the computation of simple *summary statistics* which capture some property of interest in the data, and *visualization*. EDA can be thought of as an assumption free, purely algorithmic practice.

In this text we present EDA techniques along the following lines:

- How we explore: with summary-statistics, or visually?
- How many variables analyzed simultaneously: univariate, bivariate, or multivariate?
- What type of variable: categorical or continuous?

5.1 Summary Statistics

5.1.1 Categorical Data

Categorical variables do not admit any mathematical operations on them. We cannot sum them, or even sort them. We can only **count** them. As such, summaries of categorical variables will always start with the counting of the frequency of each category.

5.1.1.1 Summary of Univariate Categorical Data

```
# Make some data
gender <- c(rep('Boy', 10), rep('Girl', 12))
drink <- c(rep('Coke', 5), rep('Sprite', 3), rep('Coffee', 6), rep('Tea', 7), rep('Water', 1))
age <-  sample(c('Young', 'Old'), size = length(gender), replace = TRUE)
# Count frequencies
table(gender)
```

```
## gender
##  Boy Girl
##   10   12
```

```
table(drink)
```

```
## drink
## Coffee   Coke Sprite    Tea  Water
##      6      5      3      7      1
```

```
table(age)
```

```
## age
##   Old Young
##    10    12
```

If instead of the level counts you want the proportions, you can use `prop.table`

```
prop.table(table(gender))
```

```
## gender
##       Boy      Girl
## 0.4545455 0.5454545
```

5.1.1.2 Summary of Bivariate Categorical Data

```
library(magrittr)
cbind(gender, drink) %>% head # bind vectors into matrix and inspect
```

```
##      gender drink
## [1,] "Boy"  "Coke"
## [2,] "Boy"  "Coke"
## [3,] "Boy"  "Coke"
## [4,] "Boy"  "Coke"
## [5,] "Boy"  "Coke"
## [6,] "Boy"  "Sprite"
```

```
table1 <- table(gender, drink) # count frequencies of bivariate combinations
table1
```

```
##       drink
## gender Coffee Coke Sprite Tea Water
##   Boy       2    5      3   0     0
##   Girl      4    0      0   7     1
```

5.1.1.3 Summary of Multivariate Categorical Data

You may be wondering how does R handle tables with more than two dimensions. It is indeed not trivial to report this in a human-readable way. R offers several solutions: `table` is easier to compute with, and `ftable` is human readable.

```
table2.1 <- table(gender, drink, age) # A machine readable table.
table2.1
```

```
## , , age = Old
##
##       drink
## gender Coffee Coke Sprite Tea Water
##   Boy       0    3      1   0     0
##   Girl      1    0      0   5     0
##
## , , age = Young
##
##       drink
## gender Coffee Coke Sprite Tea Water
##   Boy       2    2      2   0     0
##   Girl      3    0      0   2     1
```

```
table.2.2 <- ftable(gender, drink, age) # A human readable table.
table.2.2
```

```
##               age Old Young
## gender drink
```

```
## Boy    Coffee       0     2
##        Coke         3     2
##        Sprite       1     2
##        Tea          0     0
##        Water        0     0
## Girl   Coffee       1     3
##        Coke         0     0
##        Sprite       0     0
##        Tea          5     2
##        Water        0     1
```

If you want proportions instead of counts, you need to specify the denominator, i.e., the margins. Think: what is the margin in each of the following outputs?

```
prop.table(table1, margin = 1)
```

```
##       drink
## gender     Coffee       Coke     Sprite        Tea      Water
##   Boy  0.20000000 0.50000000 0.30000000 0.00000000 0.00000000
##   Girl 0.33333333 0.00000000 0.00000000 0.58333333 0.08333333
```

```
prop.table(table1, margin = 2)
```

```
##       drink
## gender    Coffee      Coke    Sprite       Tea     Water
##   Boy  0.3333333 1.0000000 1.0000000 0.0000000 0.0000000
##   Girl 0.6666667 0.0000000 0.0000000 1.0000000 1.0000000
```

5.1.2 Continous Data

Continuous variables admit many more operations than categorical. We can compute sums, means, quantiles, and more.

5.1.2.1 Summary of Univariate Continuous Data

We distinguish between several types of summaries, each capturing a different property of the data.

5.1.2.2 Summary of Location

Capture the "location" of the data. These include:

Definition 5.1 (Average)**.** The mean, or average, of a sample $x := (x_1, \ldots, x_n)$, denoted $\bar{x}$ is defined as

$$\bar{x} := n^{-1} \sum x_i.$$

The sample mean is **non robust**. A single large observation may inflate the mean indefinitely. For this reason, we define several other summaries of location, which are more robust, i.e., less affected by "contaminations" of the data.

We start by defining the sample quantiles, themselves **not** a summary of location.

Definition 5.2 (Quantiles)**.** The α quantile of a sample x, denoted x_α, is (non uniquely) defined as a value above $100\alpha\%$ of the sample, and below $100(1-\alpha)\%$.

We emphasize that sample quantiles are non-uniquely defined. See `?quantile` for the 9(!) different definitions that R provides.

Using the sample quantiles, we can now define another summary of location, the **median**.

Definition 5.3 (Median)**.** The median of a sample x, denoted $x_{0.5}$ is the $\alpha = 0.5$ quantile of the sample.

A whole family of summaries of locations is the **alpha trimmed mean**.

Definition 5.4 (Alpha Trimmed Mean)**.** The α trimmed mean of a sample x, denoted $\bar{x}_\alpha$ is the average of the sample after removing the α proportion of largest and α proportion of smallest observations.

The simple mean and median are instances of the alpha trimmed mean: $\bar{x}_0$ and $\bar{x}_{0.5}$ respectively.

Here are the R implementations:

```
x <- rexp(100) # generate some random data
mean(x) # simple mean
```

```
## [1] 1.017118
```

```
median(x) # median
```

```
## [1] 0.5805804
```

```
mean(x, trim = 0.2) # alpha trimmed mean with alpha=0.2
```

```
## [1] 0.7711528
```

5.1.2.3 Summary of Scale

The *scale* of the data, sometimes known as *spread*, can be thought of its variability.

Definition 5.5 (Standard Deviation)**.** The standard deviation of a sample x, denoted $S(x)$, is defined as

$$S(x) := \sqrt{(n-1)^{-1} \sum (x_i - \bar{x})^2}.$$

For reasons of robustness, we define other, more robust, measures of scale.

Definition 5.6 (MAD)**.** The Median Absolute Deviation from the median, denoted as $MAD(x)$, is defined as

$$MAD(x) := c\,|x - x_{0.5}|_{0.5}.$$

where c is some constant, typically set to $c = 1.4826$ so that MAD and $S(x)$ have the same large sample limit.

Definition 5.7 (IQR)**.** The Inter Quantile Range of a sample x, denoted as $IQR(x)$, is defined as

$$IQR(x) := x_{0.75} - x_{0.25}.$$

Here are the R implementations

```
sd(x) # standard deviation
```

```
## [1] 0.9981981
```

```
mad(x) # MAD
```

```
## [1] 0.6835045
```

```
IQR(x) # IQR
```

```
## [1] 1.337731
```

5.1.2.4 Summary of Asymmetry

Summaries of asymmetry, also known as *skewness*, quantify the departure of the x from a symmetric sample.

Definition 5.8 (Yule)**.** The Yule measure of assymetry, denoted $Yule(x)$ is defined as

$$Yule(x) := \frac{1/2\,(x_{0.75} + x_{0.25}) - x_{0.5}}{1/2\,IQR(x)}.$$

Here is an R implementation

```
yule <- function(x){
  numerator <- 0.5 * (quantile(x,0.75) + quantile(x,0.25))-median(x)
  denominator <- 0.5* IQR(x)
  c(numerator/denominator, use.names=FALSE)
}
yule(x)
```

```
## [1] 0.2080004
```

5.1.2.5 Summary of Bivariate Continuous Data

When dealing with bivariate, or multivariate data, we can obviously compute univariate summaries for each variable separately. This is not the topic of this section, in which we want to summarize the association **between** the variables, and not within them.

Definition 5.9 (Covariance)**.** The covariance between two samples, x and y, of same length n, is defined as

$$Cov(x,y) := (n-1)^{-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

We emphasize this is not the covariance you learned about in probability classes, since it is not the covariance between two *random variables* but rather, between two *samples*. For this reasons, some authors call it the *empirical covariance*, or *sample covariance*.

Definition 5.10 (Pearson's Correlation Coefficient)**.** Peasrson's correlation coefficient, a.k.a. Pearson's moment product correlation, or simply, the correlation, denoted `r(x,y)`, is defined as

$$r(x,y) := \frac{Cov(x,y)}{S(x)S(y)}.$$

If you find this definition enigmatic, just think of the correlation as the covariance between x and y after transforming each to the unitless scale of z-scores.

Definition 5.11 (Z-Score)**.** The z-scores of a sample x are defined as the mean-centered, scale normalized observations:

$$z_i(x) := \frac{x_i - \bar{x}}{S(x)}.$$

We thus have that $r(x,y) = Cov(z(x), z(y))$.

5.1.2.6 Summary of Multivariate Continuous Data

The covariance is a simple summary of association between two variables, but it certainly may not capture the whole "story" when dealing with more than two variables. The most common summary of multivariate relation, is the **covariance matrix**, but we warn that only the simplest multivariate relations are fully summarized by this matrix.

Definition 5.12 (Sample Covariance Matrix)**.** Given n observations on p variables, denote $x_{i,j}$ the i'th observation of the j'th variable. The *sample covariance matrix*, denoted $\hat{\Sigma}$ is defined as

$$\hat{\Sigma}_{k,l} = (n-1)^{-1} \sum_i [(x_{i,k} - \bar{x}_k)(x_{i,l} - \bar{x}_l)],$$

where $\bar{x}_k := n^{-1} \sum_i x_{i,k}$. Put differently, the k,l'th entry in $\hat{\Sigma}$ is the sample covariance between variables k and l.

Remark. $\hat{\Sigma}$ is clearly non robust. How would you define a robust covariance matrix?

5.2 Visualization

Summarizing the information in a variable to a single number clearly conceals much of the story in the sample. This is like inspecting a person using a caricature, instead of a picture. Visualizing the data, when possible, is more informative.

5.2.1 Categorical Data

Recalling that with categorical variables we can only count the frequency of each level, the plotting of such variables are typically variations on the *bar plot*.

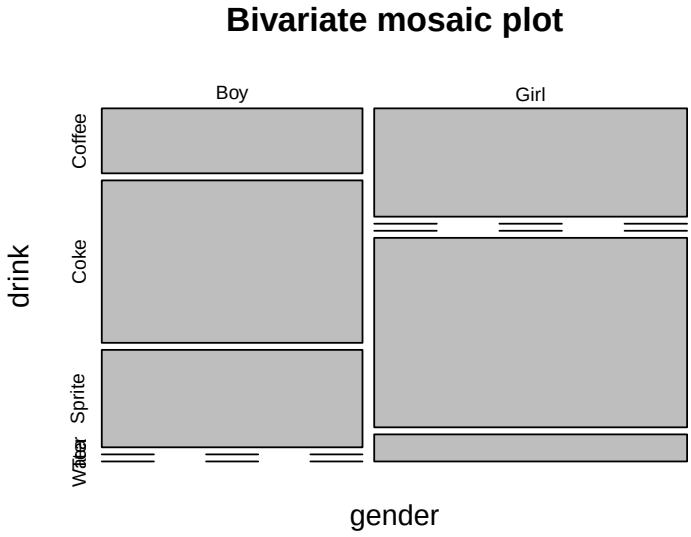
5.2.1.1 Visualizing Univariate Categorical Data

```
barplot(table(age))
```

5.2.1.2 Visualizing Bivariate Categorical Data

There are several generalizations of the barplot, aimed to deal with the visualization of bivariate categorical data. They are sometimes known as the *clustered bar plot* and the *stacked bar plot*. In this text, we advocate the use of the *mosaic plot* which is also the default in R.

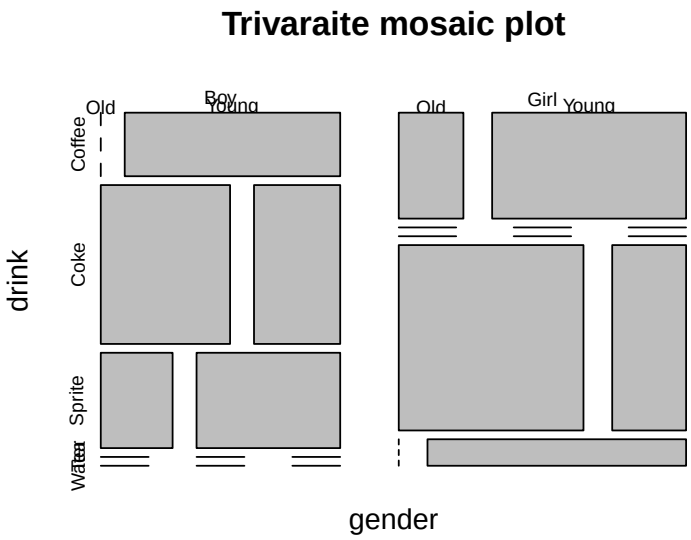
```
plot(table1, main='Bivariate mosaic plot')
```



5.2.1.3 Visualizing Multivariate Categorical Data

The *mosaic plot* is not easy to generalize to more than two variables, but it is still possible (at the cost of interpretability).

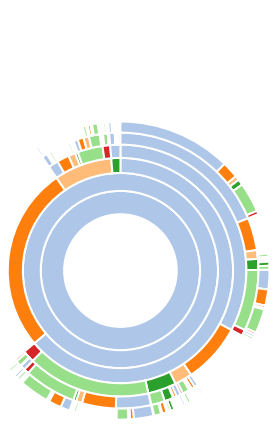
```
plot(table2.1, main='Trivaraite mosaic plot')
```



When one of the variables is a (discrete) time variable, then the plot has a notion dynamics in time. For this see the Alluvian plot 5.3.1.

If the variables represent a hierarchy, consider a **Sunburst Plot**:

```
library(sunburstR)
# read in sample visit-sequences.csv data provided in source
# https://gist.github.com/kerryrodden/7090426#file-visit-sequences-csv
sequences <- read.csv(
  system.file("examples/visit-sequences.csv",package="sunburstR")
  ,header=F
  ,stringsAsFactors = FALSE
)
sunburst(sequences) # In the HTML version of the book this plot is interactive.
```

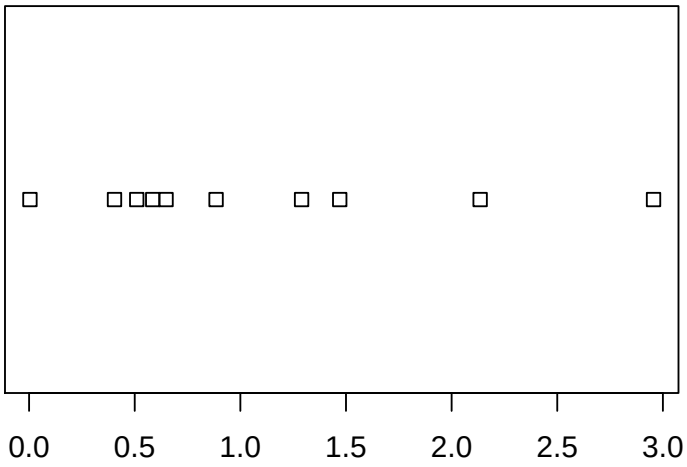


5.2.2 Continuous Data

5.2.2.1 Visualizing Univariate Continuous Data

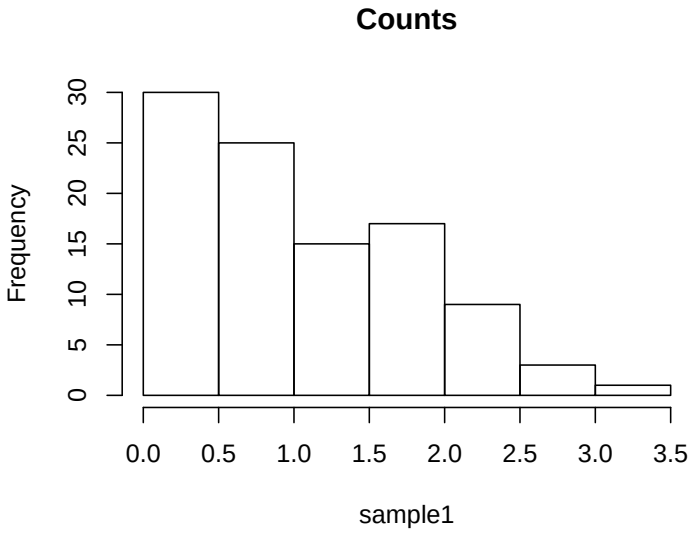
Unlike categorical variables, there are endlessly many way to visualize continuous variables. The simplest way is to look at the raw data via the `stripchart`.

```
sample1 <- rexp(10)
stripchart(sample1)
```



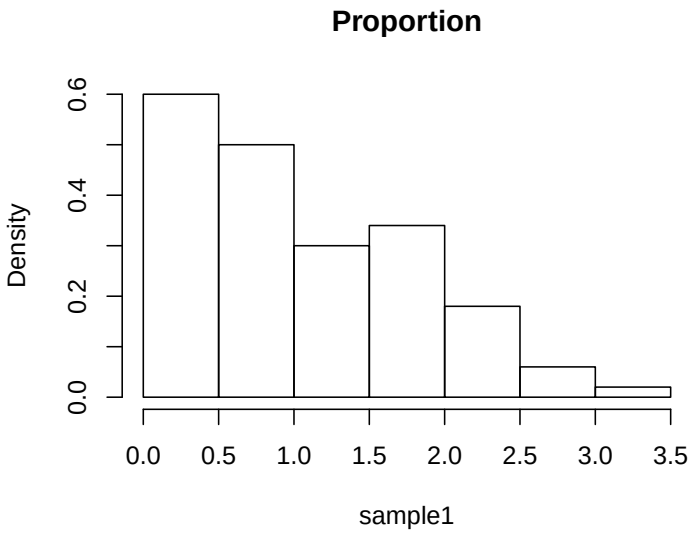
Clearly, if there are many observations, the `stripchart` will be a useless line of black dots. We thus bin them together, and look at the frequency of each bin; this is the *histogram*. R's `histogram` function has very good defaults to choose the number of bins. Here is a histogram showing the counts of each bin.

```
sample1 <- rexp(100)
hist(sample1, freq=T, main='Counts')
```



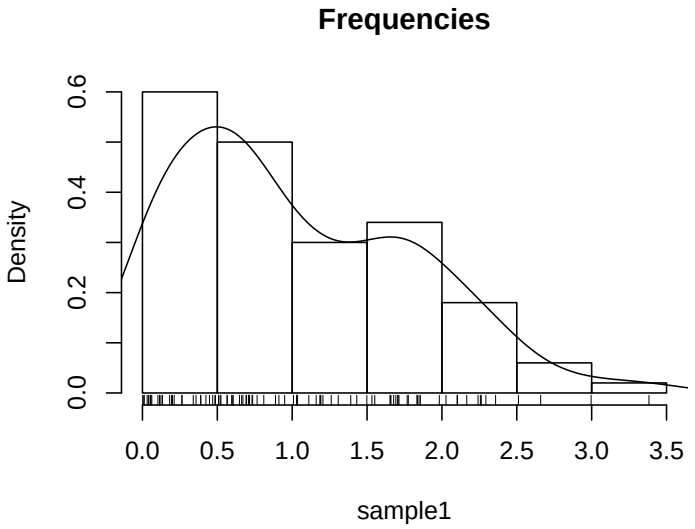
The bin counts can be replaced with the proportion of each bin using the `freq` argument.

```
hist(sample1, freq=F, main='Proportion')
```



The bins of a histogram are non overlapping. We can adopt a sliding window approach, instead of binning. This is the *density plot* which is produced with the `density` function, and added to an existing plot with the `lines` function. The `rug` function adds the original data points as ticks on the axes, and is strongly recommended to detect artifacts introduced by the binning of the histogram, or the smoothing of the density plot.

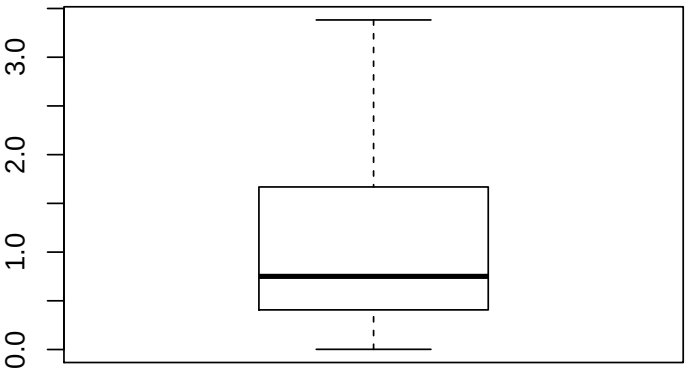
```
hist(sample1, freq=F, main='Frequencies')
lines(density(sample1))
rug(sample1)
```



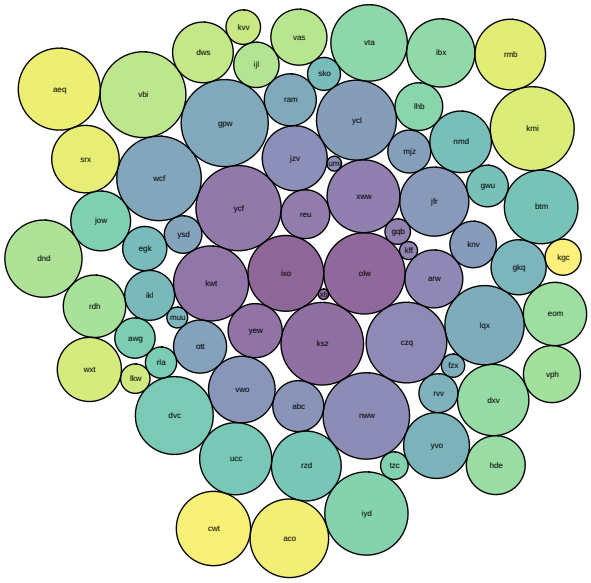
Remark. Why would it make no sense to make a table, or a barplot, of continuous data?

One particularly useful visualization, due to John W. Tukey, is the *boxplot*. The boxplot is designed to capture the main phenomena in the data, and simultaneously point to outlines.

```
boxplot(sample1)
```



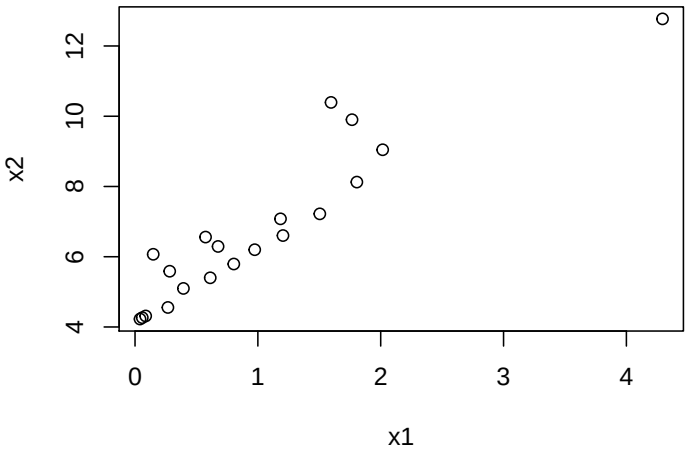
Another way to deal with a massive amount of data points, is to emphasize important points, and conceal non-important. This is the purpose of **circle-packing** (example from r-graph gallery):



5.2.2.2 Visualizing Bivariate Continuous Data

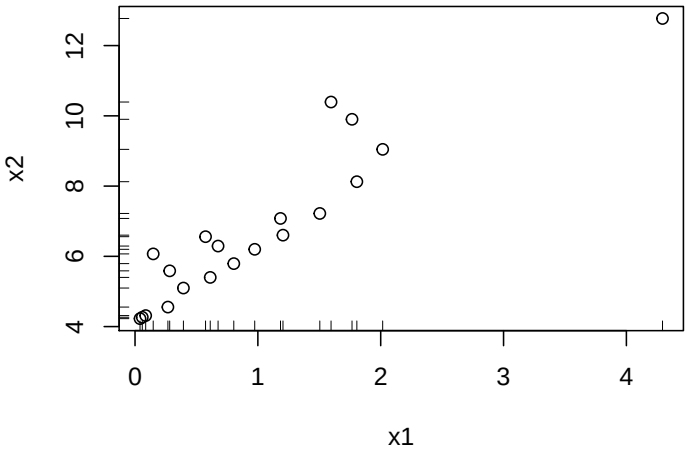
The bivariate counterpart of the `stipchart` is the celebrated scatter plot.

```
n <- 20
x1 <- rexp(n)
x2 <- 2* x1 + 4 + rexp(n)
plot(x2~x1)
```

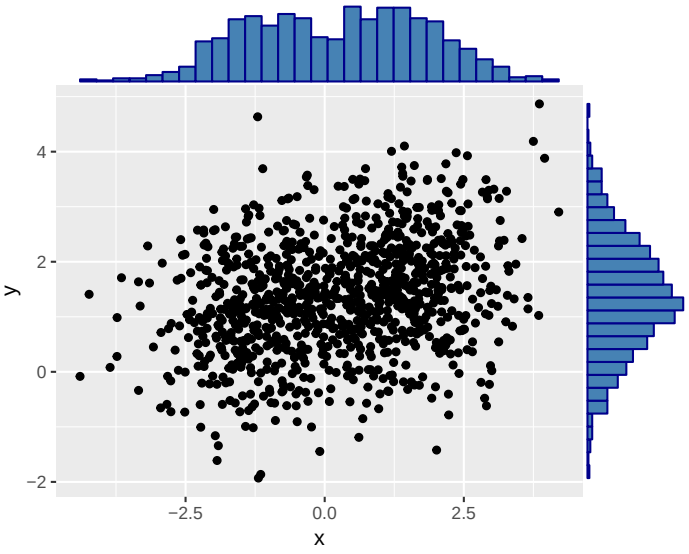


A scatter-plot may be augmented with marginal univariate visualization. See, for instance, the *rug* function to add the raw data on the margins:

```
plot(x2~x1)
rug(x1,side = 1)
rug(x2,side = 2)
```

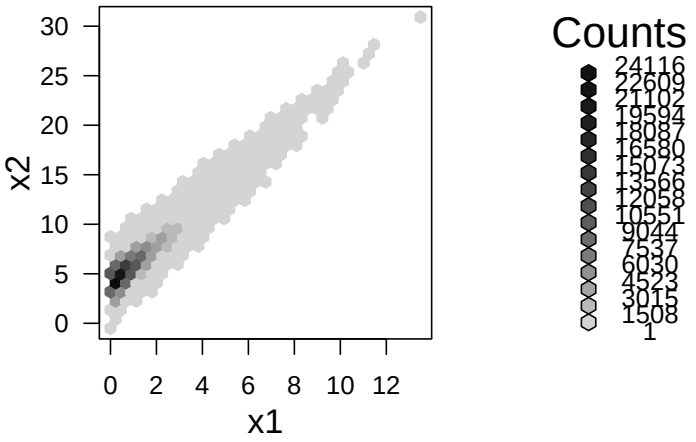


A fancier version may use a histogram on the margins:



Like the univariate `stripchart`, the scatter plot will be an uninformative mess in the presence of a lot of data. A nice bivariate counterpart of the univariate histogram is the *hexbin plot*, which tessellates the plane with hexagons, and reports their frequencies.

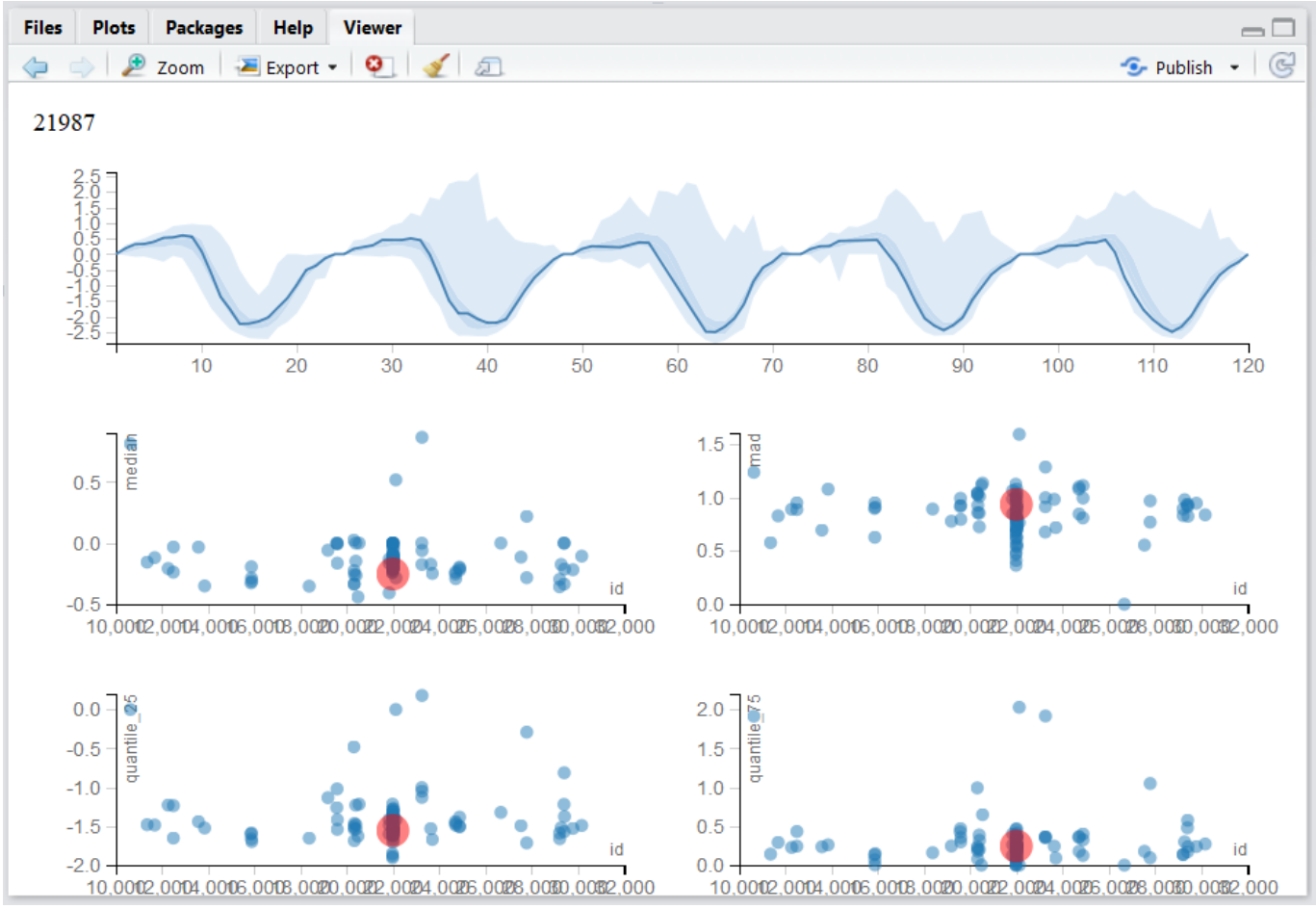
```
library(hexbin) # load required library
n <- 2e5
x1 <- rexp(n)
x2 <- 2* x1 + 4 + rnorm(n)
plot(hexbin(x = x1, y = x2))
```



5.2.2.3 Visualizing Multivariate Continuous Data

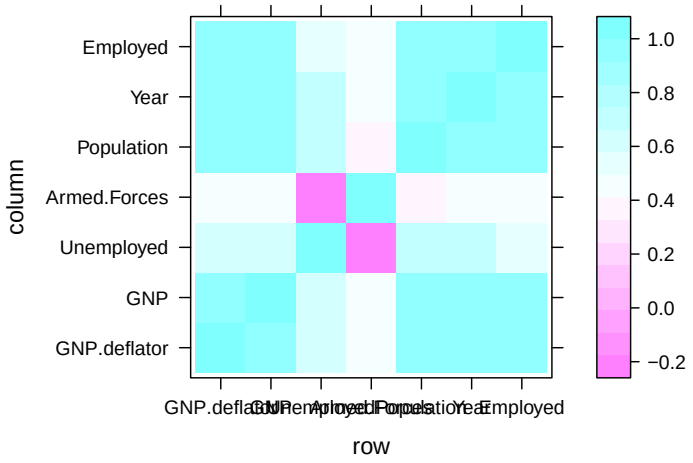
Visualizing multivariate data is a tremendous challenge given that we cannot grasp 4 dimensional spaces, nor can the computer screen present more than 2 dimensional spaces. We thus have several options: (i) To project the data to 2D. This is discussed in the Dimensionality Reduction Section 11.1. (ii) To visualize not the raw data, but rather its summaries, like the covariance matrix.

Our own Multinav package adopts an interactive approach. For each (multivaraite) observation a simple univariate summary may be computed and visualized. These summaries may be compared, and the original (multivariate) observation inspected upon demand. Contact Efrat for more details.



An alternative approach starts with the a covariance matrix, $\hat{\Sigma}$, that can be visualized as an image. Note the use of the :: operator, which is used to call a function from some package, without loading the whole package. We will use the :: operator when we want to emphasize the package of origin of a function.

```
covariance <- cov(longley) # The covariance of the longley dataset
correlations <- cor(longley) # The correlations of the longley dataset
lattice::levelplot(correlations)
```



If we believe the covariance has some structure, we can do better then viewing the raw correlations. In temporal, and spatial data, we believe correlations decay as some function of distances. We can thus view correlations as a function of the distance between observations. This is known as a *variogram*. Note that for a variogram to be informative, is it implied that correlations are merely a function of distances (and not locations themselves). This is formally known as *stationary* and *isotropic* correlations.

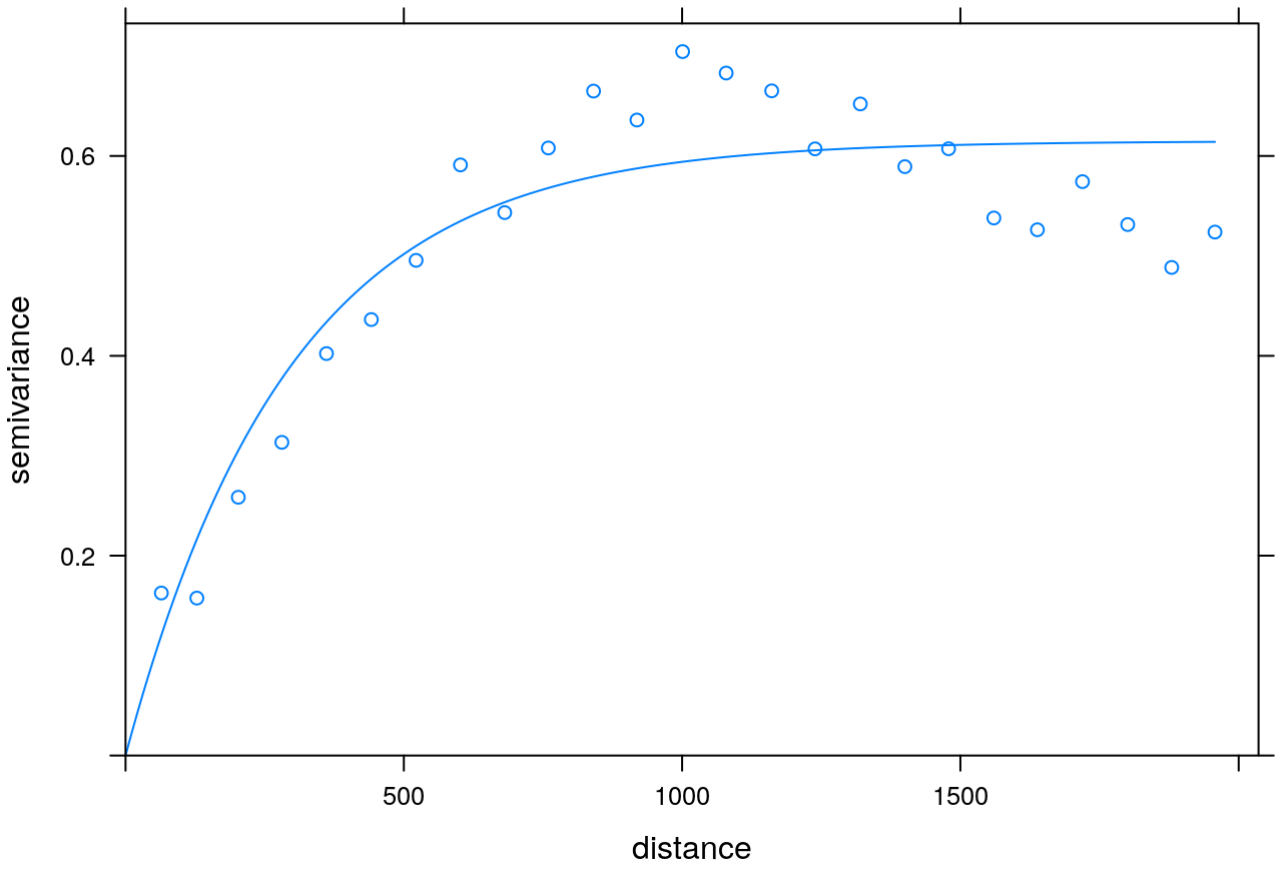
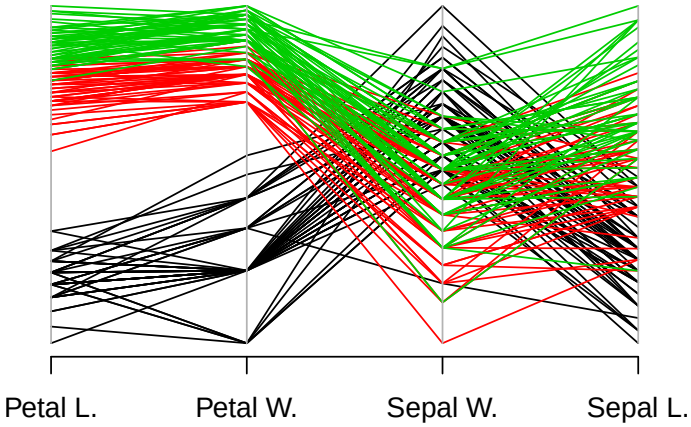


Figure 5.1: Variogram: plotting correlation as a function of spatial distance. Courtesy of Ron Sarafian.

5.2.2.4 Parallel Coordinate Plots

In a parallel coordinate plot, we plot a multivariate observation as a function of its coordinates. In the following example, we visualize the celebrated Iris dataset. In this dataset, for each of 50 iris flowers, Edgar Anderson measured 4 characteristics.

```
ir <- rbind(iris3[,,1], iris3[,,2], iris3[,,3])
MASS::parcoord(log(ir)[, c(3, 4, 2, 1)], col = 1 + (0:149)%/%50)
```



5.3 Mixed Type Data

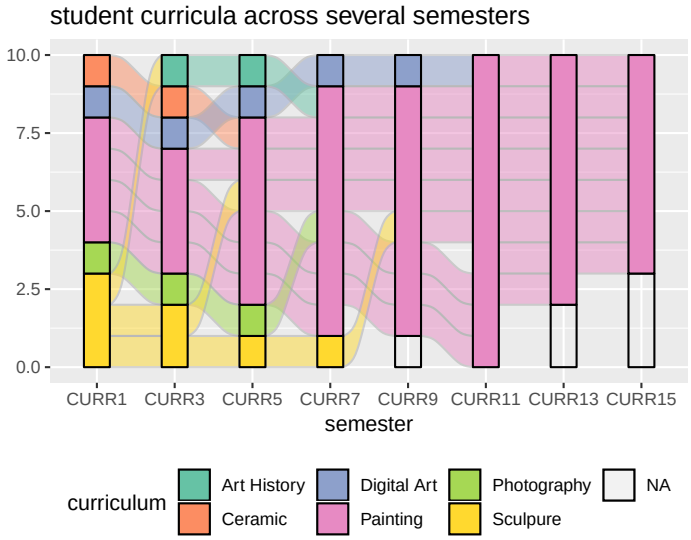
Most real data sets will be of mixed type: both categorical and w. One approach to view such data, is to visualize the continuous variables separately, for each level of the categorical variables. There are, however, interesting dedicated visualization for such data.

5.3.1 Alluvian Diagram

An Alluvian plot is a type of Parallel Coordinate Plot for multivariate categorical data. It is particularly interesting when the x axis is a discretized time variable, and it is used to visualize flow.

The following example, from the **ggalluvial** package Vignette by Jason Cory Brunson, demonstrates the flow of students between different majors, as semesters evolve.

```
library(ggalluvial)
data(majors)
majors$curriculum <- as.factor(majors$curriculum)
ggplot(majors,
       aes(x = semester, stratum = curriculum, alluvium = student,
           fill = curriculum, label = curriculum)) +
  scale_fill_brewer(type = "qual", palette = "Set2") +
  geom_flow(stat = "alluvium", lode.guidance = "rightleft",
            color = "darkgray") +
  geom_stratum() +
  theme(legend.position = "bottom") +
  ggtitle("student curricula across several semesters")
```



Things to note:

- We used the **galluvian** package of the **ggplot2** ecosystem. More on **ggplot2** in the Plotting Chapter.
- Time is on the x axis. Categories are color coded.

5.4 Bibliographic Notes

Like any other topic in this book, you can consult Venables and Ripley (2013). The seminal book on EDA, written long before R was around, is Tukey (1977). For an excellent text on robust statistics see Wilcox (2011).

5.5 Practice Yourself

1. Read about the Titanic data set using `?Titanic`. Inspect it with the `table` and with the `ftable` commands. Which do you prefer?
2. Inspect the Titanic data with a plot. Start with `plot(Titanic)` Try also `lattice::dotplot`. Which is the passenger category with most survivors? Which plot do you prefer? Which scales better to more categories?
3. Read about the women data using `?women`.
    1. Compute the average of each variable. What is the average of the heights?
    2. Plot a histogram of the heights. Add ticks using `rug`.
    3. Plot a boxplot of the weights.
    4. Plot the heights and weights using a scatter plot. Add ticks using `rug`.
4. Choose α to define a new symmetry measure: $1/2(x_\alpha + x_{1-\alpha}) - x_{0.5}$. Write a function that computes it, and apply it on women's heights data.
5. Compute the covariance matrix of women's heights and weights. Compute the correlation matrix. View the correlation matrix as an image using `lattice::levelplot`.
6. Pick a dataset with two LONG continous variables from `?datasets`. Plot it using `hexbin::hexbin`.

Chapter 6

Linear Models

6.1 Problem Setup

Example 6.1 (Bottle Cap Production)**.** Consider a randomized experiment designed to study the effects of temperature and pressure on the diameter of manufactured a bottle cap.

Example 6.2 (Rental Prices)**.** Consider the prediction of rental prices given an appartment's attributes.

Both examples require some statistical model, but they are very different. The first is a *causal inference* problem: we want to design an intervention so that we need to recover the causal effect of temperature and pressure. The second is a prediction problem, a.k.a. a forecasting problem, in which we don't care about the causal effects, we just want good predictions.

In this chapter we discuss the causal problem in Example 6.1. This means that when we assume a model, we assume it is the actual *data generating process*, i.e., we assume the *sampling distribution* is well specified. The second type of problems is discussed in the Supervised Learning Chapter 10.

Here are some more examples of the types of problems we are discussing.

Example 6.3 (Plant Growth)**.** Consider the treatment of various plants with various fertilizers to study the fertilizer's effect on growth.

Example 6.4 (Return to Education)**.** Consider the study of return to education by analyzing the incomes of individuals with different education years.

Example 6.5 (Drug Effect)**.** Consider the study of the effect of a new drug for hemophilia, by analyzing the level of blood coagulation after the administration of various amounts of the new drug.

Let's present the linear model. We assume that a response[1] variable is the sum of effects of some factors[2]. Denoting the response variable by y, the factors by $x = (x_1, \dots, x_p)$, and the effects by $\beta := (\beta_1, \dots, \beta_p)$ the linear model assumption implies that the expected response is the sum of the factors effects:

$$E[y] = x_1\beta_1 + \cdots + x_p\beta_p = \sum_{j=1}^{p} x_j\beta_j = x'\beta. \tag{6.1}$$

Clearly, there may be other factors that affect the the caps' diameters. We thus introduce an error term[3], denoted by ε, to capture the effects of all unmodeled factors and measurement error[4]. The implied generative process of a sample of $i = 1, \dots, n$ observations it thus

[1] The "response" is also know as the "dependent" variable in the statistical literature, or the "labels" in the machine learning literature.

[2] The "factors" are also known as the "independent variable", or "the design", in the statistical literature, and the "features", or "attributes" in the machine learning literature.

[3] The "error term" is also known as the "noise", or the "common causes of variability".

[4] You may philosophize if the measurement error is a mere instance of unmodeled factors or not, but this has no real implication for our purposes.

$$y_i = x_i'\beta + \varepsilon_i = \sum_j x_{i,j}\beta_j + \varepsilon_i, i = 1, \ldots, n. \tag{6.2}$$

or in matrix notation

$$y = X\beta + \varepsilon. \tag{6.3}$$

Let's demonstrate Eq.(6.2). In our bottle-caps example [6.1], we may produce bottle caps at various temperatures. We design an experiment where we produce bottle-caps at varying temperatures. Let x_i be the temperate at which bottle-cap i was manufactured. Let y_i be its measured diameter. By the linear model assumption, the expected diameter varies linearly with the temperature: $\mathbb{E}[y_i] = \beta_0 + x_i\beta_1$. This implies that β_1 is the (expected) change in diameter due to a unit change in temperature.
Remark. In Galton's classical regression problem, where we try to seek the relation between the heights of sons and fathers then $p = 1$, y_i is the height of the i'th father, and x_i the height of the i'th son. This is a prediction problem, more than it is a causal-inference problem.

There are many reasons linear models are very popular:

1. Before the computer age, these were pretty much the only models that could actually be computed[5]. The whole Analysis of Variance (ANOVA) literature is an instance of linear models, that relies on sums of squares, which do not require a computer to work with.
2. For purposes of prediction, where the actual data generating process is not of primary importance, they are popular because they simply work. Why is that? They are simple so that they do not require a lot of data to be computed. Put differently, they may be biased, but their variance is small enough to make them more accurate than other models.
3. For non continuous predictors, **any** functional relation can be cast as a linear model.
4. For the purpose of *screening*, where we only want to show the existence of an effect, and are less interested in the magnitude of that effect, a linear model is enough.
5. If the true generative relation is not linear, but smooth enough, then the linear function is a good approximation via Taylor's theorem.

There are still two matters we have to attend: (i) How the estimate β? (ii) How to perform inference?

In the simplest linear models the estimation of β is done using the method of least squares. A linear model with least squares estimation is known as Ordinary Least Squares (OLS). The OLS problem:

$$\hat{\beta} := argmin_\beta\{\sum_i (y_i - x_i'\beta)^2\}, \tag{6.4}$$

and in matrix notation

$$\hat{\beta} := argmin_\beta\{\|y - X\beta\|_2^2\}. \tag{6.5}$$

Remark. Personally, I prefer the matrix notation because it is suggestive of the geometry of the problem. The reader is referred to Friedman et al. (2001), Section 3.2, for more on the geometry of OLS.

Different software suits, and even different R packages, solve Eq.(6.4) in different ways so that we skip the details of how exactly it is solved. These are discussed in Chapters 17 and 18.

The last matter we need to attend is how to do inference on $\hat{\beta}$. For that, we will need some assumptions on ε. A typical set of assumptions is the following:

[5]By "computed" we mean what statisticians call "fitted", or "estimated", and computer scientists call "learned".

1. **Independence**: we assume ε_i are independent of everything else. Think of them as the measurement error of an instrument: it is independent of the measured value and of previous measurements.
2. **Centered**: we assume that $E[\varepsilon] = 0$, meaning there is no systematic error, sometimes it called The "Linearity assumption".
3. **Normality**: we will typically assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, but we will later see that this is not really required.

We emphasize that these assumptions are only needed for inference on $\hat{\beta}$ and not for the estimation itself, which is done by the purely algorithmic framework of OLS.

Given the above assumptions, we can apply some probability theory and linear algebra to get the distribution of the estimation error:

$$\hat{\beta} - \beta \sim \mathcal{N}(0, (X'X)^{-1}\sigma^2). \tag{6.6}$$

The reason I am not too strict about the normality assumption above, is that Eq.(6.6) is approximately correct even if ε is not normal, provided that there are many more observations than factors ($n \gg p$).

6.2 OLS Estimation in R

We are now ready to estimate some linear models with R. We will use the `whiteside` data from the **MASS** package, recording the outside temperature and gas consumption, before and after an apartment's insulation.

```
library(MASS) # load the package
library(data.table) # for some data manipulations
data(whiteside) # load the data
head(whiteside) # inspect the data
```

```
##    Insul Temp Gas
## 1 Before -0.8 7.2
## 2 Before -0.7 6.9
## 3 Before  0.4 6.4
## 4 Before  2.5 6.0
## 5 Before  2.9 5.8
## 6 Before  3.2 5.8
```

We do the OLS estimation on the pre-insulation data with `lm` function, possibly the most important function in R.

```
whiteside <- data.table(whiteside)
lm.1 <- lm(Gas~Temp, data=whiteside[Insul=='Before']) # OLS estimation
```

Things to note:

- We used the tilde syntax `Gas~Temp`, reading "gas as linear function of temperature".
- The `data` argument tells R where to look for the variables Gas and Temp. We used `Insul=='Before'` to subset observations before the insulation.
- The result is assigned to the object `lm.1`.

Like any other language, spoken or programmable, there are many ways to say the same thing. Some more elegant than others...

```
lm.1 <- lm(y=Gas, x=Temp, data=whiteside[whiteside$Insul=='Before',])
lm.1 <- lm(y=whiteside[whiteside$Insul=='Before',]$Gas,x=whiteside[whiteside$Insul=='Before',]$Temp)
lm.1 <- whiteside[whiteside$Insul=='Before',] %>% lm(Gas~Temp, data=.)
```

The output is an object of class `lm`.

```
class(lm.1)
```

```
## [1] "lm"
```

Objects of class `lm` are very complicated. They store a lot of information which may be used for inference, plotting, etc. The `str` function, short for "structure", shows us the various elements of the object.

```
str(lm.1)
```

```
## List of 12
##  $ coefficients : Named num [1:2] 6.854 -0.393
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "Temp"
##  $ residuals    : Named num [1:26] 0.0316 -0.2291 -0.2965 0.1293 0.0866 ...
##   ..- attr(*, "names")= chr [1:26] "1" "2" "3" "4" ...
##  $ effects      : Named num [1:26] -24.2203 -5.6485 -0.2541 0.1463 0.0988 ...
##   ..- attr(*, "names")= chr [1:26] "(Intercept)" "Temp" "" "" ...
##  $ rank         : int 2
##  $ fitted.values: Named num [1:26] 7.17 7.13 6.7 5.87 5.71 ...
##   ..- attr(*, "names")= chr [1:26] "1" "2" "3" "4" ...
##  $ assign       : int [1:2] 0 1
##  $ qr           :List of 5
##   ..$ qr   : num [1:26, 1:2] -5.099 0.196 0.196 0.196 0.196 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:26] "1" "2" "3" "4" ...
##   .. .. ..$ : chr [1:2] "(Intercept)" "Temp"
##   .. ..- attr(*, "assign")= int [1:2] 0 1
##   ..$ qraux: num [1:2] 1.2 1.35
##   ..$ pivot: int [1:2] 1 2
##   ..$ tol  : num 1e-07
##   ..$ rank : int 2
##   ..- attr(*, "class")= chr "qr"
##  $ df.residual  : int 24
##  $ xlevels      : Named list()
##  $ call         : language lm(formula = Gas ~ Temp, data = whiteside[Insul == "Before"])
##  $ terms        :Classes 'terms', 'formula'  language Gas ~ Temp
##   .. ..- attr(*, "variables")= language list(Gas, Temp)
##   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:2] "Gas" "Temp"
##   .. .. .. ..$ : chr "Temp"
##   .. ..- attr(*, "term.labels")= chr "Temp"
##   .. ..- attr(*, "order")= int 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(Gas, Temp)
##   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. ..- attr(*, "names")= chr [1:2] "Gas" "Temp"
##  $ model        :'data.frame':   26 obs. of  2 variables:
##   ..$ Gas : num [1:26] 7.2 6.9 6.4 6 5.8 5.8 5.6 4.7 5.8 5.2 ...
##   ..$ Temp: num [1:26] -0.8 -0.7 0.4 2.5 2.9 3.2 3.6 3.9 4.2 4.3 ...
##   ..- attr(*, "terms")=Classes 'terms', 'formula'  language Gas ~ Temp
##   .. .. ..- attr(*, "variables")= language list(Gas, Temp)
##   .. .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. .. ..$ : chr [1:2] "Gas" "Temp"
##   .. .. .. .. ..$ : chr "Temp"
##   .. .. ..- attr(*, "term.labels")= chr "Temp"
##   .. .. ..- attr(*, "order")= int 1
##   .. .. ..- attr(*, "intercept")= int 1
##   .. .. ..- attr(*, "response")= int 1
##   .. .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
```

```
##   .. .. ..- attr(*, "predvars")= language list(Gas, Temp)
##   .. .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. .. ..- attr(*, "names")= chr [1:2] "Gas" "Temp"
##  - attr(*, "class")= chr "lm"
```

In RStudio it is particularly easy to extract objects. Just write `your.object$` and press `tab` after the `$` for auto-completion.

If we only want $\hat{\beta}$, it can also be extracted with the `coef` function.

```
coef(lm.1)
```

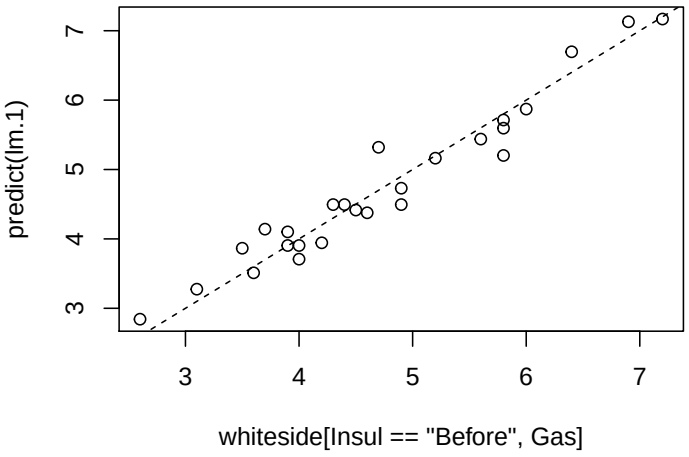
```
## (Intercept)        Temp
##   6.8538277  -0.3932388
```

Things to note:

- R automatically adds an `(Intercept)` term. This means we estimate $Gas = \beta_0 + \beta_1 Temp + \varepsilon$ and not $Gas = \beta_1 Temp + \varepsilon$. This makes sense because we are interested in the contribution of the temperature to the variability of the gas consumption about its **mean**, and not about zero.
- The effect of temperature, i.e., $\hat{\beta}_1$, is -0.39. The negative sign means that the higher the temperature, the less gas is consumed. The magnitude of the coefficient means that for a unit increase in the outside temperature, the gas consumption decreases by 0.39 units.

We can use the `predict` function to make predictions, but we emphasize that if the purpose of the model is to make predictions, and not interpret coefficients, better skip to the Supervised Learning Chapter 10.

```
plot(predict(lm.1)~whiteside[Insul=='Before',Gas])
abline(0,1, lty=2)
```



The model seems to fit the data nicely. A common measure of the goodness of fit is the *coefficient of determination*, more commonly known as the R^2.

Definition 6.1 (R2)**.** The coefficient of determination, denoted R^2, is defined as

$$R^2 := 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{6.7}$$

where $\hat{y}_i$ is the model's prediction, $\hat{y}_i = x_i \hat{\beta}$.

It can be easily computed

```
R2 <- function(y, y.hat){
  numerator <- (y-y.hat)^2 %>% sum
  denominator <- (y-mean(y))^2 %>% sum
  1-numerator/denominator
```

```
}
R2(y=whiteside[Insul=='Before',Gas], y.hat=predict(lm.1))
```

```
## [1] 0.9438081
```

This is a nice result implying that about 94% of the variability in gas consumption can be attributed to changes in the outside temperature.

Obviously, R does provide the means to compute something as basic as R^2, but I will let you find it for yourselves.

6.3 Inference

To perform inference on $\hat{\beta}$, in order to test hypotheses and construct confidence intervals, we need to quantify the uncertainly in the reported $\hat{\beta}$. This is exactly what Eq.(6.6) gives us.

Luckily, we don't need to manipulate multivariate distributions manually, and everything we need is already implemented. The most important function is `summary` which gives us an overview of the model's fit. We emphasize that that fitting a model with `lm` is an assumption free algorithmic step. Inference using `summary` is **not** assumption free, and requires the set of assumptions leading to Eq.(6.6).

```
summary(lm.1)
```

```
##
## Call:
## lm(formula = Gas ~ Temp, data = whiteside[Insul == "Before"])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62020 -0.19947  0.06068  0.16770  0.59778
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.85383    0.11842   57.88   <2e-16 ***
## Temp        -0.39324    0.01959  -20.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2813 on 24 degrees of freedom
## Multiple R-squared:  0.9438, Adjusted R-squared:  0.9415
## F-statistic: 403.1 on 1 and 24 DF,  p-value: < 2.2e-16
```

Things to note:

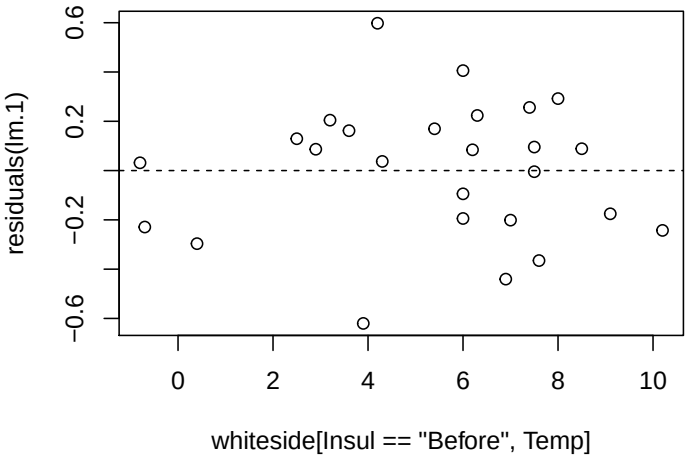
- The estimated $\hat{\beta}$ is reported in the 'Coefficients' table, which has point estimates, standard errors, t-statistics, and the p-values of a two-sided hypothesis test for each coefficient $H_{0,j} : \beta_j = 0, j = 1, \ldots, p$.
- The R^2 is reported at the bottom. The "Adjusted R-squared" is a variation that compensates for the model's complexity.
- The original call to `lm` is saved in the `Call` section.
- Some summary statistics of the residuals $(y_i - \hat{y}_i)$ in the `Residuals` section.
- The "residuals standard error"[6] is $\sqrt{(n-p)^{-1}\sum_i (y_i - \hat{y}_i)^2}$. The denominator of this expression is the *degrees of freedom*, $n - p$, which can be thought of as the hardness of the problem.

As the name suggests, `summary` is merely a summary. The full `summary(lm.1)` object is a monstrous object. Its various elements can be queried using `str(sumary(lm.1))`.

Can we check the assumptions required for inference? Some. Let's start with the linearity assumption. If we were wrong, and the data is not arranged about a linear line, the residuals will have some shape. We thus plot the residuals as a function of the predictor to diagnose shape.

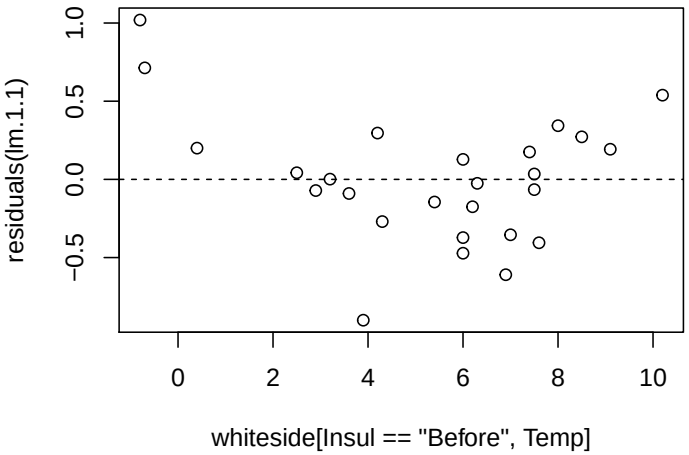
[6] Sometimes known as the Root Mean Squared Error (RMSE).

```
plot(residuals(lm.1)~whiteside[Insul=='Before',Temp])
abline(0,0, lty=2)
```



I can't say I see any shape. Let's fit a **wrong** model, just to see what "shape" means.

```
lm.1.1 <- lm(Gas~I(Temp^2), data=whiteside[Insul=='Before',])
plot(residuals(lm.1.1)~whiteside[Insul=='Before',Temp]); abline(0,0, lty=2)
```



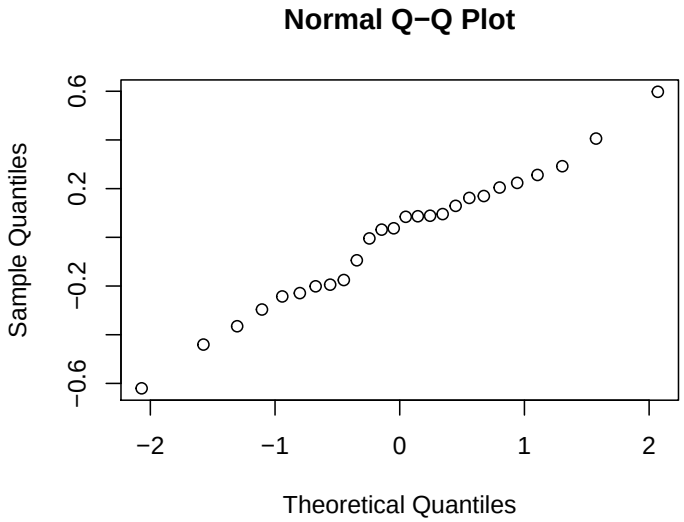
Things to note:

- We used `I(Temp)^2` to specify the model $Gas = \beta_0 + \beta_1 Temp^2 + \varepsilon$.
- The residuals have a "belly". Because they are not a cloud around the linear trend, and we have the wrong model.

To the next assumption. We assumed ε_i are independent of everything else. The residuals, $y_i - \hat{y}_i$ can be thought of a sample of ε_i. When diagnosing the linearity assumption, we already saw their distribution does not vary with the x's, `Temp` in our case. They may be correlated with themselves; a positive departure from the model, may be followed by a series of positive departures etc. Diagnosing these *auto-correlations* is a real art, which is not part of our course.

The last assumption we required is normality. As previously stated, if $n \gg p$, this assumption can be relaxed. If n is in the order of p, we need to verify this assumption. My favorite tool for this task is the *qqplot*. A qqplot compares the quantiles of the sample with the respective quantiles of the assumed distribution. If quantiles align along a line, the assumed distribution if OK. If quantiles depart from a line, then the assumed distribution does not fit the sample.

```
qqnorm(resid(lm.1))
```

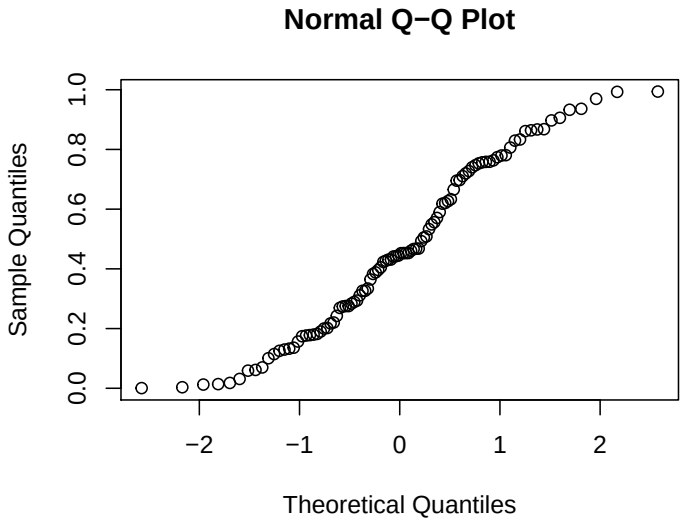


Things to note:

- The `qqnorm` function plots a qqplot against a normal distribution. For non-normal distributions try `qqplot`.
- `resid(lm.1)` extracts the residuals from the linear model, i.e., the vector of $y_i - x_i'\hat{\beta}$.

Judging from the figure, the normality assumption is quite plausible. Let's try the same on a non-normal sample, namely a uniformly distributed sample, to see how that would look.

```
qqnorm(runif(100))
```



6.3.1 Testing a Hypothesis on a Single Coefficient

The first inferential test we consider is a hypothesis test on a single coefficient. In our gas example, we may want to test that the temperature has no effect on the gas consumption. The answer for that is given immediately by `summary(lm.1)`

```
summary.lm1 <- summary(lm.1)
coefs.lm1 <- summary.lm1$coefficients
coefs.lm1
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  6.8538277 0.11842341  57.87561 2.717533e-27
## Temp        -0.3932388 0.01958601 -20.07754 1.640469e-16
```

We see that the p-value for $H_{0,1} : \hat{\beta}_1 = 0$ against a two sided alternative is effectively 0, so that β_1 is unlikely to be 0.

6.3.2 Constructing a Confidence Interval on a Single Coefficient

Since the `summary` function gives us the standard errors of $\hat{\beta}$, we can immediately compute $\hat{\beta}_j \pm 2\sqrt{Var[\hat{\beta}_j]}$ to get ourselves a (roughly) 95% confidence interval. In our example the interval is

```
coefs.lm1[2,1] + c(-2,2) * coefs.lm1[2,2]
```

```
## [1] -0.4324108 -0.3540668
```

6.3.3 Multiple Regression

Remark. Multiple regression is not to be confused with *multivariate regression* discussed in Chapter 9.

The `swiss` dataset encodes the fertility at each of Switzerland's 47 French speaking provinces, along other socio-economic indicators. Let's see if these are statistically related:

```
head(swiss)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

```
lm.5 <- lm(data=swiss, Fertility~Agriculture+Examination+Education+Education+Catholic+Infant.Mortality)
summary(lm.5)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Examination + Education +
##     Education + Catholic + Infant.Mortality, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Things to note:

- The ~ syntax allows to specify various predictors separated by the + operator.
- The summary of the model now reports the estimated effect, i.e., the regression coefficient, of each of the variables.

Clearly, naming each variable explicitly is a tedious task if there are many. The use of `Fertility~.` in the next example reads: "Fertility as a function of all other variables in the `swiss` data.frame".

```
lm.5 <- lm(data=swiss, Fertility~.)
summary(lm.5)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

6.3.4 ANOVA (*)

Our next example[7] contains a hypothetical sample of 60 participants who are divided into three stress reduction treatment groups (mental, physical, and medical) and three age groups groups. The stress reduction values are represented on a scale that ranges from 1 to 10. The values represent how effective the treatment programs were at reducing participant's stress levels, with larger effects indicating higher effectiveness.

```
twoWay <- read.csv('data/dataset_anova_twoWay_comparisons.csv')
head(twoWay)
```

```
##   Treatment   Age StressReduction
## 1    mental young              10
## 2    mental young               9
## 3    mental young               8
## 4    mental   mid               7
## 5    mental   mid               6
## 6    mental   mid               5
```

How many observations per group?

[7] The example is taken from http://rtutorialseries.blogspot.co.il/2011/02/r-tutorial-series-two-way-anova-with.html

```
table(twoWay$Treatment, twoWay$Age)
```

```
##
##            mid old young
##   medical    3   3     3
##   mental     3   3     3
##   physical   3   3     3
```

Since we have two factorial predictors, this multiple regression is nothing but a *two way ANOVA*. Let's fit the model and inspect it.

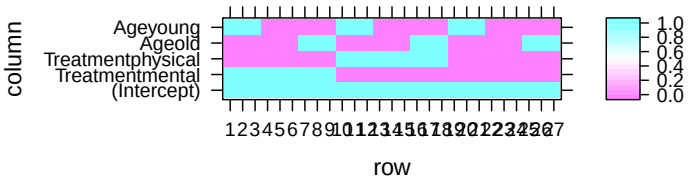
```
lm.2 <- lm(StressReduction~.,data=twoWay)
summary(lm.2)
```

```
##
## Call:
## lm(formula = StressReduction ~ ., data = twoWay)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -1     -1      0      1      1
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.0000     0.3892  10.276 7.34e-10 ***
## Treatmentmental     2.0000     0.4264   4.690 0.000112 ***
## Treatmentphysical   1.0000     0.4264   2.345 0.028444 *
## Ageold             -3.0000     0.4264  -7.036 4.65e-07 ***
## Ageyoung            3.0000     0.4264   7.036 4.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9045 on 22 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.8926
## F-statistic:    55 on 4 and 22 DF,  p-value: 3.855e-11
```

Things to note:

- The `StressReduction~.` syntax is read as "Stress reduction as a function of everything else".
- All the (main) effects and the intercept seem to be significant.
- The data has 2 factors, but the coefficients table has 4 predictors. This is because `lm` noticed that `Treatment` and `Age` are factors. Each level of each factor is thus encoded as a different (dummy) variable. The numerical values of the factors are meaningless. Instead, R has constructed a dummy variable for each level of each factor. The names of the effect are a concatenation of the factor's name, and its level. You can inspect these dummy variables with the `model.matrix` command.

```
model.matrix(lm.2) %>% lattice::levelplot()
```



If you don't want the default dummy coding, look at `?contrasts`.

If you are more familiar with the ANOVA literature, or that you don't want the effects of each level separately, but rather, the effect of **all** the levels of each factor, use the `anova` command.

```
anova(lm.2)
```

```
## Analysis of Variance Table
##
## Response: StressReduction
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Treatment  2     18   9.000      11 0.0004883 ***
## Age        2    162  81.000      99     1e-11 ***
## Residuals 22     18   0.818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Things to note:

- The ANOVA table, unlike the `summary` function, tests if **any** of the levels of a factor has an effect, and not one level at a time.
- The significance of each factor is computed using an F-test.
- The degrees of freedom, encoding the number of levels of a factor, is given in the `Df` column.
- The StressReduction seems to vary for different ages and treatments, since both factors are significant.

If you are extremely more comfortable with the ANOVA literature, you could have replaced the `lm` command with the `aov` command all along.

```
lm.2.2 <- aov(StressReduction~.,data=twoWay)
class(lm.2.2)
```

```
## [1] "aov" "lm"
```

```
summary(lm.2.2)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Treatment    2     18    9.00      11 0.000488 ***
## Age          2    162   81.00      99    1e-11 ***
## Residuals   22     18    0.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Things to note:

- The `lm` function has been replaced with an `aov` function.
- The output of `aov` is an `aov` class object, which extends the `lm` class.
- The summary of an `aov` does not like the summary of an `lm` object, but rather, like an ANOVA table.

As in any two-way ANOVA, we may want to ask if different age groups respond differently to different treatments. In the statistical parlance, this is called an *interaction*, or more precisely, an *interaction of order 2*.

```
lm.3 <- lm(StressReduction~Treatment+Age+Treatment:Age-1,data=twoWay)
```

The syntax `StressReduction~Treatment+Age+Treatment:Age-1` tells R to include main effects of Treatment, Age, and their interactions. Here are other ways to specify the same model.

```
lm.3 <- lm(StressReduction ~ Treatment * Age - 1,data=twoWay)
lm.3 <- lm(StressReduction~(.)^2 - 1,data=twoWay)
```

The syntax `Treatment * Age` means "mains effects with second order interactions". The syntax `(.)^2` means "everything with second order interactions"

Let's inspect the model

```
summary(lm.3)
```

```
##
## Call:
## lm(formula = StressReduction ~ Treatment + Age + Treatment:Age -
##     1, data = twoWay)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
##     -1     -1      0      1      1
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## Treatmentmedical            4.000e+00  5.774e-01   6.928 1.78e-06 ***
## Treatmentmental             6.000e+00  5.774e-01  10.392 4.92e-09 ***
## Treatmentphysical           5.000e+00  5.774e-01   8.660 7.78e-08 ***
## Ageold                     -3.000e+00  8.165e-01  -3.674  0.00174 **
## Ageyoung                    3.000e+00  8.165e-01   3.674  0.00174 **
## Treatmentmental:Ageold      4.246e-16  1.155e+00   0.000  1.00000
## Treatmentphysical:Ageold    1.034e-15  1.155e+00   0.000  1.00000
## Treatmentmental:Ageyoung   -3.126e-16  1.155e+00   0.000  1.00000
## Treatmentphysical:Ageyoung  5.128e-16  1.155e+00   0.000  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 18 degrees of freedom
## Multiple R-squared:  0.9794, Adjusted R-squared:  0.9691
## F-statistic:    95 on 9 and 18 DF,  p-value: 2.556e-13
```

Things to note:

- There are still 5 main effects, but also 4 interactions. This is because when allowing a different average response for every $Treatment * Age$ combination, we are effectively estimating $3 * 3 = 9$ cell means, even if they are not parametrized as cell means, but rather as main effect and interactions.
- The interactions do not seem to be significant.
- The assumptions required for inference are clearly not met in this example, which is there just to demonstrate R's capabilities.

Asking if all the interactions are significant, is asking if the different age groups have the same response to different treatments. Can we answer that based on the various interactions? We might, but it is possible that no single interaction is significant, while the combination is. To test for all the interactions together, we can simply check if the model without interactions is (significantly) better than a model with interactions. I.e., compare `lm.2` to `lm.3`. This is done with the `anova` command.

```
anova(lm.2,lm.3, test='F')
```

```
## Analysis of Variance Table
##
## Model 1: StressReduction ~ Treatment + Age
## Model 2: StressReduction ~ Treatment + Age + Treatment:Age - 1
##   Res.Df RSS Df   Sum of Sq F Pr(>F)
## 1     22  18
## 2     18  18  4 -3.5527e-15
```

We see that `lm.3` is **not** (significantly) better than `lm.2`, so that we can conclude that there are no interactions: different ages have the same response to different treatments.

6.3.5 Testing a Hypothesis on a Single Contrast (*)

Returning to the model without interactions, `lm.2`.

```
coef(summary(lm.2))
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)              4  0.3892495 10.276186 7.336391e-10
## Treatmentmental          2  0.4264014  4.690416 1.117774e-04
## Treatmentphysical        1  0.4264014  2.345208 2.844400e-02
## Ageold                  -3  0.4264014 -7.035624 4.647299e-07
```

```
## Ageyoung                  3  0.4264014  7.035624 4.647299e-07
```

We see that the effect of the various treatments is rather similar. It is possible that all treatments actually have the same effect. Comparing the effects of factor levels is called a *contrast*. Let's test if the medical treatment, has in fact, the same effect as the physical treatment.

```
library(multcomp)
my.contrast <- matrix(c(-1,0,1,0,0), nrow =  1)
lm.4 <- glht(lm.2, linfct=my.contrast)
summary(lm.4)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = StressReduction ~ ., data = twoWay)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0  -3.0000     0.7177   -4.18 0.000389 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Things to note:

- A contrast is a linear function of the coefficients. In our example $H_0 : \beta_1 - \beta_3 = 0$, which justifies the construction of `my.contrast`.
- We used the `glht` function (generalized linear hypothesis test) from the package **multcompt**.
- The contrast is significant, i.e., the effect of a medical treatment, is different than that of a physical treatment.

6.4 Bibliographic Notes

Like any other topic in this book, you can consult Venables and Ripley (2013) for more on linear models. For the theory of linear models, I like Greene (2003).

6.5 Practice Yourself

1. Inspect women's heights and weights with `?women`.
    1. What is the change in weight per unit change in height? Use the `lm` function.
    2. Is the relation of height on weight significant? Use `summary`.
    3. Plot the residuals of the linear model with `plot` and `resid`.
    4. Plot the predictions of the model using `abline`.
    5. Inspect the normality of residuals using `qqnorm`.
    6. Inspect the design matrix using `model.matrix`.
2. Write a function that takes an `lm` class object, and returns the confidence interval on the first coefficient. Apply it on the height and weight data.
3. Use the `ANOVA` function to test the significance of the effect of height.
4. Read about the "mtcars" dataset using `? mtcars`.Inspect the dependency of the fuel consumption (mpg) in the weight (wt) and the 1/4 mile time (qsec).
    1. Make a pairs scatter plot with `plot(mtcars[,c("mpg","wt","qsec")])` Does the connection look linear?
    2. Fit a multiple linear regression with `lm`. Call it `model1`.
    3. Try to add the transmission (am) as independent variable. Let R know this is a categorical variable with `factor(am)`. Call it `model2`.

4. Compare the "Adjusted R-squared" measure of the two models (we can't use the regular R2 to compare two models with a different number of variables).
5. Do the coefficients significant?
6. Inspect the normality of residuals and the linearity assumptions.
7. Now Inspect the hypothesis that the effect of weight is different between the transmission types with adding interaction to the model `wt*factor(am)`.
8. According to this model, what is the addition of one unit of weight in a manual transmission to the fuel consumption (-2.973-4.141=-7.11)?

Chapter 7

Generalized Linear Models

Example 7.1. Consider the relation between cigarettes smoked, and the occurance of lung cancer. Do we expect the probability of cancer to be linear in the number of cigarettes? Probably not. Do we expect the variability of events to be constant about the trend? Probably not.

Example 7.2. Consider the relation between the travel times to the distance travelled. Do you agree that the longer the distance travelled, then not only the travel times get longer, but they also get more variable?

7.1 Problem Setup

In the Linear Models Chapter 6, we assumed the generative process to be linear in the effects of the predictors x. We now write that same linear model, slightly differently:

$$y|x \sim \mathcal{N}(x'\beta, \sigma^2).$$

This model not allow for the non-linear relations of Example 7.1, nor does it allow for the distrbituion of ε to change with x, as in Example 7.2. *Generalize linear models* (GLM), as the name suggests, are a generalization of the linear models in Chapter 6 that allow that[1].

For Example 7.1, we would like something in the lines of

$$y|x \sim Binom(1, p(x))$$

For Example 7.2, we would like something in the lines of

$$y|x \sim \mathcal{N}(x'\beta, \sigma^2(x)),$$

or more generally

$$y|x \sim \mathcal{N}(\mu(x), \sigma^2(x)),$$

or maybe not Gaussian

$$y|x \sim Pois(\lambda(x)).$$

Even more generally, for some distribution $F(\theta)$, with a parameter θ, we would like to assume that the data is generated via

$$y|x \sim F(\theta(x)) \tag{7.1}$$

Possible examples include

[1] Do not confuse *generalized linear models* with *non-linear regression*, or *generalized least squares*. These are different things, that we do not discuss.

$$y|x \sim Poisson(\lambda(x)) \tag{7.2}$$
$$y|x \sim Exp(\lambda(x)) \tag{7.3}$$
$$y|x \sim \mathcal{N}(\mu(x), \sigma^2(x)) \tag{7.4}$$

GLMs allow models of the type of Eq.(7.1), while imposing some constraints on F and on the relation $\theta(x)$. GLMs assume the data distribution F to be in a "well-behaved" family known as the *Natural Exponential Family* of distributions. This family includes the Gaussian, Gamma, Binomial, Poisson, and Negative Binomial distributions. These five include as special cases the exponential, chi-squared, Rayleigh, Weibull, Bernoulli, and geometric distributions.

GLMs also assume that the distribution's parameter, θ, is some simple function of a linear combination of the effects. In our cigarettes example this amounts to assuming that each cigarette has an additive effect, but not on the probability of cancer, but rather, on some simple function of it. Formally

$$g(\theta(x)) = x'\beta,$$

and we recall that

$$x'\beta = \beta_0 + \sum_j x_j\beta_j.$$

The function g is called the *link* function, its inverse, g^{-1} is the *mean function*. We thus say that "the effects of each cigarette is linear **in link scale**". This terminology will later be required to understand R's output.

7.2 Logistic Regression

The best known of the GLM class of models is the *logistic regression* that deals with Binomial, or more precisely, Bernoulli-distributed data. The link function in the logistic regression is the *logit function*

$$g(t) = log\left(\frac{t}{(1-t)}\right) \tag{7.5}$$

implying that under the logistic model assumptions

$$y|x \sim Binom\left(1, p = \frac{e^{x'\beta}}{1+e^{x'\beta}}\right). \tag{7.6}$$

Before we fit such a model, we try to justify this construction, in particular, the enigmatic link function in Eq.(7.5). Let's look at the simplest possible case: the comparison of two groups indexed by x: $x = 0$ for the first, and $x = 1$ for the second. We start with some definitions.

Definition 7.1 (Odds)**.** The *odds*, of a binary random variable, y, is defined as

$$\frac{P(y=1)}{P(y=0)}.$$

Odds are the same as probabilities, but instead of telling me there is a 66% of success, they tell me the odds of success are "2 to 1". If you ever placed a bet, the language of "odds" should not be unfamiliar to you.

Definition 7.2 (Odds Ratio)**.** The *odds ratio* between two binary random variables, y_1 and y_2, is defined as the ratio between their odds. Formally:

$$OR(y_1, y_2) := \frac{P(y_1=1)/P(y_1=0)}{P(y_2=1)/P(y_2=0)}.$$

Odds ratios (OR) compare between the probabilities of two groups, only that it does not compare them in probability scale, but rather in odds scale. You can also think of ORs as a measure of distance between two Brenoulli distributions. ORs have better mathematical properties than other candidate distance measures, such as $P(y_1 = 1) - P(y_2 = 1)$.

Under the logit link assumption formalized in Eq.(7.6), the OR between two conditions indexed by $y|x = 1$ and $y|x = 0$, returns:

$$OR(y|x = 1, y|x = 0) = \frac{P(y = 1|x = 1)/P(y = 0|x = 1)}{P(y = 1|x = 0)/P(y = 0|x = 0)} = e^{\beta_1}. \tag{7.7}$$

The last equality demystifies the choice of the link function in the logistic regression: **it allows us to interpret β of the logistic regression as a measure of change of binary random variables, namely, as the (log) odds-ratios due to a unit increase in x.**
Remark. Another popular link function is the normal quantile function, a.k.a., the Gaussian inverse CDF, leading to *probit regression* instead of logistic regression.

7.2.1 Logistic Regression with R

Let's get us some data. The `PlantGrowth` data records the weight of plants under three conditions: control, treatment1, and treatment2.

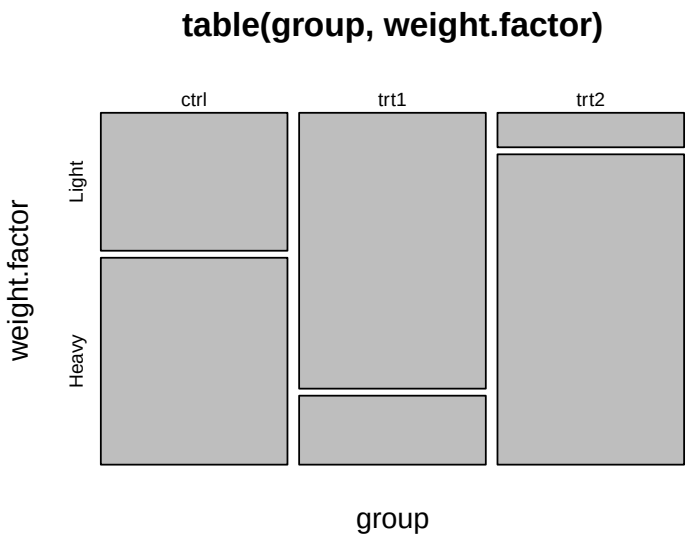
```
head(PlantGrowth)
```

```
##   weight group
## 1   4.17  ctrl
## 2   5.58  ctrl
## 3   5.18  ctrl
## 4   6.11  ctrl
## 5   4.50  ctrl
## 6   4.61  ctrl
```

We will now `attach` the data so that its contents is available in the workspace (don't forget to `detach` afterwards, or you can expect some conflicting object names). We will also use the `cut` function to create a binary response variable for Light, and Heavy plants (we are doing logistic regression, so we need a two-class response). As a general rule of thumb, when we discretize continuous variables, we lose information. For pedagogical reasons, however, we will proceed with this bad practice.

Look at the following output and think: how many `group` effects do we expect? What should be the sign of each effect?

```
attach(PlantGrowth)
weight.factor<- cut(weight, 2, labels=c('Light', 'Heavy')) # binarize weights
plot(table(group, weight.factor))
```



Let's fit a logistic regression, and inspect the output.

```
glm.1<- glm(weight.factor~group, family=binomial)
summary(glm.1)
```

```
##
## Call:
## glm(formula = weight.factor ~ group, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1460  -0.6681   0.4590   0.8728   1.7941
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4055     0.6455   0.628   0.5299
## grouptrt1    -1.7918     1.0206  -1.756   0.0792 .
## grouptrt2     1.7918     1.2360   1.450   0.1471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.054  on 29  degrees of freedom
## Residual deviance: 29.970  on 27  degrees of freedom
## AIC: 35.97
##
## Number of Fisher Scoring iterations: 4
```

Things to note:

- The `glm` function is our workhorse for all GLM models.
- The `family` argument of `glm` tells R the respose variable is brenoulli, thus, performing a logistic regression.
- The `summary` function is content aware. It gives a different output for `glm` class objects than for other objects, such as the `lm` we saw in Chapter 6. In fact, what `summary` does is merely call `summary.glm`.
- As usual, we get the coefficients table, but recall that they are to be interpreted as (log) odd-ratios, i.e., in "link scale". To return to probabilities ("response scale"), we will need to undo the logistic transformation.
- As usual, we get the significance for the test of no-effect, versus a two-sided alternative. P-values are asymptotic, thus, only approximate (and can be very bad approximations in small samples).
- The residuals of `glm` are slightly different than the `lm` residuals, and called *Deviance Residuals*.
- For help see `?glm`, `?family`, and `?summary.glm`.

Like in the linear models, we can use an ANOVA table to check if treatments have any effect, and not one treatment

at a time. In the case of GLMs, this is called an *analysis of deviance* table.

```
anova(glm.1, test='LRT')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: weight.factor
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                     29     41.054
## group  2   11.084        27     29.970 0.003919 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Things to note:

- The `anova` function, like the `summary` function, are content-aware and produce a different output for the `glm` class than for the `lm` class. All that `anova` does is call `anova.glm`.
- In GLMs there is no canonical test (like the F test for `lm`). `LRT` implies we want an approximate Likelihood Ratio Test. We thus specify the type of test desired with the `test` argument.
- The distribution of the weights of the plants does vary with the treatment given, as we may see from the significance of the `group` factor.
- Readers familiar with ANOVA tables, should know that we computed the GLM equivalent of a type I sum-of-squares. Run `drop1(glm.1, test='Chisq')` for a GLM equivalent of a type III sum-of-squares.
- For help see `?anova.glm`.

Let's predict the probability of a heavy plant for each treatment.

```
predict(glm.1, type='response')
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
## 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2
##  19  20  21  22  23  24  25  26  27  28  29  30
## 0.2 0.2 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
```

Things to note:

- Like the `summary` and `anova` functions, the `predict` function is aware that its input is of `glm` class. All that `predict` does is call `predict.glm`.
- In GLMs there are many types of predictions. The `type` argument controls which type is returned. Use `type=response` for predictions in probability scale; use 'type=link' for predictions in log-odds scale.
- How do I know we are predicting the probability of a heavy plant, and not a light plant? Just run `contrasts(weight.factor)` to see which of the categories of the factor `weight.factor` is encoded as 1, and which as 0.
- For help see `?predict.glm`.

Let's detach the data so it is no longer in our workspace, and object names do not collide.

```
detach(PlantGrowth)
```

We gave an example with a factorial (i.e. discrete) predictor. We can do the same with multiple continuous predictors.

```
data('Pima.te', package='MASS') # Loads data
head(Pima.te)
```

```
##   npreg glu bp skin  bmi   ped age type
## 1     6 148 72   35 33.6 0.627  50  Yes
## 2     1  85 66   29 26.6 0.351  31   No
## 3     1  89 66   23 28.1 0.167  21   No
```

```
## 4     3  78 50   32 31.0 0.248  26  Yes
## 5     2 197 70   45 30.5 0.158  53  Yes
## 6     5 166 72   19 25.8 0.587  51  Yes
```

```
glm.2<- step(glm(type~., data=Pima.te, family=binomial))
```

```
## Start:  AIC=301.79
## type ~ npreg + glu + bp + skin + bmi + ped + age
##
##         Df Deviance    AIC
## - skin   1   286.22 300.22
## - bp     1   286.26 300.26
## - age    1   286.76 300.76
## <none>       285.79 301.79
## - npreg  1   291.60 305.60
## - ped    1   292.15 306.15
## - bmi    1   293.83 307.83
## - glu    1   343.68 357.68
##
## Step:  AIC=300.22
## type ~ npreg + glu + bp + bmi + ped + age
##
##         Df Deviance    AIC
## - bp     1   286.73 298.73
## - age    1   287.23 299.23
## <none>       286.22 300.22
## - npreg  1   292.35 304.35
## - ped    1   292.70 304.70
## - bmi    1   302.55 314.55
## - glu    1   344.60 356.60
##
## Step:  AIC=298.73
## type ~ npreg + glu + bmi + ped + age
##
##         Df Deviance    AIC
## - age    1   287.44 297.44
## <none>       286.73 298.73
## - npreg  1   293.00 303.00
## - ped    1   293.35 303.35
## - bmi    1   303.27 313.27
## - glu    1   344.67 354.67
##
## Step:  AIC=297.44
## type ~ npreg + glu + bmi + ped
##
##         Df Deviance    AIC
## <none>       287.44 297.44
## - ped    1   294.54 302.54
## - bmi    1   303.72 311.72
## - npreg  1   304.01 312.01
## - glu    1   349.80 357.80
```

```
summary(glm.2)
```

```
##
## Call:
## glm(formula = type ~ npreg + glu + bmi + ped, family = binomial,
##     data = Pima.te)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9845  -0.6462  -0.3661   0.5977   2.5304
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.552177   1.096207  -8.714  < 2e-16 ***
## npreg        0.178066   0.045343   3.927  8.6e-05 ***
## glu          0.037971   0.005442   6.978  3.0e-12 ***
## bmi          0.084107   0.021950   3.832 0.000127 ***
## ped          1.165658   0.444054   2.625 0.008664 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 420.30  on 331  degrees of freedom
## Residual deviance: 287.44  on 327  degrees of freedom
## AIC: 297.44
##
## Number of Fisher Scoring iterations: 5
```

Things to note:

- We used the `~.` syntax to tell R to fit a model with all the available predictors.
- Since we want to focus on significant predictors, we used the `step` function to perform a *step-wise* regression, i.e. sequentially remove non-significant predictors. The function reports each model it has checked, and the variable it has decided to remove at each step.
- The output of `step` is a single model, with the subset of selected predictors.

7.3 Poisson Regression

Poisson regression means we fit a model assuming $y|x \sim Poisson(\lambda(x))$. Put differently, we assume that for each treatment, encoded as a combinations of predictors x, the response is Poisson distributed with a rate that depends on the predictors.

The typical link function for Poisson regression is the logarithm: $g(t) = log(t)$. This means that we assume $y|x \sim Poisson(\lambda(x) = e^{x'\beta})$. Why is this a good choice? We again resort to the two-group case, encoded by $x = 1$ and $x = 0$, to understand this model: $\lambda(x = 1) = e^{\beta_0+\beta_1} = e^{\beta_0}\ e^{\beta_1} = \lambda(x = 0)\ e^{\beta_1}$. We thus see that this link function implies that a change in x **multiples** the rate of events by e^{β_1}.

For our example[2] we inspect the number of infected high-school kids, as a function of the days since an outbreak.

```
cases <-
structure(list(Days = c(1L, 2L, 3L, 3L, 4L, 4L, 4L, 6L, 7L, 8L,
8L, 8L, 8L, 12L, 14L, 15L, 17L, 17L, 17L, 18L, 19L, 19L, 20L,
23L, 23L, 23L, 24L, 24L, 25L, 26L, 27L, 28L, 29L, 34L, 36L, 36L,
42L, 42L, 43L, 43L, 44L, 44L, 44L, 44L, 45L, 46L, 48L, 48L, 49L,
49L, 53L, 53L, 53L, 54L, 55L, 56L, 56L, 58L, 60L, 63L, 65L, 67L,
67L, 68L, 71L, 71L, 72L, 72L, 72L, 73L, 74L, 74L, 74L, 75L, 75L,
80L, 81L, 81L, 81L, 81L, 88L, 88L, 90L, 93L, 93L, 94L, 95L, 95L,
95L, 96L, 96L, 97L, 98L, 100L, 101L, 102L, 103L, 104L, 105L,
106L, 107L, 108L, 109L, 110L, 111L, 112L, 113L, 114L, 115L),
    Students = c(6L, 8L, 12L, 9L, 3L, 3L, 11L, 5L, 7L, 3L, 8L,
    4L, 6L, 8L, 3L, 6L, 3L, 2L, 2L, 6L, 3L, 7L, 7L, 2L, 2L, 8L,
    3L, 6L, 5L, 7L, 6L, 4L, 4L, 3L, 3L, 5L, 3L, 3L, 3L, 5L, 3L,
    5L, 6L, 3L, 3L, 3L, 3L, 2L, 3L, 1L, 3L, 3L, 5L, 4L, 4L, 3L,
```

[2]Taken from http://www.theanalysisfactor.com/generalized-linear-models-in-r-part-6-poisson-regression-count-variables/

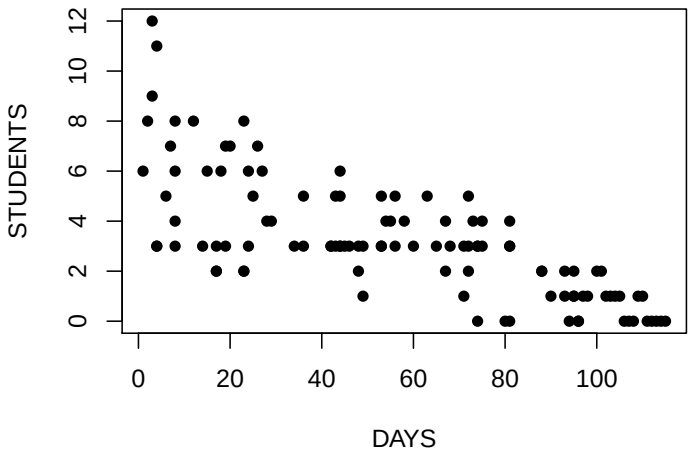
```
    5L, 4L, 3L, 5L, 3L, 4L, 2L, 3L, 3L, 1L, 3L, 2L, 5L, 4L, 3L,
    0L, 3L, 3L, 4L, 0L, 3L, 3L, 4L, 0L, 2L, 2L, 1L, 1L, 2L, 0L,
    2L, 1L, 1L, 0L, 0L, 1L, 1L, 2L, 2L, 1L, 1L, 1L, 1L, 0L, 0L,
    0L, 1L, 1L, 0L, 0L, 0L, 0L, 0L)), .Names = c("Days", "Students"
), class = "data.frame", row.names = c(NA, -109L))
attach(cases)
head(cases)
```

```
##   Days Students
## 1    1        6
## 2    2        8
## 3    3       12
## 4    3        9
## 5    4        3
## 6    4        3
```

Look at the following plot and think:

- Can we assume that the errors have constant variace?
- What is the sign of the effect of time on the number of sick students?
- Can we assume a linear effect of time?

```
plot(Days, Students, xlab = "DAYS", ylab = "STUDENTS", pch = 16)
```



We now fit a model to check for the change in the rate of events as a function of the days since the outbreak.

```
glm.3 <- glm(Students ~ Days, family = poisson)
summary(glm.3)
```

```
##
## Call:
## glm(formula = Students ~ Days, family = poisson)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.00482  -0.85719  -0.09331   0.63969   1.73696
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.990235   0.083935   23.71   <2e-16 ***
## Days        -0.017463   0.001727  -10.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 215.36  on 108  degrees of freedom
## Residual deviance: 101.17  on 107  degrees of freedom
## AIC: 393.11
##
## Number of Fisher Scoring iterations: 5
```

Things to note:

- We used `family=poisson` in the `glm` function to tell R that we assume a Poisson distribution.
- The coefficients table is there as usual. When interpreting the table, we need to recall that the effect, i.e. the $\hat{\beta}$, are **multiplicative** due to the assumed link function.
- Each day **decreases** the rate of events by a factor of about $e^{\beta_1} = 0.02$.
- For more information see `?glm` and `?family`.

7.4 Extensions

As we already implied, GLMs are a very wide class of models. We do not need to use the default link function,but more importantly, we are not constrained to Binomial, or Poisson distributed response. For exponential, gamma, and other response distributions, see `?glm` or the references in the Bibliographic Notes section.

7.5 Bibliographic Notes

The ultimate reference on GLMs is McCullagh (1984). For a less technical exposition, we refer to the usual Venables and Ripley (2013).

7.6 Practice Yourself

1. Try using `lm` for analyzing the plant growth data in `weight.factor` as a function of `group` in the `PlantGrowth` data.
2. Generate some synthetic data for a logistic regression:
   a. Generate two predictor variables of length 100. They can be random from your favorite distribution.
   b. Fix `beta<- c(-1,2)`, and generate the response with:`rbinom(n=100,size=1,prob=exp(x %*% beta)/(1+exp(x %*% beta)))`. Think: why is this the model implied by the logistic regression?
   c. Fit a Logistic regression to your synthetic data using `glm`.
   d. Are the estimated coefficients similar to the true ones you used?
   e. What is the estimated probability of an event at `x=1,1`? Use `predict.glm` but make sure to read the documentation on the `type` argument.
3. Read about the `epil` dataset using `? MASS::epil`. Inspect the dependency of the number of seizures (y) in the age of the patient (`age`) and the treatment (`trt`).
   1. Fit a Poisson regression with `glm` and `family = "poisson"`.
   2. Are the coefficients significant?
   3. Does the treatment reduce the frequency of the seizures?
   4. According to this model, what would be the number of seizures for 20 years old patient with progabide treatment?

See DataCamp's Generalized Linear Models in R for more self practice.

Chapter 8

Linear Mixed Models

Example 8.1 (Dependent Samples on the Mean)**.** Consider inference on a population's mean. Supposdly, more observations imply more infotmation on the mean. This, however, is not the case if samples are completely dependant. More observations do not add any new information. From this example one may think that dependence is a bad thing. This is a false intuitiont: negative correlations imply oscilations about the mean, so they are actually more informative on the mean than independent observations.

Example 8.2 (Repeated Measures)**.** Consider a prospective study, i.e., data that originates from selecting a set of subjects and making measurements on them over time. Also assume that some subjects received some treatment, and other did not. When we want to infer on the population from which these subjects have been sampled, we need to recall that some series of observations came from the same subject. If we were to ignore the subject of origin, and treat each observation as an independent sample point, we will think we have more information in our data than we actually do. For a rough intuition, think of a case where observatiosn within subject are perfectly dependent.

The sources of variability, i.e. noise, are known in the statistical literature as "random effects". Specifying these sources determines the correlation structure in our measurements. In the simplest linear models of Chapter 6, we thought of the variability as a measurement error, independent of anything else. This, however, is rarely the case when time or space are involved.

The variability in our data is rarely the object of interest. It is merely the source of uncertainty in our measurements. The effects we want to infer on are assumingly non-random, thus known as "fixed-effects". A model which has several sources of variability, i.e. random-effects, and several deterministic effects to study, i.e. fixed-effects, is known as a "mixed effects" model. If the model is also linear, it is known as a *linear mixed model* (LMM). Here are some examples of such models.

Example 8.3 (Fixed and Random Machine Effect)**.** Consider the problem of testing for a change in the distribution of diamteters of manufactured bottle caps. We want to study the (fixed) effect of time: before versus after. Bottle caps are produced by several machines. Clearly there is variablity in the diameters within-machine and between-machines. Given many measurements on many bottle caps from many machines, we could standardize measurements by removing each machine's average. This implies the within-machine variability is the only source of variability we care about, because the substration of the machine effect, removed information on the between-machine variability.
Alternatively, we could treat the between-machine variability as another source of noise/uncertainty when inferring on the temporal fixed effect.

Example 8.4 (Fixed and Random Subject Effect)**.** Consider an experimenal design where each subject is given 2 types of diets, and his health condition is recorded. We could standardize over subjects by removing the subject-wise average, before comparing diets. This is what a paired t-test does. This also implies the within-subject variability is the only source of variability we care about. Alternatively, for inference on the population of "all subjects" we need to adress the between-subject variability, and not only the within-subject variability.

The unifying theme of the above examples, is that the variability in our data has several sources. Which are the sources of variability that need to concern us? This is a delicate matter which depends on your goals. As a rule of thumb, we will suggest the following view: **If information of an effect will be available at the time of prediction, treat it as a fixed effect. If it is not, treat it as a random-effect.**

LMMs are so fundamental, that they have earned many names:

- **Mixed Effects**: Because we may have both *fixed effects* we want to estimate and remove, and *random effects* which contribute to the variability to infer against.
- **Variance Components**: Because as the examples show, variance has more than a single source (like in the Linear Models of Chapter 6).
- **Hirarchial Models**: Because as Example 8.4 demonstrates, we can think of the sampling as hierarchical– first sample a subject, and then sample its response.
- **Multilevel Analysis**: For the same reasons it is also known as Hierarchical Models.
- **Repeated Measures**: Because we make several measurements from each unit, like in Example 8.4.
- **Longitudinal Data**: Because we follow units over time, like in Example 8.4.
- **Panel Data**: Is the term typically used in econometric for such longitudinal data.
- **MANOVA**: Many of the problems that may be solved with a multivariate analysis of variance (MANOVA), may be solved with an LMM for reasons we detail in 9.
- **Structured Prediction**: In the machine learning literature, predicting outcomes with structure, such as correlated vectors, is known as Structured Learning. Because LMMs merely specify correlations, using a LMM for making predictions may be thought of as an instance of structured prediction.

Whether we are aiming to infer on a generative model's parameters, or to make predictions, there is no "right" nor "wrong" approach. Instead, there is always some implied measure of error, and an algorithm may be good, or bad, with respect to this measure (think of false and true positives, for instance). This is why we care about dependencies in the data: ignoring the dependence structure will probably yield inefficient algorithms. Put differently, if we ignore the statistical dependence in the data we will probably me making more errors than possible/optimal.

We now emphasize:

1. Like in previous chapters, by "model" we refer to the assumed generative distribution, i.e., the sampling distribution.
2. LMMs are a way to infer against the right level of variability. Using a naive linear model (which assumes a single source of variability) instead of a mixed effects model, probably means your inference is overly anti-conservative. Put differently, the uncertainty in your estimates is higher than the linear model from Chapter 6 may suggest.
3. In a LMM we will specify the dependence structure via the hierarchy in the sampling scheme (e.g. caps within machine, students within class, etc.). Not all dependency models can be specified in this way. Dependency structures that are not hierarchical include temporal dependencies (AR, ARIMA, ARCH and GARCH), spatial, Markov Chains, and more. To specify dependency structures that are no hierarchical, see Chapter 8 in (the excellent) Weiss (2005).
4. If you are using the model merely for predictions, and not for inference on the fixed effects or variance components, then stating the generative distribution may be be useful, but not necessarily. See the Supervised Learning Chapter 10 for more on prediction problems. Also recall that machine learning from non-independent observations (such as LMMs) is a delicate matter that is rarely treated in the literature.

8.1 Problem Setup

$$y|x,u = x'\beta + z'u + \varepsilon \tag{8.1}$$

where x are the factors with fixed effects, β, which we may want to study. The factors z, with effects u, are the random effects which contribute to variability. In our repeated measures example (8.2) the treatment is a fixed effect, and the subject is a random effect. In our bottle-caps example (8.3) the time (before vs. after) is a fixed effect, and the machines may be either a fixed or a random effect (depending on the purpose of inference). In our diet example (8.4) the diet is the fixed effect and the family is a random effect.

Notice that we state $y|x,z$ merely as a convenient way to do inference on $y|x$, instead of directly specifying $Var[y|x]$. This is exactly the power of LMMs: we specify the covariance not via the matrix $Var[y,z]$, but rather via the sampling hierarchy.

Given a sample of n observations (y_i, x_i, z_i) from model (8.1), we will want to estimate (β, u). Under some assumption on the distribution of ε and z, we can use *maximum likelihood* (ML). In the context of LMMs, however, ML is typically replaced with *restricted maximum likelihood* (ReML), because it returns unbiased estimates of $Var[y|x]$ and ML does not.

8.1.1 Non-Linear Mixed Models

The idea of random-effects can also be implemented for non-linear mean models. Formally, this means that $y|x,z = f(x,z,\varepsilon)$ for some non-linear f. This is known as *non-linead-mixed-models*, which will not be discussed in this text.

8.1.2 Generalized Linear Mixed Models (GLMM)

You can marry the ideas of random effects, with non-linear link functions, and non-Gaussian distribution of the response. These are known as Generalized Linear Mixed Models. Wikidot has a nice comparison of several software suits for GLMMs. Also consider the mcglm R pacakge (Bonat, 2018).

8.2 Mixed Models with R

We will fit mixed models with the `lmer` function from the **lme4** package, written by the mixed-models Guru Douglas Bates. We start with a small simulation demonstrating the importance of acknowledging your sources of variability. Our demonstration consists of fitting a linear model that assumes independence, when data is clearly dependent.

```
# Simulation parameters
n.groups <- 4 # number of groups
n.repeats <- 2 # sample per group
groups <- rep(1:n.groups, each=n.repeats) %>% as.factor
n <- length(groups)
z0 <- rnorm(n.groups,0,10) # generate group effects
(z <- z0[as.numeric(groups)]) # generate and inspect random group effects
```

```
## [1]   8.901364   8.901364  -4.318889  -4.318889   9.708611   9.708611
## [7] -10.693773 -10.693773
```

```
epsilon <- rnorm(n,0,1) # generate measurement error

# Generate data
beta0 <- 2 # set global mean
y <- beta0 + z + epsilon # generate synthetic sample
```

We can now fit the linear and mixed models.

```
lm.5 <- lm(y~1)  # fit a linear model assuming independence
library(lme4)
lme.5 <- lmer(y~1|groups) # fit a mixed-model that deals with the group dependence
```

The summary of the linear model

```
summary.lm.5 <- summary(lm.5)
summary.lm.5
```

```
##
## Call:
## lm(formula = y ~ 1)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.949  -7.275   1.629   8.668  10.005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.317      3.500   0.948    0.375
##
## Residual standard error: 9.898 on 7 degrees of freedom
```

The summary of the mixed-model

```
summary.lme.5 <- summary(lme.5)
summary.lme.5
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 | groups
##
## REML criterion at convergence: 41
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.15395 -0.50048  0.04306  0.55891  0.99797
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  groups   (Intercept) 111.962  10.581
##  Residual               2.012   1.418
## Number of obs: 8, groups:  groups, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    3.317      5.314   0.624
```

Look at the standard error of the global mean, i.e., the intercept: for `lm` it is 3.4996374, and for `lme` it is 5.3143284. Why this difference? Because `lm` treats the group effect[1] as a fixed while the mixed model treats the group effect as a source of noise/uncertainty. Clearly, inference using `lm` underestimates our uncertainty in the estimated population mean (β_0).

Now let's adopt the paired t-test view, which removes the group mean, so that it implicitly ignores the between-group variability. Which is the model compatible with this view?

```
diffs <- tapply(y, groups, diff)
diffs # Q:what is this estimating? A: epsilon+epsilon.
```

```
##         1         2         3         4
## -1.411024 -1.598983 -1.493730  3.052394
```

```
sd(diffs) #
```

```
## [1] 2.278119
```

So we see that a paired t-test infers only against the within-group variability. Q:Is this a good think? A: depends…

8.2.1 A Single Random Effect

We will use the `Dyestuff` data from the **lme4** package, which encodes the yield, in grams, of a coloring solution (`dyestuff`), produced in 6 batches using 5 different preparations.

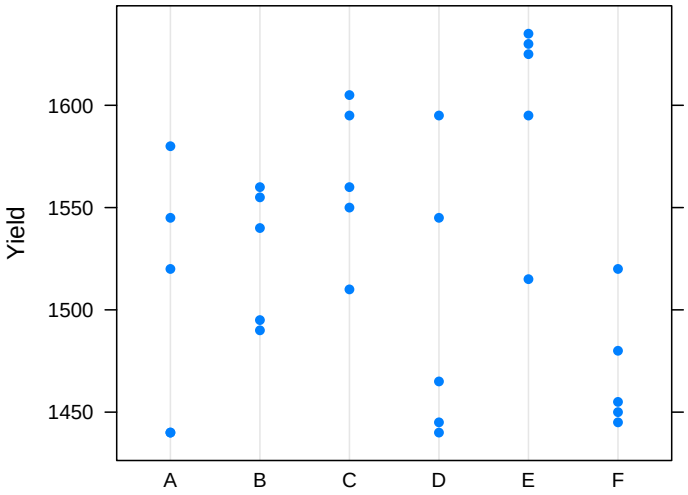
[1] A.k.a. the *cluster effect*.

```
data(Dyestuff, package='lme4')
attach(Dyestuff)
head(Dyestuff)
```

```
##   Batch Yield
## 1     A  1545
## 2     A  1440
## 3     A  1440
## 4     A  1520
## 5     A  1580
## 6     B  1540
```

And visually

```
lattice::dotplot(Yield~Batch)
```



If we want to do inference on the (global) mean yield, we need to account for the two sources of variability: the within-batch variability, and the between-batch variability We thus fit a mixed model, with an intercept and random batch effect.

```
lme.1<- lmer( Yield ~ 1  | Batch  , Dyestuff )
summary(lme.1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Yield ~ 1 | Batch
##    Data: Dyestuff
##
## REML criterion at convergence: 319.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4117 -0.7634  0.1418  0.7792  1.8296
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Batch    (Intercept) 1764     42.00
##  Residual             2451     49.51
## Number of obs: 30, groups:  Batch, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1527.50      19.38    78.8
```

Things to note:

- The syntax `Yield ~ 1 | Batch` tells R to fit a model with a global intercept (`1`) and a random Batch effect (`|Batch`). More on that later.
- As usual, `summary` is content aware and has a different behavior for `lme` class objects.
- The output distinguishes between random effects (u), a source of variability, and fixed effect (β), which we want to study. The mean of the random effect is not reported because it is unassumingly 0.
- Were we not interested in the variance components, and only in the coefficients or predictions, an (almost) equivalent `lm` formulation is `lm(Yield ~ Batch)`.

Some utility functions let us query the `lme` object. The function `coef` will work, but will return a cumbersome output. Better use `fixef` to extract the fixed effects, and `ranef` to extract the random effects. The model matrix (of the fixed effects alone), can be extracted with `model.matrix`, and predictions made with `predict`. Note, however, that predictions with mixed-effect models are better treated as prediction problems as in the Supervised Learning Chapter 10, but are a very delicate matter.

```
detach(Dyestuff)
```

8.2.2 Multiple Random Effects

Let's make things more interesting by allowing more than one random effect. One-way ANOVA can be thought of as the fixed-effects counterpart of the single random effect.

In the `Penicillin` data, we measured the diameter of spread of an organism, along the plate used (a to x), and penicillin type (A to F). We will now try to infer on the diameter of typical organism, and compute its variability over plates and Penicillin types.

```
head(Penicillin)
```

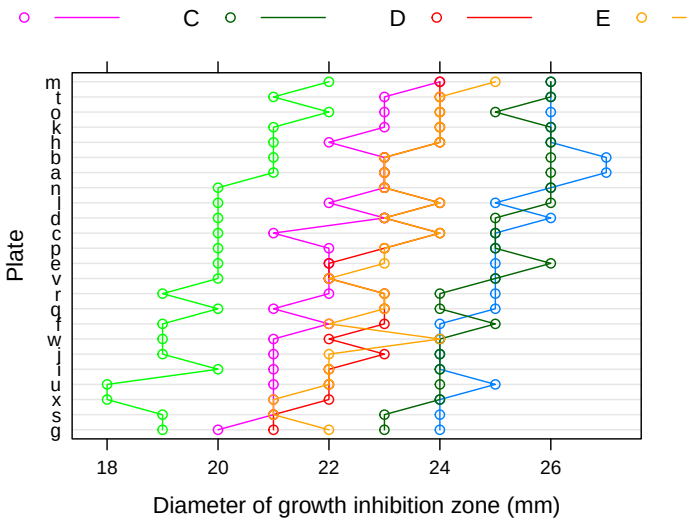
```
##   diameter plate sample
## 1       27     a      A
## 2       23     a      B
## 3       26     a      C
## 4       23     a      D
## 5       23     a      E
## 6       21     a      F
```

One sample per combination:

```
attach(Penicillin)
table(sample, plate) # how many observations per plate & type?
```

```
##       plate
## sample a b c d e f g h i j k l m n o p q r s t u v w x
##      A 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      B 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      C 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      D 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      E 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##      F 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

And visually:



Let's fit a mixed-effects model with a random plate effect, and a random sample effect:

```
lme.2 <- lmer ( diameter ~  1  + (1|plate )+(1|sample) , Penicillin )
fixef(lme.2) # Fixed effects
```

```
## (Intercept)
##    22.97222
```

```
ranef(lme.2) # Random effects
```

```
## $plate
##   (Intercept)
## a  0.80454389
## b  0.80454389
## c  0.18167120
## d  0.33738937
## e  0.02595303
## f -0.44120149
## g -1.37551052
## h  0.80454389
## i -0.75263783
## j -0.75263783
## k  0.96026206
## l  0.49310755
## m  1.42741658
## n  0.49310755
## o  0.96026206
## p  0.02595303
## q -0.28548332
## r -0.28548332
## s -1.37551052
## t  0.96026206
## u -0.90835601
## v -0.28548332
## w -0.59691966
## x -1.21979235
##
## $sample
##   (Intercept)
## A  2.18705819
## B -1.01047625
## C  1.93789966
```

```
## D -0.09689498
## E -0.01384214
## F -3.00374447
##
## with conditional variances for "plate" "sample"
```

Things to note:

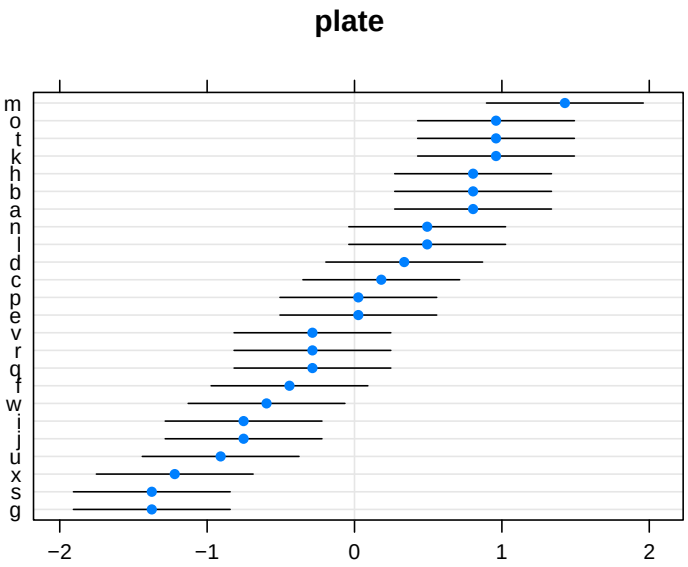
- The syntax `1+ (1| plate ) + (1| sample )` fits a global intercept (mean), a random plate effect, and a random sample effect.
- Were we not interested in the variance components, an (almost) equivalent `lm` formulation is `lm(diameter ~ plate + sample)`.
- The output of `ranef` is somewhat controversial. Think about it: Why would we want to plot the estimates of a random variable?

Since we have two random effects, we may compute the variability of the global mean (the only fixed effect) as we did before. Perhaps more interestingly, we can compute the variability in the response, for a particular plate or sample type.

```
random.effect.lme2 <- ranef(lme.2, condVar = TRUE)
qrr2 <- lattice::dotplot(random.effect.lme2, strip = FALSE)
```

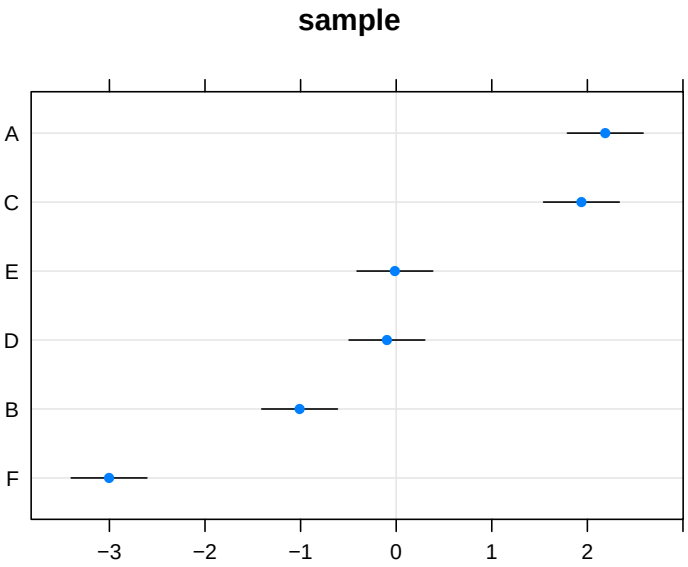
Variability in response for each plate, over various sample types:

```
print(qrr2[[1]])
```



Variability in response for each sample type, over the various plates:

```
print(qrr2[[2]])
```



Things to note:

- The `condVar` argument of the `ranef` function tells R to compute the variability in response conditional on each random effect at a time.
- The `dotplot` function, from the **lattice** package, is only there for the fancy plotting.

We used the penicillin example to demonstrate the incorporation of two random-effects. We could have, however, compared between penicillin types. For this matter, penicillin types are fixed effects to infer on, and not part of the uncertainty in the mean diameter. The appropriate model is the following:

```
lme.2.2 <- lmer( diameter ~  1  + sample + (1|plate) , Penicillin )
```

I may now ask myself: does the `sample`, i.e. penicillin, have any effect? This is what the ANOVA table typically gives us. The next table can be thought of as a "repeated measures ANOVA":

```
anova(lme.2.2)
```

```
## Analysis of Variance Table
##        Df Sum Sq Mean Sq F value
## sample  5 449.22  89.844  297.09
```

Ugh! No p-values. Why is this? Because Doug Bates, the author of **lme4** makes a strong argument against current methods of computing p-values in mixed models. If you insist on an p-value, you may recur to other packages that provide that, at your own caution:

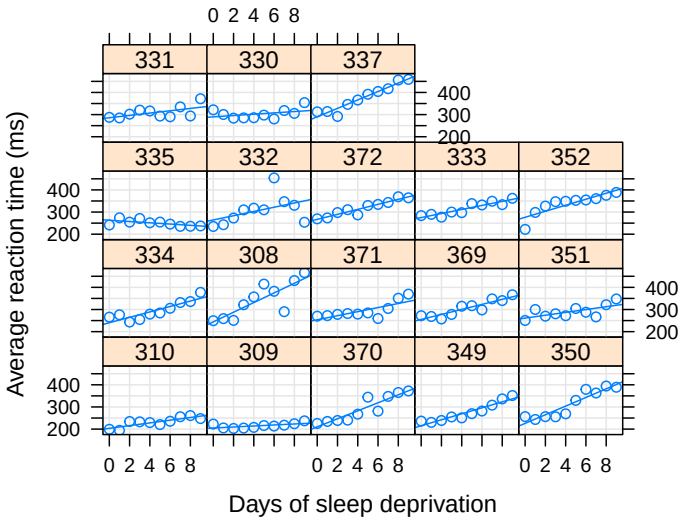
```
car::Anova(lme.2.2)
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: diameter
##         Chisq Df Pr(>Chisq)
## sample 1485.5  5  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

… and yes; the penicillin type has a significant effect on the diameter.

8.2.3 A Full Mixed-Model

In the `sleepstudy` data, we recorded the reaction times to a series of tests (`Reaction`), after various subject (`Subject`) underwent various amounts of sleep deprivation (`Day`).



We now want to estimate the (fixed) effect of the days of sleep deprivation on response time, while allowing each subject to have his/hers own effect. Put differently, we want to estimate a *random slope* for the effect of `day`. The fixed `Days` effect can be thought of as the average slope over subjects.

```
lme.3 <- lmer ( Reaction ~ Days + ( Days | Subject ) , data= sleepstudy )
```

Things to note:

- `~Days` specifies the fixed effect.
- We used the `Days|Subect` syntax to tell R we want to fit the model `~Days` within each subject.
- Were we fitting the model for purposes of prediction only, an (almost) equivalent `lm` formulation is `lm(Reaction~Days*Subject)`.

The fixed day effect is:

```
fixef(lme.3)
```

```
## (Intercept)        Days
##   251.40510    10.46729
```

The variability in the average response (intercept) and day effect is

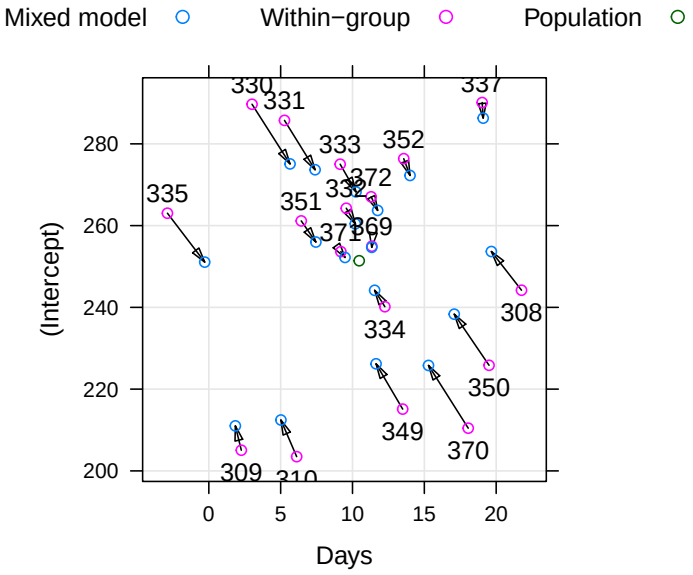
```
ranef(lme.3)
```

```
## $Subject
##     (Intercept)        Days
## 308   2.2575329   9.1992737
## 309 -40.3942719  -8.6205161
## 310 -38.9563542  -5.4495796
## 330  23.6888704  -4.8141448
## 331  22.2585409  -3.0696766
## 332   9.0387625  -0.2720535
## 333  16.8389833  -0.2233978
## 334  -7.2320462   1.0745075
## 335  -0.3326901 -10.7524799
## 337  34.8865253   8.6290208
## 349 -25.2080191   1.1730997
## 350 -13.0694180   6.6142185
## 351   4.5777099  -3.0152825
## 352  20.8614523   3.5364062
## 369   3.2750882   0.8722876
## 370 -25.6110745   4.8222518
## 371   0.8070591  -0.9881730
## 372  12.3133491   1.2842380
##
```

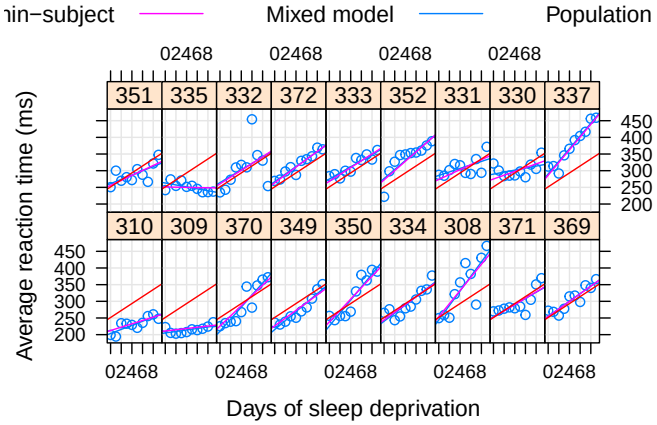
```
## with conditional variances for "Subject"
```

Did we really need the whole `lme` machinery to fit a within-subject linear regression and then average over subjects? The answer is yes. The assumptions on the distribution of random effect, namely, that they are normally distributed, allows us to pool information from one subject to another. In the words of John Tukey: "we borrow strength over subjects". Is this a good thing? If the normality assumption is true, it certainly is. If, on the other hand, you have a lot of samples per subject, and you don't need to "borrow strength" from one subject to another, you can simply fit within-subject linear models without the mixed-models machinery.

To demonstrate the "strength borrowing", here is a comparison of the lme, versus the effects of fitting a linear model to each subject separately.



Here is a comparison of the random-day effect from `lme` versus a subject-wise linear model. They are not the same.




```
detach(Penicillin)
```

8.3 Serial Correlations

As previously stated, a hierarchical model is a very convenient way to state correlations. The hierarchical sampling scheme will always yield correlations in blocks. What is the correlation does not have a block structure? Like a smooth temporal decay for time-series, or a smooth spatial decay for geospatial data?

One way to go about, is to find a dedicated package. For instance, in the Spatio-Temporal Data task view, or the Ecological and Environmental task view. Fans of vector-auto-regression should have a look at the vars package.

Instead, we will show how to solve this matter using the **nlme** package. This is because **nlme** allows to specify both

a block-covariance structure using the mixed-models framework, and the smooth parametric covariances we find in temporal and spatial data.

The `nlme::Ovary` data is panel data of number of ovarian follicles in different mares (female horse), at various times.

with an AR(1) temporal correlation, alongside random-effects, we take an example from the help of `nlme::corAR1`.

```
library(nlme)
head(nlme::Ovary)
```

```
## Grouped Data: follicles ~ Time | Mare
##   Mare        Time follicles
## 1    1 -0.13636360        20
## 2    1 -0.09090910        15
## 3    1 -0.04545455        19
## 4    1  0.00000000        16
## 5    1  0.04545455        13
## 6    1  0.09090910        10
```

```
fm1Ovar.lme <- nlme::lme(fixed=follicles ~ sin(2*pi*Time) + cos(2*pi*Time),
                  data = Ovary,
                  random = pdDiag(~sin(2*pi*Time)),
                  correlation=corAR1() )
summary(fm1Ovar.lme)
```

```
## Linear mixed-effects model fit by REML
##  Data: Ovary
##        AIC     BIC   logLik
##   1563.448 1589.49 -774.724
##
## Random effects:
##  Formula: ~sin(2 * pi * Time) | Mare
##  Structure: Diagonal
##         (Intercept) sin(2 * pi * Time) Residual
## StdDev:    2.858385           1.257977 3.507053
##
## Correlation Structure: AR(1)
##  Formula: ~1 | Mare
##  Parameter estimate(s):
##       Phi
## 0.5721866
## Fixed effects: follicles ~ sin(2 * pi * Time) + cos(2 * pi * Time)
##                        Value Std.Error  DF   t-value p-value
## (Intercept)        12.188089 0.9436602 295 12.915760  0.0000
## sin(2 * pi * Time) -2.985297 0.6055968 295 -4.929513  0.0000
## cos(2 * pi * Time) -0.877762 0.4777821 295 -1.837159  0.0672
##  Correlation:
##                    (Intr) s(*p*T
## sin(2 * pi * Time)  0.000
## cos(2 * pi * Time) -0.123  0.000
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.34910093 -0.58969626 -0.04577893  0.52931186  3.37167486
##
## Number of Observations: 308
## Number of Groups: 11
```

Things to note:

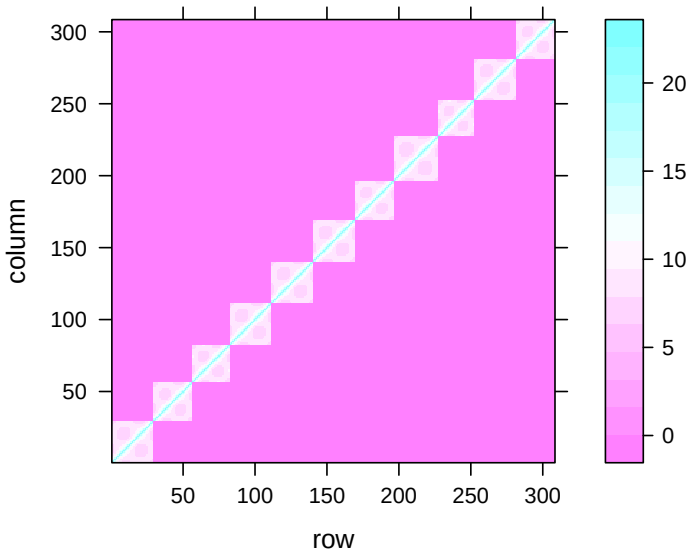
- The fitting is done with the `nlme::lme` function, and not `lme4::lmer` (which does not allow for non blocked

covariance models).

- `sin(2*pi*Time) + cos(2*pi*Time)` is a fixed effect that captures seasonality.
- The temporal covariance, is specified using the `correlations=` argument.
- AR(1) was assumed by calling `correlation=corAR1()`. See `nlme::corClasses` for a list of supported correlation structures.
- From the summary, we see that a `Mare` random effect has also been added. Where is it specified? It is implied by the `random=` argument. Read `?lme` for further details.

We can now inspect the contrivance implied by our model's specification:

```
the.cov <- mgcv::extract.lme.cov(fm1Ovar.lme, data = Ovary)
lattice::levelplot(the.cov)
```



8.4 Extensions

8.4.1 Cluster Robust Standard Errors

As previously stated, random effects are nothing more than a convenient way to specify dependencies within a level of a random effect, i.e., within a group/cluster. This is also the motivation underlying *cluster robust* inference, which is immensely popular with econometricians, but less so elsewhere. What is the difference between the two?

Mixed models framework is a bona-fide generalization of cluster robust inference. This author thus recommends using the **lme4** and **nlme** packages for mixed models to deal with correlations within cluster.

For a longer comparison between the two approaches, see Michael Clarck's guide.

8.4.2 Linear Models for Panel Data

nlme and **lme4** will probably provide you with all the functionality you need for panel data. If, however, you are trained as an econometrist, prefer the econometric parlance, and are not using non-linead models, then the plm package is just for you. In particular, it allows for cluster-robust covariance estimates, and Durbin–Wu–Hausman test for random effects. The **plm** package vignette also has a comparison to the **nlme** package.

8.4.3 Testing Hypotheses on Correlations

After working so hard to model the correlations in observation, we may want to test if it was all required. Douglas Bates, the author of **nlme** and **lme4** wrote a famous cautionary note, found here, on hypothesis testing in mixed models. Many practitioners, however, do not adopt Doug's view. Many of the popular tests, particularly the ones in

the econometric literature, can be found in the **plm** package (see Section 6 in the package vignette). These include tests for poolability, Hausman test, tests for serial correlations, tests for cross-sectional dependence, and unit root tests.

8.5 Relation to Other Estimators

8.5.1 Fixed Effects in the Econometric Literature

Fixed effects in the statistical literature, as discussed herein, are different than those in the econometric literature. See Section 7 of the **plm** package vignette for a comparison.

8.5.2 Relation to Generalized Least Squares (GLS)

GLS is the solution to a decorrelated least squares problem:

$$\hat{\beta}_{GLS} := argmin_{\beta}\{(X'\beta - y)'\Sigma^{-1}(X'\beta - y)'\}.$$

This estimator can be viewed as a least squares estimator that accounts for correlations in the data. It is also a maximum likelihood estimator under a Gaussian error assumption. Viewed as the latter, then linear mixed models under a Gaussian error assumption, collapses to a GLS estimator.

8.5.3 Relation to Conditional Gaussian Fields

In the geo-spatial literature, geo-located measurements are typically assumed to be sampled from a *Gaussian Random Field*. All the models discussed in this chapter can be stated in terms of these random fields. In the random field nomenclature, the fixed effects are known as the *drift*, or the *mean field*, and the covariance in errors is known as the *correlation function*. In other fields of literature the correlation function is known as a *charachteristic function*, *radial basis functions*, or *kernel*. Assuming stationarity, these simplify to the *power spectrum* via the *Wiener–Khinchin theorem*. The predictions of such models may be found under the names of *linear projection operators*, *best linear unbiased prediction*, *Kriging*, *radial basis function interpolators*.

8.5.4 Relation to Empirical Risk Minimization (ERM)

ERM is more general than mixed-models estimation since it allows loss functions that are not the (log) likelihood. ERM is less general than LMM, in that ERM (typically) does not account for correlations in the data.

8.5.5 Relation to M-Estimation

M-estimation is term in the statistical literature for ERM.

8.5.6 Relation to Generalize Estimating Equations (GEE)

The first order condition of the LMM problem returns a set of (non-linear) estimating equations. In this sense, GEE can be seen as more general than LMM in that the GEE need not be the derivative of the (log) likelihood.

8.5.7 Relation to MANOVA

Multivariate analysis of variance (MANOVA) deals with the estimation of effect on **vector valued** outcomes. Put differently: in ANOVA the response, y, is univariate. In MANOVA, the outcome is multivariate. MANOVA is useful when there are correlations among the entries of y. Otherwise- one may simply solve many ANOVA problems, instead of a single MANOVA.

Now assume that the outcome of a MANOVA is measurements of an individual at several time periods. The measurements are clearly correlated, so that MANOVA may be useful. But one may also treat the subject as a random effect, with a univariate response. We thus see that this seemingly MANOVA problem can be solved with the mixed models framework.

What MANOVA problems cannot be solved with mixed models? There may be cases where the covariance of the multivariate outcome, y, is very complicated. If the covariance in y may not be stated using a combination of random and fixed effects, then the covariance has to be stated explicitly. It is also possible to consider mixed-models with multivariate outcomes, i.e., a *mixed MANOVA*, or *hirarchial MANOVA*. The R functions we present herein permit this.

8.5.8 Relation to Seemingly Unrelated Equations (SUR)

SUR is the econometric term for MANOVA.

8.6 Bibliographic Notes

Most of the examples in this chapter are from the documentation of the **lme4** package (Bates et al., 2015). For a general and very applied treatment, see Pinero and Bates (2000). As usual, a hands on view can be found in Venables and Ripley (2013), and also in an excellent blog post by Kristoffer Magnusson For a more theoretical view see Weiss (2005) or Searle et al. (2009). Sometimes it is unclear if an effect is random or fixed; on the difference between the two types of inference see the classics: Eisenhart (1947), Kempthorne (1975), and the more recent Rosset and Tibshirani (2018). For more on predictions in linear mixed models see Robinson (1991), Rabinowicz and Rosset (2018), and references therein. See Michael Clarck's guide for various ways of dealing with correlations within groups. For the geo-spatial view and terminology of correlated data, see Christakos (2000), Diggle et al. (1998), Allard (2013), and Cressie and Wikle (2015).

8.7 Practice Yourself

1. Computing the variance of the sample mean given dependent correlations. How does it depend on the covariance between observations? When is the sample most informative on the population mean?
2. Return to the `Penicillin` data set. Instead of fitting an LME model, fit an LM model with `lm`. I.e., treat all random effects as fixed.
    a. Compare the effect estimates.
    b. Compare the standard errors.
    c. Compare the predictions of the two models.
3. [Very Advanced!] Return to the `Penicillin` data and use the `gls` function to fit a generalized linear model, equivalent to the LME model in our text.
4. Read about the "oats" dataset using `? MASS::oats`.Inspect the dependency of the yield (Y) in the Varieties (V) and the Nitrogen treatment (N).
    1. Fit a linear model, does the effect of the treatment significant? The interaction between the Varieties and Nitrogen is significant?
    2. An expert told you that could be a variance between the different blocks (B) which can bias the analysis. fit a LMM for the data.
    3. Do you think the blocks should be taken into account as "random effect" or "fixed effect"?
5. Return to the temporal correlation in Section 8.3, and replace the AR(1) covariance, with an ARMA covariance. Visualize the data's covariance matrix, and compare the fitted values.

See DataCamps' Hierarchical and Mixed Effects Models for more self practice.

Chapter 9

Multivariate Data Analysis

The term "multivariate data analysis" is so broad and so overloaded, that we start by clarifying what is discussed and what is not discussed in this chapter. Broadly speaking, we will discuss statistical *inference*, and leave more "exploratory flavored" matters like clustering, and visualization, to the Unsupervised Learning Chapter 11.

We start with an example.

Example 9.1. Consider the problem of a patient monitored in the intensive care unit. At every minute the monitor takes p physiological measurements: blood pressure, body temperature, etc. The total number of minutes in our data is n, so that in total, we have $n \times p$ measurements, arranged in a matrix. We also know the typical measurements for this patient when healthy: μ_0.

Formally, let y be single (random) measurement of a p-variate random vector. Denote $\mu := E[y]$. Here is the set of problems we will discuss, in order of their statistical difficulty.

- **Signal detection**: a.k.a. *multivariate hypothesis testing*, i.e., testing if μ equals μ_0 and for $\mu_0 = 0$ in particular. In our example: "are the current measurement different than a typical one?"
- **Signal counting**: Counting the number of elements in μ that differ from μ_0, and for $\mu_0 = 0$ in particular. In our example: "how many measurements differ than their typical values?"
- **Signal identification**: a.k.a. *multiple testing*, i.e., testing which of the elements in μ differ from μ_0 and for $\mu_0 = 0$ in particular. In the ANOVA literature, this is known as a **post-hoc** analysis. In our example: "which measurements differ than their typical values?"
- **Signal estimation**: Estimating the magnitudes of the departure of μ from μ_0, and for $\mu_0 = 0$ in particular. If estimation follows a *signal detection* or *signal identification* stage, this is known as a *selective estimation* problem. In our example: "what is the value of the measurements that differ than their typical values?"
- **Multivariate Regression**: a.k.a. *MANOVA* in statistical literature, and *structured learning* in the machine learning literature. In our example: "what factors affect the physiological measurements?"

Example 9.2. Consider the problem of detecting regions of cognitive function in the brain using fMRI. Each measurement is the activation level at each location in a brain's region. If the region has a cognitive function, the mean activation differs than $\mu_0 = 0$ when the region is evoked.

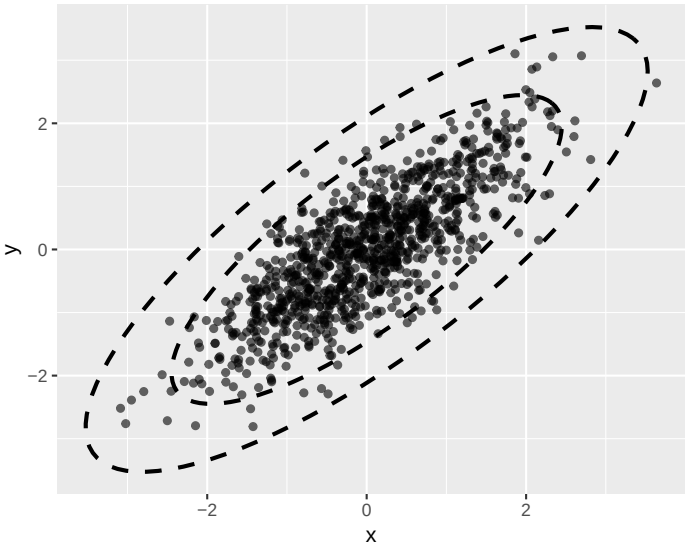
Example 9.3. Consider the problem of detecting cancer encoding regions in the genome. Each measurement is the vector of the genetic configuration of an individual. A cancer encoding region will have a different (multivariate) distribution between sick and healthy. In particular, μ of sick will differ from μ of healthy.

Example 9.4. Consider the problem of the simplest multiple regression. The estimated coefficient, $\hat{\beta}$ are a random vector. Regression theory tells us that its covariance is $(X'X)^{-1}\sigma^2$, and null mean of β. We thus see that inference on the vector of regression coefficients, is nothing more than a multivaraite inference problem.

Remark. In the above, "signal" is defined in terms of μ. It is possible that the signal is not in the location, μ, but rather in the covariance, Σ. We do not discuss these problems here, and refer the reader to Nadler (2008).

Another possible question is: does a multivariate analysis gives us something we cannot get from a mass-univariate analysis (i.e., a multivariate analysis on each variable separately). In Example 9.1 we could have just performed multiple univariate tests, and sign an alarm when any of the univariate detectors was triggered. The reason we want a multivariate detector, and not multiple univariate detectors is that it is possible that each measurement alone is borderline, but together, the signal accumulates. In our ICU example is may mean that the pulse is borderline, the body temperature is borderline, etc. Analyzed simultaneously, it is clear that the patient is in distress.

The next figure[1] illustrates the idea that some bi-variate measurements may seem ordinary univariately, while very anomalous when examined bi-variately.
Remark. The following figure may also be used to demonstrate the difference between Euclidean Distance and Mahalanobis Distance.



9.1 Signal Detection

Signal detection deals with the detection of the departure of μ from some μ_0, and especially, $\mu_0 = 0$. This problem can be thought of as the multivariate counterpart of the univariate hypothesis t-test.

9.1.1 Hotelling's T2 Test

The most fundamental approach to signal detection is a mere generalization of the t-test, known as *Hotelling's T^2 test*.

Recall the univariate t-statistic of a data vector x of length n:

$$t^2(x) := \frac{(\bar{x} - \mu_0)^2}{Var[\bar{x}]} = (\bar{x} - \mu_0)Var[\bar{x}]^{-1}(\bar{x} - \mu_0), \tag{9.1}$$

where $Var[\bar{x}] = S^2(x)/n$, and $S^2(x)$ is the unbiased variance estimator $S^2(x) := (n-1)^{-1}\sum(x_i - \bar{x})^2$.

Generalizing Eq(9.1) to the multivariate case: μ_0 is a p-vector, $\bar{x}$ is a p-vector, and $Var[\bar{x}]$ is a $p \times p$ matrix of the covariance between the p coordinated of $\bar{x}$. When operating with vectors, the squaring becomes a quadratic form, and the division becomes a matrix inverse. We thus have

$$T^2(x) := (\bar{x} - \mu_0)'Var[\bar{x}]^{-1}(\bar{x} - \mu_0), \tag{9.2}$$

[1] My thanks to Efrat Vilneski for the figure.

which is the definition of Hotelling's T^2 test statistic. We typically denote the covariance between coordinates in x with $\hat{\Sigma}(x)$, so that $\widehat{\Sigma}_{k,l} := \widehat{Cov}[x_k, x_l] = (n-1)^{-1}\sum(x_{k,i} - \bar{x}_k)(x_{l,i} - \bar{x}_l)$. Using the Σ notation, Eq.(9.2) becomes

$$T^2(x) := n(\bar{x} - \mu_0)'\hat{\Sigma}(x)^{-1}(\bar{x} - \mu_0), \tag{9.3}$$

which is the standard notation of Hotelling's test statistic.

For inference, we need the null distribution of Hotelling's test statistic. For this we introduce some vocabulary[2]:

1. **Low Dimension**: We call a problem *low dimensional* if $n \gg p$, i.e. $p/n \approx 0$. This means there are many observations per estimated parameter.
2. **High Dimension**: We call a problem *high dimensional* if $p/n \to c$, where $c \in (0, 1)$. This means there are more observations than parameters, but not many.
3. **Very High Dimension**: We call a problem *very high dimensional* if $p/n \to c$, where $1 < c < \infty$. This means there are less observations than parameter.

Hotelling's T^2 test can only be used in the low dimensional regime. For some intuition on this statement, think of taking $n = 20$ measurements of $p = 100$ physiological variables. We seemingly have 20 observations, but there are 100 unknown quantities in μ. Would you trust your conclusion that $\bar{x}$ is different than μ_0 based on merely 20 observations.

The above criticism is formalized in Bai and Saranadasa (1996). For modern applications, Hotelling's T^2 is not recommended, since many modern alternatives have been made available. See Rosenblatt et al. (2016) and references for a review.

9.1.2 Various Types of Signal to Detect

In the previous, we assumed that the signal is a departure of μ from some μ_0. For vactor-valued data y, that is distributed F, we may define "signal" as any departure from some F_0. This is the multivaraite counterpart of goodness-of-fit (GOF) tests.

Even when restricting "signal" to departures of μ from μ_0, we may try to detect various types of signal:

1. **Dense Signal**: when the departure is in all coordinates of μ.
2. **Sparse Signal**: when the departure is in a subset of coordinates of μ.

A manufactoring motivation is consistent with a dense signal: if a manufacturing process has failed, we expect a change in many measurements (i.e. coordinates of μ). A brain-imaging motivation is consistent with a dense signal: if a region encodes cognitive function, we expect a change in many brain locations (i.e. coordinates of μ.) A genetic motivation is consistent with a sparse signal: if susceptibility of disease is genetic, only a small subset of locations in the genome will encode it.

Hotelling's T^2 statistic is designed for dense signal. The following is a simple statistic designed for sparse signal.

9.1.3 Simes' Test

Hotelling's T^2 statistic has currently two limitations: It is designed for dense signals, and it requires estimating the covariance, which is a very difficult problem.

An algorithm, that is sensitive to sparse signal and allows statistically valid detection under a wide range of covariances (even if we don't know the covariance) is known as *Simes' Test*. The statistic is defined vie the following algorithm:

1. Compute p variable-wise p-values: $p_1, \ldots, p_j$.
2. Denote $p_{(1)}, \ldots, p_{(j)}$ the sorted p-values.
3. Simes' statistic is $p_{Simes} := min_j\{p_{(j)} \times p/j\}$.
4. Reject the "no signal" null hypothesis at significance α if $p_{Simes} < \alpha$.

[2]This vocabulary is not standard in the literature, so when you read a text, you need to verify yourself what the author means.

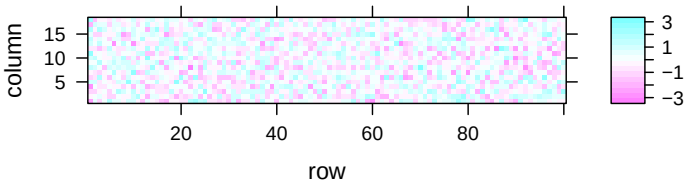
9.1.4 Signal Detection with R

Let's generate some data with no signal.

```
library(mvtnorm)
n <- 100 # observations
p <- 18 # parameter dimension
mu <- rep(0,p) # no signal
x <- rmvnorm(n = n, mean = mu)
dim(x)
```

```
## [1] 100  18
```

```
lattice::levelplot(x)
```



Now make our own Hotelling function.

```
hotellingOneSample <- function(x, mu0=rep(0,ncol(x))){
  n <- nrow(x)
  p <- ncol(x)
  stopifnot(n > 5* p)
  bar.x <- colMeans(x)
  Sigma <- var(x)
  Sigma.inv <- solve(Sigma)
  T2 <- n * (bar.x-mu0) %*% Sigma.inv %*% (bar.x-mu0)
  p.value <- pchisq(q = T2, df = p, lower.tail = FALSE)
  return(list(statistic=T2, pvalue=p.value))
}
hotellingOneSample(x)
```

```
## $statistic
##          [,1]
## [1,] 17.22438
##
## $pvalue
##           [,1]
## [1,] 0.5077323
```

Things to note:

- `stopifnot(n > 5 * p)` is a little verification to check that the problem is indeed low dimensional. Otherwise, the χ^2 approximation cannot be trusted.
- `solve` returns a matrix inverse.
- `%*%` is the matrix product operator (see also `crossprod()`).
- A function may return only a single object, so we wrap the statistic and its p-value in a `list` object.

Just for verification, we compare our home made Hotelling's test, to the implementation in the **rrcov** package. The statistic is clearly OK, but our χ^2 approximation of the distribution leaves room to desire. Personally, I would never trust a Hotelling test if n is not much greater than p, in which case I would use a high-dimensional adaptation (see Bibliography).

```
rrcov::T2.test(x)
```

```
##
##  One-sample Hotelling test
##
```

```
## data:  x
## T2 = 17.22400, F = 0.79259, df1 = 18, df2 = 82, p-value = 0.703
## alternative hypothesis: true mean vector is not equal to (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
##
## sample estimates:
##                       [,1]       [,2]       [,3]       [,4]      [,5]
## mean x-vector -0.01746212 0.03776332 0.1006145 -0.2083005 0.1026982
##                       [,6]         [,7]       [,8]       [,9]       [,10]
## mean x-vector -0.05220043 -0.009497987 -0.1139856 0.02851701 -0.03089953
##                      [,11]      [,12]      [,13]      [,14]      [,15]
## mean x-vector -0.02457798 -0.1270753 0.04717076 0.01683591 0.03085023
##                    [,16]       [,17]     [,18]
## mean x-vector 0.1499485 -0.07630663 0.1004852
```

Let's do the same with Simes':

```
Simes <- function(x){
  p.vals <- apply(x, 2, function(z) t.test(z)$p.value) # Compute variable-wise pvalues
  p <- ncol(x)
  p.Simes <- p * min(sort(p.vals)/seq_along(p.vals)) # Compute the Simes statistic
  return(c(pvalue=p.Simes))
}
Simes(x)
```

```
##    pvalue
## 0.6398998
```

And now we verify that both tests can indeed detect signal when present. Are p-values small enough to reject the "no signal" null hypothesis?

```
mu <- rep(x = 10/p,times=p) # inject signal
x <- rmvnorm(n = n, mean = mu)
hotellingOneSample(x)
```

```
## $statistic
##          [,1]
## [1,] 686.8046
##
## $pvalue
##                [,1]
## [1,] 3.575926e-134
```

```
Simes(x)
```

```
##       pvalue
## 2.765312e-10
```

... yes. All p-values are very small, so that all statistics can detect the non-null distribution.

9.2 Signal Counting

There are many ways to approach the *signal counting* problem. For the purposes of this book, however, we will not discuss them directly, and solve the signal counting problem as a signal identification problem: if we know **where** μ departs from μ_0, we only need to count coordinates to solve the signal counting problem.
Remark. In the sparsity or multiple-testing literature, what we call "signal counting" is known as "adapting to sparsit", or "adaptivity".

9.3 Signal Identification

The problem of *signal identification* is also known as *selective testing*, or more commonly as *multiple testing*.

In the ANOVA literature, an identification stage will typically follow a detection stage. These are known as the *omnibus F test*, and *post-hoc* tests, respectively. In the multiple testing literature there will typically be no preliminary detection stage. It is typically assumed that signal is present, and the only question is "where?"

The first question when approaching a multiple testing problem is "what is an error"? Is an error declaring a coordinate in μ to be different than μ_0 when it is actually not? Is an error an overly high proportion of falsely identified coordinates? The former is known as the *family wise error rate* (FWER), and the latter as the *false discovery rate* (FDR).
Remark. These types of errors have many names in many communities. See the Wikipedia entry on ROC for a table of the (endless) possible error measures.

9.3.1 Signal Identification in R

One (of many) ways to do signal identification involves the `stats::p.adjust` function. The function takes as inputs a p-vector of the variable-wise **p-values**. Why do we start with variable-wise p-values, and not the full data set?

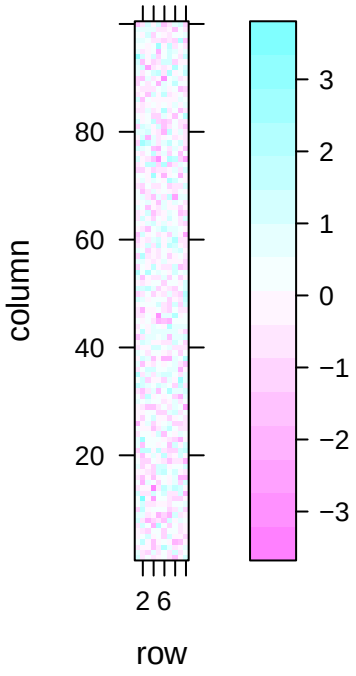
a. Because we want to make inference variable-wise, so it is natural to start with variable-wise statistics.
b. Because we want to avoid dealing with covariances if possible. Computing variable-wise p-values does not require estimating covariances.
c. So that the identification problem is decoupled from the variable-wise inference problem, and may be applied much more generally than in the setup we presented.

We start be generating some high-dimensional multivariate data and computing the coordinate-wise (i.e. hypothesis-wise) p-value.

```
library(mvtnorm)
n <- 1e1
p <- 1e2
mu <- rep(0,p)
x <- rmvnorm(n = n, mean = mu)
dim(x)
```

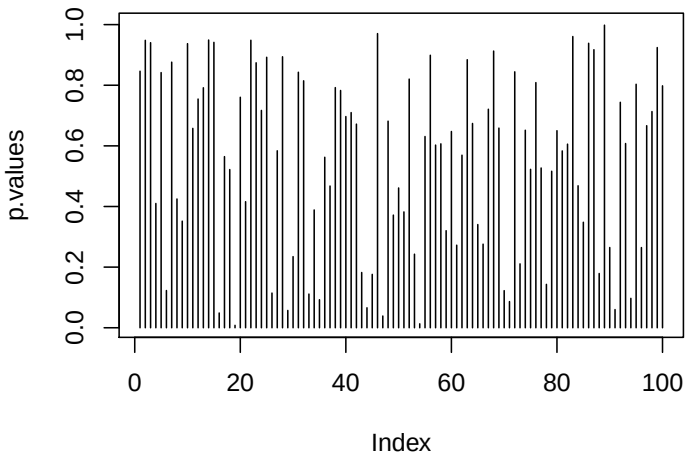
```
## [1]  10 100
```

```
lattice::levelplot(x)
```



We now compute the pvalues of each coordinate. We use a coordinate-wise t-test. Why a t-test? Because for the purpose of demonstration we want a simple test. In reality, you may use any test that returns valid p-values.

```
t.pval <- function(y) t.test(y)$p.value
p.values <- apply(X = x, MARGIN = 2, FUN = t.pval)
plot(p.values, type='h')
```



Things to note:

- `t.pval` is a function that merely returns the p-value of a t.test.
- We used the `apply` function to apply the same function to each column of `x`.
- `MARGIN=2` tells `apply` to compute over columns and not rows.
- The output, `p.values`, is a vector of 100 p-values.

We are now ready to do the identification, i.e., find which coordinate of μ is different than $\mu_0 = 0$. The workflow for identification has the same structure, regardless of the desired error guarantees:

1. Compute an `adjusted p-value`.
2. Compare the adjusted p-value to the desired error level.

If we want $FWER \leq 0.05$, meaning that we allow a 5% probability of making any mistake, we will use the `method="holm"` argument of `p.adjust`.

```
alpha <- 0.05
p.values.holm <- p.adjust(p.values, method = 'holm' )
which(p.values.holm < alpha)
```

```
## integer(0)
```

If we want $FDR \leq 0.05$, meaning that we allow the proportion of false discoveries to be no larger than 5%, we use the `method="BH"` argument of `p.adjust`.

```
alpha <- 0.05
p.values.BH <- p.adjust(p.values, method = 'BH' )
which(p.values.BH < alpha)
```

```
## integer(0)
```

We now inject some strong signal in μ just to see that the process works. We will artificially inject signal in the first 10 coordinates.

```
mu[1:10] <- 2 # inject signal in first 10 variables
x <- rmvnorm(n = n, mean = mu) # generate data
p.values <- apply(X = x, MARGIN = 2, FUN = t.pval)
p.values.BH <- p.adjust(p.values, method = 'BH' )
which(p.values.BH < alpha)
```

```
##  [1]  1  2  3  4  5  6  7  9 10 55
```

Indeed- we are now able to detect that the first coordinates carry signal, because their respective coordinate-wise null hypotheses have been rejected.

9.4 Signal Estimation (*)

The estimation of the elements of μ is a seemingly straightforward task. This is not the case, however, if we estimate only the elements that were selected because they were significant (or any other data-dependent criterion). Clearly, estimating only significant entries will introduce a bias in the estimation. In the statistical literature, this is known as *selection bias*. Selection bias also occurs when you perform inference on regression coefficients after some model selection, say, with a lasso, or a forward search[3].

Selective inference is a complicated and active research topic so we will not offer any off-the-shelf solution to the matter. The curious reader is invited to read Rosenblatt and Benjamini (2014), Javanmard and Montanari (2014), or Will Fithian's PhD thesis (Fithian, 2015) for more on the topic.

9.5 Biblipgraphic Notes

For a general introduction to multivariate data analysis see Anderson-Cook (2004). For an R oriented introduction, see Everitt and Hothorn (2011). For more on the difficulties with high dimensional problems, see Bai and Saranadasa (1996). For some cutting edge solutions for testing in high-dimension, see Rosenblatt et al. (2016) and references therein. Simes' test is not very well known. It is introduced in Simes (1986), and proven to control the type I error of detection under a PRDS type of dependence in Benjamini and Yekutieli (2001). For more on multiple testing, and signal identification, see Efron (2012). For more on the choice of your error rate see Rosenblatt (2013). For an excellent review on graphical models see Kalisch and Bühlmann (2014). Everything you need on graphical models, Bayesian belief networks, and structure learning in R, is collected in the Task View.

9.6 Practice Yourself

1. Generate multivariate data with:

```
set.seed(3)
mean<-rexp(50,6)
multi<-  rmvnorm(n = 100, mean = mean)
```

   a. Use Hotelling's test to determine if μ equals $\mu_0 = 0$. Can you detect the signal?
   b. Perform t.test on each variable and extract the p-value. Try to identify visually the variables which depart from μ_0.
   c. Use `p.adjust` to identify in which variables there are any departures from $\mu_0 = 0$. Allow 5% probability of making any false identification.
   d. Use `p.adjust` to identify in which variables there are any departures from $\mu_0 = 0$. Allow a 5% proportion of errors within identifications.

2. Generate multivariate data from two groups: `rmvnorm(n = 100, mean = rep(0,10))` for the first, and `rmvnorm(n = 100, mean = rep(0.1,10))` for the second.

   a. Do we agree the groups differ?
   b. Implement the two-group Hotelling test described in Wikipedia: (https://en.wikipedia.org/wiki/Hotelling%27s_T-squared_distribution#Two-sample_statistic).
   c. Verify that you are able to detect that the groups differ.
   d. Perform a two-group t-test on each coordinate. On which coordinates can you detect signal while controlling the FWER? On which while controlling the FDR? Use `p.adjust`.

3. Return to the previous problem, but set `n=9`. Verify that you cannot compute your Hotelling statistic.

[3] You might find this shocking, but it does mean that you cannot trust the `summary` table of a model that was selected from a multitude of models.

Chapter 10

Supervised Learning

Machine learning is very similar to statistics, but it is certainly not the same. As the name suggests, in machine learning we want machines to learn. This means that we want to replace hard-coded expert algorithm, with data-driven self-learned algorithm.

There are many learning setups, that depend on what information is available to the machine. The most common setup, discussed in this chapter, is *supervised learning*. The name takes from the fact that by giving the machine data samples with known inputs (a.k.a. features) and desired outputs (a.k.a. labels), the human is effectively supervising the learning. If we think of the inputs as predictors, and outcomes as predicted, it is no wonder that supervised learning is very similar to statistical prediction. When asked "are these the same?" I like to give the example of internet fraud. If you take a sample of fraud "attacks", a statistical formulation of the problem is highly unlikely. This is because fraud events are not randomly drawn from some distribution, but rather, arrive from an adversary learning the defenses and adapting to it. This instance of supervised learning is more similar to game theory than statistics.

Other types of machine learning problems include (Sammut and Webb, 2011):

- **Unsupervised learning**: See Chapter 11.
- **Semi supervised learning**: Where only part of the samples are labeled. A.k.a. *co-training*, *learning from labeled and unlabeled data*, *transductive learning*.
- **Active learning**: Where the machine is allowed to query the user for labels. Very similar to *adaptive design of experiments*.
- **Learning on a budget**: A version of active learning where querying for labels induces variable costs.
- **Weak learning**: A version of supervised learning where the labels are given not by an expert, but rather by some heuristic rule. Example: mass-labeling cyber attacks by a rule based software, instead of a manual inspection.
- **Reinforcement learning**:
  Similar to active learning, in that the machine may query for labels. Different from active learning, in that the machine does not receive labels, but *rewards*.
- **Structure learning**: When predicting objects with structure such as dependent vectors, graphs, images, tensors, etc.
- **Manifold learning**: An instance of unsupervised learning, where the goal is to reduce the dimension of the data by embedding it into a lower dimensional manifold. A.k.a. *support estimation*.
- **Similarity Learning**: Where we try to learn how to measure similarity between objects (like faces, texts, images, etc.).
- **Metric Learning**: Like *similarity learning*, only that the similarity has to obey the definition of a *metric*.
- **Learning to learn**: Deals with the carriage of "experience" from one learning problem to another. A.k.a. *cummulative learning*, *knowledge transfer*, and *meta learning*.

10.1 Problem Setup

We now present the *empirical risk minimization* (ERM) approach to supervised learning, a.k.a. *M-estimation* in the statistical literature.
Remark. We do not discuss purely algorithmic approaches such as K-nearest neighbour and *kernel smoothing* due to space constraints. For a broader review of supervised learning, see the Bibliographic Notes.

Example 10.1 (Rental Prices)**.** Consider the problem of predicting if a mail is spam or not based on its attributes: length, number of exclamation marks, number of recipients, etc.

Given n samples with inputs x from some space $\mathcal{X}$ and desired outcome, y, from some space $\mathcal{Y}$. In our example, y is the spam/no-spam label, and x is a vector of the mail's attributes. Samples, (x, y) have some distribution we denote P. We want to learn a function that maps inputs to outputs, i.e., that classifies to spam given. This function is called a *hypothesis*, or *predictor*, denoted f, that belongs to a hypothesis class $\mathcal{F}$ such that $f : \mathcal{X} \to \mathcal{Y}$. We also choose some other function that fines us for erroneous prediction. This function is called the *loss*, and we denote it by $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$.
Remark. The *hypothesis* in machine learning is only vaguely related the *hypothesis* in statistical testing, which is quite confusing.
Remark. The *hypothesis* in machine learning is not a bona-fide *statistical model* since we don't assume it is the data generating process, but rather some function which we choose for its good predictive performance.

The fundamental task in supervised (statistical) learning is to recover a hypothesis that minimizes the average loss in the sample, and not in the population. This is know as the *risk minimization problem*.

Definition 10.1 (Risk Function)**.** The *risk function*, a.k.a. *generalization error*, or *test error*, is the population average loss of a predictor f:

$$R(f) := \mathbb{E}_P[l(f(x), y)]. \tag{10.1}$$

The best predictor, is the risk minimizer:

$$f^* := argmin_f\{R(f)\}. \tag{10.2}$$

Another fundamental problem is that we do not know the distribution of all possible inputs and outputs, P. We typically only have a sample of $(x_i, y_i), i = 1, \ldots, n$. We thus state the *empirical* counterpart of (10.2), which consists of minimizing the average loss. This is known as the *empirical risk miminization* problem (ERM).

Definition 10.2 (Empirical Risk)**.** The *empirical risk function*, a.k.a. *in-sample error*, or *train error*, is the sample average loss of a predictor f:

$$R_n(f) := 1/n \sum_i l(f(x_i), y_i). \tag{10.3}$$

A good candidate proxy for f^* is its empirical counterpart, $\hat{f}$, known as the *empirical risk minimizer*:

$$\hat{f} := argmin_f\{R_n(f)\}. \tag{10.4}$$

To make things more explicit:

- f may be a linear function of the attributes, so that it may be indexed simply with its coefficient vector β.
- l may be a squared error loss: $l(f(x), y) := (f(x) - y)^2$.

Under these conditions, the best predictor $f^* \in \mathcal{F}$ from problem (10.2) is to

$$f^* := argmin_\beta\{\mathbb{E}_{P(x,y)}[(x'\beta - y)^2]\}. \tag{10.5}$$

When using a linear hypothesis with squared loss, we see that the empirical risk minimization problem collapses to an ordinary least-squares problem:

$$\hat{f} := argmin_\beta\{1/n \sum_i (x_i'\beta - y_i)^2\}. \tag{10.6}$$

When data samples are assumingly independent, then maximum likelihood estimation is also an instance of ERM, when using the (negative) log likelihood as the loss function.

If we don't assume any structure on the hypothesis, f, then $\hat{f}$ from (10.4) will interpolate the data, and $\hat{f}$ will be a very bad predictor. We say, it will *overfit* the observed data, and will have bad performance on new data.

We have several ways to avoid overfitting:

1. Restrict the hypothesis class $\mathcal{F}$ (such as linear functions).
2. Penalize for the complexity of f. The penalty denoted by $\|f\|$.
3. Unbiased risk estimation: $R_n(f)$ is not an unbiased estimator of $R(f)$. Why? Think of estimating the mean with the sample minimum… Because $R_n(f)$ is downward biased, we may add some correction term, or compute $R_n(f)$ on different data than the one used to recover $\hat{f}$.

Almost all ERM algorithms consist of some combination of all the three methods above.

10.1.1 Common Hypothesis Classes

Some common hypothesis classes, $\mathcal{F}$, with restricted complexity, are:

1. **Linear hypotheses**: such as linear models, GLMs, and (linear) support vector machines (SVM).
2. **Neural networks**: a.k.a. *feed-forward* neural nets, *artificial* neural nets, and the celebrated class of *deep* neural nets.
3. **Tree**: a.k.a. *decision rules*, is a class of hypotheses which can be stated as "if-then" rules.
4. **Reproducing Kernel Hilbert Space**: a.k.a. RKHS, is a subset of "the space of all functions[1]" that is both large enough to capture very complicated relations, but small enough so that it is less prone to overfitting, and also surprisingly simple to compute with.

10.1.2 Common Complexity Penalties

The most common complexity penalty applies to classes that have a finite dimensional parametric representation, such as the class of linear predictors, parametrized via its coefficients β. In such classes we may penalize for the norm of the parameters. Common penalties include:

1. **Ridge penalty**: penalizing the l_2 norm of the parameter. I.e. $\|f\| = \|\beta\|_2^2 = \sum_j \beta_j^2$.
2. **LASSO penalty**: penalizing the l_1 norm of the parameter. I.e., $\|f\| = \|\beta\|_1 = \sum_j |\beta_j|$
3. **Elastic net**: a combination of the lasso and ridge penalty. I.e. ,$\|f\| = \alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1$.
4. **Function Norms**: If the hypothesis class $\mathcal{F}$ does not admit a finite dimensional representation, the penalty is no longer a function of the parameters of the function. We may, however, penalize not the parametric representation of the function, but rather the function itself $\|f\| = \sqrt{\int f(t)^2 dt}$.

[1] It is even a subset of the Hilbert space, itself a subset of the space of all functions.

10.1.3 Unbiased Risk Estimation

The fundamental problem of overfitting, is that the empirical risk, $R_n(\hat{f})$, is downward biased to the population risk, $R(\hat{f})$. We can remove this bias in two ways: (a) purely algorithmic *resampling* approaches, and (b) theory driven estimators.

1. **Train-Validate-Test**: The simplest form of algorithmic validation is to split the data. A *train* set to train/estimate/learn $\hat{f}$. A *validation* set to compute the out-of-sample expected loss, $R(\hat{f})$, and pick the best performing predictor. A *test* sample to compute the out-of-sample performance of the selected hypothesis. This is a very simple approach, but it is very "data inefficient", thus motivating the next method.
2. **V-Fold Cross Validation**: By far the most popular algorithmic unbiased risk estimator; in *V-fold CV* we "fold" the data into V non-overlapping sets. For each of the V sets, we learn $\hat{f}$ with the non-selected fold, and assess $R(\hat{f})$) on the selected fold. We then aggregate results over the V folds, typically by averaging.
3. **AIC**: Akaike's information criterion (AIC) is a theory driven correction of the empirical risk, so that it is unbiased to the true risk. It is appropriate when using the likelihood loss.
4. **Cp**: Mallow's Cp is an instance of AIC for likelihood loss under normal noise.

Other theory driven unbiased risk estimators include the *Bayesian Information Criterion* (BIC, aka SBC, aka SBIC), the *Minimum Description Length* (MDL), *Vapnic's Structural Risk Minimization* (SRM), the *Deviance Information Criterion* (DIC), and the *Hannan-Quinn Information Criterion* (HQC).

Other resampling based unbiased risk estimators include resampling **without replacement** algorithms like *delete-d cross validation* with its many variations, and **resampling with replacement**, like the *bootstrap*, with its many variations.

10.1.4 Collecting the Pieces

An ERM problem with regularization will look like

$$\hat{f} := argmin_{f \in \mathcal{F}} \{R_n(f) + \lambda \|f\|\}. \tag{10.7}$$

Collecting ideas from the above sections, a typical supervised learning pipeline will include: choosing the hypothesis class, choosing the penalty function and level, unbiased risk estimator. We emphasize that choosing the penalty function, $\|f\|$ is not enough, and we need to choose how "hard" to apply it. This if known as the *regularization level*, denoted by λ in Eq.(10.7).

Examples of such combos include:

1. Linear regression, no penalty, train-validate test.
2. Linear regression, no penalty, AIC.
3. Linear regression, l_2 penalty, V-fold CV. This combo is typically known as *ridge regression*.
4. Linear regression, l_1 penalty, V-fold CV. This combo is typically known as *LASSO regression*.
5. Linear regression, l_1 and l_2 penalty, V-fold CV. This combo is typically known as *elastic net regression*.
6. Logistic regression, l_2 penalty, V-fold CV.
7. SVM classification, l_2 penalty, V-fold CV.
8. Deep network, no penalty, V-fold CV.
9. Unrestricted, $\|\partial^2 f\|_2$, V-fold CV. This combo is typically known as a *smoothing spline*.

For fans of statistical hypothesis testing we will also emphasize: Testing and prediction are related, but are not the same:

- In the current chapter, we do not claim our models, f, are generative. I.e., we do not claim that there is some causal relation between x and y. We only claim that x predicts y.
- It is possible that we will want to ignore a significant predictor, and add a non-significant one (Foster and Stine, 2004).
- Some authors will use hypothesis testing as an initial screening for candidate predictors. This is a useful heuristic, but that is all it is– a heuristic. It may also fail miserably if predictors are linearly dependent (a.k.a. multi-collinear).

10.2 Supervised Learning in R

At this point, we have a rich enough language to do supervised learning with R.

In these examples, I will use two data sets from the **ElemStatLearn** package, that accompanies the seminal book by Friedman et al. (2001). I use the `spam` data for categorical predictions, and `prostate` for continuous predictions. In `spam` we will try to decide if a mail is spam or not. In `prostate` we will try to predict the size of a cancerous tumor. You can now call `?prostate` and `?spam` to learn more about these data sets.

Some boring pre-processing.

```
# Preparing prostate data
data("prostate", package = 'ElemStatLearn')
prostate <- data.table::data.table(prostate)
prostate.train <- prostate[train==TRUE, -"train"]
prostate.test <- prostate[train!=TRUE, -"train"]
y.train <- prostate.train$lcavol
X.train <- as.matrix(prostate.train[, -'lcavol'] )
y.test <- prostate.test$lcavol
X.test <- as.matrix(prostate.test[, -'lcavol'] )

# Preparing spam data:
data("spam", package = 'ElemStatLearn')
n <- nrow(spam)
train.prop <- 0.66
train.ind <- sample(x = c(TRUE,FALSE),
                    size = n,
                    prob = c(train.prop,1-train.prop),
                    replace=TRUE)
spam.train <- spam[train.ind,]
spam.test <- spam[!train.ind,]

y.train.spam <- spam.train$spam
X.train.spam <- as.matrix(spam.train[,names(spam.train)!='spam'] )
y.test.spam <- spam.test$spam
X.test.spam <-  as.matrix(spam.test[,names(spam.test)!='spam'])

spam.dummy <- spam
spam.dummy$spam <- as.numeric(spam$spam=='spam')
spam.train.dummy <- spam.dummy[train.ind,]
spam.test.dummy <- spam.dummy[!train.ind,]
```

We also define some utility functions that we will require down the road.

```
l2 <- function(x) x^2 %>% sum %>% sqrt
l1 <- function(x) abs(x) %>% sum
MSE <- function(x) x^2 %>% mean
missclassification <- function(tab) sum(tab[c(2,3)])/sum(tab)
```

10.2.1 Linear Models with Least Squares Loss

The simplest approach to supervised learning, is simply with OLS: a linear predictor, squared error loss, and train-test risk estimator. Notice the better in-sample MSE than the out-of-sample. That is overfitting in action.

```
ols.1 <- lm(lcavol~. ,data = prostate.train)
# Train error:
MSE( predict(ols.1)-prostate.train$lcavol)
```

```
## [1] 0.4383709
```

```
# Test error:
MSE( predict(ols.1, newdata=prostate.test)- prostate.test$lcavol)
```

```
## [1] 0.5084068
```

Things to note:

- I use the `newdata` argument of the `predict` function to make the out-of-sample predictions required to compute the test-error.
- The test error is larger than the train error. That is overfitting in action.

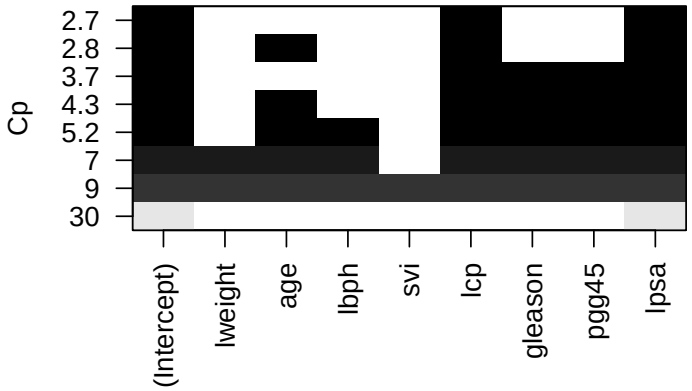
We now implement a V-fold CV, instead of our train-test approach. The assignment of each observation to each fold is encoded in `fold.assignment`. The following code is extremely inefficient, but easy to read.

```
folds <- 10
fold.assignment <- sample(1:folds, nrow(prostate), replace = TRUE)
errors <- NULL

for (k in 1:folds){
  prostate.cross.train <- prostate[fold.assignment!=k,] # train subset
  prostate.cross.test <-  prostate[fold.assignment==k,] # test subset
  .ols <- lm(lcavol~. ,data = prostate.cross.train) # train
  .predictions <- predict(.ols, newdata=prostate.cross.test)
  .errors <-  .predictions-prostate.cross.test$lcavol # save prediction errors in the fold
  errors <- c(errors, .errors) # aggregate error over folds.
}

# Cross validated prediction error:
MSE(errors)
```

```
## [1] 0.5742128
```

Let's try all possible variable subsets, and choose the best performer with respect to the Cp criterion, which is an unbiased risk estimator. This is done with `leaps::regsubsets`. We see that the best performer has 3 predictors.

```
regfit.full <- prostate.train %>%
  leaps::regsubsets(lcavol~.,data = ., method = 'exhaustive') # best subset selection
plot(regfit.full, scale = "Cp")
```



subset-1.bb

Things to note:

- The plot shows us which is the variable combination which is the best, i.e., has the smallest Cp.
- Scanning over all variable subsets is impossible when the number of variables is large.

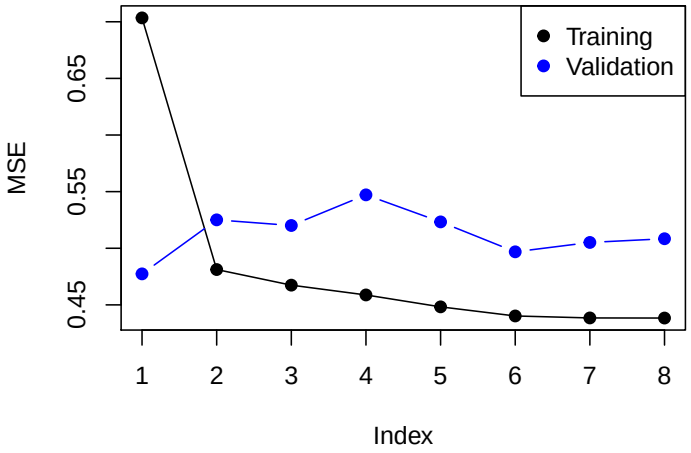
Instead of the Cp criterion, we now compute the train and test errors for all the possible predictor subsets[2]. In the resulting plot we can see overfitting in action.

[2]Example taken from https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/ch6.html

```
model.n <- regfit.full %>% summary %>% length
X.train.named <- model.matrix(lcavol ~ ., data = prostate.train )
X.test.named <- model.matrix(lcavol ~ ., data = prostate.test )

val.errors <- rep(NA, model.n)
train.errors <- rep(NA, model.n)
for (i in 1:model.n) {
    coefi <- coef(regfit.full, id = i) # exctract coefficients of i'th model

    pred <-  X.train.named[, names(coefi)] %*% coefi # make in-sample predictions
    train.errors[i] <- MSE(y.train - pred) # train errors

    pred <-  X.test.named[, names(coefi)] %*% coefi # make out-of-sample predictions
    val.errors[i] <- MSE(y.test - pred) # test errors
}
```

Plotting results.

```
plot(train.errors, ylab = "MSE", pch = 19, type = "o")
points(val.errors, pch = 19, type = "b", col="blue")
legend("topright",
       legend = c("Training", "Validation"),
       col = c("black", "blue"),
       pch = 19)
```



Checking all possible models is computationally very hard. *Forward selection* is a greedy approach that adds one variable at a time.

```
ols.0 <- lm(lcavol~1 ,data = prostate.train)
model.scope <- list(upper=ols.1, lower=ols.0)
step(ols.0, scope=model.scope, direction='forward', trace = TRUE)
```

```
## Start:  AIC=30.1
## lcavol ~ 1
##
##           Df Sum of Sq     RSS     AIC
## + lpsa     1    54.776  47.130 -19.570
## + lcp      1    48.805  53.101 -11.578
## + svi      1    35.829  66.077   3.071
## + pgg45    1    23.789  78.117  14.285
## + gleason  1    18.529  83.377  18.651
## + lweight  1     9.186  92.720  25.768
## + age      1     8.354  93.552  26.366
## <none>                 101.906  30.097
```

```
## + lbph     1     0.407 101.499  31.829
##
## Step:  AIC=-19.57
## lcavol ~ lpsa
##
##           Df Sum of Sq    RSS     AIC
## + lcp      1   14.8895 32.240 -43.009
## + svi      1    5.0373 42.093 -25.143
## + gleason  1    3.5500 43.580 -22.817
## + pgg45    1    3.0503 44.080 -22.053
## + lbph     1    1.8389 45.291 -20.236
## + age      1    1.5329 45.597 -19.785
## <none>                 47.130 -19.570
## + lweight  1    0.4106 46.719 -18.156
##
## Step:  AIC=-43.01
## lcavol ~ lpsa + lcp
##
##           Df Sum of Sq    RSS     AIC
## <none>                 32.240 -43.009
## + age      1   0.92315 31.317 -42.955
## + pgg45    1   0.29594 31.944 -41.627
## + gleason  1   0.21500 32.025 -41.457
## + lbph     1   0.13904 32.101 -41.298
## + lweight  1   0.05504 32.185 -41.123
## + svi      1   0.02069 32.220 -41.052
```

```
##
## Call:
## lm(formula = lcavol ~ lpsa + lcp, data = prostate.train)
##
## Coefficients:
## (Intercept)         lpsa          lcp
##     0.08798      0.53369      0.38879
```

Things to note:

- By default `step` add variables according to the AIC criterion, which is a theory-driven unbiased risk estimator.
- We need to tell `step` which is the smallest and largest models to consider using the `scope` argument.
- `direction='forward'` is used to "grow" from a small model. For "shrinking" a large model, use `direction='backward'`, or the default `direction='stepwise'`.

We now learn a linear predictor on the `spam` data using, a least squares loss, and train-test risk estimator.

```
# Train the predictor
ols.2 <- lm(spam~., data = spam.train.dummy)

# make in-sample predictions
.predictions.train <- predict(ols.2) > 0.5
# inspect the confusion matrix
(confusion.train <- table(prediction=.predictions.train, truth=spam.train.dummy$spam))
```

```
##           truth
## prediction    0    1
##      FALSE 1778  227
##      TRUE    66  980
```

```
# compute the train (in sample) misclassification
missclassification(confusion.train)
```

```
## [1] 0.09603409
```

```
# make out-of-sample prediction
.predictions.test <- predict(ols.2, newdata = spam.test.dummy) > 0.5
# inspect the confusion matrix
(confusion.test <- table(prediction=.predictions.test, truth=spam.test.dummy$spam))
```

```
##           truth
## prediction   0   1
##      FALSE 884 139
##      TRUE   60 467
```

```
# compute the train (in sample) misclassification
missclassification(confusion.test)
```

```
## [1] 0.1283871
```

Things to note:

- I can use `lm` for categorical outcomes. `lm` will simply dummy-code the outcome.
- A linear predictor trained on 0's and 1's will predict numbers. Think of these numbers as the probability of 1, and my prediction is the most probable class: `predicts()>0.5`.
- The train error is smaller than the test error. This is overfitting in action.

The `glmnet` package is an excellent package that provides ridge, LASSO, and elastic net regularization, for all GLMs, so for linear models in particular.

```
suppressMessages(library(glmnet))

means <- apply(X.train, 2, mean)
sds <- apply(X.train, 2, sd)
X.train.scaled <- X.train %>% sweep(MARGIN = 2, STATS = means, FUN = `-`) %>%
  sweep(MARGIN = 2, STATS = sds, FUN = `/`)
ridge.2 <- glmnet(x=X.train.scaled, y=y.train, family = 'gaussian', alpha = 0)

# Train error:
MSE( predict(ridge.2, newx =X.train.scaled)- y.train)
```

```
## [1] 1.006028
```

```
# Test error:
X.test.scaled <- X.test %>% sweep(MARGIN = 2, STATS = means, FUN = `-`) %>%
  sweep(MARGIN = 2, STATS = sds, FUN = `/`)
MSE(predict(ridge.2, newx = X.test.scaled)- y.test)
```

```
## [1] 0.7678264
```

Things to note:

- The `alpha=0` parameters tells R to do ridge regression. Setting $alpha = 1$ will do LASSO, and any other value, with return an elastic net with appropriate weights.
- The `family='gaussian'` argument tells R to fit a linear model, with least squares loss.
- Features for regularized predictors should be z-scored before learning.
- We use the `sweep` function to z-score the predictors: we learn the z-scoring from the train set, and apply it to both the train and the test.
- The test error is **smaller** than the train error. This may happen because risk estimators are random. Their variance may mask the overfitting.

We now use the LASSO penalty.

```
lasso.1 <- glmnet(x=X.train.scaled, y=y.train, , family='gaussian', alpha = 1)

# Train error:
MSE( predict(lasso.1, newx =X.train.scaled)- y.train)
```

```
## [1] 0.5525279
```

```
# Test error:
MSE( predict(lasso.1, newx = X.test.scaled)- y.test)
```

```
## [1] 0.5211263
```

We now use `glmnet` for classification.

```
means.spam <- apply(X.train.spam, 2, mean)
sds.spam <- apply(X.train.spam, 2, sd)
X.train.spam.scaled <- X.train.spam %>% sweep(MARGIN = 2, STATS = means.spam, FUN = `-`) %>%
  sweep(MARGIN = 2, STATS = sds.spam, FUN = `/`) %>% as.matrix

logistic.2 <- cv.glmnet(x=X.train.spam.scaled, y=y.train.spam, family = "binomial", alpha = 0)
```

Things to note:

- We used `cv.glmnet` to do an automatic search for the optimal level of regularization (the `lambda` argument in `glmnet`) using V-fold CV.
- Just like the `glm` function, `'family='binomial'` is used for logistic regression.
- We z-scored features so that they all have the same scale.
- We set `alpha=0` for an l_2 penalization of the coefficients of the logistic regression.

```
# Train confusion matrix:
.predictions.train <- predict(logistic.2, newx = X.train.spam.scaled, type = 'class')
(confusion.train <- table(prediction=.predictions.train, truth=spam.train$spam))
```

```
##           truth
## prediction email spam
##      email  1778  167
##      spam     66 1040
```

```
# Train misclassification error
missclassification(confusion.train)
```

```
## [1] 0.0763684
```

```
# Test confusion matrix:
X.test.spam.scaled <- X.test.spam %>% sweep(MARGIN = 2, STATS = means.spam, FUN = `-`) %>%
  sweep(MARGIN = 2, STATS = sds.spam, FUN = `/`) %>% as.matrix

.predictions.test <- predict(logistic.2, newx = X.test.spam.scaled, type='class')
(confusion.test <- table(prediction=.predictions.test, truth=y.test.spam))
```

```
##           truth
## prediction email spam
##      email   885  110
##      spam     59  496
```

```
# Test misclassification error:
missclassification(confusion.test)
```

```
## [1] 0.1090323
```

10.2.2 SVM

A support vector machine (SVM) is a linear hypothesis class with a particular loss function known as a hinge loss. We learn an SVM with the `svm` function from the **e1071** package, which is merely a wrapper for the libsvm C library; the most popular implementation of SVM today.

```
library(e1071)
svm.1 <- svm(spam~., data = spam.train, kernel='linear')
```

```
# Train confusion matrix:
.predictions.train <- predict(svm.1)
(confusion.train <- table(prediction=.predictions.train, truth=spam.train$spam))
```

```
##           truth
## prediction email spam
##      email  1774  106
##      spam     70 1101
```

```
missclassification(confusion.train)
```

```
## [1] 0.057686
```

```
# Test confusion matrix:
.predictions.test <- predict(svm.1, newdata = spam.test)
(confusion.test <- table(prediction=.predictions.test, truth=spam.test$spam))
```

```
##           truth
## prediction email spam
##      email   876   75
##      spam     68  531
```

```
missclassification(confusion.test)
```

```
## [1] 0.09225806
```

We can also use SVM for regression.

```
svm.2 <- svm(lcavol~., data = prostate.train, kernel='linear')

# Train error:
MSE( predict(svm.2)- prostate.train$lcavol)
```

```
## [1] 0.4488577
```

```
# Test error:
MSE( predict(svm.2, newdata = prostate.test)- prostate.test$lcavol)
```

```
## [1] 0.5547759
```

Things to note:

- The use of `kernel='linear'` forces the predictor to be linear. Various hypothesis classes may be used by changing the `kernel` argument.

10.2.3 Neural Nets

Neural nets (non deep) can be fitted, for example, with the `nnet` function in the **nnet** package. We start with a nnet regression.

```
library(nnet)
nnet.1 <- nnet(lcavol~., size=20, data=prostate.train, rang = 0.1, decay = 5e-4, maxit = 1000, trace=FALSE)

# Train error:
MSE( predict(nnet.1)- prostate.train$lcavol)
```

```
## [1] 1.177099
```

```
# Test error:
MSE( predict(nnet.1, newdata = prostate.test)- prostate.test$lcavol)
```

```
## [1] 1.21175
```

And nnet classification.

```
nnet.2 <- nnet(spam~., size=5, data=spam.train, rang = 0.1, decay = 5e-4, maxit = 1000, trace=FALSE)

# Train confusion matrix:
.predictions.train <- predict(nnet.2, type='class')
(confusion.train <- table(prediction=.predictions.train, truth=spam.train$spam))
```

```
##           truth
## prediction email spam
##      email  1806   59
##      spam     38 1148
```

```
missclassification(confusion.train)
```

```
## [1] 0.03179285
```

```
# Test confusion matrix:
.predictions.test <- predict(nnet.2, newdata = spam.test, type='class')
(confusion.test <- table(prediction=.predictions.test, truth=spam.test$spam))
```

```
##           truth
## prediction email spam
##      email   897   64
##      spam     47  542
```

```
missclassification(confusion.test)
```

```
## [1] 0.0716129
```

10.2.3.1 Deep Neural Nets

Deep-Neural-Networks are undoubtedly the "hottest" topic in machine-learning and artificial intelligence. This real is too vast to be covered in this text. We merely refer the reader to the tensorflow package documentation as a starting point.

10.2.4 Classification and Regression Trees (CART)

A CART, is not a linear hypothesis class. It partitions the feature space $\mathcal{X}$, thus creating a set of if-then rules for prediction or classification. It is thus particularly useful when you believe that the predicted classes may change abruptly with small changes in x.

10.2.4.1 The rpart Package

This view clarifies the name of the function `rpart`, which *recursively partitions* the feature space.

We start with a regression tree.

```
library(rpart)
tree.1 <- rpart(lcavol~., data=prostate.train)

# Train error:
MSE( predict(tree.1)- prostate.train$lcavol)
```

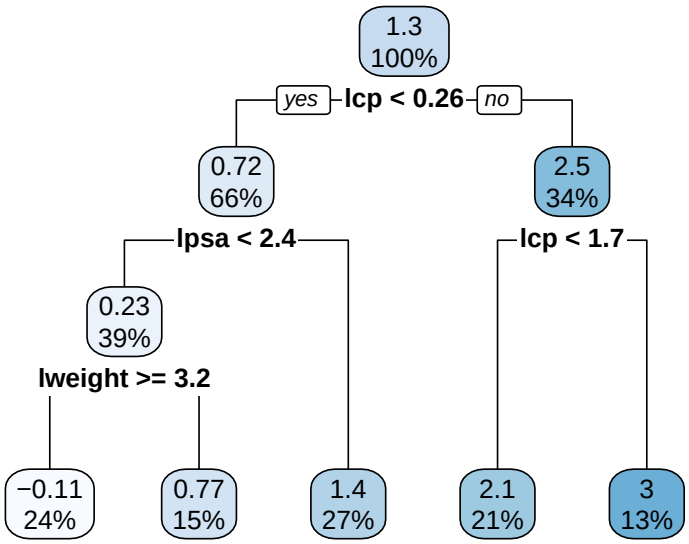
```
## [1] 0.4909568
```

```
# Test error:
MSE( predict(tree.1, newdata = prostate.test)- prostate.test$lcavol)
```

```
## [1] 0.5623316
```

We can use the **rpart.plot** package to visualize and interpret the predictor.

```
rpart.plot::rpart.plot(tree.1)
```



Trees are very prone to overfitting. To avoid this, we reduce a tree's complexity by *pruning* it. This is done with the `rpart::prune` function (not demonstrated herein).

We now fit a classification tree.

```
tree.2 <- rpart(spam~., data=spam.train)

# Train confusion matrix:
.predictions.train <- predict(tree.2, type='class')
(confusion.train <- table(prediction=.predictions.train, truth=spam.train$spam))
```

```
##           truth
## prediction email spam
##      email  1785  217
##      spam     59  990
```

```
missclassification(confusion.train)
```

```
## [1] 0.09046214
```

```
# Test confusion matrix:
.predictions.test <- predict(tree.2, newdata = spam.test, type='class')
(confusion.test <- table(prediction=.predictions.test, truth=spam.test$spam))
```

```
##           truth
## prediction email spam
##      email   906  125
##      spam     38  481
```

```
missclassification(confusion.test)
```

```
## [1] 0.1051613
```

10.2.4.2 The caret Package

In the **rpart** package [10.2.4.1] we grow a tree with one function, and then prune it with another.
The **caret** implementation of trees does both with a single function. We demonstrate the package in the context of trees, but it is actually a very convenient wrapper for many learning algorithms; 237(!) learning algorithms to be precise.

```
library(caret)
# Control some training parameters
train.control <- trainControl(method = "cv",
                              number = 10)

tree.3 <- train(lcavol~., data=prostate.train,
                method='rpart',
                trControl=train.control)
tree.3
```

```
## CART
##
## 67 samples
##  8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 61, 60, 59, 60, 60, 61, ...
## Resampling results across tuning parameters:
##
##   cp          RMSE       Rsquared   MAE
##   0.04682924  0.9118374  0.5026786  0.7570798
##   0.14815712  0.9899308  0.4690557  0.7972803
##   0.44497285  1.1912870  0.3264172  1.0008574
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.04682924.
```

```
# Train error:
MSE( predict(tree.3)- prostate.train$lcavol)
```

```
## [1] 0.6188435
```

```
# Test error:
MSE( predict(tree.3, newdata = prostate.test)- prostate.test$lcavol)
```

```
## [1] 0.545632
```

Things to note:

- A tree was trained because of the `method='rpart'` argument. Many other predictive models are available. See here.
- The pruning of the tree was done automatically by the `caret::train()` function.
- The method of pruning is controlled by a control object, generated with the `caret::trainControl()` function. In our case, `method = "cv"` for cross-validation, and `number = 10` for 10-folds.
- The train error is larger than the test error. This is possible because the tree is not an ERM on the train data. Rather, it is an ERM on the variations of the data generated by the cross-validation process.

10.2.5 K-nearest neighbour (KNN)

KNN is not an ERM problem. In the KNN algorithm, a prediction at some x is made based on the y is it neighbors. This means that:

- KNN is an Instance Based learning algorithm where we do not learn the values of some parametric function, but rather, need the original sample to make predictions. This has many implications when dealing with "BigData".
- It may only be applied in spaces with known/defined metric. It is thus harder to apply in the presence of missing values, or in "string-spaces", "genome-spaces", etc. where no canonical metric exists.

KNN is so fundamental that we show how to fit such a hypothesis class, even if it not an ERM algorithm. Is KNN any good? I have never seen a learning problem where KNN beats other methods. Others claim differently.

```
library(class)
knn.1 <- knn(train = X.train.spam.scaled, test = X.test.spam.scaled, cl =y.train.spam, k = 1)

# Test confusion matrix:
.predictions.test <- knn.1
(confusion.test <- table(prediction=.predictions.test, truth=spam.test$spam))
```

```
##           truth
## prediction email spam
##      email   856   86
##      spam     88  520
```

```
missclassification(confusion.test)
```

```
## [1] 0.1122581
```

10.2.6 Linear Discriminant Analysis (LDA)

LDA is equivalent to least squares classification 10.2.1. This means that we actually did LDA when we used `lm` for binary classification (feel free to compare the confusion matrices). There are, however, some dedicated functions to fit it which we now introduce.

```
library(MASS)
lda.1 <- lda(spam~., spam.train)

# Train confusion matrix:
.predictions.train <- predict(lda.1)$class
(confusion.train <- table(prediction=.predictions.train, truth=spam.train$spam))
```

```
##           truth
## prediction email spam
##      email  1776  227
##      spam     68  980
```

```
missclassification(confusion.train)
```

```
## [1] 0.09668961
```

```
# Test confusion matrix:
.predictions.test <- predict(lda.1, newdata = spam.test)$class
(confusion.test <- table(prediction=.predictions.test, truth=spam.test$spam))
```

```
##           truth
## prediction email spam
##      email   884  138
##      spam     60  468
```

```
missclassification(confusion.test)
```

```
## [1] 0.1277419
```

10.2.7 Naive Bayes

Naive-Bayes can be thought of LDA, i.e. linear regression, where predictors are assume to be uncorrelated. Predictions may be very good and certainly very fast, even if this assumption is not true.

```
library(e1071)
nb.1 <- naiveBayes(spam~., data = spam.train)

# Train confusion matrix:
```

```
.predictions.train <- predict(nb.1, newdata = spam.train)
(confusion.train <- table(prediction=.predictions.train, truth=spam.train$spam))
```

```
##           truth
## prediction email spam
##      email  1025   55
##      spam    819 1152
```

```
missclassification(confusion.train)
```

```
## [1] 0.2864635
```

```
# Test confusion matrix:
.predictions.test <- predict(nb.1, newdata = spam.test)
(confusion.test <- table(prediction=.predictions.test, truth=spam.test$spam))
```

```
##           truth
## prediction email spam
##      email   484   42
##      spam    460  564
```

```
missclassification(confusion.test)
```

```
## [1] 0.323871
```

10.2.8 Random Forrest

A Random Forrest is one of the most popular supervised learning algorithms. It it an extremely successful algorithm, with very few tuning parameters, and easily parallelizable (thus salable to massive datasets).

```
# Control some training parameters
train.control <- trainControl(method = "cv", number = 10)
rf.1 <- caret::train(lcavol~., data=prostate.train,
                method='rf',
                trControl=train.control)
rf.1
```

```
## Random Forest
##
## 67 samples
##  8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 62, 59, 60, 60, 59, 61, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##   2     0.7885535  0.6520820  0.6684168
##   5     0.7782809  0.6687843  0.6550590
##   8     0.7894338  0.6665277  0.6626417
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 5.
```

```
# Train error:
MSE( predict(rf.1)- prostate.train$lcavol)
```

```
## [1] 0.1340291
```

```r
# Test error:
MSE( predict(rf.1, newdata = prostate.test)- prostate.test$lcavol)
```

```
## [1] 0.5147782
```

Some of the many many many packages that learn random-forests include: randomForest, ranger.

10.2.9 Boosting

The fundamental idea behind **Boosting** is to construct a predictor, as the sum of several "weak" predictors. These weak predictors, are not trained on the same data. Instead, each predictor is trained on the residuals of the previous. Think of it this way: The first predictor targets the strongest signal. The second targets what the first did not predict. Etc. At some point, the residuals cannot be predicted anymore, and the learning will stabilize. Boosting is typically, but not necessarily, implemented as a sum of trees (@(trees)).

10.2.9.1 The gbm Package

TODO

10.2.9.2 The xgboost Package

TODO

10.3 Bibliographic Notes

The ultimate reference on (statistical) machine learning is Friedman et al. (2001). For a softer introduction, see James et al. (2013). A statistician will also like Ripley (2007). For a very algorithmic view, see the seminal Leskovec et al. (2014) or Conway and White (2012). For a much more theoretical reference, see Mohri et al. (2012), Vapnik (2013), Shalev-Shwartz and Ben-David (2014). Terminology taken from Sammut and Webb (2011). For an R oriented view see Lantz (2013). For review of other R sources for machine learning see Jim Savege's post, or the official Task View. For a review of resampling based unbiased risk estimation (i.e. cross validation) see the exceptional review of Arlot et al. (2010). If you want to know about Deep-Nets in R see here.

10.4 Practice Yourself

1. In 7.6 we fit a GLM for the `MASS::epil` data (Poisson family). We assume that the number of seizures (y) depending on the age of the patient (`age`) and the treatment (`trt`).
   1. What was the MSE of the model?
   2. Now, try the same with a ridge penalty using `glmnet` (`alpha=0`).
   3. Do the same with a LASSO penalty (`alpha=1`).
   4. Compare the test MSE of the three models. Which is the best ?
2. Read about the `Glass` dataset using `library(e1071)` and `?Glass`.
   1. Divide the dataset to train set and test set.
   2. Apply the various predictors from this chapter, and compare them using the proportion of missclassified.

See DataCamp's Supervised Learning in R: Classification, and Supervised Learning in R: Regression for more self practice.

Chapter 11

Unsupervised Learning

This chapter deals with machine learning problems which are unsupervised. This means the machine has access to a set of inputs, x, but the desired outcome, y is not available. Clearly, learning a relation between inputs and outcomes makes no sense, but there are still a lot of problems of interest. In particular, we may want to find a compact representation of the inputs, be it for visualization of further processing. This is the problem of *dimensionality reduction*. For the same reasons we may want to group similar inputs. This is the problem of *clustering*.

In the statistical terminology, with some exceptions, this chapter can be thought of as multivariate **exploratory** statistics. For multivariate **inference**, see Chapter 9.

11.1 Dimensionality Reduction

Example 11.1. Consider the heights and weights of a sample of individuals. The data may seemingly reside in 2 dimensions but given the height, we have a pretty good guess of a persons weight, and vice versa. We can thus state that heights and weights are not really two dimensional, but roughly lay on a 1 dimensional subspace of $\mathbb{R}^2$.

Example 11.2. Consider the correctness of the answers to a questionnaire with p questions. The data may seemingly reside in a p dimensional space, but assuming there is a thing as "skill", then given the correctness of a person's reply to a subset of questions, we have a good idea how he scores on the rest. Put differently, we don't really need a 200 question questionnaire– 100 is more than enough. If skill is indeed a one dimensional quality, then the questionnaire data should organize around a single line in the p dimensional cube.

Example 11.3. Consider n microphones recording an individual. The digitized recording consists of p samples. Are the recordings really a shapeless cloud of n points in $\mathbb{R}^p$? Since they all record the same sound, one would expect the n p-dimensional points to arrange around the source sound bit: a single point in $\mathbb{R}^p$. If microphones have different distances to the source, volumes may differ. We would thus expect the n points to arrange about a **line** in $\mathbb{R}^p$.

11.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is such a basic technique, it has been rediscovered and renamed independently in many fields. It can be found under the names of Discrete Karhunen–Loève Transform; Hotteling Transform; Proper Orthogonal Decomposition; Eckart–Young Theorem; Schmidt–Mirsky Theorem; Empirical Orthogonal Functions; Empirical Eigenfunction Decomposition; Empirical Component Analysis; Quasi-Harmonic Modes; Spectral Decomposition; Empirical Modal Analysis, and possibly more[1]. The many names are quite interesting as they offer an insight into the different problems that led to PCA's (re)discovery.

Return to the BMI problem in Example 11.1. Assume you wish to give each individual a "size score", that is a **linear** combination of height and weight: PCA does just that. It returns the linear combination that has the largest variability, i.e., the combination which best distinguishes between individuals.

[1] http://en.wikipedia.org/wiki/Principal_component_analysis

The variance maximizing motivation above was the one that guided Hotelling (1933). But 30 years before him, Pearson (1901) derived the same procedure with a different motivation in mind. Pearson was also trying to give each individual a score. He did not care about variance maximization, however. He simply wanted a small set of coordinates in some (linear) space that approximates the original data well.

Before we proceed, we give an example to fix ideas. Consider the crime rate data in `USArrests`, which encodes reported murder events, assaults, rapes, and the urban population of each american state.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

Following Hotelling's motivation, we may want to give each state a "crimilality score". We first remove the `UrbanPop` variable, which does not encode crime levels. We then z-score each variable with `scale`, and call PCA for a sequence of $1, \ldots, 3$ criminality scores that best separate between states.

```
USArrests.1 <- USArrests[,-3] %>% scale
pca.1 <- prcomp(USArrests.1, scale = TRUE)
pca.1
```

```
## Standard deviations (1, .., p=3):
## [1] 1.5357670 0.6767949 0.4282154
##
## Rotation (n x k) = (3 x 3):
##                PC1        PC2        PC3
## Murder  -0.5826006  0.5339532 -0.6127565
## Assault -0.6079818  0.2140236  0.7645600
## Rape    -0.5393836 -0.8179779 -0.1999436
```

Things to note and terminology:

- Distinguishing between states, i.e., finding the variance maximizing scores, should be indifferent to the **average** of each variable. We also don't want the score to be sensitive to the measurement **scale**. Formally, we say we want the scores to be *affine invariant*. We thus perform PCA in the z-score scale of each variable, obtained with the `scale` function.

- PCA is performed with the `prcomp` function. It returns the contribution (weight) of the original variables, to the new crimeness score. These weights are called the *loadings* . Z-scored loadings are known as *Rotations*, which is also the term in the `prcomp` output. If you are confused between loadings and rotations, see this Cross Validated entry.

- The number of possible scores, is the same as the number of original variables in the data.

- The new scores are called the *principal components*, labeled `PC1`,…,`PC3` in our output. They are computed by summing the original variables weighted by their loadings.

- The loadings/rotation on PC1 tell us that the best separation between states is along the average crime rate. Why is this? Because all the 3 crime variables have a similar loading on PC1.

- The other PCs are slightly harder to interpret, but it is an interesting exercise.

If we now represent each state, not with its original 4 variables, but only with the first 2 PCs (for example), we have reduced the dimensionality of the data.

11.1.1.1 Mathematics of PCA

What is the mathematical problem that is actually solved with PCA? Finding a linear combination (v) of the original variables (x), so that the new score/index ($v'x$) best separates individuals. Best separation implies that the variance of $v'x$ is maximal. Clearly, $Var[v'x]$ may explode if any v is allowed, so we need to pick v from some "fair" set. It is most convenient, mathematically, to constrain the l_2 norm to some constant: $\|v\|_2^2 = \sum v_j^2 = 1$. The first "best separating score", known as the first *principal component* (PC), is thus

$$v_1'x \quad s.t. \quad v_1 = argmax_v\{Var[v'x], \text{ and } \|v\|_2 = 1\}.$$

The second PC, is the same, only that it is required to be orthogonal to the first PC:

$$v_2'x \quad s.t. \quad v_2 = argmax_v\{Var[v'x], \text{ and } \|v\|_2 = 1, \text{ and } v'v_1 = 0\}.$$

The construction of the next PCs follows the same lines: find a linear transformation of the data that best separates observations and is orthogonal to the previous PCs.

11.1.1.2 How Hard is the PCA Problem?

Estimating all the PCs in the data is well defined algebraically if $n > p$, in which case p PCs are computable. This is the algebraic part of the problem, which is rather easy, and solved with SVD.

If viewing PCA as inference tool, we may ask about its statistical performance. It turns out that PCA has the same statistical difficulty as estimating a covariance matrix. As we already saw in the Multivariate Statistics Chapter (9), estimating covariances is a hard task, we thus recommend: don't trust your PCs if n is not much larger than p, and see the bibliographic notes for further details.

11.1.2 Dimensionality Reduction Preliminaries

Before presenting methods other than PCA, we need some terminology.

- **Variable**: A.k.a. *dimension*, or *feature*, or *column*. A vector of p measurements in their raw scale.
- **Feature Mapping**: A.k.a. *variable transformation*, or *data augmentation*. A measurement in a new, transformed, scale.
- **Data**: A.k.a. *sample*, *observations*. Will typically consist of n, vectors of dimension p. We typically denote the data as a $n \times p$ matrix X.
- **Data space**: A.k.a. *feature space*. The space of all possible values of X. We denote with $\mathcal{X}$.
- **Network**: A representation of the similarities (or dissimilarities) between the n units in the data. We denote with $\mathcal{G}$, and may be encoded in an $n \times n$ matrix.
- **Manifold**: A generalization of a linear space, which is regular enough so that, **locally**, it has all the properties of a linear space. We will denote an arbitrary manifold by $\mathcal{M}$, and by $\mathcal{M}_q$ a q dimensional[2] manifold.
- **Embedding**: Informally speaking: a "shape preserving" mapping (see figure below). We denote an embedding of the data X, into a manifold $\mathcal{M}$ by $X \mapsto \mathcal{M}$.
- **Embedding Function**: If the embedding is not only an algorithm, but rather, has a functional form representation, we call it an *embedding function* f. Given such a function, we are not restricted to embeddings of the original data, X, but may also embed new data points from $\mathcal{X}$: $f : \mathcal{X} \mapsto \mathcal{M}$.
- **Generative Model**: Known to statisticians as the **sampling distribution**. The assumed stochastic process that generated the observed X.

There are many motivations for dimensionality reduction:

1. **Scoring**: Give each observation an interpretable, simple score (Hotelling's motivation).

[2] You are probably used to thinking of the **dimension** of linear spaces. We will not rigorously define what is the dimension of a manifold, but you may think of it as the number of free coordinates needed to navigate along the manifold.

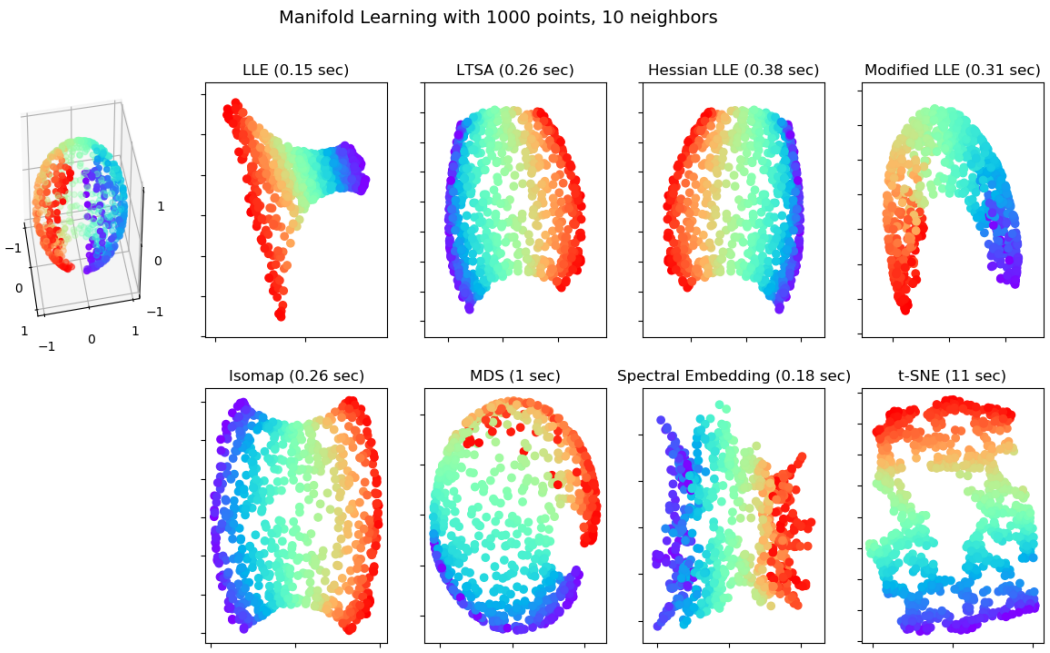


Figure 11.1: Various embedding algorithms. No embedding of the sphere to the plane is perfect. This is obviously not new. Maps makers have known this for centuries!Source: http://sci kit-learn.org/stable/auto_examples/manifold/plot_manifold_sphere.html#sphx-glr-auto-examples-manifold-plot-manifold-sphere-py

2. **Latent structure**: Recover unobservable information from indirect measurements. E.g: Blind signal reconstruction, CT scan, cryo-electron microscopy, etc.
3. **Signal to Noise**: Denoise measurements before further processing like clustering, supervised learning, etc.
4. **Compression**: Save on RAM ,CPU, and communication when operating on a lower dimensional representation of the data.

11.1.3 Latent Variable Generative Approaches

All generative approaches to dimensionality reduction will include a set of latent/unobservable variables, which we can try to recover from the observables X. The unobservable variables will typically have a lower dimension than the observables, thus, dimension is reduced. We start with the simplest case of linear Factor Analysis.

11.1.3.1 Factor Analysis (FA)

FA originates from the psychometric literature. We thus revisit the IQ (actually g-factor[3]) Example 11.2:

Example 11.4. Assume n respondents answer p quantitative questions: $x_i \in \mathbb{R}^p, i = 1, \ldots, n$. Also assume, their responses are some linear function of a single personality attribute, s_i. We can think of s_i as the subject's "intelligence". We thus have

$$x_i = s_i A + \varepsilon_i \tag{11.1}$$

And in matrix notation:

[3]https://en.wikipedia.org/wiki/G_factor_(psychometrics)

$$X = SA + \varepsilon, \tag{11.2}$$

where A is the $q \times p$ matrix of factor loadings, and S the $n \times q$ matrix of latent personality traits. In our particular example where $q = 1$, the problem is to recover the unobservable intelligence scores, $s_1, \ldots, s_n$, from the observed answers X.

We may try to estimate SA by assuming some distribution on S and ε and apply maximum likelihood. Under standard assumptions on the distribution of S and ε, recovering S from $\widehat{SA}$ is still impossible as there are infinitely many such solutions. In the statistical parlance we say the problem is *non identifiable*, and in the applied mathematics parlance we say the problem is *ill posed*.
Remark. The non-uniqueness (non-identifiability) of the FA solution under variable rotation is never mentioned in the PCA context. Why is this? This is because the methods solve different problems. The reason the solution to PCA is well defined is that PCA does not seek a single S but rather a **sequence** of S_q with dimensions growing from $q = 1$ to $q = p$.

The FA terminology is slightly different than PCA:

- **Factors**: The unobserved attributes S. Akin to the *principal components* in PCA.
- **Loading**: The A matrix; the contribution of each factor to the observed X.
- **Rotation**: An arbitrary orthogonal re-combination of the factors, S, and loadings, A, which changes the interpretation of the result.

The FA literature offers several heuristics to "fix" the identifiability problem of FA. These are known as *rotations*, and go under the names of *Varimax*, *Quartimax*, *Equimax*, *Oblimin*, *Promax*, and possibly others.

Because of their great similarity, FA is often confused with PCA. For a discussion of the similarities and dissimilarities, see this excellent StackExchange Q.

11.1.3.2 Independent Component Analysis (ICA)

Like FA, *independent compoent analysis* (ICA) is a family of latent space models, thus, a *meta-method*. It assumes data is generated as some function of the latent variables S. In many cases this function is assumed to be linear in S so that ICA is compared, if not confused, with PCA and even more so with FA.

The fundamental idea of ICA is that S has a joint distribution of **non-Gaussian**, **independent** variables. This independence assumption, solves the the non-uniqueness of S in FA. As such, it can be thought of as a type of rotation in FA. Then again, if the assumed distribution of S is both non-Gaussian, and well justified, then ICA is well defined, and more than just an arbitrary rotation of FA.

Being a generative model, estimation of S can then be done using maximum likelihood, or other estimation principles.

ICA is a popular technique in signal processing, where A is actually the signal, such as sound in Example 11.3. Recovering A is thus recovering the original signals mixing in the recorded X.

11.1.4 Purely Algorithmic Approaches

We now discuss dimensionality reduction approaches that are not stated via their generative model, but rather, directly as an algorithm. This does not mean that they cannot be cast via their generative model, but rather they were not motivated as such.

11.1.4.1 Multidimensional Scaling (MDS)

MDS can be thought of as a variation on PCA, that begins with the $n \times n$ graph of distances between data points $\mathcal{G}$; In contrast to PCA which operates on the original $n \times p$ data X. The term *graph* is typically used in this context, but saying *network* instead of *graph* is more appropriate. This is because a graph encodes connections (topology) and

networks encode distances (geometry). Put differently, a graph can be encoded in a matrix of zeroes and ones, and a network with real numbers.

MDS aims at embedding $\mathcal{G}$ into the plane, typically for visualization, while preserving the original distances. Basic results in graph/network theory suggest that the geometry of a graph cannot be preserved when embedding it into lower dimensions (Graham, 1988). The different types of MDSs, such as *Classical MDS*, and *Sammon Mappings*, differ in the *stress function* that penalizes for the geometric distortion caused by the embedding.

11.1.4.2 Local Multidimensional Scaling (Local MDS)

Example 11.5. Consider data of coordinates on the globe. At short distances, constructing a dissimilarity graph with Euclidean distances will capture the true distance between points. At long distances, however, the Euclidean distances as grossly inappropriate. A more extreme example is coordinates on the brain's cerebral cortex. Being a highly folded surface, the Euclidean distance between points is far from the true geodesic distances along the cortex's surface[4].

Local MDS is aimed at solving the case where Euclidean distances are inappropriate. Instead of using the graph of Euclidean distances between any two points, $\mathcal{G} = X'X$, local MDS computes $\mathcal{G}$ starting with the Euclidean distance between pairs of nearest points. Longer distances are solved as a shortest path problem. This is akin to computing the distance between Jerusalem to Beijing by computing Euclidean distances between Jerusalem-Bagdad, Bagdad-Teheran, Teheran-Ashgabat, Ashgabat-Tashkent,and so on. Because the geographical-distance between these cities is well approximated with the Euclidean distance, summing local distanes is better than operating with the Euclidean distance between Jerusalem and Beijing.

After computing $\mathcal{G}$, local MDS ends with the usual MDS for the embedding. Because local MDS ends with a regular MDS, it can be seen as a non-linear embedding into a linear manifold $\mathcal{M}$.

11.1.4.3 Isometric Feature Mapping (IsoMap)

Like localMDS, only that the embedding, and not only the computation of the distances, is local.

11.1.4.4 Local Linear Embedding (LLE)

Very similar to IsoMap 11.1.4.3.

11.1.4.5 t-SNE

t-SNE is a recently popularized visualization method for high dimentional data. t-SNE starts by computing a proximity graph, $\mathcal{G}$. Computation of distances in the graph assumes a Gaussian decay of distances. Put differently: only the nearest observations have a non-trivial similarity. This stage is similar (in spirit) to the growing of $\mathcal{G}$ in 11.1.4.2.

The second stage in t-SNE consists of finding a mapping to 2D (or 3D), which conserves distances in $\mathcal{G}$. The uniquness of t-SNE copared to other space embeddings is in the way distances are computed in the target 2D (or 3D) space.

11.1.4.6 Force Directed Graph Drawing

This class of algorithms star with a proximty graph $\mathcal{G}$, to define a set of phisically motivated "forces", operating between data-points. Think of $\mathcal{G}$ as governing a set of springs between data points. These springs have some steady-state: this is what you see in the visualization.

[4]Then again, it is possible that the true distances are the white matter fibers connecting going within the cortex, in which case, Euclidean distances are more appropriate than geodesic distances. We put that aside for now.

11.1.4.7 Kernel PCA (kPCA)

Returning to the BMI example (11.1); what if we want to learn scores that best separate between individuals, but unlike PCA, are non-linear in the original features. Kernel PCA does just that, only that it restricts the possible scores to simple functions of the original variables. The allowed functions resides in a function space called Reproducing Kernel Hilbert Space (RKHS), this giving kPCA it's name.

11.1.4.8 Sparse PCA (sPCA)

sPCA is a type of PCA where the loadings are sparse. This means that PCs are linear combinations of a small number of variables. This makes sPCA easier to interpret. Note that the *varimax* rotation of factor-analysis has a similar goal: create factors with minimally contributing variables, so that they are easy to explain.

11.1.4.9 Sparse kernel PCA (skPCA)

A marriage between sPCA and kPCA: generate scores that are non linear transformations of a minimal number of variables, each.

11.1.4.10 Correspondence Analysis (CA)

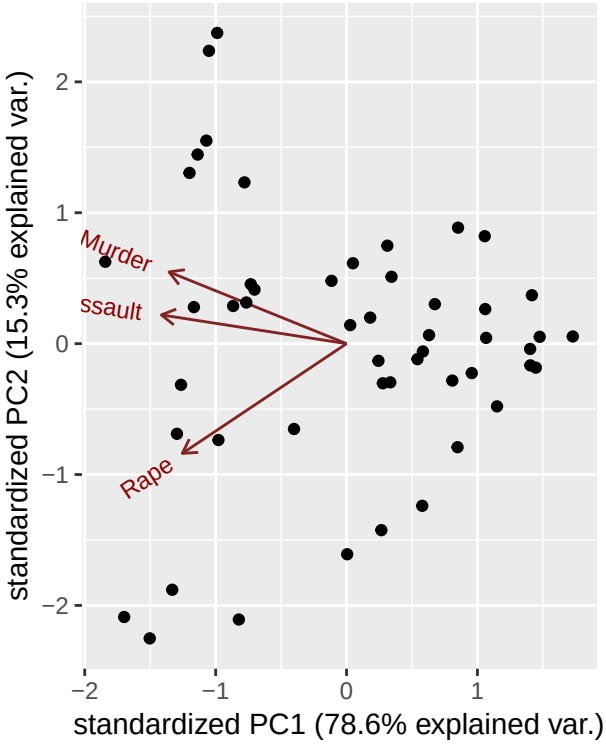
What if x is not continuous, i.e., $\mathcal{X} \neq \mathbb{R}^p$? We could dummy-code x, and then use plain PCA. A more principled view, when x is categorical, is that of Correspondence Analysis.

11.1.5 Dimensionality Reduction in R

11.1.5.1 PCA

We already saw the basics of PCA in 11.1.1. The fitting is done with the `prcomp` function. The *bi-plot* is a useful way to visualize the output of PCA.

```
# library(devtools)
# install_github("vqv/ggbiplot")
ggbiplot::ggbiplot(pca.1)
```

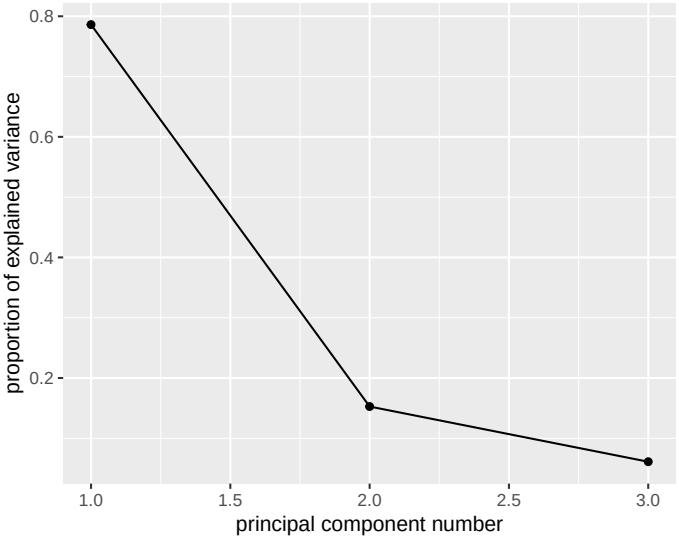


Things to note:

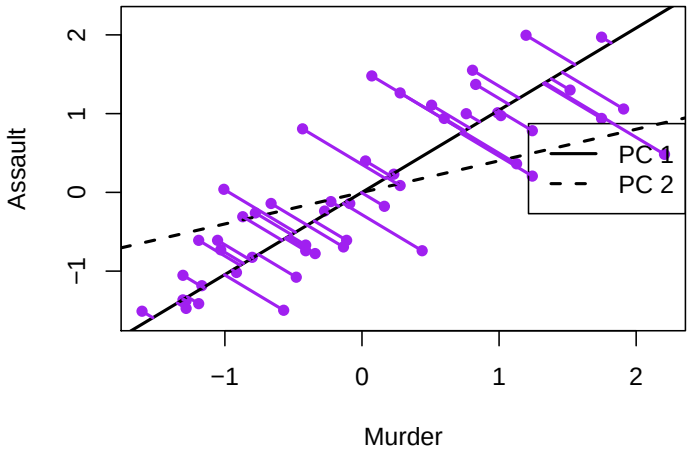
- We used the `ggbiplot` function from the **ggbiplot** (available from github, but not from CRAN), because it has a nicer output than `stats::biplot`.
- The bi-plot also plots the loadings as arrows. The coordinates of the arrows belong to the weight of each of the original variables in each PC. For example, the x-value of each arrow is the loadings on the first PC (on the x-axis). Since the weights of Murder, Assault, and Rape are almost the same, we conclude that PC1 captures the average crime rate in each state.
- The bi-plot plots each data point along its PCs.

The *scree plot* depicts the quality of the approximation of X as q grows, i.e., as we add more and more PCs to explain the data. This is depicted using the proportion of variability in X that is removed by each added PC. It is customary to choose q as the first PC that has a relative low contribution to the approximation of X. This is known as the "knee heuristic".

```
ggbiplot::ggscreeplot(pca.1)
```



See how the first PC captures the variability in the Assault levels and Murder levels, with a single score.



More implementations of PCA:

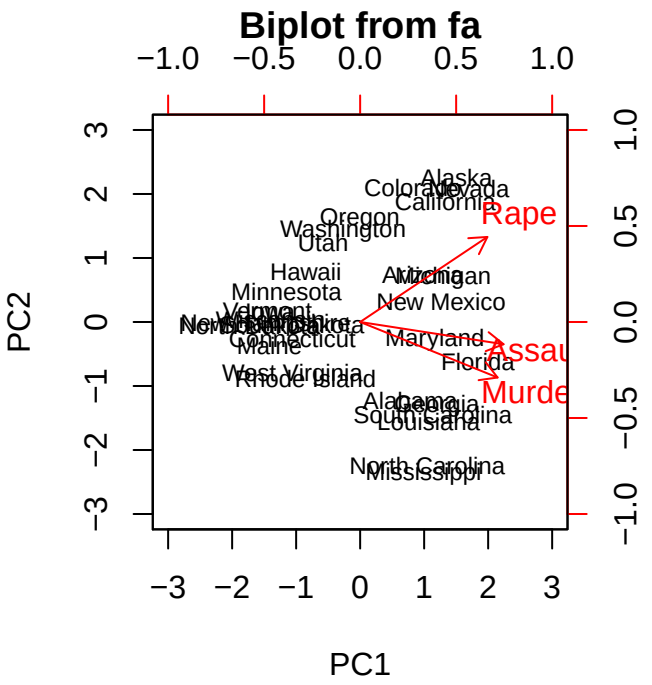
```
# FAST solutions:
gmodels::fast.prcomp()

# More detail in output:
FactoMineR::PCA()

# For flexibility in algorithms and visualization:
ade4::dudi.pca()

# Another one...
amap::acp()
```

11.1.5.2 FA

```
fa.1 <- psych::principal(USArrests.1, nfactors = 2, rotate = "none")
fa.1
```

```
## Principal Components Analysis
## Call: psych::principal(r = USArrests.1, nfactors = 2, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##          PC1   PC2   h2     u2 com
## Murder  0.89 -0.36 0.93 0.0688 1.3
## Assault 0.93 -0.14 0.89 0.1072 1.0
## Rape    0.83  0.55 0.99 0.0073 1.7
##
##                        PC1  PC2
## SS loadings           2.36 0.46
## Proportion Var        0.79 0.15
## Cumulative Var        0.79 0.94
## Proportion Explained  0.84 0.16
## Cumulative Proportion 0.84 1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.05
##  with the empirical chi square  0.87  with prob <  NA
##
## Fit based upon off diagonal values = 0.99
```

```
biplot(fa.1, labels =  rownames(USArrests.1))
```




```
# Numeric comparison with PCA:
fa.1$loadings
```

```
##
## Loadings:
##         PC1    PC2
## Murder   0.895 -0.361
## Assault  0.934 -0.145
## Rape     0.828  0.554
##
##                  PC1   PC2
## SS loadings    2.359 0.458
## Proportion Var 0.786 0.153
## Cumulative Var 0.786 0.939
```

```
pca.1$rotation
```

```
##                PC1        PC2        PC3
## Murder  -0.5826006  0.5339532 -0.6127565
## Assault -0.6079818  0.2140236  0.7645600
## Rape    -0.5393836 -0.8179779 -0.1999436
```

Things to note:

- We perform FA with the `psych::principal` function. The `Principal Component Analysis` title is due to the fact that FA without rotations, is equivalent to PCA.
- The first factor (`fa.1$loadings`) has different weights than the first PC (`pca.1$rotation`) because of normalization. They are the same, however, in that the first PC, and the first factor, capture average crime levels.

Graphical model fans will like the following plot, where the contribution of each variable to each factor is encoded in the width of the arrow.

```
qgraph::qgraph(fa.1)
```

Let's add a rotation (Varimax), and note that the rotation has indeed changed the loadings of the variables, thus the interpretation of the factors.

```
fa.2 <- psych::principal(USArrests.1, nfactors = 2, rotate = "varimax")

fa.2$loadings
```

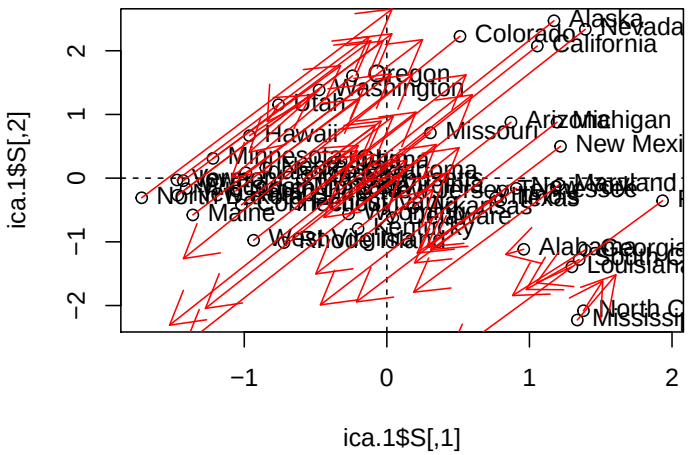
```
##
## Loadings:
##          RC1   RC2
## Murder  0.930 0.257
## Assault 0.829 0.453
## Rape    0.321 0.943
##
##                  RC1   RC2
## SS loadings    1.656 1.160
## Proportion Var 0.552 0.387
## Cumulative Var 0.552 0.939
```

Things to note:

- FA with a rotation is no longer equivalent to PCA.
- The rotated factors are now called *rotated componentes*, and reported in `RC1` and `RC2`.

11.1.5.3 ICA

```
ica.1 <- fastICA::fastICA(USArrests.1, n.com=2) # Also performs projection pursuit

plot(ica.1$S)
abline(h=0, v=0, lty=2)
text(ica.1$S, pos = 4, labels = rownames(USArrests.1))

# Compare with PCA (first two PCs):
arrows(x0 = ica.1$S[,1], y0 = ica.1$S[,2], x1 = pca.1$x[,2], y1 = pca.1$x[,1], col='red', pch=19, cex=0.5)
```



Things to note:

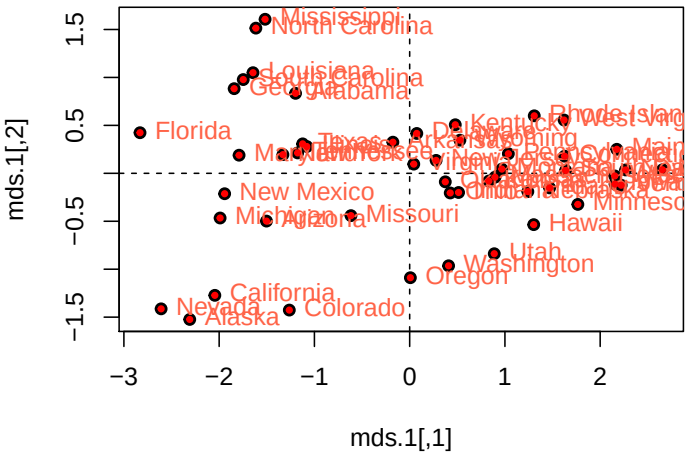
- ICA is fitted with `fastICA::fastICA`.
- The ICA components, like any other rotated components, are different than the PCA components.

11.1.5.4 MDS

Classical MDS, also compared with PCA.

```
# We first need a dissimarity matrix/graph:
state.disimilarity <- dist(USArrests.1)
```

```
mds.1 <- cmdscale(state.disimilarity)

plot(mds.1, pch = 19)
abline(h=0, v=0, lty=2)
USArrests.2 <- USArrests[,1:2] %>%  scale
text(mds.1, pos = 4, labels = rownames(USArrests.2), col = 'tomato')

# Compare with PCA (first two PCs):
points(pca.1$x[,1:2], col='red', pch=19, cex=0.5)
```

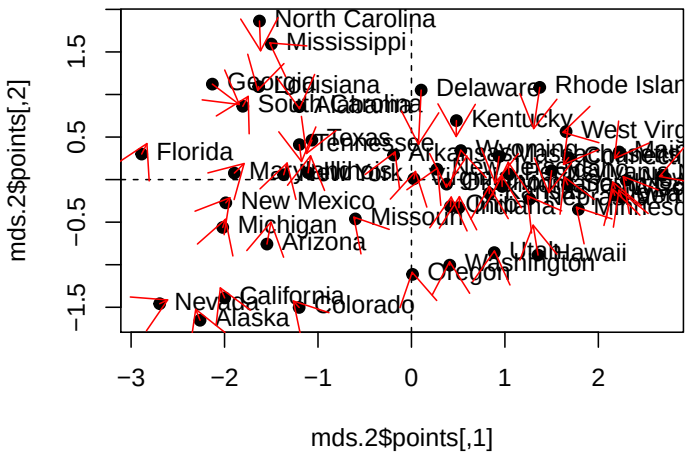


Things to note:

- We first compute a dissimilarity graph with `dist`. See the `cluster::daisy` function for more dissimilarity measures.
- We learn the MDS embedding with `cmdscale`.
- The embedding of PCA is the same as classical MDS with Euclidean distances.

Let's try other strain functions for MDS, like Sammon's strain, and compare it with the PCs.

```
mds.2 <- MASS::sammon(state.disimilarity, trace = FALSE)
plot(mds.2$points, pch = 19)
abline(h=0, v=0, lty=2)
text(mds.2$points, pos = 4, labels = rownames(USArrests.2))

# Compare with PCA (first two PCs):
arrows(
  x0 = mds.2$points[,1], y0 = mds.2$points[,2],
  x1 = pca.1$x[,1], y1 = pca.1$x[,2],
  col='red', pch=19, cex=0.5)
```



Things to note:

- `MASS::sammon` does the embedding.
- Sammon strain is different than PCA.

11.1.5.5 t-SNE

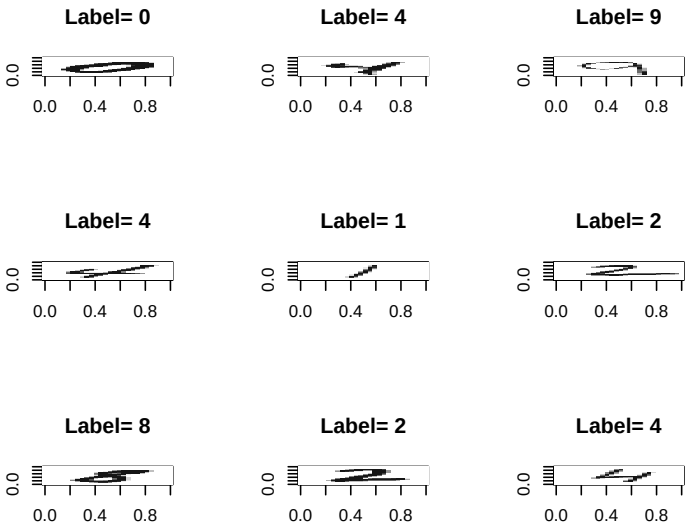
For a native R implementation: tsne package. For an R wrapper for C libraries: Rtsne package.

The MNIST image bank of hand-written images has its own data format. The import process is adapted from David Dalpiaz:

```
show_digit <- function(arr784, col = gray(12:1 / 12), ...) {
  image(matrix(as.matrix(arr784[-785]), nrow = 28)[, 28:1], col = col, ...)
}

# load image files
load_image_file <- function(filename) {
  ret <- list()
  f <- file(filename, 'rb')
  readBin(f, 'integer', n = 1, size = 4, endian = 'big')
  n    <- readBin(f, 'integer', n = 1, size = 4, endian = 'big')
  nrow <- readBin(f, 'integer', n = 1, size = 4, endian = 'big')
  ncol <- readBin(f, 'integer', n = 1, size = 4, endian = 'big')
  x <- readBin(f, 'integer', n = n * nrow * ncol, size = 1, signed = FALSE)
  close(f)
  data.frame(matrix(x, ncol = nrow * ncol, byrow = TRUE))
}

# load label files
load_label_file <- function(filename) {
  f <- file(filename, 'rb')
  readBin(f, 'integer', n = 1, size = 4, endian = 'big')
  n <- readBin(f, 'integer', n = 1, size = 4, endian = 'big')
  y <- readBin(f, 'integer', n = n, size = 1, signed = FALSE)
  close(f)
  y
}

# load images
train <- load_image_file("data/train-images-idx3-ubyte")
test  <- load_image_file("data/t10k-images-idx3-ubyte")

# load labels
train$y = as.factor(load_label_file("data/train-labels-idx1-ubyte"))
test$y  = as.factor(load_label_file("data/t10k-labels-idx1-ubyte"))
```

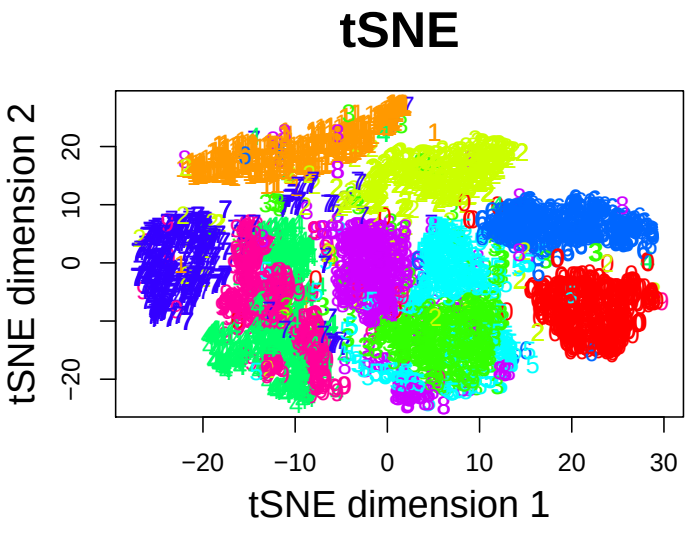
Inspect some digits:

```
par(mfrow=c(3,3))
ind <- sample(1:nrow(train),9)
for(i in 1:9){
  show_digit(train[ind[i],], main=paste('Label= ',train$y[ind[i]], sep=''))  }
```



The analysis is adapted from Shruti Marwaha. I start by dowloading and importing the famous digit data of .

```
numTrain <- 5e3 # Subset data for speed
rows <- sample(1:nrow(train), numTrain)
train.sub <- train[rows, -which(names(train)=='y')] %>% as.matrix
train.sub.labs <- train[rows, which(names(train)=='y')]

tsne <- Rtsne::Rtsne(train.sub, dims = 2, perplexity=30, verbose=FALSE, max_iter = 500)

colors <- rainbow(length(unique(train.sub.labs)))
names(colors) <- unique(train.sub.labs)
par(mgp=c(2.5,1,0))
par(mfrow=c(1,1))
plot(tsne$Y, t='n',
     main="tSNE",
     xlab="tSNE dimension 1",
     ylab="tSNE dimension 2",
     "cex.main"=2,
     "cex.lab"=1.5)
text(tsne$Y, labels=train.sub.labs, col=colors[train.sub.labs])
```



11.1.5.6 Force Embedding

I am unaware of an R implementation of force-embedding. Maybe because of the interactive nature of the algorithm, that is not suited for R. Force embedding is much more natural to interactive GUIs. Here is a fun javascript implementation.

11.1.5.7 Sparse PCA

```
# Compute similarity graph
state.similarity <- MASS::cov.rob(USArrests.1)$cov

spca1 <- elasticnet::spca(state.similarity, K=2, type="Gram", sparse="penalty", trace=FALSE, para=c(0.06,0.1
spca1$loadings
```

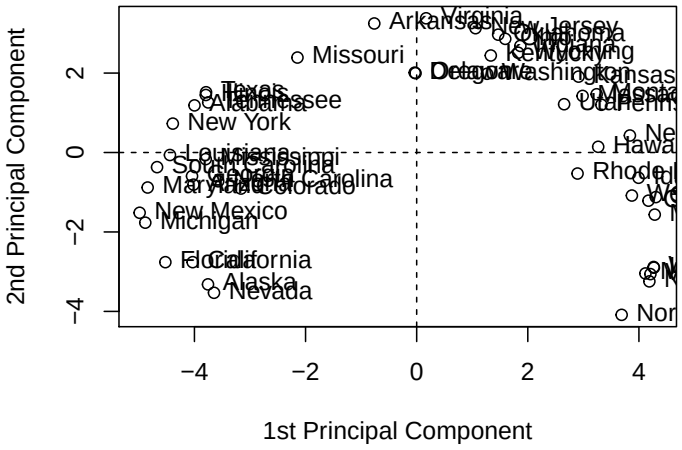
```
##                PC1 PC2
## Murder  -0.1626431   1
## Assault -0.8200474   0
## Rape    -0.5486979   0
```

Things to note:

- I used the `spca` function in the **elasticnet** package.
- Is the solutions sparse? Yes! PC2 depends on a single variable only: `Murder`.

11.1.5.8 Kernel PCA

```
library(kernlab)
kpc <- kpca(~.,data=as.data.frame(USArrests.1), kernel="rbfdot", kpar=list(sigma=0.2), features=2)

plot(rotated(kpc),
     xlab="1st Principal Component",
     ylab="2nd Principal Component")

abline(h=0, v=0, lty=2)
text(rotated(kpc), pos = 4, labels = rownames(USArrests.2))
```



Things to note:

- kPCA is implemented in `kernlab::kpca`. `kernel=` governs the class of variable mappings allowed for the scoring.
- See `?'kpca-class'` or `?rotated` for help on available utility functions.
- `rotated` projects the data on its principal components (the above "scores").
- `kpar=list(sigma=0.2)` provides parameters specific to each type of kernel. See `?kpca`.
- `features=2` is the number of principal components (scores) to learn.

- You may notice the "Horseshoe" pattern of the kPCA embedding, a.k.a. "Croissants", or the *Guttman Effect*. The horseshoe is a recurring and old phenomenon. This phenomenon implies that the data has "structure". Alas, it is possible that this structure is merely due to the sampling scheme that introduced correlations. See J. De Leeuw's online paper for a review.

11.1.5.9 MCA

See Izenman (2008).

11.2 Clustering

Example 11.6. Consider the tagging of your friends' pictures on Facebook. If you tagged some pictures, Facebook may try to use a supervised approach to automatically label photos. If you never tagged pictures, a supervised approach is impossible. It is still possible, however, to group simiar pictures together.

Example 11.7. Consider the problem of spam detection. It would be nice if each user could label several thousands emails, to apply a supervised learning approach to spam detection. This is an unrealistic demand, so a pre-clustering stage is useful: the user only needs to tag a couple dozens of homogenous clusters, before solving the supervised learning problem.

In clustering problems, we seek to group observations that are similar.

There are many motivations for clustering:

1. **Understanding**: The most common use of clustering is probably as a an exploratory step, to identify homogeneous groups in the data.
2. **Dimensionality reduction**: Clustering may be seen as a method for dimensionality reduction. Unlike the approaches in the Dimensionality Reduction Section 11.1, it does not compress **variables** but rather **observations**. Each group of homogeneous observations may then be represented as a single prototypical observation of the group.
3. **Pre-Labelling**: Clustering may be performed as a pre-processing step for supervised learning, when labeling all the samples is impossible due to "budget" constraints, like in Example 11.7. This is sometimes known as *pre-clustering*.

Clustering, like dimensionality reduction, may rely on some latent variable generative model, or on purely algorithmic approaches.

11.2.1 Latent Variable Generative Approaches

11.2.1.1 Finite Mixture

Example 11.8. Consider the distribution of heights. Heights have a nice bell shaped distribution within each gender. If genders have not been recorded, heights will be distributed like a *mixture* of males and females. The gender in this example, is a *latent* variable taking $K = 2$ levels: male and female.

A *finite mixture* is the marginal distribution of K distinct classes, when the class variable is *latent*. This is useful for clustering: We can assume the number of classes, K, and the distribution of each class. We then use maximum likelihood to fit the mixture distribution, and finally, cluster by assigning observations to the most probable class.

11.2.2 Purely Algorithmic Approaches

11.2.2.1 K-Means

The *K-means* algorithm is possibly the most popular clustering algorithm. The goal behind K-means clustering is finding a representative point for each of K clusters, and assign each data point to one of these clusters. As each

cluster has a representative point, this is also a *prototype method*. The clusters are defined so that they minimize the average Euclidean distance between all points to the center of the cluster.

In K-means, the clusters are first defined, and then similarities computed. This is thus a *top-down* method.

K-means clustering requires the raw features X as inputs, and not only a similarity graph. This is evident when examining the algorithm below.

The k-means algorithm works as follows:

1. Choose the number of clusters K.
2. Arbitrarily assign points to clusters.
3. While clusters keep changing:
    1. Compute the cluster centers as the average of their points.
    2. Assign each point to its closest cluster center (in Euclidean distance).
4. Return Cluster assignments and means.

Remark. If trained as a statistician, you may wonder- what population quantity is K-means actually estimating? The estimand of K-means is known as the *K principal points*. Principal points are points which are *self consistent*, i.e., they are the mean of their neighbourhood.

11.2.2.2 K-Means++

K-means++ is a fast version of K-means thanks to a smart initialization.

11.2.2.3 K-Medoids

If a Euclidean distance is inappropriate for a particular set of variables, or that robustness to corrupt observations is required, or that we wish to constrain the cluster centers to be actual observations, then the *K-Medoids* algorithm is an adaptation of K-means that allows this. It is also known under the name *partition around medoids* (PAM) clustering, suggesting its relation to graph partitioning.

The k-medoids algorithm works as follows.

1. Given a dissimilarity graph.
2. Choose the number of clusters K.
3. Arbitrarily assign points to clusters.
4. While clusters keep changing:
    1. Within each cluster, set the center as the data point that minimizes the sum of distances to other points in the cluster.
    2. Assign each point to its closest cluster center.
5. Return Cluster assignments and centers.

Remark. If trained as a statistician, you may wonder- what population quantity is K-medoids actually estimating? The estimand of K-medoids is the median of their neighbourhood. A delicate matter is that quantiles are not easy to define for **multivariate** variables so that the "multivaraitre median", may be a more subtle quantity than you may think. See Small (1990).

11.2.2.4 Hirarchial Clustering

Hierarchical clustering algorithms take dissimilarity graphs as inputs. Hierarchical clustering is a class of greedy *graph-partitioning* algorithms. Being hierarchical by design, they have the attractive property that the evolution of the clustering can be presented with a *dendogram*, i.e., a tree plot.
A particular advantage of these methods is that they do not require an a-priori choice of the number of cluster (K).

Two main sub-classes of algorithms are *agglomerative*, and *divisive*.

Agglomerative clustering algorithms are **bottom-up** algorithm which build clusters by joining smaller clusters. To decide which clusters are joined at each iteration some measure of closeness between clusters is required.

- **Single Linkage**: Cluster distance is defined by the distance between the two **closest** members.
- **Complete Linkage**: Cluster distance is defined by the distance between the two **farthest** members.

- **Group Average**: Cluster distance is defined by the **average** distance between members.
- **Group Median**: Like Group Average, only using the median.

Divisive clustering algorithms are **top-down** algorithm which build clusters by splitting larger clusters.

11.2.2.5 Fuzzy Clustering

Can be thought of as a purely algorithmic view of the finite-mixture in Section 11.2.1.1.

11.2.3 Clustering in R

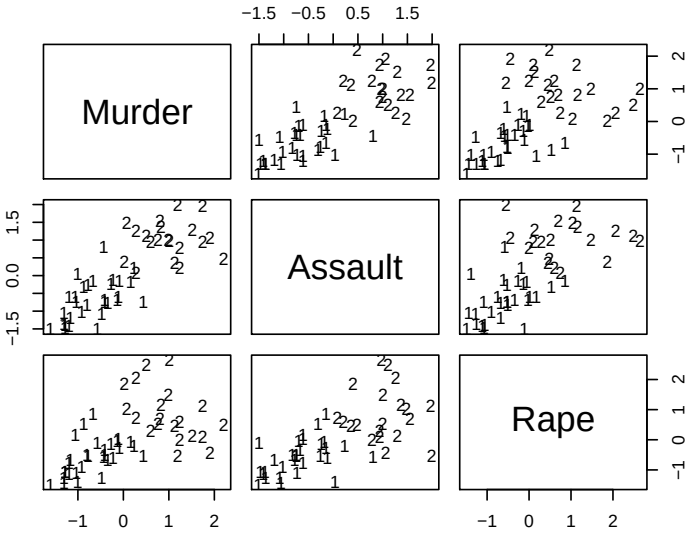
11.2.3.1 K-Means

The following code is an adaptation from David Hitchcock.

```
k <- 2
kmeans.1 <- stats::kmeans(USArrests.1, centers = k)
head(kmeans.1$cluster) # cluster asignments
```

```
##    Alabama     Alaska    Arizona   Arkansas California   Colorado
##          2          2          2          1          2          2
```

```
pairs(USArrests.1, panel=function(x,y) text(x,y,kmeans.1$cluster))
```



Things to note:

- The `stats::kmeans` function does the clustering.
- The cluster assignment is given in the `cluster` element of the `stats::kmeans` output.
- The visual inspection confirms that similar states have been assigned to the same cluster.

11.2.3.2 K-Means ++

K-Means++ is a smart initialization for K-Means. The following code is taken from the r-help mailing list.

```
# Write my own K-means++ function.
kmpp <- function(X, k) {

  n <- nrow(X)
  C <- numeric(k)
  C[1] <- sample(1:n, 1)
```

```
  for (i in 2:k) {
    dm <- pracma::distmat(X, X[C, ])
    pr <- apply(dm, 1, min); pr[C] <- 0
    C[i] <- sample(1:n, 1, prob = pr)
  }

  kmeans(X, X[C, ])
}

kmeans.2 <- kmpp(USArrests.1, k)
head(kmeans.2$cluster)
```

```
##    Alabama     Alaska    Arizona   Arkansas California   Colorado
##          2          2          2          1          2          2
```

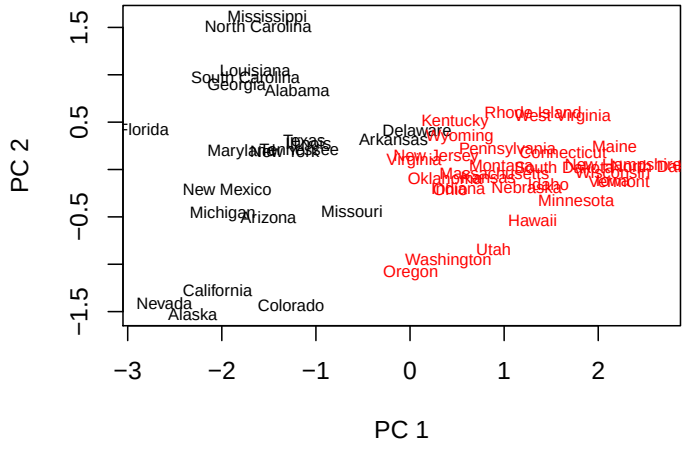
11.2.3.3 K-Medoids

Start by growing a distance graph with `dist` and then partition using `pam`.

```
state.disimilarity <- dist(USArrests.1)
kmed.1 <- cluster::pam(x= state.disimilarity, k=2)
head(kmed.1$clustering)
```

```
##    Alabama     Alaska    Arizona   Arkansas California   Colorado
##          1          1          1          1          1          1
```

```
plot(pca.1$x[,1], pca.1$x[,2], xlab="PC 1", ylab="PC 2", type ='n', lwd=2)
text(pca.1$x[,1], pca.1$x[,2], labels=rownames(USArrests.1), cex=0.7, lwd=2, col=kmed.1$cluster)
```



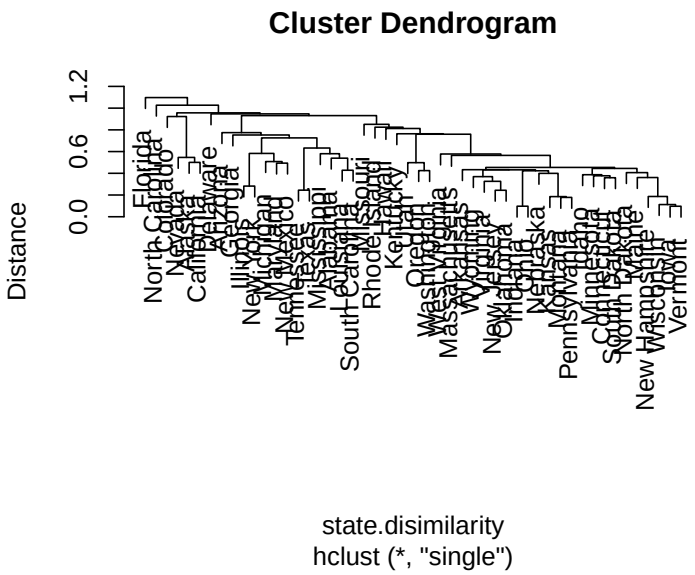
Things to note:

- K-medoids starts with the computation of a dissimilarity graph, done by the `dist` function.
- The clustering is done by the `cluster::pam` function.
- Inspecting the output confirms that similar states have been assigned to the same cluster.
- Many other similarity measures can be found in `proxy::dist()`.
- See `cluster::clara()` for a big-data implementation of PAM.

11.2.3.4 Hirarchial Clustering

We start with agglomerative clustering with single-linkage.

```
hirar.1 <- hclust(state.disimilarity, method='single')
plot(hirar.1, labels=rownames(USArrests.1), ylab="Distance")
```

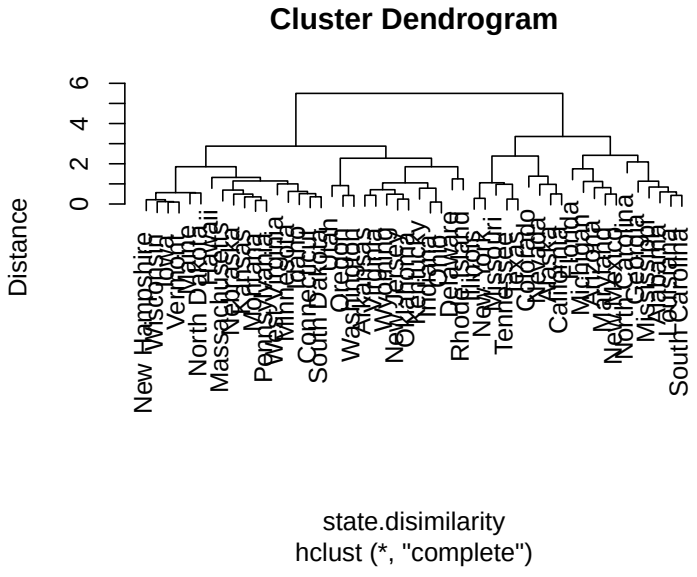


Things to note:

- The clustering is done with the `hclust` function.
- We choose the single-linkage distance using the `method='single'` argument.
- We did not need to a-priori specify the number of clusters, K, since all the possible K's are included in the output tree.
- The `plot` function has a particular method for `hclust` class objects, and plots them as dendograms.

We try other types of linkages, to verify that the indeed affect the clustering. Starting with complete linkage.

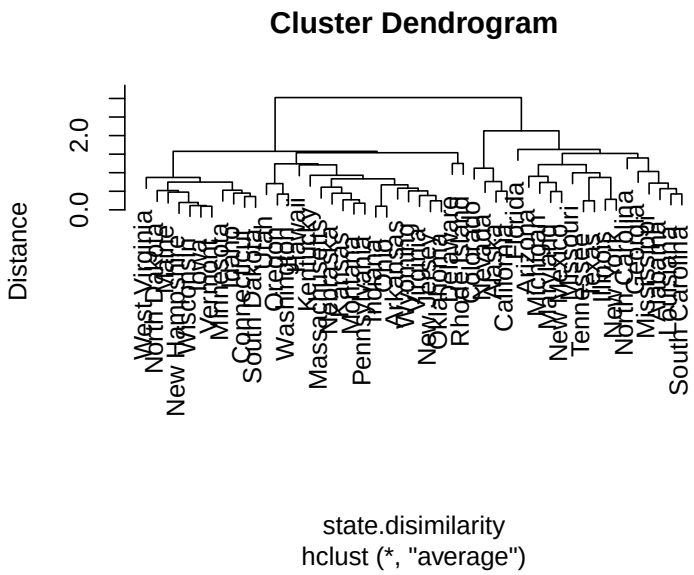
```
hirar.2 <- hclust(state.disimilarity, method='complete')
plot(hirar.2, labels=rownames(USArrests.1), ylab="Distance")
```



linkage-1.bb

Now with average linkage.

```
hirar.3 <- hclust(state.disimilarity, method='average')
plot(hirar.3, labels=rownames(USArrests.1), ylab="Distance")
```



linkage-1.bb

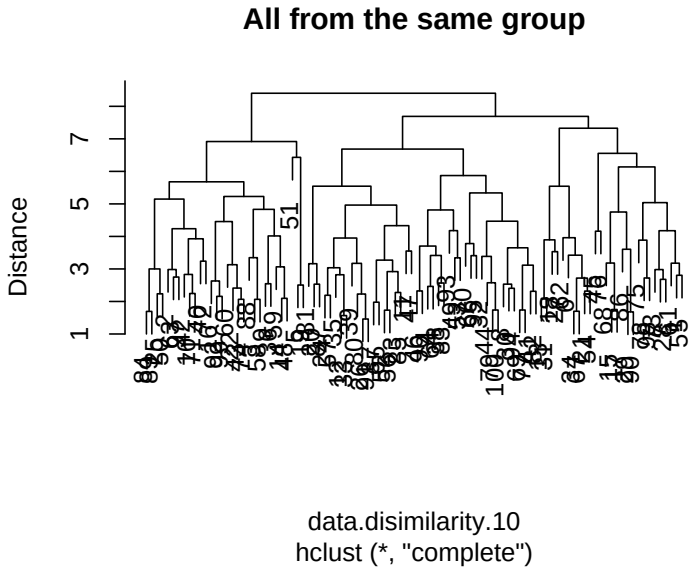
If we know how many clusters we want, we can use `cuttree` to get the class assignments.

```
cut.2.2 <- cutree(hirar.2, k=2)
head(cut.2.2)
```

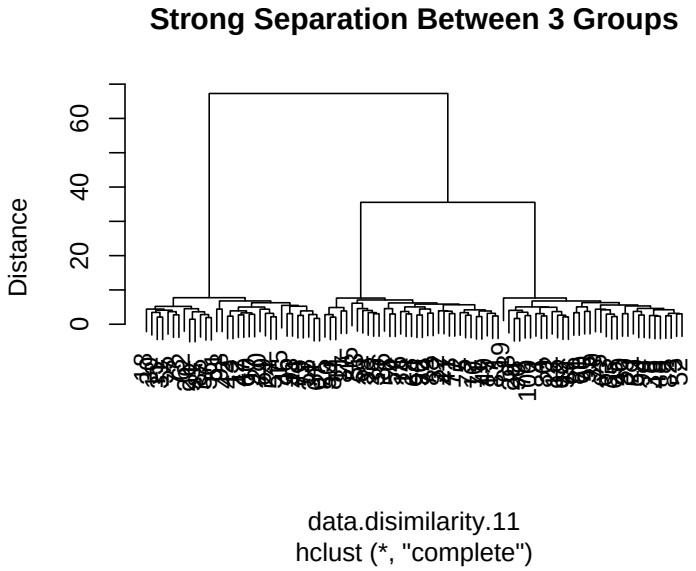
```
##    Alabama     Alaska    Arizona   Arkansas California   Colorado
##          1          1          1          2          1          1
```

How to choose the number of clusters? Just like the Scree Plot above, we can use a "knee heuristic". Because the length of a tree's branch is proportional to distances, long branches imply inhomogenous groups, while short branches imply homogeneous groups. Here is a little simulation to demonstrate this:

```
n.groups <- 3
data.p <- 10
data.n <- 100

# data with no separation between groups
the.data.10 <- mvtnorm::rmvnorm(n = data.n, mean = rep(0,data.p))
data.disimilarity.10 <- dist(the.data.10)
hirar.10 <- hclust(data.disimilarity.10, method = "complete")
plot(hirar.10, ylab="Distance", main='All from the same group')
```



```
# data with strong separation between groups
the.data.11 <-the.data.10 +sample(c(0,10,20), data.n, replace=TRUE) # Shift each group
data.disimilarity.11 <- dist(the.data.11)
hirar.11 <- hclust(data.disimilarity.11, method = "complete")
plot(hirar.11, ylab="Distance", main=paste('Strong Separation Between',n.groups, 'Groups'))
```



11.3 Bibliographic Notes

For some theory of PCA see my Dimensionality Reduction Class Notes and references therein. For a SUPERB, interactive, visual demonstration of dimensionality reduction, see Christopher Olah's blog. An excellent reference on multivariate analysis (exploratory and inference) is Izenman (2008). For t-SNE see the creator's site: Laurens van der Maaten. For an excellent book on kernel methods (RKHS) see Shawe-Taylor and Cristianini (2004). For more on everything, see Friedman et al. (2001). For a softer introduction (to everything), see James et al. (2013).

11.4 Practice Yourself

1. Generate data from multivariate Gaussian data with `mvtnorm::rmvnorm()`. Clearly this data has no structure at all: it is a p-dimensional shapeless cloud of n points.
    1. Now try various dimensionality reduction algorithms such as PCA, MDS, kPCA, sPCA. How does the sphere map to the plane? How does the mapping depend on n? And on p?
    2. Map the data to a p-dimensional unit sphere by dividing each observation with its l_2 norm: `map2sphere <- function(x) x/sqrt(sum(x^2))`. Repeat the previous embeddings. Does this structureless data embeds itself with structure?
    3. Introduce artificial "structure" in the data and repeat the previous exercise. Use the Factor Analysis generative model in Eq.((11.2)) to generate p dimensional data along a one-dimensional line. Can you see that observations arrange themselves along a single line in after your plane embedding?
2. Read about the Iris dataset using `?iris`. "Forget" the `Species` column to make the problem unsupervised.
    1. Make pairs of scatter plots. Can you identify the clusters in the data?
    2. Perform K-means with `centers=3`. To extract the clustering results (cluster of each instance) use `kmeans$clusters`. Now recall the `Species` column to verify your clustering.
    3. Perform hierarchical clustering with `hclust`, `method="single"` and `method="average"`.Extract the clustering results with `cutree`. Compare the accuracy of the two linkage methods.
    4. Perform PCA on the data with `prcomp` function.
    5. Print the Rotation matrix.

6. Print the PCA's vectors with `pca$x`. These vectors are the new values for each instance in the dataset after the rotation.
7. Let's look at the first component (PC1) with `plot(pca$x[,1])` (i.e reduce the dimensionality from 4 to 1 features). Can you identify visually the three clusters (species)?

8. Determine the color of the points to be the truth species with `col=iris$Species`.

See DataCap's Unsupervised Learning in R, Cluster Analysis in R, Dimensionality Reduction in R, and Advanced Dimensionality Reduction in R for more self practice.

Chapter 12

Plotting

Whether you are doing EDA, or preparing your results for publication, you need plots. R has many plotting mechanisms, allowing the user a tremendous amount of flexibility, while abstracting away a lot of the tedious details. To be concrete, many of the plots in R are simply impossible to produce with Excel, SPSS, or SAS, and would take a tremendous amount of work to produce with Python, Java and lower level programming languages.

In this text, we will focus on two plotting packages. The basic **graphics** package, distributed with the base R distribution, and the **ggplot2** package.

Before going into the details of the plotting packages, we start with some philosophy. The **graphics** package originates from the mainframe days. Computers had no graphical interface, and the output of the plot was immediately sent to a printer. Once a plot has been produced with the **graphics** package, just like a printed output, it cannot be queried nor changed, except for further additions.

The philosophy of R is that **everyting is an object**. The **graphics** package does not adhere to this philosophy, and indeed it was soon augmented with the **grid** package (R Core Team, 2016), that treats plots as objects. **grid** is a low level graphics interface, and users may be more familiar with the **lattice** package built upon it (Sarkar, 2008).

lattice is very powerful, but soon enough, it was overtaken in popularity by the **ggplot2** package (Wickham, 2009). **ggplot2** was the PhD project of Hadley Wickham, a name to remember… Two fundamental ideas underlay **ggplot2**: (i) everything is an object, and (ii), plots can be described by a simple grammar, i.e., a language to describe the building blocks of the plot. The grammar in **ggplot2** are is the one stated by Wilkinson (2006). The objects and grammar of **ggplot2** have later evolved to allow more complicated plotting and in particular, interactive plotting.

Interactive plotting is a very important feature for EDA, and reporting. The major leap in interactive plotting was made possible by the advancement of web technologies, such as JavaScript and D3.JS. Why is this? Because an interactive plot, or report, can be seen as a web-site. Building upon the capabilities of JavaScript and your web browser to provide the interactivity, greatly facilitates the development of such plots, as the programmer can rely on the web-browsers capabilities for interactivity.

12.1 The graphics System

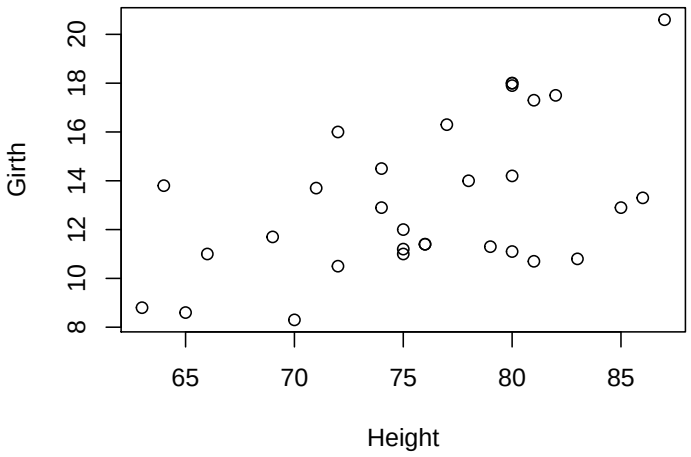
The R code from the Basics Chapter 3 is a demonstration of the **graphics** package and plotting system. We make a quick review of the basics.

12.1.1 Using Existing Plotting Functions

12.1.1.1 Scatter Plot

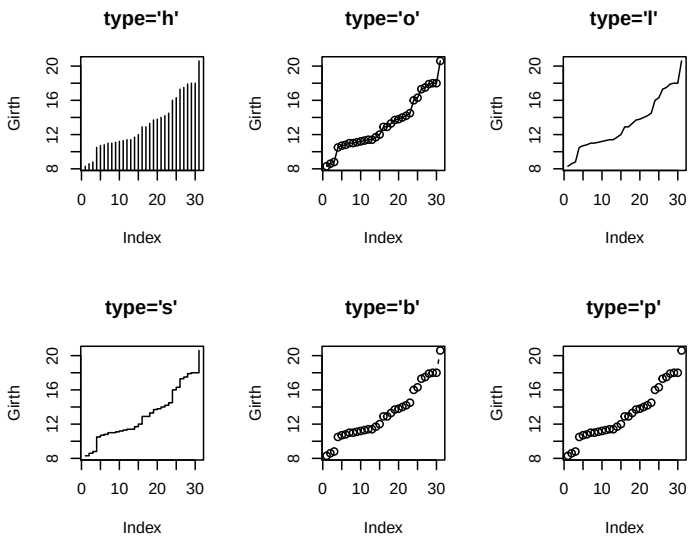
A simple scatter plot.

```
attach(trees)
plot(Girth ~ Height)
```



Various types of plots.

```
par.old <- par(no.readonly = TRUE)
par(mfrow=c(2,3))
plot(Girth, type='h', main="type='h'")
plot(Girth, type='o', main="type='o'")
plot(Girth, type='l', main="type='l'")
plot(Girth, type='s', main="type='s'")
plot(Girth, type='b', main="type='b'")
plot(Girth, type='p', main="type='p'")
```



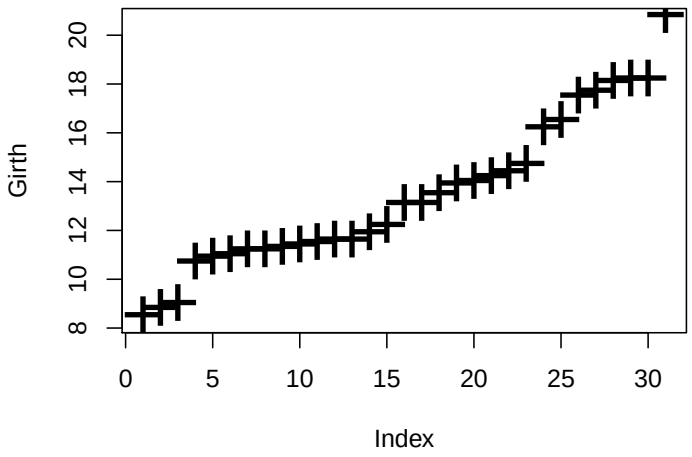

```
par(par.old)
```

Things to note:

- The `par` command controls the plotting parameters. `mfrow=c(2,3)` is used to produce a matrix of plots with 2 rows and 3 columns.
- The `par.old` object saves the original plotting setting. It is restored after plotting using `par(par.old)`.
- The `type` argument controls the type of plot.
- The `main` argument controls the title.
- See `?plot` and `?par` for more options.

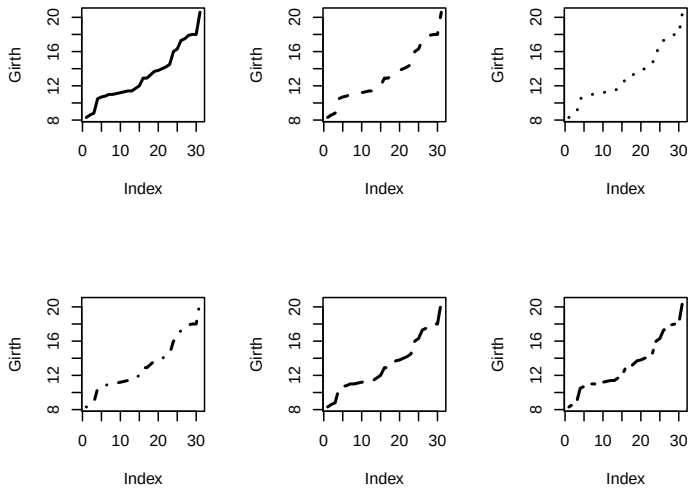
Control the plotting characters with the `pch` argument, and size with the `cex` argument.

```
plot(Girth, pch='+', cex=3)
```



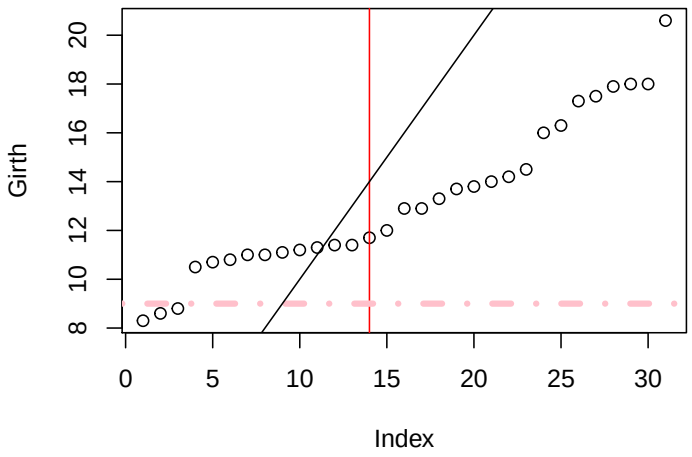
Control the line's type with `lty` argument, and width with `lwd`.

```
par(mfrow=c(2,3))
plot(Girth, type='l', lty=1, lwd=2)
plot(Girth, type='l', lty=2, lwd=2)
plot(Girth, type='l', lty=3, lwd=2)
plot(Girth, type='l', lty=4, lwd=2)
plot(Girth, type='l', lty=5, lwd=2)
plot(Girth, type='l', lty=6, lwd=2)
```

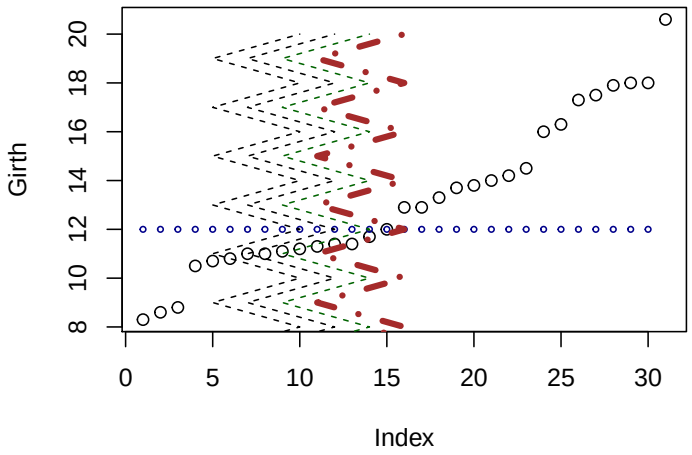


Add line by slope and intercept with `abline`.

```
plot(Girth)
abline(v=14, col='red') # vertical line at 14.
abline(h=9, lty=4,lwd=4, col='pink') # horizontal line at 9.
abline(a = 0, b=1) # linear line with intercept a=0, and slope b=1.
```




```
plot(Girth)
points(x=1:30, y=rep(12,30), cex=0.5, col='darkblue')
lines(x=rep(c(5,10), 7), y=7:20, lty=2 )
lines(x=rep(c(5,10), 7)+2, y=7:20, lty=2 )
lines(x=rep(c(5,10), 7)+4, y=7:20, lty=2 , col='darkgreen')
lines(x=rep(c(5,10), 7)+6, y=7:20, lty=4 , col='brown', lwd=4)
```



Things to note:

- `points` adds points on an existing plot.
- `lines` adds lines on an existing plot.
- `col` controls the color of the element. It takes names or numbers as argument.
- `cex` controls the scale of the element. Defaults to `cex=1`.

Add other elements.

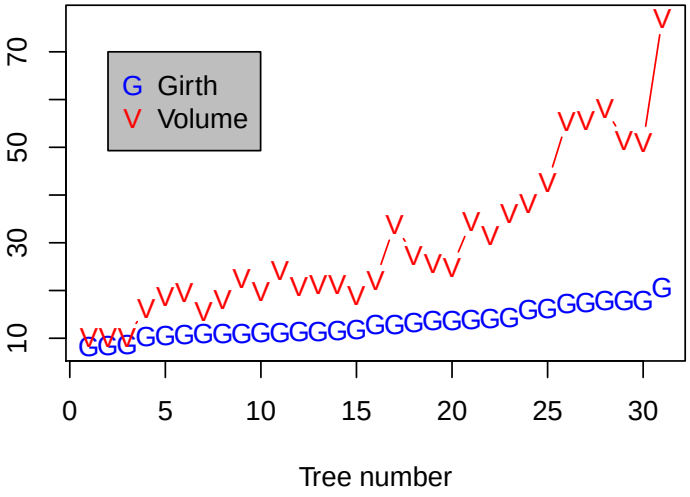
```
plot(Girth)
segments(x0=rep(c(5,10), 7), y0=7:20, x1=rep(c(5,10), 7)+2, y1=(7:20)+2 ) # line segments
arrows(x0=13,y0=16,x1=16,y1=17) # arrows
rect(xleft=10, ybottom=12,  xright=12, ytop=16) # rectangle
polygon(x=c(10,11,12,11.5,10.5), y=c(9,9.5,10,10.5,9.8), col='grey') # polygon
title(main='This plot makes no sense', sub='Or does it?')
mtext('Printing in the margins', side=2) # math text
mtext(expression(alpha==log(f[i])), side=4)
```

Things to note:

- The following functions add the elements they are named after: `segments`, `arrows`, `rect`, `polygon`, `title`.
- `mtext` adds mathematical text, which needs to be wrapped in `expression()`. For more information for mathematical annotation see `?plotmath`.

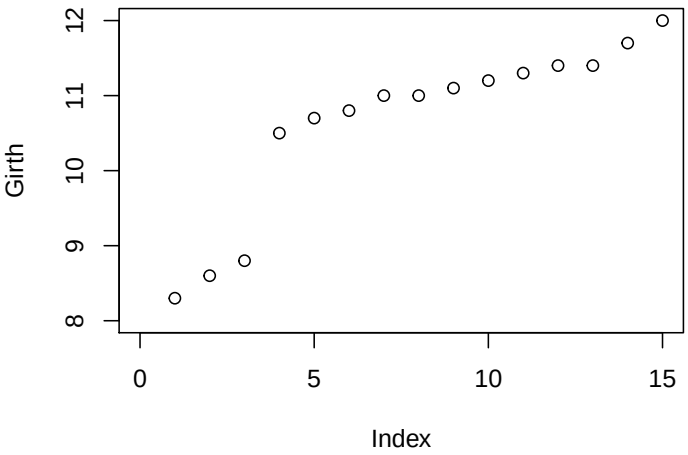
Add a legend.

```
plot(Girth, pch='G',ylim=c(8,77), xlab='Tree number', ylab='', type='b', col='blue')
points(Volume, pch='V', type='b', col='red')
legend(x=2, y=70, legend=c('Girth', 'Volume'), pch=c('G','V'), col=c('blue','red'), bg='grey')
```



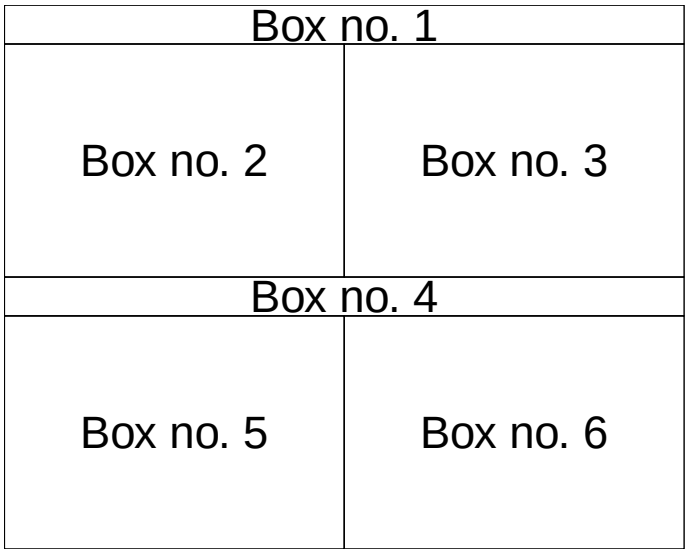
Adjusting Axes with `xlim` and `ylim`.

```
plot(Girth, xlim=c(0,15), ylim=c(8,12))
```



Use `layout` for complicated plot layouts.

```
A<-matrix(c(1,1,2,3,4,4,5,6), byrow=TRUE, ncol=2)
layout(A,heights=c(1/14,6/14,1/14,6/14))

oma.saved <- par("oma")
par(oma = rep.int(0, 4))
par(oma = oma.saved)
o.par <- par(mar = rep.int(0, 4))
for (i in seq_len(6)) {
    plot.new()
    box()
    text(0.5, 0.5, paste('Box no.',i), cex=3)
}
```



Always detach.

```
detach(trees)
```

12.1.2 Exporting a Plot

The pipeline for exporting graphics is similar to the export of data. Instead of the `write.table` or `save` functions, we will use the `pdf`, `tiff`, `png`, functions. Depending on the type of desired output.

Check and set the working directory.

```
getwd()
setwd("/tmp/")
```

Export tiff.

```
tiff(filename='graphicExample.tiff')
plot(rnorm(100))
dev.off()
```

Things to note:

- The `tiff` function tells R to open a .tiff file, and write the output of a plot.
- Only a single (the last) plot is saved.
- `dev.off` to close the tiff device, and return the plotting to the R console (or RStudio).

If you want to produce several plots, you can use a counter in the file's name. The counter uses the printf format string.

```
tiff(filename='graphicExample%d.tiff') #Creates a sequence of files
plot(rnorm(100))
boxplot(rnorm(100))
hist(rnorm(100))
dev.off()
```

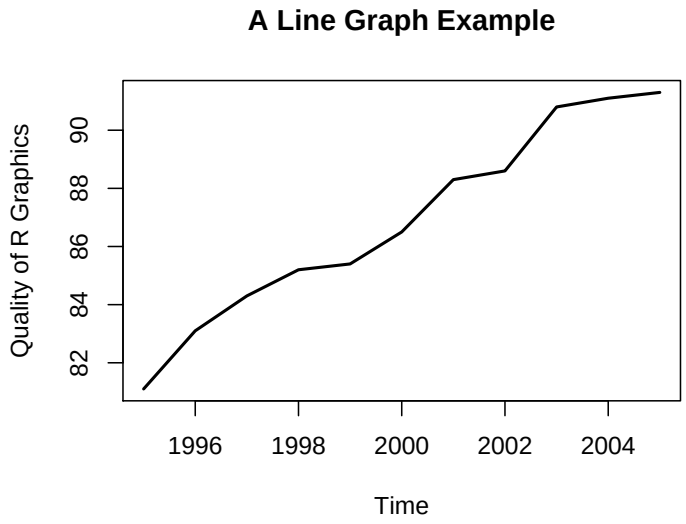
To see the list of all open devices use `dev.list()`. To close **all** device, (not only the last one), use `graphics.off()`.

See `?pdf` and `?jpeg` for more info.

12.1.3 Fancy graphics Examples

12.1.3.1 Line Graph

```
x = 1995:2005
y = c(81.1, 83.1, 84.3, 85.2, 85.4, 86.5, 88.3, 88.6, 90.8, 91.1, 91.3)
plot.new()
plot.window(xlim = range(x), ylim = range(y))
abline(h = -4:4, v = -4:4, col = "lightgrey")
lines(x, y, lwd = 2)
title(main = "A Line Graph Example",
        xlab = "Time",
        ylab = "Quality of R Graphics")
axis(1)
axis(2)
box()
```

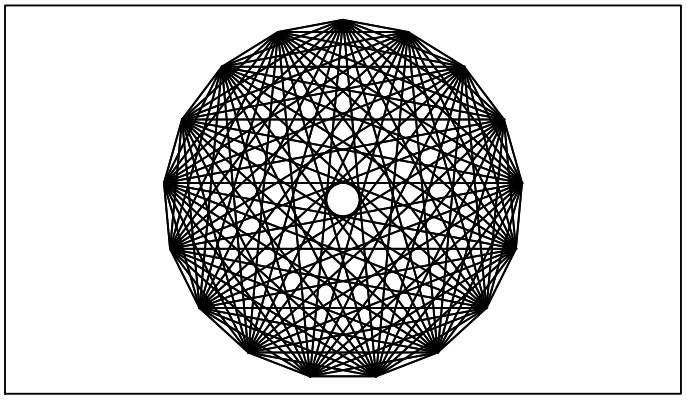


Things to note:

- `plot.new` creates a new, empty, plotting device.
- `plot.window` determines the limits of the plotting region.
- `axis` adds the axes, and `box` the framing box.
- The rest of the elements, you already know.

12.1.3.2 Rosette

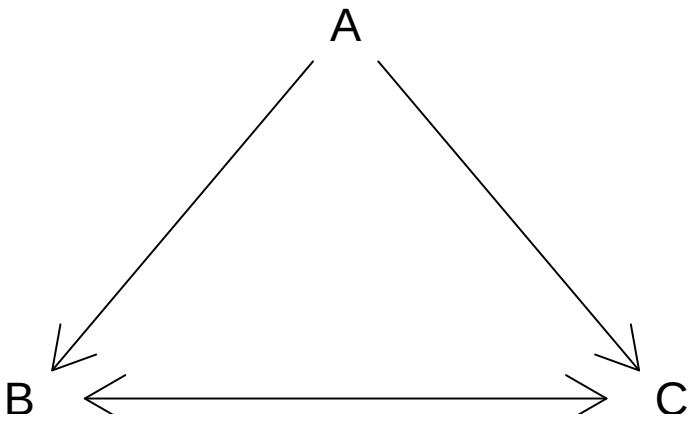
```
n = 17
theta = seq(0, 2 * pi, length = n + 1)[1:n]
x = sin(theta)
y = cos(theta)
v1 = rep(1:n, n)
v2 = rep(1:n, rep(n, n))
plot.new()
plot.window(xlim = c(-1, 1), ylim = c(-1, 1), asp = 1)
segments(x[v1], y[v1], x[v2], y[v2])
box()
```



12.1.3.3 Arrows

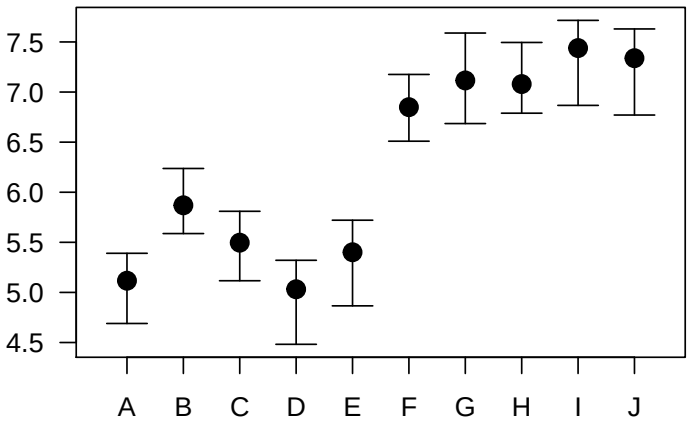
```
plot.new()
plot.window(xlim = c(0, 1), ylim = c(0, 1))
arrows(.05, .075, .45, .9, code = 1)
arrows(.55, .9, .95, .075, code = 2)
```

```
arrows(.1, 0, .9, 0, code = 3)
text(.5, 1, "A", cex = 1.5)
text(0, 0, "B", cex = 1.5)
text(1, 0, "C", cex = 1.5)
```



12.1.3.4 Arrows as error bars

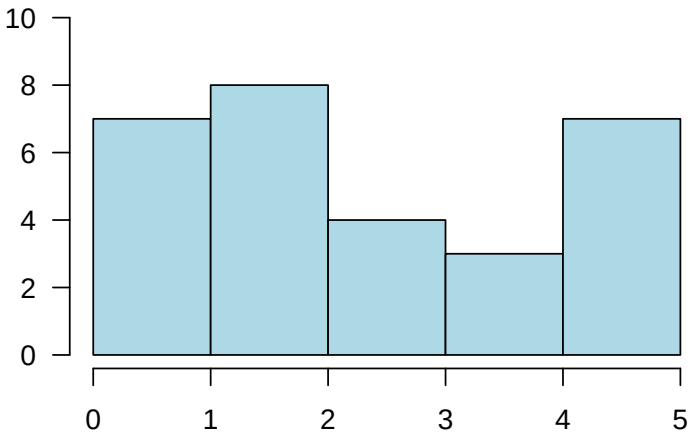
```
x = 1:10
y = runif(10) + rep(c(5, 6.5), c(5, 5))
yl = y - 0.25 - runif(10)/3
yu = y + 0.25 + runif(10)/3
plot.new()
plot.window(xlim = c(0.5, 10.5), ylim = range(yl, yu))
arrows(x, yl, x, yu, code = 3, angle = 90, length = .125)
points(x, y, pch = 19, cex = 1.5)
axis(1, at = 1:10, labels = LETTERS[1:10])
axis(2, las = 1)
box()
```



12.1.3.5 Histogram

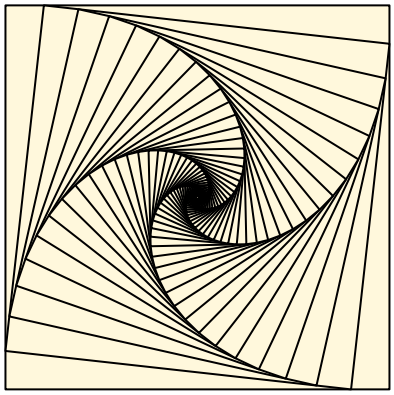
A histogram is nothing but a bunch of rectangle elements.

```
plot.new()
plot.window(xlim = c(0, 5), ylim = c(0, 10))
rect(0:4, 0, 1:5, c(7, 8, 4, 3), col = "lightblue")
axis(1)
axis(2, las = 1)
```



12.1.3.5.1 Spiral Squares

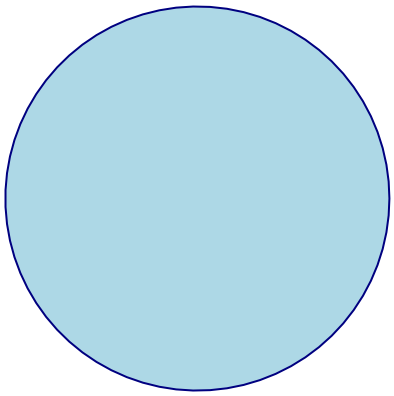
```
plot.new()
plot.window(xlim = c(-1, 1), ylim = c(-1, 1), asp = 1)
x = c(-1, 1, 1, -1)
y = c( 1, 1, -1, -1)
polygon(x, y, col = "cornsilk")
vertex1 = c(1, 2, 3, 4)
vertex2 = c(2, 3, 4, 1)
for(i in 1:50) {
    x = 0.9 * x[vertex1] + 0.1 * x[vertex2]
    y = 0.9 * y[vertex1] + 0.1 * y[vertex2]
    polygon(x, y, col = "cornsilk")
}
```



12.1.3.6 Circles

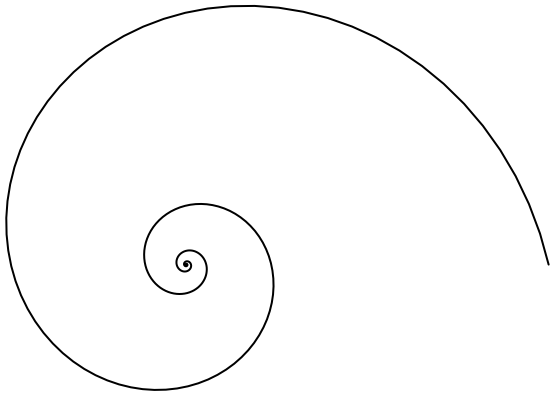
Circles are just dense polygons.

```
R = 1
xc = 0
yc = 0
n = 72
t = seq(0, 2 * pi, length = n)[1:(n-1)]
x = xc + R * cos(t)
y = yc + R * sin(t)
plot.new()
plot.window(xlim = range(x), ylim = range(y), asp = 1)
polygon(x, y, col = "lightblue", border = "navyblue")
```



12.1.3.7 Spiral

```
k = 5
n = k * 72
theta = seq(0, k * 2 * pi, length = n)
R = .98^(1:n - 1)
x = R * cos(theta)
y = R * sin(theta)
plot.new()
plot.window(xlim = range(x), ylim = range(y), asp = 1)
lines(x, y)
```



12.2 The ggplot2 System

The philosophy of **ggplot2** is very different from the **graphics** device. Recall, in **ggplot2**, a plot is a object. It can be queried, it can be changed, and among other things, it can be plotted.

ggplot2 provides a convenience function for many plots: `qplot`. We take a non-typical approach by ignoring `qplot`, and presenting the fundamental building blocks. Once the building blocks have been understood, mastering `qplot` will be easy.

The following is taken from UCLA's idre.

A **ggplot2** object will have the following elements:

- **Data** the data frame holding the data to be plotted.
- **Aes** defines the mapping between variables to their visualization.
- **Geoms** are the objects/shapes you add as layers to your graph.
- **Stats** are statistical transformations when you are not plotting the raw data, such as the mean or confidence intervals.

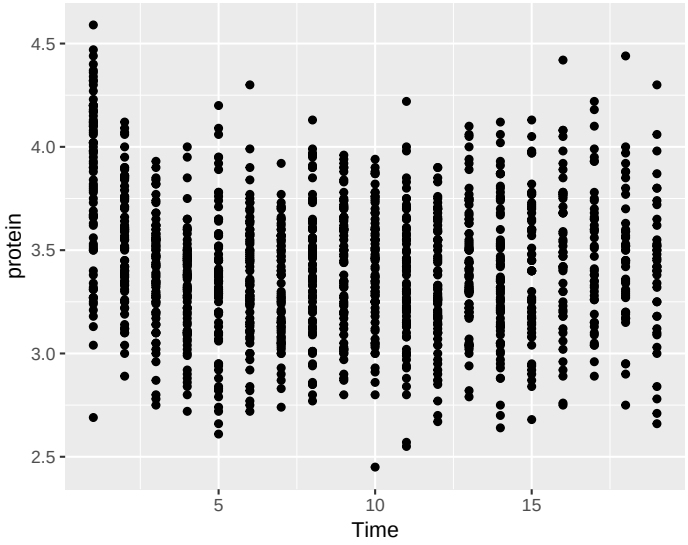
- **Faceting** splits the data into subsets to create multiple variations of the same graph (paneling).

The `nlme::Milk` dataset has the protein level of various cows, at various times, with various diets.

```
library(nlme)
data(Milk)
head(Milk)
```

```
## Grouped Data: protein ~ Time | Cow
##   protein Time Cow   Diet
## 1    3.63    1 B01 barley
## 2    3.57    2 B01 barley
## 3    3.47    3 B01 barley
## 4    3.65    4 B01 barley
## 5    3.89    5 B01 barley
## 6    3.73    6 B01 barley
```

```
library(ggplot2)
ggplot(data = Milk, aes(x=Time, y=protein)) +
  geom_point()
```

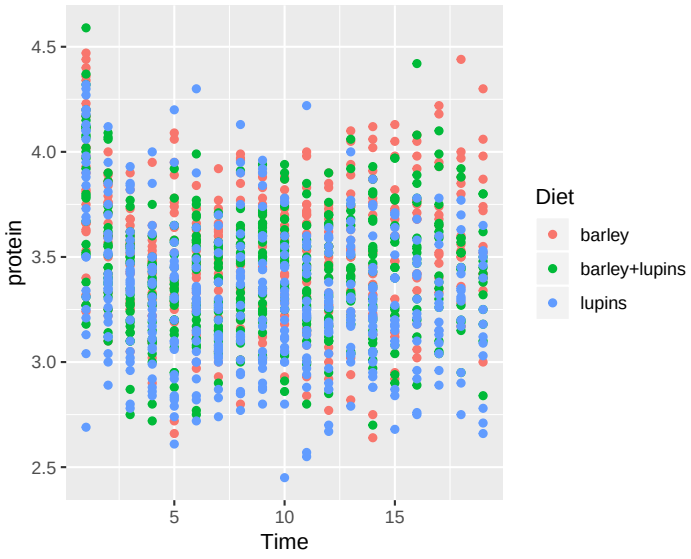


Things to note:

- The `ggplot` function is the constructor of the **ggplot2** object. If the object is not assigned, it is plotted.
- The `aes` argument tells R that the `Time` variable in the `Milk` data is the x axis, and protein is y.
- The `geom_point` defines the **Geom**, i.e., it tells R to plot the points as they are (and not lines, histograms, etc.).
- The **ggplot2** object is build by compounding its various elements separated by the `+` operator.
- All the variables that we will need are assumed to be in the `Milk` data frame. This means that (a) the data needs to be a data frame (not a matrix for instance), and (b) we will not be able to use variables that are not in the `Milk` data frame.

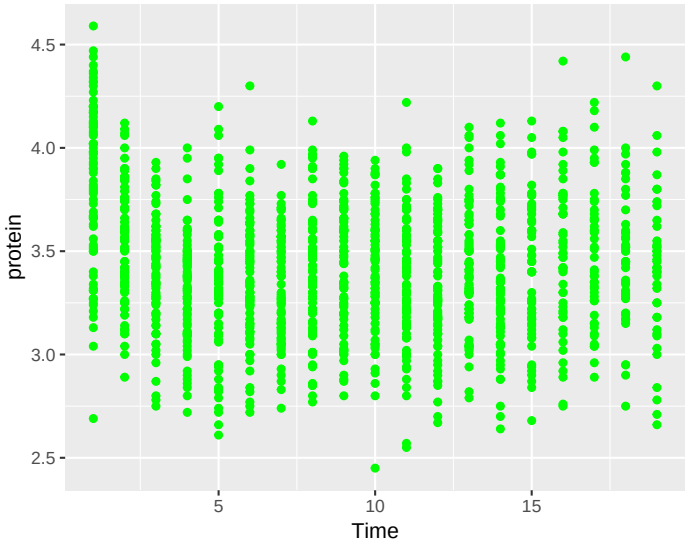
Let's add some color.

```
ggplot(data = Milk, aes(x=Time, y=protein)) +
  geom_point(aes(color=Diet))
```



The `color` argument tells R to use the variable `Diet` as the coloring. A legend is added by default. If we wanted a fixed color, and not a variable dependent color, `color` would have been put outside the `aes` function.

```
ggplot(data = Milk, aes(x=Time, y=protein)) +
  geom_point(color="green")
```

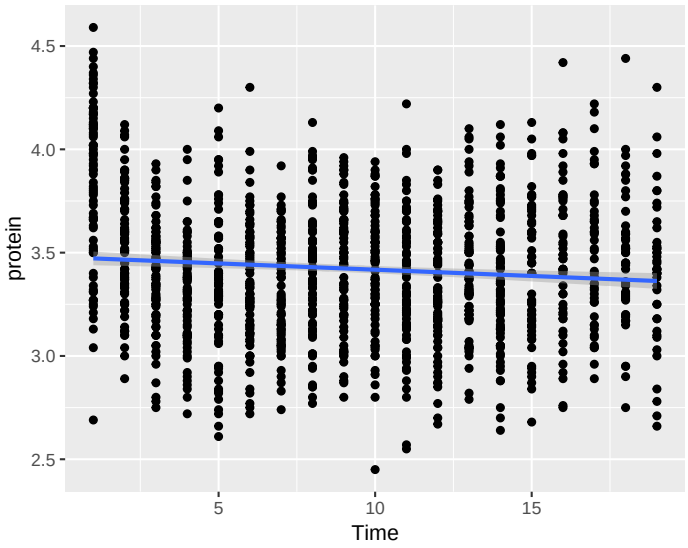


Let's save the **ggplot2** object so we can reuse it. Notice it is not plotted.

```
p <- ggplot(data = Milk, aes(x=Time, y=protein)) +
  geom_point()
```

We can change^{In the Object-Oriented Programming lingo, this is known as mutating} existing plots using the `+` operator. Here, we add a smoothing line to the plot `p`.

```
p + geom_smooth(method = 'gam')
```

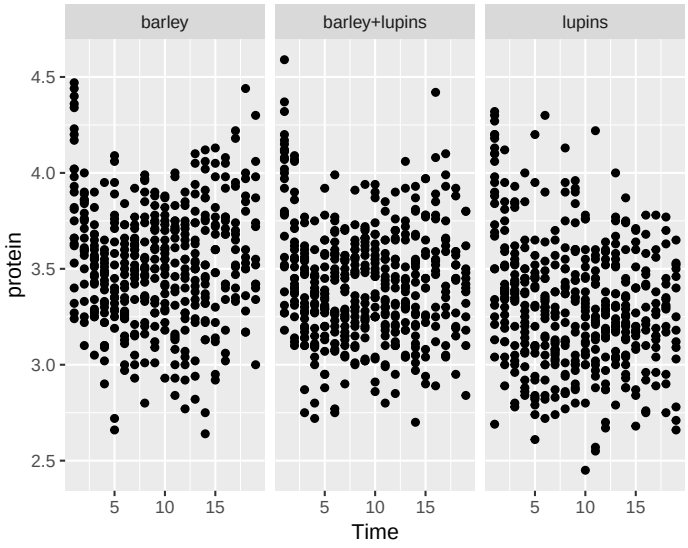


Things to note:

- The smoothing line is a layer added with the `geom_smooth()` function.
- Lacking arguments of its own, the new layer will inherit the `aes` of the original object, x and y variables in particular.

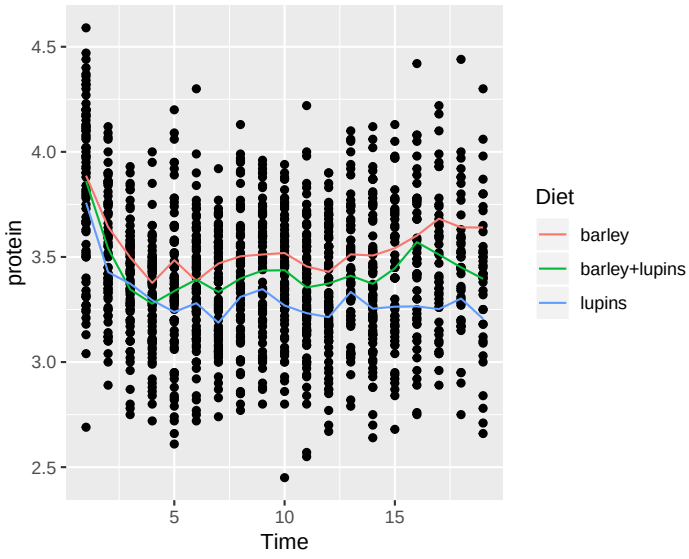
To split the plot along some variable, we use faceting, done with the `facet_wrap` function.

```
p + facet_wrap(~Diet)
```



Instead of faceting, we can add a layer of the mean of each `Diet` subgroup, connected by lines.

```
p + stat_summary(aes(color=Diet), fun.y="mean", geom="line")
```



Things to note:

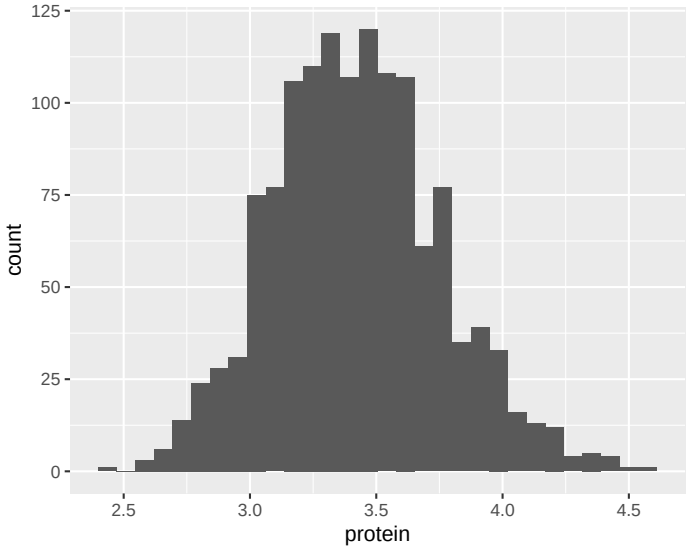
- `stat_summary` adds a statistical summary.
- The summary is applied along `Diet` subgroups, because of the `color=Diet` aesthetic, which has already split the data.
- The summary to be applied is the mean, because of `fun.y="mean"`.
- The group means are connected by lines, because of the `geom="line"` argument.

What layers can be added using the **geoms** family of functions?

- `geom_bar`: bars with bases on the x-axis.
- `geom_boxplot`: boxes-and-whiskers.
- `geom_errorbar`: T-shaped error bars.
- `geom_histogram`: histogram.
- `geom_line`: lines.
- `geom_point`: points (scatterplot).
- `geom_ribbon`: bands spanning y-values across a range of x-values.
- `geom_smooth`: smoothed conditional means (e.g. loess smooth).

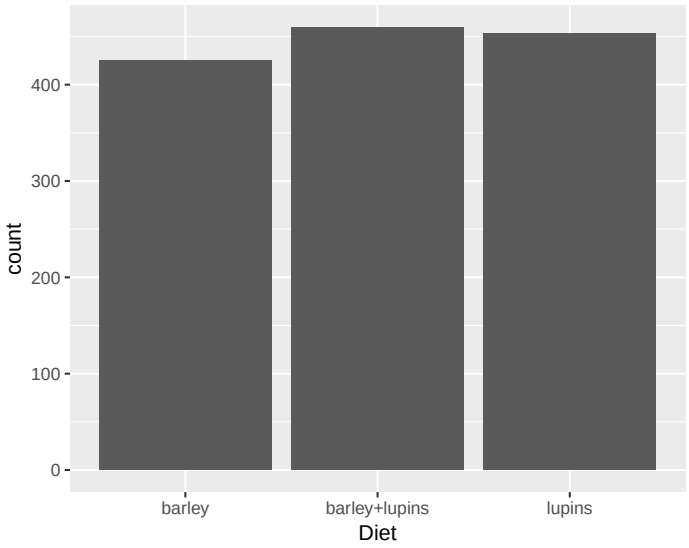
To demonstrate the layers added with the `geoms_*` functions, we start with a histogram.

```
pro <- ggplot(Milk, aes(x=protein))
pro + geom_histogram(bins=30)
```



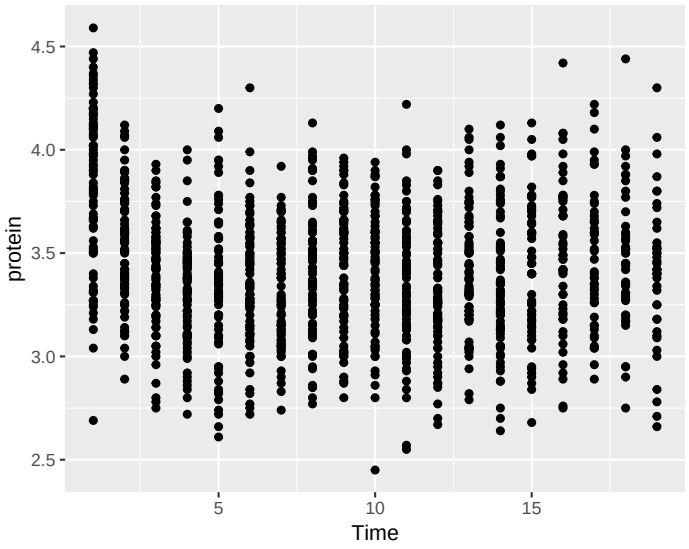
A bar plot.

```
ggplot(Milk, aes(x=Diet)) +
  geom_bar()
```



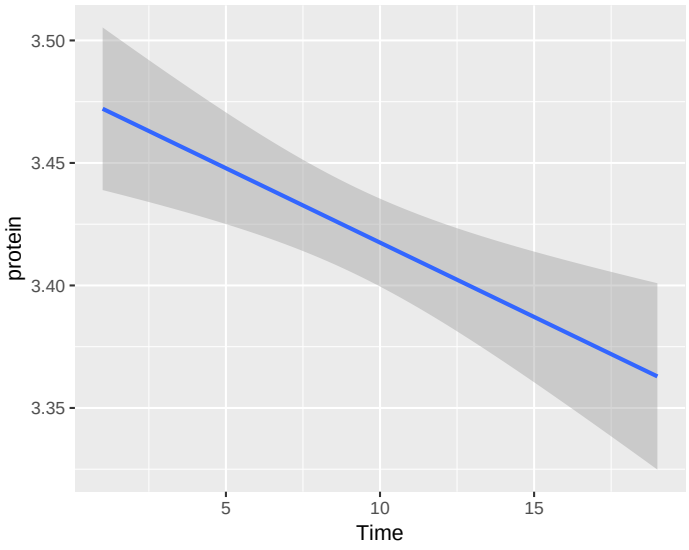
A scatter plot.

```
tp <- ggplot(Milk, aes(x=Time, y=protein))
tp + geom_point()
```



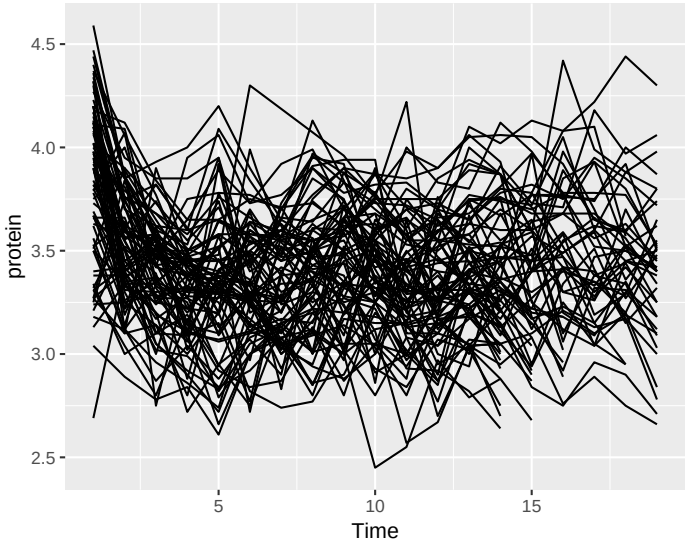
A smooth regression plot, reusing the `tp` object.

```
tp + geom_smooth(method='gam')
```



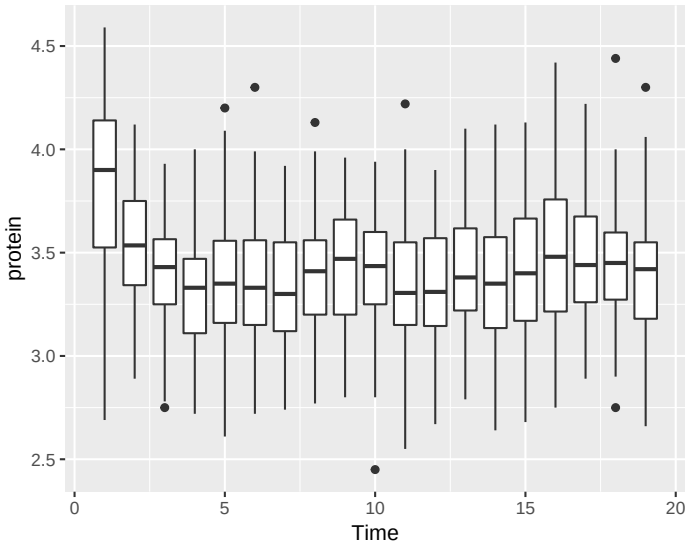
And now, a simple line plot, reusing the `tp` object, and connecting lines along `Cow`.

```
tp + geom_line(aes(group=Cow))
```



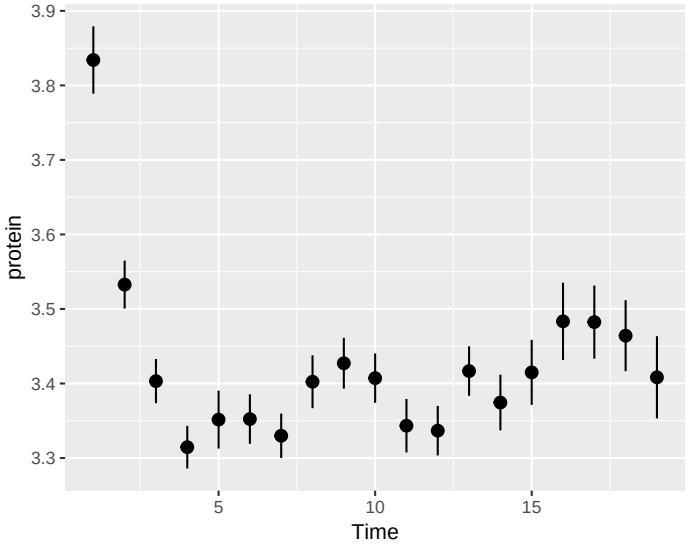
The line plot is completely incomprehensible. Better look at boxplots along time (even if omitting the `Cow` information).

```
tp + geom_boxplot(aes(group=Time))
```



We can do some statistics for each subgroup. The following will compute the mean and standard errors of `protein` at each time point.

```
ggplot(Milk, aes(x=Time, y=protein)) +
  stat_summary(fun.data = 'mean_se')
```

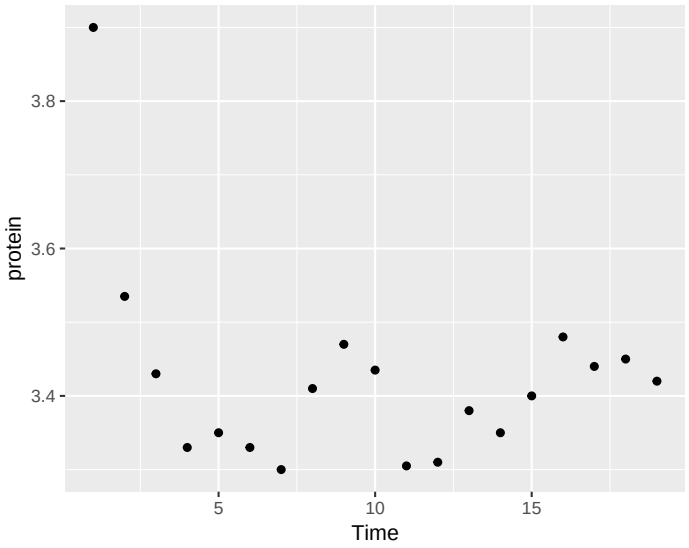


Some popular statistical summaries, have gained their own functions:

- `mean_cl_boot`: mean and bootstrapped confidence interval (default 95%).
- `mean_cl_normal`: mean and Gaussian (t-distribution based) confidence interval (default 95%).
- `mean_dsl`: mean plus or minus standard deviation times some constant (default constant=2).
- `median_hilow`: median and outer quantiles (default outer quantiles = 0.025 and 0.975).

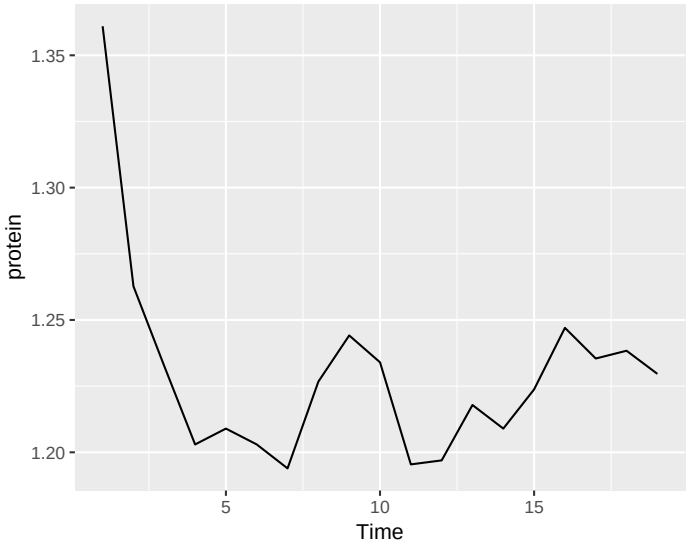
For less popular statistical summaries, we may specify the statistical function in `stat_summary`. The median is a first example.

```
ggplot(Milk, aes(x=Time, y=protein)) +
  stat_summary(fun.y="median", geom="point")
```



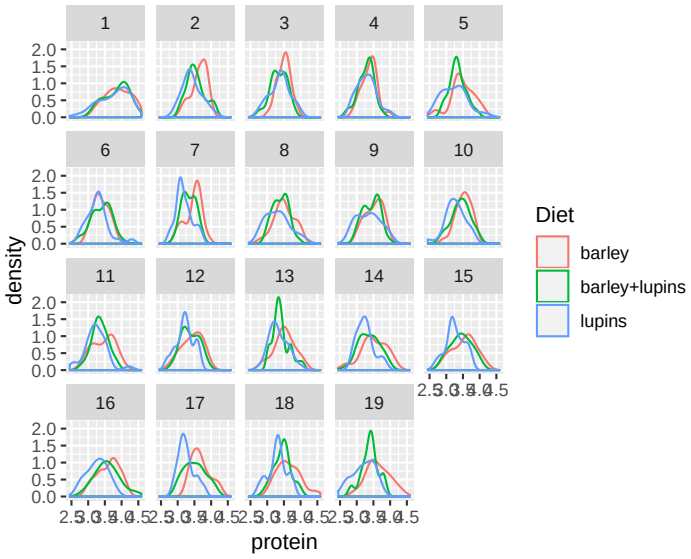
We can also define our own statistical summaries.

```
medianlog <- function(y) {median(log(y))}
ggplot(Milk, aes(x=Time, y=protein)) +
  stat_summary(fun.y="medianlog", geom="line")
```



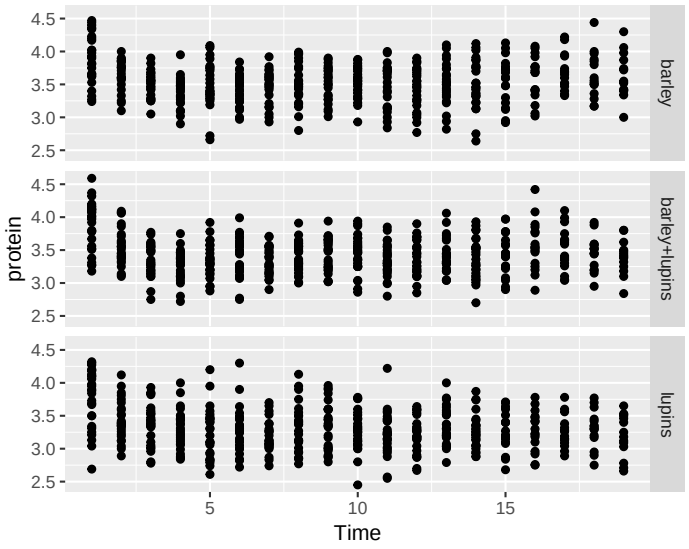
Faceting allows to split the plotting along some variable. `face_wrap` tells R to compute the number of columns and rows of plots automatically.

```
ggplot(Milk, aes(x=protein, color=Diet)) +
  geom_density() +
  facet_wrap(~Time)
```



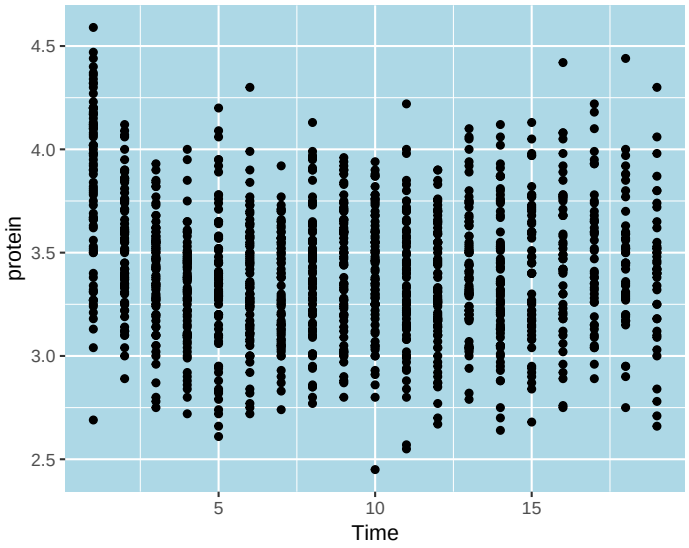
`facet_grid` forces the plot to appear allow rows or columns, using the `~` syntax.

```
ggplot(Milk, aes(x=Time, y=protein)) +
  geom_point() +
  facet_grid(Diet~.) # `.~Diet` to split along columns and not rows.
```

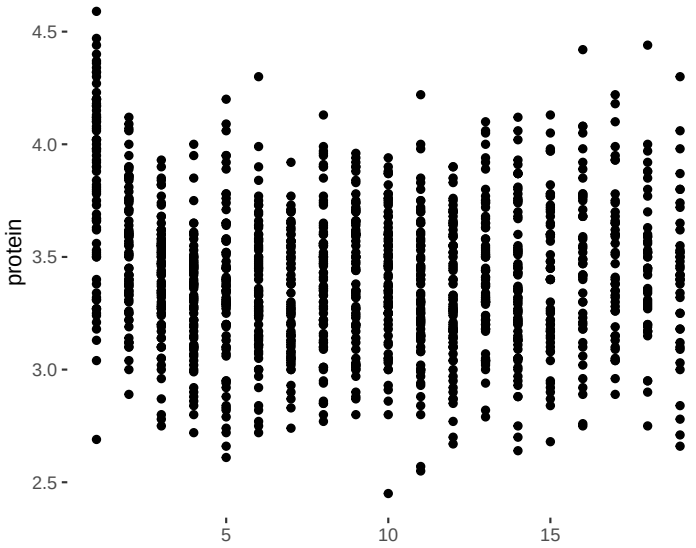


To control the looks of the plot, **ggplot2** uses **themes**.

```
ggplot(Milk, aes(x=Time, y=protein)) +
  geom_point() +
  theme(panel.background=element_rect(fill="lightblue"))
```




```
ggplot(Milk, aes(x=Time, y=protein)) +
  geom_point() +
  theme(panel.background=element_blank(),
        axis.title.x=element_blank())
```



Saving plots can be done using `ggplot2::ggsave`, or with `pdf` like the **graphics** plots:

```
pdf(file = 'myplot.pdf')
print(tp) # You will need an explicit print command!
dev.off()
```

Remark. If you are exporting a PDF for publication, you will probably need to embed your fonts in the PDF. In this case, use `cairo_pdf()` instead of `pdf()`.

Finally, what every user of **ggplot2** constantly uses, is the (excellent!) online documentation at http://docs.ggplot2.org.

12.2.1 Extensions of the ggplot2 System

Because **ggplot2** plots are R objects, they can be used for computations and altered. Many authors, have thus extended the basic **ggplot2** functionality. A list of **ggplot2** extensions is curated by Daniel Emaasit at http://www.ggplot2-exts.org. The RStudio team has its own list of recommended packages at RStartHere.

12.3 Interactive Graphics

As already mentioned, the recent and dramatic advancement in interactive visualization was made possible by the advances in web technologies, and the D3.JS JavaScript library in particular. This is because it allows developers to rely on existing libraries designed for web browsing, instead of re-implementing interactive visualizations. These libraries are more visually pleasing, and computationally efficient, than anything they could have developed themselves.

The htmlwidgets package does not provide visualization, but rather, it facilitates the creation of new interactive visualizations. This is because it handles all the technical details that are required to use R output within JavaScript visualization libraries.

For a list of interactive visualization tools that rely on **htmlwidgets** see their (amazing) gallery, and the RStartsHere page. In the following sections, we discuss a selected subset.

12.3.1 Plotly

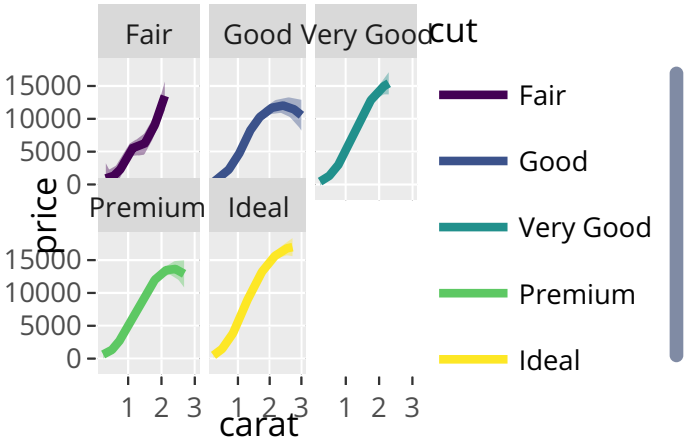
You can create nice interactive graphs using `plotly::plot_ly`:

```
library(plotly)
set.seed(100)
d <- diamonds[sample(nrow(diamonds), 1000), ]
```

```
plot_ly(data = d, x = ~carat, y = ~price, color = ~carat, size = ~carat, text = ~paste("Clarity: ", clarity)
```

More conveniently, any **ggplot2** graph can be made interactive using `plotly::ggplotly`:

```
p <- ggplot(data = d, aes(x = carat, y = price)) +
  geom_smooth(aes(colour = cut, fill = cut), method = 'loess') +
  facet_wrap(~ cut) # make ggplot
ggplotly(p) # from ggplot to plotly
```



How about exporting **plotly** objects? Well, a **plotly** object is nothing more than a little web site: an HTML file. When showing a **plotly** figure, RStudio merely servers you as a web browser. You could, alternatively, export this HTML file to send your colleagues as an email attachment, or embed it in a web site. To export these, use the `plotly::export` or the `htmlwidgets::saveWidget` functions.

For more on **plotly** see https://plot.ly/r/.

12.4 Other R Interfaces to JavaScript Plotting

Plotly is not the only interactive plotting framework in R that relies o JavaScript for interactivity. Here are some more interactive and beautiful charting libraries.

- Highcharts, like Plotly [12.3.1], is a popular collection of JavaScript plotting libraries, with great emphasis on aesthetics. The package highcharter is an R wrapper for dispatching plots to highcharts. For a demo of the capabilities of Highcarts, see here.
- Rbokeh is a R wrapper for the popular Bokeh JavaScript charting libraries.
- r2d3: a R wrapper to the D3 plotting libraries.

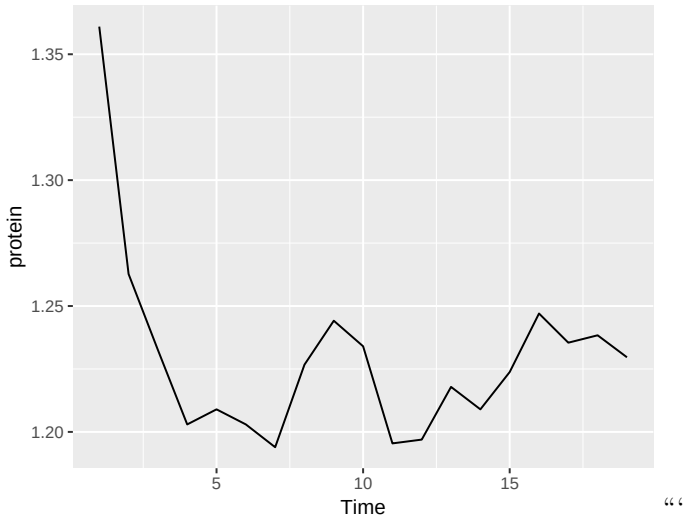
12.5 Bibliographic Notes

For the **graphics** package, see R Core Team (2016). For **ggplot2** see Wickham (2009). For the theory underlying **ggplot2**, i.e. the Grammar of Graphics, see Wilkinson (2006). A video by one of my heroes, Brian Caffo, discussing **graphics** vs. **ggplot2**.

12.6 Practice Yourself

1. Go to the Fancy Graphics Section 12.1.3. Try parsing the commands in your head.
2. Recall the `medianlog` example and replace the `medianlog` function with a harmonic mean.

```
medianlog <- function(y) {median(log(y))}
ggplot(Milk, aes(x=Time, y=protein)) +
  stat_summary(fun.y="medianlog", geom="line")
```



“‘

3. Write a function that creates a boxplot from scratch. See how I built a line graph in Section 12.1.3.
4. Export my plotly example using the RStudio interface and send it to yourself by email.

ggplot2:

1. Read about the “oats” dataset using `? MASS::oats`.
    1. Inspect, visually, the dependency of the yield (Y) in the Varieties (V) and the Nitrogen treatment (N).
    2. Compute the mean and the standard error of the yield for every value of Varieties and Nitrogen treatment.
    3. Change the axis labels to be informative with `labs` function and give a title to the plot with `ggtitle` function.
2. Read about the “mtcars” data set using `? mtcars`.
    1. Inspect, visually, the dependency of the Fuel consumption (mpg) in the weight (wt)
    2. Inspect, visually, the assumption that the Fuel consumption also depends on the number of cylinders.
    3. Is there an interaction between the number of cylinders to the weight (i.e. the slope of the regression line is different between the number of cylinders)? Use `geom_smooth`.

See DataCamp’s Data Visualization with ggplot2 for more self practice.

Chapter 13

Reports

If you have ever written a report, you are probably familiar with the process of preparing your figures in some software, say R, and then copy-pasting into your text editor, say MS Word. While very popular, this process is both tedious, and plain painful if your data has changed and you need to update the report. Wouldn't it be nice if you could produce figures and numbers from within the text of the report, and everything else would be automated? It turns out it is possible. There are actually several systems in R that allow this. We start with a brief review.

1. **Sweave**: *LaTeX* is a markup language that compiles to *Tex* programs that compile, in turn, to documents (typically PS or PDFs). If you never heard of it, it may be because you were born the the MS Windows+MS Word era. You should know, however, that *LaTeX* was there much earlier, when computers were mainframes with text-only graphic devices. You should also know that *LaTeX* is still very popular (in some communities) due to its very rich markup syntax, and beautiful output. *Sweave* (Leisch, 2002) is a compiler for *LaTeX* that allows you do insert R commands in the *LaTeX* source file, and get the result as part of the outputted PDF. It's name suggests just that: it allows to weave S[1] output into the document, thus, Sweave.
2. **knitr**: *Markdown* is a text editing syntax that, unlike *LaTeX*, is aimed to be human-readable, but also compilable by a machine. If you ever tried to read HTML or *LaTeX* source files, you may understand why human-readability is a desirable property. There are many *markdown* compilers. One of the most popular is Pandoc, written by the Berkeley philosopher(!) Jon MacFarlane. The availability of Pandoc gave Yihui Xie, a name to remember, the idea that it is time for Sweave to evolve. Yihui thus wrote **knitr** (Xie, 2015), which allows to write human readable text in *Rmarkdown*, a superset of *markdown*, compile it with R and the compile it with Pandoc. Because Pandoc can compile to PDF, but also to HTML, and DOCX, among others, this means that you can write in Rmarkdown, and get output in almost all text formats out there.
3. **bookdown**: **Bookdown** (Xie, 2016) is an evolution of **knitr**, also written by Yihui Xie, now working for RStudio. The text you are now reading was actually written in **bookdown**. It deals with the particular needs of writing large documents, and cross referencing in particular (which is very challenging if you want the text to be human readable).
4. **Shiny**: Shiny is essentially a framework for quick web-development. It includes (i) an abstraction layer that specifies the layout of a web-site which is our report, (ii) the command to start a web server to deliver the site. For more on Shiny see Chang et al. (2017).

13.1 knitr

13.1.1 Installation

To run **knitr** you will need to install the package.

```
install.packages('knitr')
```

It is also recommended that you use it within RStudio (version>0.96), where you can easily create a new `.Rmd` file.

[1] Recall, S was the original software from which R evolved.

13.1.2 Pandoc Markdown

Because **knitr** builds upon *Pandoc markdown*, here is a simple example of markdown text, to be used in a `.Rmd` file, which can be created using the *File-> New File -> R Markdown* menu of RStudio.

Underscores or asterisks for `_italics1_` and `*italics2*` return *italics1* and *italics2*. Double underscores or asterisks for `__bold1__` and `**bold2**` return **bold1** and **bold2**. Subscripts are enclosed in tildes, `like~this~` (like$_{\text{this}}$), and superscripts are enclosed in carets `like^this^` (like$^{\text{this}}$).

For links use `[text](link)`, like `[my site](www.john-ros.com)`. An image is the same as a link, starting with an exclamation, like this `![image caption](image path)`.

An itemized list simply starts with hyphens preceeded by a blank line (don't forget that!):

```
- bullet
- bullet
    - second level bullet
    - second level bullet
```

Compiles into:

- bullet
- bullet
  - second level bullet
  - second level bullet

An enumerated list starts with an arbitrary number:

```
1. number
1. number
    1. second level number
    1. second level number
```

Compiles into:

1. number
2. number
   1. second level number
   2. second level number

For more on markdown see https://bookdown.org/yihui/bookdown/markdown-syntax.html.

13.1.3 Rmarkdown

Rmarkdown, is an extension of *markdown* due to RStudio, that allows to incorporate R expressions in the text, that will be evaluated at the time of compilation, and the output automatically inserted in the outputted text. The output can be a `.PDF`, `.DOCX`, `.HTML` or others, thanks to the power of *Pandoc*.

The start of a code chunk is indicated by three backticks and the end of a code chunk is indicated by three backticks. Here is an example.

````
```{r  eval=FALSE}
rnorm(10)
```
````

This chunk will compile to the following output (after setting `eval=FALSE` to `eval=TRUE`):

```r
rnorm(10)
```

```
##  [1] -0.36276795  2.43272890  0.59209119 -0.57620077  0.40662824
##  [6] -0.04525463  0.44090002  2.23537620 -0.58615484  0.59936025
```

Things to note:

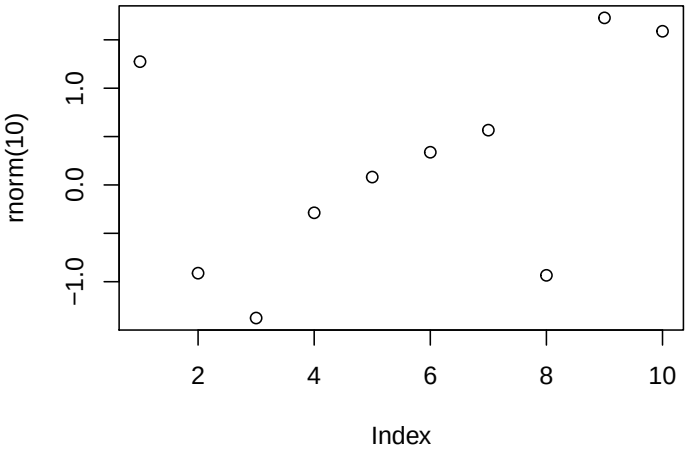
- The evaluated expression is added in a chunk of highlighted text, before the R output.
- The output is prefixed with `##`.
- The `eval=` argument is not required, since it is set to `eval=TRUE` by default. It does demonstrate how to set the options of the code chunk.

In the same way, we may add a plot:

````
```{r  eval=FALSE}
plot(rnorm(10))
```
````

which compiles into

```r
plot(rnorm(10))
```



Some useful code chunk options include:

- `eval=FALSE`: to return code only, without output.
- `echo=FALSE`: to return output, without code.
- `cache=`: to save results so that future compilations are faster.
- `results='hide'`: to plot figures, without text output.
- `collapse=TRUE`: if you want the whole output after the whole code, and not interleaved.
- `warning=FALSE`: to supress watning. The same for `message=FALSE`, and `error=FALSE`.

You can also call r expressions inline. This is done with a single tick and the `r` argument. For instance:

> `` `r rnorm(1)` `` is a random Gaussian

will output

> -1.0374466 is a random Gaussian.

13.1.4 BibTex

BibTex is both a file format and a compiler. The bibtex compiler links documents to a reference database stored in the `.bib` file format.

Bibtex is typically associated with Tex and LaTex typesetting, but it also operates within the markdown pipeline.

Just store your references in a `.bib` file, add a `bibliography: yourFile.bib` in the YML preamble of your Rmarkdown file, and call your references from the Rmarkdown text using `@referencekey`. Rmarkdow will take care of creating the bibliography, and linking to it from the text.

13.1.5 Compiling

Once you have your `.Rmd` file written in RMarkdown, **knitr** will take care of the compilation for you. You can call the `knitr::knitr` function directly from some `.R` file, or more conveniently, use the RStudio (0.96) Knit button above

the text editing window. The location of the output file will be presented in the console.

13.2 bookdown

As previously stated, **bookdown** is an extension of **knitr** intended for documents more complicated than simple reports– such as books. Just like **knitr**, the writing is done in **RMarkdown**. Being an extension of **knitr**, **bookdown** does allow some markdowns that are not supported by other compilers. In particular, it has a more powerful cross referencing system.

13.3 Shiny

Shiny (Chang et al., 2017) is different than the previous systems, because it sets up an interactive web-site, and not a static file. The power of Shiny is that the layout of the web-site, and the settings of the web-server, is made with several simple R commands, with no need for web-programming. Once you have your app up and running, you can setup your own Shiny server on the web, or publish it via Shinyapps.io. The freemium versions of the service can deal with a small amount of traffic. If you expect a lot of traffic, you will probably need the paid versions.

13.3.1 Installation

To setup your first Shiny app, you will need the **shiny** package. You will probably want RStudio, which facilitates the process.

```
install.packages('shiny')
```

Once installed, you can run an example app to get the feel of it.

```
library(shiny)
runExample("01_hello")
```

Remember to press the **Stop** button in RStudio to stop the web-server, and get back to RStudio.

13.3.2 The Basics of Shiny

Every Shiny app has two main building blocks.

1. A user interface, specified via the `ui.R` file in the app's directory.
2. A server side, specified via the `server.R` file, in the app's directory.

You can run the app via the **RunApp** button in the RStudio interface, of by calling the app's directory with the `shinyApp` or `runApp` functions– the former designed for single-app projects, and the latter, for multiple app projects.

```
shiny::runApp("my_app") # my_app is the app's directory.
```

The site's layout, is specified in the `ui.R` file using one of the *layout functions*. For instance, the function `sidebarLayout`, as the name suggest, will create a sidebar. More layouts are detailed in the layout guide.

The active elements in the UI, that control your report, are known as *widgets*. Each widget will have a unique `inputId` so that it's values can be sent from the UI to the server. More about widgets, in the widget gallery.

The `inputId` on the UI are mapped to `input` arguments on the server side. The value of the `mytext inputId` can be queried by the server using `input$mytext`. These are called *reactive values*. The way the server "listens" to the UI, is governed by a set of functions that must wrap the `input` object. These are the `observe`, `reactive`, and `reactive*` class of functions.

With `observe` the server will get triggered when any of the reactive values change. With `observeEvent` the server will only be triggered by specified reactive values. Using `observe` is easier, and `observeEvent` is more prudent programming.

A `reactive` function is a function that gets triggered when a reactive element changes. It is defined on the server side, and reside within an `observe` function.

We now analyze the `1_Hello` app using these ideas. Here is the `ui.R` file.

```
library(shiny)

shinyUI(fluidPage(

  titlePanel("Hello Shiny!"),

  sidebarLayout(
    sidebarPanel(
      sliderInput(inputId = "bins",
                  label = "Number of bins:",
                  min = 1,
                  max = 50,
                  value = 30)
    ),

    mainPanel(
      plotOutput(outputId = "distPlot")
    )
  )
))
```

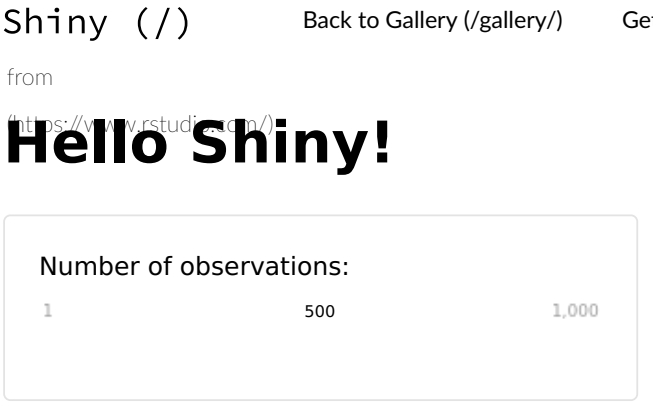
Here is the `server.R` file:

```
library(shiny)

shinyServer(function(input, output) {

  output$distPlot <- renderPlot({
    x    <- faithful[, 2]  # Old Faithful Geyser data
    bins <- seq(min(x), max(x), length.out = input$bins + 1)

    hist(x, breaks = bins, col = 'darkgray', border = 'white')
  })
})
```

Things to note:

- `ShinyUI` is a (deprecated) wrapper for the UI.
- `fluidPage` ensures that the proportions of the elements adapt to the window side, thus, are fluid.
- The building blocks of the layout are a title, and the body. The title is governed by `titlePanel`, and the body is governed by `sidebarLayout`. The `sidebarLayout` includes the `sidebarPanel` to control the sidebar, and the `mainPanel` for the main panel.
- `sliderInput` calls a widget with a slider. Its `inputId` is `bins`, which is later used by the server within the `renderPlot` reactive function.
- `plotOutput` specifies that the content of the `mainPanel` is a plot (`textOutput` for text). This expectation is satisfied on the server side with the `renderPlot` function (`renderText`).
- `shinyServer` is a (deprecated) wrapper function for the server.
- The server runs a function with an `input` and an `output`. The elements of `input` are the `inputId`s from the UI. The elements of the `output` will be called by the UI using their `outputId`.

This is the output.



Here is another example, taken from the RStudio Shiny examples.

ui.R:

```
library(shiny)

fluidPage(

  titlePanel("Tabsets"),

  sidebarLayout(
    sidebarPanel(
      radioButtons(inputId = "dist",
                   label = "Distribution type:",
                   c("Normal" = "norm",
                     "Uniform" = "unif",
                     "Log-normal" = "lnorm",
                     "Exponential" = "exp")),
      br(), # add a break in the HTML page.

      sliderInput(inputId = "n",
                  label = "Number of observations:",
                   value = 500,
                   min = 1,
                   max = 1000)
    ),

    mainPanel(
      tabsetPanel(type = "tabs",
        tabPanel(title = "Plot", plotOutput(outputId = "plot")),
        tabPanel(title = "Summary", verbatimTextOutput(outputId = "summary")),
        tabPanel(title = "Table", tableOutput(outputId = "table"))
      )
    )
  )
)
```

server.R:

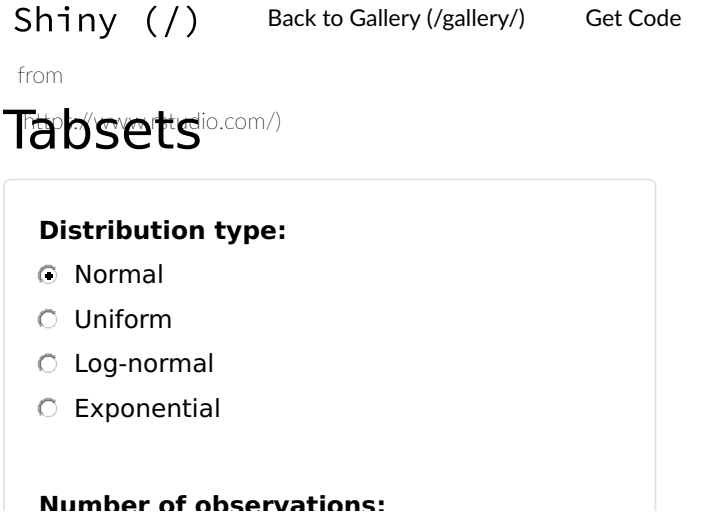
```
library(shiny)

# Define server logic for random distribution application
function(input, output) {

  data <- reactive({
    dist <- switch(input$dist,
```

```
                    norm = rnorm,
                    unif = runif,
                    lnorm = rlnorm,
                    exp = rexp,
                    rnorm)

    dist(input$n)
  })

  output$plot <- renderPlot({
    dist <- input$dist
    n <- input$n

    hist(data(), main=paste('r', dist, '(', n, ')', sep=''))
  })

  output$summary <- renderPrint({
    summary(data())
  })

  output$table <- renderTable({
    data.frame(x=data())
  })

}
```

Things to note:

- We reused the `sidebarLayout`.
- As the name suggests, `radioButtons` is a widget that produces radio buttons, above the `sliderInput` widget. Note the different `inputId`s.
- Different widgets are separated in `sidebarPanel` by commas.
- `br()` produces extra vertical spacing (break).
- `tabsetPanel` produces tabs in the main output panel. `tabPanel` governs the content of each panel. Notice the use of various output functions (`plotOutput`,`verbatimTextOutput`, `tableOutput`) with corresponding `outputId`s.
- In `server.R` we see the usual `function(input,output)`.
- The `reactive` function tells the server the trigger the function whenever `input` changes.
- The `output` object is constructed outside the `reactive` function. See how the elements of `output` correspond to the `outputId`s in the UI.

This is the output:



13.3.3 Beyond the Basics

Now that we have seen the basics, we may consider extensions to the basic report.

13.3.3.1 Widgets

- `actionButton` Action Button.
- `checkboxGroupInput` A group of check boxes.
- `checkboxInput` A single check box.
- `dateInput` A calendar to aid date selection.
- `dateRangeInput` A pair of calendars for selecting a date range.
- `fileInput` A file upload control wizard.
- `helpText` Help text that can be added to an input form.
- `numericInput` A field to enter numbers.
- `radioButtons` A set of radio buttons.
- `selectInput` A box with choices to select from.
- `sliderInput` A slider bar.
- `submitButton` A submit button.
- `textInput` A field to enter text.

See examples here.

13.3.3.2 Output Elements

The `ui.R` output types.

- `htmlOutput` raw HTML.
- `imageOutput` image.
- `plotOutput` plot.
- `tableOutput` table.
- `textOutput` text.
- `uiOutput` raw HTML.
- `verbatimTextOutput` text.

The corresponding `server.R` renderers.

- `renderImage` images (saved as a link to a source file).
- `renderPlot` plots.
- `renderPrint` any printed output.
- `renderTable` data frame, matrix, other table like structures.
- `renderText` character strings.
- `renderUI` a Shiny tag object or HTML.

Your Shiny app can use any R object. The things to remember:

- The working directory of the app is the location of `server.R`.
- The code before `shinyServer` is run only once.
- The code inside '`shinyServer` is run whenever a reactive is triggered, and may thus slow things.

To keep learning, see the RStudio's tutorial, and the Biblipgraphic notes herein.

13.3.4 shinydashboard

A template for Shiny to give it s modern look.

13.4 flexdashboard

If you want to quickly write an interactive dashboard, which is simple enough to be a static HTML file and does not need an HTML server, then Shiny may be an overkill. With **flexdashboard** you can write your dashboard a single .Rmd file, which will generate an interactive dashboard as a static HTML file.

See [http://rmarkdown.rstudio.com/flexdashboard/] for more info.

13.5 Bibliographic Notes

For RMarkdown see here. For everything on **knitr** see Yihui's blog, or the book Xie (2015). For a **bookdown** manual, see Xie (2016). For a Shiny manual, see Chang et al. (2017), the RStudio tutorial, or Zev Ross's excellent guide. Video tutorials are available here.

13.6 Practice Yourself

1. Generate a report using **knitr** with your name as title, and a scatter plot of two random variables in the body. Save it as PDF, DOCX, and HTML.
2. Recall that this book is written in **bookdown**, which is a superset of **knitr**. Go to the source .Rmd file of the first chapter, and parse it in your head: (https://raw.githubusercontent.com/johnros/Rcourse/master/02-r-basics.Rmd)

Chapter 14

Sparse Representations

Analyzing "bigdata" in R is a challenge because the workspace is memory resident, i.e., all your objects are stored in RAM. As a rule of thumb, fitting models requires about 5 times the size of the data. This means that if you have 1 GB of data, you might need about 5 GB to fit a linear models. We will discuss how to compute *out of RAM* in the Memory Efficiency Chapter 15. In this chapter, we discuss efficient representations of your data, so that it takes less memory. The fundamental idea, is that if your data is *sparse*, i.e., there are many zero entries in your data, then a naive `data.frame` or `matrix` will consume memory for all these zeroes. If, however, you have many recurring zeroes, it is more efficient to save only the non-zero entries.

When we say *data*, we actually mean the `model.matrix`. The `model.matrix` is a matrix that R grows, converting all your factors to numeric variables that can be computed with. *Dummy coding* of your factors, for instance, is something that is done in your `model.matrix`. If you have a factor with many levels, you can imagine that after dummy coding it, many zeroes will be present.

The **Matrix** package replaces the `matrix` class, with several sparse representations of matrix objects.

When using sparse representation, and the **Matrix** package, you will need an implementation of your favorite model fitting algorithm (e.g. `lm`) that is adapted to these sparse representations; otherwise, R will cast the sparse matrix into a regular (non-sparse) matrix, and you will have saved nothing in RAM.
Remark. If you are familiar with MATLAB you should know that one of the great capabilities of MATLAB, is the excellent treatment of sparse matrices with the `sparse` function.

Before we go into details, here is a simple example. We will create a factor of letters with the `letters` function. Clearly, this factor can take only 26 values. This means that 25/26 of the `model.matrix` will be zeroes after dummy coding. We will compare the memory footprint of the naive `model.matrix` with the sparse representation of the same matrix.

```
library(magrittr)
reps <- 1e6 # number of samples
y<-rnorm(reps)
x<- letters %>%
  sample(reps, replace=TRUE) %>%
  factor
```

The object `x` is a factor of letters:

```
head(x)
```

```
## [1] l s q b h p
## Levels: a b c d e f g h i j k l m n o p q r s t u v w x y z
```

We dummy code `x` with the `model.matrix` function.

```
X.1 <- model.matrix(~x-1)
head(X.1)
```

```
##   xa xb xc xd xe xf xg xh xi xj xk xl xm xn xo xp xq xr xs xt xu xv xw xx
```

```
## 1  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
## 2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0
## 3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0
## 4  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 5  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## 6  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0
##   xy xz
## 1  0  0
## 2  0  0
## 3  0  0
## 4  0  0
## 5  0  0
## 6  0  0
```

We call **MatrixModels** for an implementation of `model.matrix` that supports sparse representations.

```
suppressPackageStartupMessages(library(MatrixModels))
X.2<- as(x,"sparseMatrix") %>% t # Makes sparse dummy model.matrix
head(X.2)
```

```
## 6 x 26 sparse Matrix of class "dgCMatrix"
```

```
##    [[ suppressing 26 column names 'a', 'b', 'c' ... ]]
```

```
##
## [1,] . . . . . . . . . . 1 . . . . . . . . . . . . . .
## [2,] . . . . . . . . . . . . . . . . . . . . 1 . . . . . . .
## [3,] . . . . . . . . . . . . . . . . . . . . 1 . . . . . . . .
## [4,] . 1 . . . . . . . . . . . . . . . . . . . . . . . . . .
## [5,] . . . . . . . 1 . . . . . . . . . . . . . . . . . . . .
## [6,] . . . . . . . . . . . . . . . . . . . . 1 . . . . . . . .
```

Notice that the matrices have the same dimensions:

```
dim(X.1)
```

```
## [1] 1000000      26
```

```
dim(X.2)
```

```
## [1] 1000000      26
```

The memory footprint of the matrices, given by the `pryr::object_size` function, are very very different.

```
pryr::object_size(X.1)
```

```
## 272 MB
```

```
pryr::object_size(X.2)
```

```
## 12 MB
```

Things to note:

- The sparse representation takes a whole lot less memory than the non sparse.
- The `as(,"sparseMatrix")` function grows the dummy variable representation of the factor `x`.
- The **pryr** package provides many facilities for inspecting the memory footprint of your objects and code.

With a sparse representation, we not only saved on RAM, but also on the computing time of fitting a model. Here is the timing of a non sparse representation:

```
system.time(lm.1 <- lm(y ~ X.1))
```

```
##    user  system elapsed
##   2.326   0.458   2.783
```

Well actually, `lm` is a wrapper for the `lm.fit` function. If we override all the overhead of `lm`, and call `lm.fit` directly, we gain some time:

```
system.time(lm.1 <- lm.fit(y=y, x=X.1))
```

```
##    user  system elapsed
##   0.948   0.076   1.025
```

We now do the same with the sparse representation:

```
system.time(lm.2 <- MatrixModels:::lm.fit.sparse(X.2,y))
```

```
##    user  system elapsed
##   0.177   0.004   0.181
```

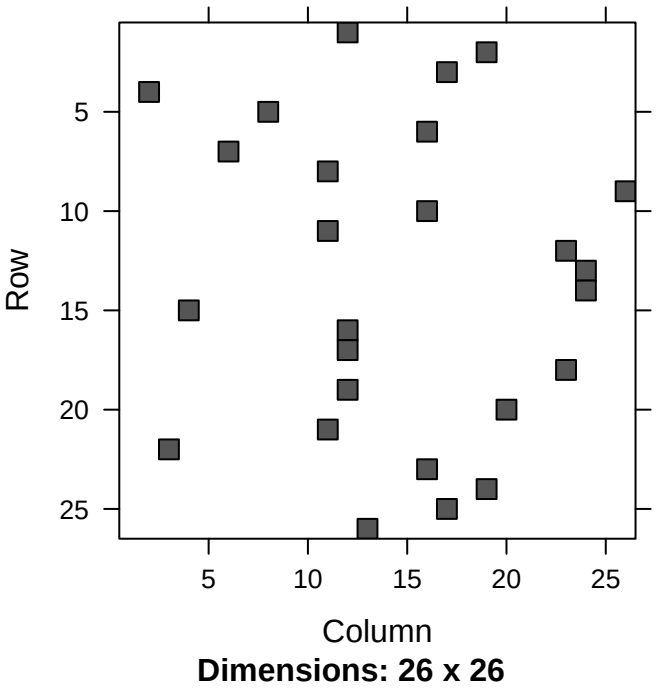
It is only left to verify that the returned coefficients are the same:

```
all.equal(lm.2, unname(lm.1$coefficients), tolerance = 1e-12)
```

```
## [1] TRUE
```

You can also visualize the non zero entries, i.e., the sparsity structure.

```
image(X.2[1:26,1:26])
```



14.1 Sparse Matrix Representations

We first distinguish between the two main goals of the efficient representation: (i) efficient writing, i.e., modification; (ii) efficient reading, i.e., access. For our purposes, we will typically want efficient reading, since the `model.matrix` will not change while a model is being fitted.

Representations designed for writing include the *dictionary of keys*, *list of lists*, and a *coordinate list*. Representations designed for efficient reading include the *compressed sparse row* and *compressed sparse column*.

14.1.1 Coordinate List Representation

A *coordinate list representation*, also known as *COO*, or *triplet represantation* is simply a list of the non zero entries. Each element in the list is a triplet of the row, column, and value, of each non-zero entry in the matrix. For instance the matrix

$$\begin{bmatrix} 0 & a_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix}$$

will be

$$\begin{bmatrix} 1 & 2 & a_2 \\ 2 & 3 & b_3 \end{bmatrix}.$$

14.1.2 Compressed Row Oriented Representation

Compressed row oriented representation, also known as *compressed sparse row*, or *CSR*. *CSR* is similar to *COO* with a compressed row vector. Instead of holding the row of each non-zero entry, the row vector holds the locations in the colum vector where a row is increased. See the next illustration.

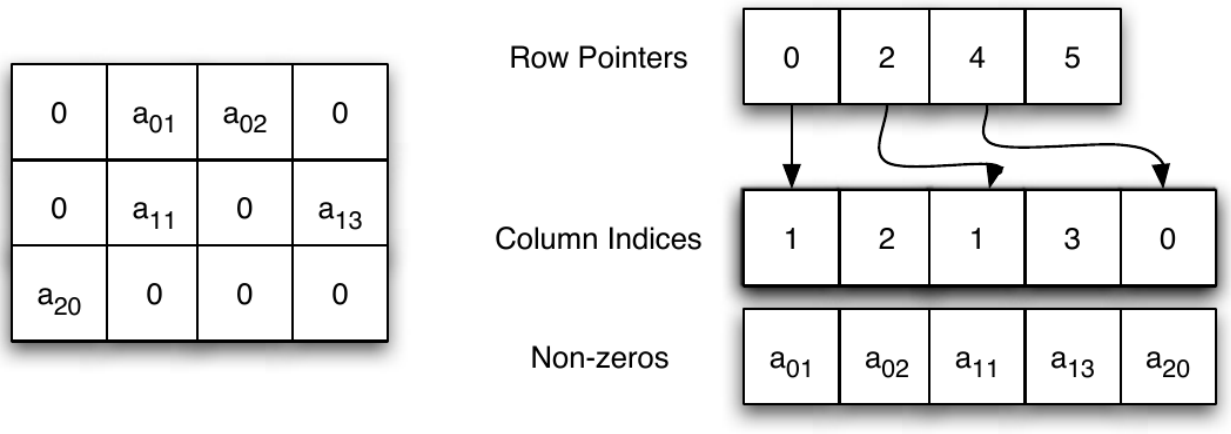


Figure 14.1: The CSR data structure. From Shah and Gilbert (2004). Remember that MATLAB is written in C, where the indexing starts at 0, and not 1.

14.1.3 Compressed Column Oriented Representation

A *compressed column oriented representation*, also known as *compressed sparse column*, or *CSC*. In *CSC* the column vector is compressed. Unlike *CSR* where the row vector is compressed. The nature of statistical applications is such, that CSC representation is typically the most economical, justifying its popularity.

14.1.4 Sparse Algorithms

We will go into the details of some algorithms in the Numerical Linear Algebra Chapter 17. For our current purposes two things need to be emphasized:

1. Working with sparse representations requires using a function that is aware of the representation you are using.
2. A mathematician may write $Ax = b \Rightarrow x = A^{-1}b$. This is a predicate of x,i.e., a property that x satisfies, which helps with its analysis. A computer, however, would **never** compute A^{-1} in order to find x, but rather use one of many endlessly many numerical algorithms. A computer will typically "search" various x's until it finds the one that fulfils the predicate.

14.2 Sparse Matrices and Sparse Models in R

14.2.1 The Matrix Package

The **Matrix** package provides facilities to deal with real (stored as double precision), logical and so-called "pattern" (binary) dense and sparse matrices. There are provisions to provide integer and complex (stored as double precision complex) matrices.

The sparse matrix classes include:

- `TsparseMatrix`: a virtual class of the various sparse matrices in triplet representation.
- `CsparseMatrix`: a virtual class of the various sparse matrices in CSC representation.
- `RsparseMatrix`: a virtual class of the various sparse matrices in CSR representation.

For matrices of real numbers, stored in *double precision*, the **Matrix** package provides the following (non virtual) classes:

- `dgTMatrix`: a **general** sparse matrix of **doubles**, in **triplet** representation.
- `dgCMatrix`: a **general** sparse matrix of **doubles**, in **CSC** representation.
- `dsCMatrix`: a **symmetric** sparse matrix of **doubles**, in **CSC** representation.
- `dtCMatrix`: a **triangular** sparse matrix of **doubles**, in **CSC** representation.

Why bother with distinguishing between the different shapes of the matrix? Because the more structure is assumed on a matrix, the more our (statistical) algorithms can be optimized. For our purposes `dgCMatrix` will be the most useful.

14.2.2 The glmnet Package

As previously stated, an efficient storage of the `model.matrix` is half of the story. We now need implementations of our favorite statistical algorithms that make use of this representation. At the time of writing, a very useful package that does that is the **glmnet** package, which allows to fit linear models, generalized linear models, with ridge, lasso, and elastic net regularization. The **glmnet** package allows all of this, using the sparse matrices of the **Matrix** package.

The following example is taken from John Myles White's blog, and compares the runtime of fitting an OLS model, using `glmnet` with both sparse and dense matrix representations.

```
suppressPackageStartupMessages(library('glmnet'))

set.seed(1)
performance <- data.frame()

for (sim in 1:10){
  n <- 10000
  p <- 500

  nzc <- trunc(p / 10)

  x <- matrix(rnorm(n * p), n, p) #make a dense matrix
  iz <- sample(1:(n * p),
               size = n * p * 0.85,
               replace = FALSE)
  x[iz] <- 0 # sparsify by injecting zeroes
  sx <- Matrix(x, sparse = TRUE) # save as a sparse object

  beta <- rnorm(nzc)
  fx <- x[, seq(nzc)] %*% beta

  eps <- rnorm(n)
  y <- fx + eps # make data
```

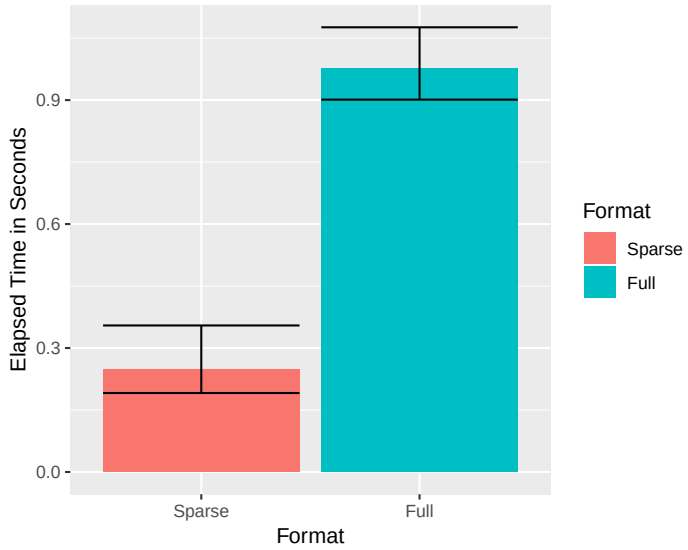
```
  # Now to the actual model fitting:
  sparse.times <- system.time(fit1 <- glmnet(sx, y)) # sparse glmnet
  full.times <- system.time(fit2 <- glmnet(x, y)) # dense glmnet

  sparse.size <- as.numeric(object.size(sx))
  full.size <- as.numeric(object.size(x))

  performance <- rbind(performance, data.frame(Format = 'Sparse',
                                                UserTime = sparse.times[1],
                                              SystemTime = sparse.times[2],
                                              ElapsedTime = sparse.times[3],
                                              Size = sparse.size))
  performance <- rbind(performance, data.frame(Format = 'Full',
                                              UserTime = full.times[1],
                                              SystemTime = full.times[2],
                                              ElapsedTime = full.times[3],
                                              Size = full.size))
}
```

Things to note:

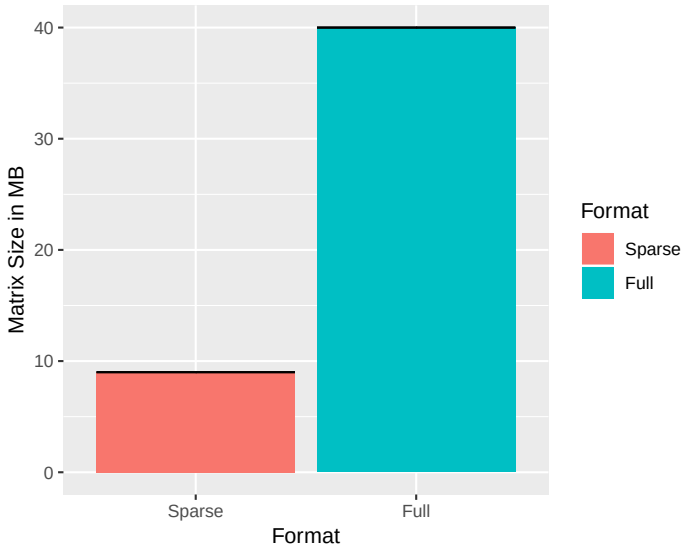
- The simulation calls `glmnet` twice. Once with the non-sparse object `x`, and once with its sparse version `sx`.
- The degree of sparsity of `sx` is 85%. We know this because we "injected" zeroes in 0.85 of the locations of `x`.
- Because `y` is continuous `glmnet` will fit a simple OLS model. We will see later how to use it to fit GLMs and use lasso, ridge, and elastic-net regularization.

We now inspect the computing time, and the memory footprint, only to discover that sparse representations make a BIG difference.

```
suppressPackageStartupMessages(library('ggplot2'))
ggplot(performance, aes(x = Format, y = ElapsedTime, fill = Format)) +
  stat_summary(fun.data = 'mean_cl_boot', geom = 'bar') +
  stat_summary(fun.data = 'mean_cl_boot', geom = 'errorbar') +
  ylab('Elapsed Time in Seconds')
```




```
ggplot(performance, aes(x = Format, y = Size / 1000000, fill = Format)) +
  stat_summary(fun.data = 'mean_cl_boot', geom = 'bar') +
  stat_summary(fun.data = 'mean_cl_boot', geom = 'errorbar') +
  ylab('Matrix Size in MB')
```



How do we perform other types of regression with the **glmnet**? We just need to use the `family` and `alpha` arguments of `glmnet::glmnet`. The `family` argument governs the type of GLM to fit: logistic, Poisson, probit, or other types of GLM. The `alpha` argument controls the type of regularization. Set to `alpha=0` for ridge, `alpha=1` for lasso, and any value in between for elastic-net regularization.

14.2.3 The MatrixModels Package

The MatrixModels package is designed to fit various models (linear, non-linear, generalized) using sparse matries. The function `MatrixModels::glm4` can easily replace `stats::glm` for all your needs. Unlike **glmnet**, the **MatrixModels** package will not offer you model regularization.

14.2.4 The SparseM Package

Basic linear algebra with sparse matrices.

14.3 Bibliographic Notes

The best place to start reading on sparse representations and algorithms is the vignettes of the **Matrix** package. Gilbert et al. (1992) is also a great read for some general background. For the theory on solving sparse linear systems see Davis (2006). For general numerical linear algebra see Gentle (2012).

14.4 Practice Yourself

1. What is the CSC representation of the following matrix:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 6 \\ 1 & 0 & 1 \end{bmatrix}$$

2. Write a function that takes two matrices in CSC and returns their matrix product.

Chapter 15

Memory Efficiency

As put by Kane et al. (2013), it was quite puzzling when very few of the competitors, for the Million dollars prize in the Netflix challenge, were statisticians. This is perhaps because the statistical community historically uses SAS, SPSS, and R. The first two tools are very well equipped to deal with big data, but are very unfriendly when trying to implement a new method. R, on the other hand, is very friendly for innovation, but was not equipped to deal with the large data sets of the Netflix challenge. A lot has changed in R since 2006. This is the topic of this chapter.

As we have seen in the Sparsity Chapter 14, an efficient representation of your data in RAM will reduce computing time, and will allow you to fit models that would otherwise require tremendous amounts of RAM. Not all problems are sparse however. It is also possible that your data does not fit in RAM, even if sparse. There are several scenarios to consider:

1. Your data fits in RAM, but is too big to compute with.
2. Your data does not fit in RAM, but fits in your local storage (HD, SSD, etc.)
3. Your data does not fit in your local storage.

If your data fits in RAM, but is too large to compute with, a solution is to replace the algorithm you are using. Instead of computing with the whole data, your algorithm will compute with parts of the data, also called *chunks*, or *batches*. These algorithms are known as *external memory algorithms* (EMA).

If your data does not fit in RAM, but fits in your local storage, you have two options. The first is to save your data in a *database management system* (DBMS). This will allow you to use the algorithms provided by your DBMS, or let R use an EMA while "chunking" from your DBMS. Alternatively, and preferably, you may avoid using a DBMS, and work with the data directly form your local storage by saving your data in some efficient manner.

Finally, if your data does not fit on you local storage, you will need some external storage solution such as a distributed DBMS, or distributed file system.
Remark. If you use Linux, you may be better of than Windows users. Linux will allow you to compute with larger datasets using its *swap file* that extends RAM using your HD or SSD. On the other hand, relying on the swap file is a BAD practice since it is much slower than RAM, and you can typically do much better using the tricks of this chapter. Also, while I LOVE Linux, I would never dare to recommend switching to Linux just to deal with memory contraints.

15.1 Efficient Computing from RAM

If our data can fit in RAM, but is still too large to compute with it (recall that fitting a model requires roughly 5-10 times more memory than saving it), there are several facilities to be used. The first, is the sparse representation discussed in Chapter 14, which is relevant when you have factors, which will typically map to sparse model matrices. Another way is to use *external memory algorithms* (EMA).

The `biglm::biglm` function provides an EMA for linear regression. The following if taken from the function's example.

```
data(trees)
ff<-log(Volume)~log(Girth)+log(Height)
```

```
chunk1<-trees[1:10,]
chunk2<-trees[11:20,]
chunk3<-trees[21:31,]

library(biglm)
a <- biglm(ff,chunk1)
a <- update(a,chunk2)
a <- update(a,chunk3)

coef(a)
```

```
## (Intercept)  log(Girth) log(Height)
##   -6.631617    1.982650    1.117123
```

Things to note:

- The data has been chunked along rows.
- The initial fit is done with the `biglm` function.
- The model is updated with further chunks using the `update` function.

We now compare it to the in-memory version of `lm` to verify the results are the same.

```
b <- lm(ff, data=trees)
rbind(coef(a),coef(b))
```

```
##      (Intercept) log(Girth) log(Height)
## [1,]   -6.631617    1.98265    1.117123
## [2,]   -6.631617    1.98265    1.117123
```

Other packages that follow these lines, particularly with classification using SVMs, are **LiblineaR**, and **RSofia**.

15.1.1 Summary Statistics from RAM

If you are not going to do any model fitting, and all you want is efficient filtering, selection and summary statistics, then a lot of my warnings above are irrelevant. For these purposes, the facilities provided by **base**, **stats**, and **dplyr** are probably enough. If the data is large, however, these facilities may be too slow. If your data fits into RAM, but speed bothers you, take a look at the **data.table** package. The syntax is less friendly than **dplyr**, but **data.table** is BLAZING FAST compared to competitors. Here is a little benchmark[1].

First, we setup the data.

```
library(data.table)

n <- 1e6 # number of rows
k <- c(200,500) # number of distinct values for each 'group_by' variable
p <- 3 # number of variables to summarize

L1 <- sapply(k, function(x) as.character(sample(1:x, n, replace = TRUE) ))
L2 <- sapply(1:p, function(x) rnorm(n) )

tbl <- data.table(L1,L2) %>%
  setnames(c(paste("v",1:length(k),sep=""), paste("x",1:p,sep="") ))

tbl_dt <- tbl
tbl_df <- tbl %>% as.data.frame
```

We compare the aggregation speeds. Here is the timing for **dplyr**.

[1] The code was contributed by Liad Shekel.

```
library(dplyr)
system.time( tbl_df %>%
               group_by(v1,v2) %>%
               summarize(
                 x1 = sum(abs(x1)),
                 x2 = sum(abs(x2)),
                 x3 = sum(abs(x3))
                 )
             )
```

```
##    user  system elapsed
##   1.457   0.032   1.489
```

And now the timing for **data.table**.

```
system.time(
  tbl_dt[ ,  .( x1 = sum(abs(x1)), x2 = sum(abs(x2)), x3 = sum(abs(x3)) ), .(v1,v2)]
  )
```

```
##    user  system elapsed
##   0.741   0.652   1.406
```

The winner is obvious. Let's compare filtering (i.e. row subsets, i.e. SQL's SELECT).

```
system.time(
  tbl_df %>% filter(v1 == "1")
  )
```

```
##    user  system elapsed
##   0.360   0.122   0.935
```

```
system.time(
  tbl_dt[v1 == "1"]
  )
```

```
##    user  system elapsed
##   0.023   0.095   0.117
```

15.2 Computing from a Database

The early solutions to oversized data relied on storing your data in some DBMS such as *MySQL*, *PostgresSQL*, *SQLite*, *H2*, *Oracle*, etc. Several R packages provide interfaces to these DBMSs, such as **sqldf**, **RDBI**, **RSQite**. Some will even include the DBMS as part of the package itself.

Storing your data in a DBMS has the advantage that you can typically rely on DBMS providers to include very efficient algorithms for the queries they support. On the downside, SQL queries may include a lot of summary statistics, but will rarely include model fitting[2]. This means that even for simple things like linear models, you will have to revert to R's facilities– typically some sort of EMA with chunking from the DBMS. For this reason, and others, we prefer to compute from efficient file structures, as described in Section 15.3.

If, however, you have a powerful DBMS around, or you only need summary statistics, or you are an SQL master, keep reading.

The package **RSQLite** includes an SQLite server, which we now setup for demonstration. The package **dplyr**, discussed in the Hadleyverse Chapter 23, will take care of translating the **dplyr** syntax, to the SQL syntax of the DBMS. The following example is taken from the **dplyr** Databases vignette.

```
library(RSQLite)
library(dplyr)
```

[2]This is slowly changing. Indeed, Microsoft's SQL Server 2016 is already providing in-database-analytics, and other will surely follow.

```
file.remove('my_db.sqlite3')
my_db <- src_sqlite(path = "my_db.sqlite3", create = TRUE)

library(nycflights13)
flights_sqlite <- copy_to(
  dest= my_db,
  df= flights,
  temporary = FALSE,
  indexes = list(c("year", "month", "day"), "carrier", "tailnum"))
```

Things to note:

- `src_sqlite` to start an empty table, managed by SQLite, at the desired path.
- `copy_to` copies data from R to the database.
- Typically, setting up a DBMS like this makes no sense, since it requires loading the data into RAM, which is precisely what we want to avoid.

We can now start querying the DBMS.

```
select(flights_sqlite, year:day, dep_delay, arr_delay)
```

```
filter(flights_sqlite, dep_delay > 240)
```

15.3 Computing From Efficient File Structrures

It is possible to save your data on your storage device, without the DBMS layer to manage it. This has several advantages:

- You don't need to manage a DBMS.
- You don't have the computational overhead of the DBMS.
- You may optimize the file structure for statistical modelling, and not for join and summary operations, as in relational DBMSs.

There are several facilities that allow you to save and compute directly from your storage:

1. **Memory Mapping**: Where RAM addresses are mapped to a file on your storage. This extends the RAM to the capacity of your storage (HD, SSD,...). Performance slightly deteriorates, but the access is typically very fast. This approach is implemented in the **bigmemory** package.

2. **Efficient Binaries**: Where the data is stored as a file on the storage device. The file is binary, with a well designed structure, so that chunking is easy. This approach is implemented in the **ff** package, and the commercial **RevoScaleR** package.

Your algorithms need to be aware of the facility you are using. For this reason each facility (**bigmemory**, **ff**, **RevoScaleR**,...) has an eco-system of packages that implement various statistical methods using that facility. As a general rule, you can see which package builds on a package using the *Reverse Depends* entry in the package description. For the **bigmemory** package, for instance, we can see that the packages **bigalgebra**, **biganalytics**, **bigFastlm**, **biglasso**, **bigpca**, **bigtabulate**, **GHap**, and **oem**, build upon it. We can expect this list to expand.

Here is a benchmark result, from Wang et al. (2015). It can be seen that **ff** and **bigmemory** have similar performance, while **RevoScaleR** (RRE in the figure) outperforms them. This has to do both with the efficiency of the binary representation, but also because **RevoScaleR** is inherently parallel. More on this in the Parallelization Chapter 16.

	Reading	Transforming	Fitting
bigmemory	968.6	105.5	1501.7
ff	1111.3	528.4	1988.0
RRE	851.7	107.5	189.4

15.3.1 bigmemory

We now demonstrate the workflow of the **bigmemory** package. We will see that **bigmemory**, with it's `big.matrix` object is a very powerful mechanism. If you deal with big numeric matrices, you will find it very useful. If you deal with big data frames, or any other non-numeric matrix, **bigmemory** may not be the appropriate tool, and you should try **ff**, or the commercial **RevoScaleR**.

```
# download.file("http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAP
# unzip(zipfile="2010_Carrier_PUF.zip")

library(bigmemory)
x <- read.big.matrix("data/2010_BSA_Carrier_PUF.csv", header = TRUE,
                     backingfile = "airline.bin",
                     descriptorfile = "airline.desc",
                     type = "integer")
dim(x)
```

```
## [1] 2801660      11
```

```
pryr::object_size(x)
```

```
## 696 B
```

```
class(x)
```

```
## [1] "big.matrix"
## attr(,"package")
## [1] "bigmemory"
```

Things to note:

- The basic building block of the **bigmemory** ecosystem, is the `big.matrix` class, we constructed with `read.big.matrix`.
- `read.big.matrix` handles the import to R, and the saving to a memory mapped file. The implementation is such that at no point does R hold the data in RAM.
- The memory mapped file will be there after the session is over. It can thus be called by other R sessions using `attach.big.matrix("airline.desc")`. This will be useful when parallelizing.
- `pryr::object_size` return the size of the object. Since `x` holds only the memory mappings, it is much smaller than the 100MB of data that it holds.

We can now start computing with the data. Many statistical procedures for the `big.matrix` object are provided by the **biganalytics** package. In particular, the `biglm.big.matrix` and `bigglm.big.matrix` functions, provide an interface from `big.matrix` objects, to the EMA linear models in `biglm::biglm` and `biglm::bigglm`.

```
library(biganalytics)
biglm.2 <- bigglm.big.matrix(BENE_SEX_IDENT_CD~CAR_LINE_HCPCS_CD, data=x)
coef(biglm.2)
```

```
##       (Intercept) CAR_LINE_HCPCS_CD
##      1.537848e+00      1.210282e-07
```

Other notable packages that operate with `big.matrix` objects include:

- **bigtabulate**: Extend the bigmemory package with 'table', 'tapply', and 'split' support for 'big.matrix' objects.
- **bigalgebra**: For matrix operation.
- **bigpca**: principle components analysis (PCA), or singular value decomposition (SVD).
- **bigFastlm**: for (fast) linear models.
- **biglasso**: extends lasso and elastic nets.
- **GHap**: Haplotype calling from phased SNP data.

15.3.2 bigstep

The bigstep package uses the **bigmemory** framework to perform stepwise model selction, when the data cannot fit into RAM.

TODO

15.4 ff

The **ff** packages replaces R's in-RAM storage mechanism with on-disk (efficient) storage. Unlike **bigmemory**, **ff** supports all of R vector types such as factors, and not only numeric. Unlike `big.matrix`, which deals with (numeric) matrices, the `ffdf` class can deal with data frames.

Here is an example. First open a connection to the file, without actually importing it using the `LaF::laf_open_csv` function.

```
.dat <- LaF::laf_open_csv(filename = "data/2010_BSA_Carrier_PUF.csv",
                    column_types = c("integer", "integer", "categorical", "categorical", "categorical", "int
                    column_names = c("sex", "age", "diagnose", "healthcare.procedure", "typeofservice", "ser
                    skip = 1)
```

Now write the data to local storage as an ff data frame, using `laf_to_ffdf`.

```
data.ffdf <- ffbase::laf_to_ffdf(laf = .dat)
head(data.ffdf)
```

```
## ffdf (all open) dim=c(2801660,6), dimorder=c(1,2) row.names=NULL
## ffdf virtual mapping
##                              PhysicalName VirtualVmode PhysicalVmode  AsIs
## sex                                   sex      integer       integer FALSE
## age                                   age      integer       integer FALSE
## diagnose                         diagnose      integer       integer FALSE
## healthcare.procedure healthcare.procedure      integer       integer FALSE
## typeofservice               typeofservice      integer       integer FALSE
## service.count               service.count      integer       integer FALSE
##                      VirtualIsMatrix PhysicalIsMatrix PhysicalElementNo
## sex                            FALSE            FALSE                 1
## age                            FALSE            FALSE                 2
## diagnose                       FALSE            FALSE                 3
## healthcare.procedure           FALSE            FALSE                 4
## typeofservice                  FALSE            FALSE                 5
## service.count                  FALSE            FALSE                 6
##                      PhysicalFirstCol PhysicalLastCol PhysicalIsOpen
## sex                                 1               1           TRUE
## age                                 1               1           TRUE
## diagnose                            1               1           TRUE
## healthcare.procedure                1               1           TRUE
## typeofservice                       1               1           TRUE
## service.count                       1               1           TRUE
```

```
## ffdf data
##           sex   age diagnose healthcare.procedure typeofservice
## 1       1     1        NA                   99213         M1B
## 2       1     1        NA                   A0425         O1A
## 3       1     1        NA                   A0425         O1A
## 4       1     1        NA                   A0425         O1A
## 5       1     1        NA                   A0425         O1A
## 6       1     1        NA                   A0425         O1A
## 7       1     1        NA                   A0425         O1A
## 8       1     1        NA                   A0425         O1A
## :           :     :        :                    :             :
## 2801653 2     6        V82                  85025         T1D
## 2801654 2     6        V82                  87186         T1H
## 2801655 2     6        V82                  99213         M1B
## 2801656 2     6        V82                  99213         M1B
## 2801657 2     6        V82                  A0429         O1A
## 2801658 2     6        V82                  G0328         T1H
## 2801659 2     6        V86                  80053         T1B
## 2801660 2     6        V88                  76856         I3B
##         service.count
## 1                   1
## 2                   1
## 3                   1
## 4                   2
## 5                   2
## 6                   3
## 7                   3
## 8                   4
## :                   :
## 2801653             1
## 2801654             1
## 2801655             1
## 2801656             1
## 2801657             1
## 2801658             1
## 2801659             1
## 2801660             1
```

We can verify that the `ffdf` data frame has a small RAM footprint.

```
pryr::object_size(data.ffdf)
```

```
## 392 kB
```

The **ffbase** package provides several statistical tools to compute with `ff` class objects. Here is simple table.

```
ffbase::table.ff(data.ffdf$age)
```

```
##
##      1      2      3      4      5      6
## 517717 495315 492851 457643 419429 418705
```

The EMA implementation of `biglm::biglm` and `biglm::bigglm` have their **ff** versions.

```
library(biglm)
mymodel.ffdf <- biglm(payment ~ factor(sex) + factor(age) + place.served,
                      data = data.ffdf)
summary(mymodel.ffdf)
```

```
## Large data regression model: biglm(payment ~ factor(sex) + factor(age) + place.served, data = data.ffdf)
## Sample size =  2801660
```

```
##                 Coef    (95%     CI)     SE      p
## (Intercept)  97.3313 96.6412 98.0214 0.3450 0.0000
## factor(sex)2 -4.2272 -4.7169 -3.7375 0.2449 0.0000
## factor(age)2  3.8067  2.9966  4.6168 0.4050 0.0000
## factor(age)3  4.5958  3.7847  5.4070 0.4056 0.0000
## factor(age)4  3.8517  3.0248  4.6787 0.4135 0.0000
## factor(age)5  1.0498  0.2030  1.8965 0.4234 0.0132
## factor(age)6 -4.8313 -5.6788 -3.9837 0.4238 0.0000
## place.served -0.6132 -0.6253 -0.6012 0.0060 0.0000
```

Things to note:

- `biglm::biglm` notices the input of of class `ffdf` and calls the appropriate implementation.
- The model formula, `payment ~ factor(sex) + factor(age) + place.served`, includes factors which cause no difficulty.
- You cannot inspect the factor coding (dummy? effect?) using `model.matrix`. This is because EMAs never really construct the whole matrix, let alone, save it in memory.

15.5 matter

Memory-efficient reading, writing, and manipulation of structured binary data on disk as vectors, matrices, arrays, lists, and data frames.

TODO

15.6 iotools

A low level facility for connecting to on-disk binary storage. Unlike **ff**, and **bigmemory**, it behaves like native R objects, with their copy-on-write policy. Unlike **reader**, it allows chunking. Unlike `read.csv`, it allows fast I/O. **iotools** is thus a potentially very powerfull facility. See Arnold et al. (2015) for details.

TODO

15.7 HDF5

Like **ff**, HDF5 is an on-disk efficient file format. The package **h5** is interface to the "HDF5" library supporting fast storage and retrieval of R-objects like vectors, matrices and arrays.

TODO

15.8 DelayedArray

An abstraction layer for operations on array objects, which supports various backend storage of arrays such as:

- In RAM: base, Matrix, DelayedArray.
- In Disk: HDF5Array, matterArray.

Link Several application packages already build upon the **DelayedArray** pacakge:

- DelayedMatrixStats: Functions that Apply to Rows and Columns of **DelayedArray** Objects.
- beachmat C++ API for (most) **DelayedMatrix** backends.

15.9 Computing from a Distributed File System

If your data is SOOO big that it cannot fit on your local storage, you will need a distributed file system or DBMS. We do not cover this topic here, and refer the reader to the **RHipe**, **RHadoop**, and **RSpark** packages and references therein.

15.10 Bibliographic Notes

An absolute SUPERB review on computing with big data is Wang et al. (2015), and references therein (Kane et al. (2013) in particular). Here is also an excellent talk by Charles DiMaggio. For an up-to-date list of the packages that deal with memory constraints, see the **Large memory and out-of-memory data** section in the High Performance Computing task view. For a list of resources to interface to DMBS, see the Databases with R task view. For more on data analysis from disk, and not from RAM, see Peter_Hickey's JSM talk.

15.11 Practice Yourself

Chapter 16

Parallel Computing

You would think that because you have an expensive multicore computer your computations will speed up. Well, no. At least not if you don't make sure they do. By default, no matter how many cores you have, the operating system will allocate each R session to a single core.

For starters, we need to distinguish between two types of parallelism:

1. **Explicit parallelism**: where the user handles the parallelisation.
2. **Implicit parallelism**: where the parallelisation is abstracted away from the user.

Clearly, implicit parallelism is more desirable, but the state of mathematical computing is such that no sufficiently general implicit parallelism framework exists. The R Consortium is currently financing a major project for a A "Unified Framework For Distributed Computing in R" so we can expect things to change soon. In the meanwhile, most of the parallel implementations are explicit.

16.1 Implicit Parallelism

We will not elaborate on implicit parallelism except mentioning the following:

- You can enjoy parallel linear algebra by replacing the linear algebra libraries with BLAS and LAPACK as described here.
- You should read the "Parallel computing: Implicit parallelism" section in the excellent High Performance Computing task view, for the latest developments in implicit parallelism.

16.2 Explicit Parallelism

R provides many frameworks for explicit parallelism. Because the parallelism is initiated by the user, we first need to decide **when to parallelize?** As a rule of thumb, you want to parallelise when you encounter a CPU bottleneck, and not a memory bottleneck. Memory bottlenecks are released with sparsity (Chapter 14), or efficient memory usage (Chapter 15).

Several ways to diagnose your bottleneck include:

- Keep your Windows Task Manager, or Linux `top` open, and look for the CPU load, and RAM loads.
- The computation takes a long time, and when you stop it pressing ESC, R is immediately responsive. If it is not immediately responsive, you have a memory bottleneck.
- Profile your code. See Hadley's guide.

For reasons detailed in Kane et al. (2013), we will present the **foreach** parallelisation package (Analytics and Weston, 2015). It will allow us to:

1. Decouple between our parallel algorithm and the parallelisation mechanism: we write parallelisable code once, and can then switch the underlying parallelisation mechanism.

2. Combine with the `big.matrix` object from Chapter 15 for *shared memory parallisation*: all the machines may see the same data, so that we don't need to export objects from machine to machine.

What do we mean by "switch the underlying parallesation mechanism"? It means there are several packages that will handle communication between machines. Some are very general and will work on any cluster. Some are more specific and will work only on a single multicore machine (not a cluster) with a particular operating system. These mechanisms include **multicore**, **snow**, **parallel**, and **Rmpi**. The compatibility between these mechanisms and **foreach** is provided by another set of packages: **doMC** , **doMPI**, **doRedis**, **doParallel**, and **doSNOW**.
Remark. I personally prefer the **multicore** mechanism, with the **doMC** adapter for **foreach**. I will not use this combo, however, because **multicore** will not work on Windows machines. I will thus use the more general **snow** and **doParallel** combo. If you do happen to run on Linux, or Unix, you will want to replace all **doParallel** functionality with **doMC**.

Let's start with a simple example, taken from "Getting Started with doParallel and foreach".

```
library(doParallel)
cl <- makeCluster(2)
registerDoParallel(cl)
result <- foreach(i=1:3) %dopar% sqrt(i)
class(result)
```

```
## [1] "list"
```

```
result
```

```
## [[1]]
## [1] 1
##
## [[2]]
## [1] 1.414214
##
## [[3]]
## [1] 1.732051
```

Things to note:

- `makeCluster` creates an object with the information our cluster. On a single machine it is very simple. On a cluster of machines, you will need to specify the i.p. addresses or other identifiers of the machines.
- `registerDoParallel` is used to inform the **foreach** package of the presence of our cluster.
- The `foreach` function handles the looping. In particular note the `%dopar` operator that ensures that looping is in parallel. `%dopar%` can be replaced by `%do%` if you want serial looping (like the `for` loop), for instance, for debugging.
- The output of the various machines is collected by `foreach` to a list object.
- In this simple example, no data is shared between machines so we are not putting the shared memory capabilities to the test.
- We can check how many workers were involved using the `getDoParWorkers()` function.
- We can check the parallelisation mechanism used with the `getDoParName()` function.

Here is a more involved example. We now try to make Bootstrap inference on the coefficients of a logistic regression. Bootstrapping means that in each iteration, we resample the data, and refit the model.

```
x <- iris[which(iris[,5] != "setosa"), c(1,5)]
trials <- 1e4
ptime <- system.time({
 r <- foreach(icount(trials), .combine=cbind) %dopar% {
 ind <- sample(100, 100, replace=TRUE)
 result1 <- glm(x[ind,2]~x[ind,1], family=binomial(logit))
 coefficients(result1)
 }
 })[3]
ptime
```

```
## elapsed
```

```
##  11.499
```

Things to note:

- As usual, we use the `foreach` function with the `%dopar%` operator to loop in parallel.
- The `icounts` function generates a counter.
- The `.combine=cbind` argument tells the `foreach` function how to combine the output of different machines, so that the returned object is not the default list.

How long would that have taken in a simple (serial) loop? We only need to replace `%dopar%` with `%do%` to test.

```
stime <- system.time({
 r <- foreach(icount(trials), .combine=cbind) %do% {
 ind <- sample(100, 100, replace=TRUE)
 result1 <- glm(x[ind,2]~x[ind,1], family=binomial(logit))
 coefficients(result1)
 }
 })[3]
stime
```

```
## elapsed
##  22.171
```

Yes. Parallelising is clearly faster.

Let's see how we can combine the power of **bigmemory** and **foreach** by creating a file mapped `big.matrix` object, which is shared by all machines. The following example is taken from Kane et al. (2013), and uses the `big.matrix` object we created in Chapter 15.

```
library(bigmemory)
x <- attach.big.matrix("airline.desc")

library(foreach)
library(doSNOW)
cl <- makeSOCKcluster(rep("localhost", 4)) # make a cluster of 4 machines
registerDoSNOW(cl) # register machines for foreach()
```

Get a "description" of the `big.matrix` object that will be used to call it from each machine.

```
xdesc <- describe(x)
```

Split the data along values of `BENE_AGE_CAT_CD`.

```
G <- split(1:nrow(x), x[, "BENE_AGE_CAT_CD"])
```

Define a function that computes quantiles of `CAR_LINE_ICD9_DGNS_CD`.

```
GetDepQuantiles <- function(rows, data) {
 quantile(data[rows, "CAR_LINE_ICD9_DGNS_CD"], probs = c(0.5, 0.9, 0.99),
 na.rm = TRUE)
}
```

We are all set up to loop, in parallel, and compute quantiles of `CAR_LINE_ICD9_DGNS_CD` for each value of `BENE_AGE_CAT_CD`.

```
qs <- foreach(g = G, .combine = rbind) %dopar% {
 require("bigmemory")
 x <- attach.big.matrix(xdesc)
 GetDepQuantiles(rows = g, data = x)
}
qs
```

```
##          50% 90% 99%
## result.1 558 793 996
## result.2 518 789 996
```

```
## result.3 514 789 996
## result.4 511 789 996
## result.5 511 790 996
## result.6 518 796 995
```

16.2.1 Caution: Implicit with Explicit Parallelism

A common problem when parallelizing is that the machines you invoked with **explicit** parallelism, invoke other machines using **implicit** parallelism. You then lose control of the number of machine being invoked, and the operating system spends most of its time managing resources, instead of doing your computations.

Modern linear algebra libraries may try to implicitly parallelize vector and matrix operations. To avoid this type of implicity parallelism within an explicit paralleliation, use the RhpcBLASctl package, and in particular *evaRhpcBLASctl::blas_set_num_threads()*. In other cases, you should consult an expert.

16.3 Bibliographic Notes

For a brief and excellent explanation on parallel computing in R see Schmidberger et al. (2009). For a full review see Chapple et al. (2016). For a blog-level introduction see ParallelR. For an up-to-date list of packages supporting parallel programming see the High Performance Computing R task view.

16.4 Practice Yourself

Chapter 17

Numerical Linear Algebra

In your algebra courses you would write $Ax = b$ and solve $x = A^{-1}b$. This is useful to understand the algebraic properties of x, but a computer would never recover x that way. Even the computation of the sample variance, $S^2(x) = (n-1)^{-1}\sum(x_i - \bar{x})^2$ is not solved that way in a computer, because of numerical and speed considerations.

In this chapter, we discuss several ways a computer solves systems of linear equations, with their application to statistics, namely, to OLS problems.

17.1 LU Factorization

Definition 17.1 (LU Factorization)**.** For some matrix A, the LU factorization is defined as

$$A = LU \tag{17.1}$$

where L is unit lower triangular and U is upper triangular.

The LU factorization is essentially the matrix notation for the Gaussian elimination you did in your introductory algebra courses.

For a square $n \times n$ matrix, the LU factorization requires $n^3/3$ operations, and stores $n^2 + n$ elements in memory.

17.2 Cholesky Factorization

Definition 17.2 (Non Negative Matrix)**.** A matrix A is said to be *non-negative* if $x'Ax \geq 0$ for all x.

Seeing the matrix A as a function, non-negative matrices can be thought of as functions that generalize the *squaring* operation.

Definition 17.3 (Cholesky Factorization)**.** For some non-negative matrix A, the Cholesky factorization is defined as

$$A = T'T \tag{17.2}$$

where T is upper triangular with positive diagonal elements.

For obvious reasons, the Cholesky factorization is known as the *square root* of a matrix.

Because Cholesky is less general than LU, it is also more efficient. It can be computed in $n^3/6$ operations, and requires storing $n(n+1)/2$ elements.

17.3 QR Factorization

Definition 17.4 (QR Factorization)**.** For some matrix A, the QR factorization is defined as

$$A = QR \tag{17.3}$$

where Q is orthogonal and R is upper triangular.

The QR factorization is very useful to solve the OLS problem as we will see in 17.6. The QR factorization takes $2n^3/3$ operations to compute. Three major methods for computing the QR factorization exist. These rely on *Householder transformations*, *Givens transformations*, and a (modified) *Gram-Schmidt procedure* (Gentle, 2012).

17.4 Singular Value Factorization

Definition 17.5 (SVD)**.** For an arbitrary $n \times m$ matrix A, the *singular valued decomposition* (SVD), is defined as

$$A = U\Sigma V' \tag{17.4}$$

where U is an orthonormal $n \times n$ matrix, V is an $m \times m$ orthonormal matrix, and Σ is diagonal.

The SVD factorization is very useful for algebraic analysis, but less so for computations. This is because it is (typically) solved via the QR factorization.

17.5 Iterative Methods

The various matrix factorizations above may be used to solve a system of linear equations, and in particular, the OLS problem. There is, however, a very different approach to solving systems of linear equations. This approach relies on the fact that solutions of linear systems of equations, can be cast as optimization problems: simply find x by minimizing $\|Ax - b\|$.

Some methods for solving (convex) optimization problems are reviewed in the Convex Optimization Chapter 18. For our purposes we will just mention that historically (this means in the `lm` function, and in the LAPACK numerical libraries) the factorization approach was preferred, and now optimization approaches are preferred. This is because the optimization approach is more numerically stable, and easier to parallelize.

17.6 Solving the OLS Problem

Recalling the OLS problem in Eq.(6.5): we wish to find β such that

$$\hat{\beta} := argmin_\beta\{\|y - X\beta\|_2^2\}. \tag{17.5}$$

The solution, $\hat{\beta}$ that solves this problem has to satisfy

$$X'X\beta = X'y. \tag{17.6}$$

Eq.(17.6) are known as the *normal equations*. The normal equations are the link between the OLS problem, and the matrix factorization discussed above.

Using the QR decomposition in the normal equations we have that

$$\hat{\beta} = R^{-1}_{(1:p,1:p)} y,$$

where $(R_{n\times p}) = (R_{(1:p,1:p)}, 0_{(p+1:n,1:p)})$ is the

17.7 Numerical Libraries for Linear Algebra

TODO. In the meanwhile: comparison of numerical libraries; installing MKL in Ubnutu; how to speed-up linear algebra in R; and another; install Open-Blas;

17.7.1 OpenBlas

17.7.2 MKL

17.8 Bibliographic Notes

For an excellent introduction to numerical algorithms in statistics, see Weihs et al. (2013). For an emphasis on numerical linear algebra, see Gentle (2012), and Golub and Van Loan (2012).

17.9 Practice Yourself

Chapter 18

Convex Optimization

TODO

18.1 Theoretical Backround

18.2 Optimizing with R

18.2.1 The optim Function

18.2.2 The nloptr Package

18.2.3 minqa Package

18.3 Bibliographic Notes

Task views

18.4 Practice Yourself

Chapter 19

RCpp

19.1 Bibliographic Notes

19.2 Practice Yourself

Chapter 20

Debugging Tools

TODO. In the meanwhile, get started with Wickham (2011), and get pro with Cotton (2017).

20.1 Bibliographic Notes

20.2 Practice Yourself

Chapter 21

Econometrics

TODO

- VAR
- Robust and Clustered SEs.
- GEE
- TSLS

21.1 Bibliographic Notes

21.2 Practice Yourself

Chapter 22

Psychometrics

TODO

22.1 Bibliographic Notes

22.2 Practice Yourself

Chapter 23

The Hadleyverse

The *Hadleyverse*, short for "Hadley Wickham's universe", is a set of packages that make it easier to handle data. If you are developing packages, you should be careful since using these packages may create many dependencies and compatibility issues. If you are analyzing data, and the portability of your functions to other users, machines, and operating systems is not of a concern, you will LOVE these packages. The term Hadleyverse refers to **all** of Hadley's packages, but here, we mention only a useful subset, which can be collectively installed via the **tidyverse** package:

- **ggplot2** for data visualization. See the Plotting Chapter 12.
- **dplyr** for data manipulation.
- **tidyr** for data tidying.
- **readr** for data import.
- **stringr** for character strings.
- **anytime** for time data.

23.1 readr

The **readr** package (Wickham et al., 2016) replaces base functions for importing and exporting data such as `read.table`. It is faster, with a cleaner syntax.

We will not go into the details and refer the reader to the official documentation here and the R for data sciecne book.

23.2 dplyr

When you think of data frame operations, think **dplyr** (Wickham and Francois, 2016). Notable utilities in the package include:

- `select()` Select columns from a data frame.
- `filter()` Filter rows according to some condition(s).
- `arrange()` Sort / Re-order rows in a data frame.
- `mutate()` Create new columns or transform existing ones.
- `group_by()` Group a data frame by some factor(s) usually in conjunction to summary.
- `summarize()` Summarize some values from the data frame or across groups.
- `inner_join(x,y,by="col")`return all rows from 'x' where there are matching values in 'x', and all columns from 'x' and 'y'. If there are multiple matches between 'x' and 'y', all combination of the matches are returned.
- `left_join(x,y,by="col")` return all rows from 'x', and all columns from 'x' and 'y'. Rows in 'x' with no match in 'y' will have 'NA' values in the new columns. If there are multiple matches between 'x' and 'y', all combinations of the matches are returned.
- `right_join(x,y,by="col")` return all rows from 'y', and all columns from 'x' and y. Rows in 'y' with no match in 'x' will have 'NA' values in the new columns. If there are multiple matches between 'x' and 'y', all combinations of the matches are returned.

- `anti_join(x,y,by="col")` return all rows from 'x' where there are not matching values in 'y', keeping just columns from 'x'.

The following example involve `data.frame` objects, but **dplyr** can handle other classes. In particular `data.tables` from the **data.table** package (Dowle and Srinivasan, 2017), which is designed for very large data sets.

dplyr can work with data stored in a database. In which case, it will convert your command to the appropriate SQL syntax, and issue it to the database. This has the advantage that (a) you do not need to know the specific SQL implementation of your database, and (b), you can enjoy the optimized algorithms provided by the database supplier. For more on this, see the databses vignette.

The following examples are taken from Kevin Markham. The `nycflights13::flights` has delay data for US flights.

```
library(nycflights13)
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515         2      830
##  2  2013     1     1      533            529         4      850
##  3  2013     1     1      542            540         2      923
##  4  2013     1     1      544            545        -1     1004
##  5  2013     1     1      554            600        -6      812
##  6  2013     1     1      554            558        -4      740
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

The data is of class `tbl_df` which is an extension of the `data.frame` class, designed for large data sets. Notice that the printing of `flights` is short, even without calling the `head` function. This is a feature of the `tbl_df` class (`print(data.frame)` would try to load all the data, thus take a long time).

```
class(flights) # a tbl_df is an extension of the data.frame class
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

Let's filter the observations from the first day of the first month. Notice how much better (i.e. readable) is the **dplyr** syntax, with piping, compared to the basic syntax.

```
flights[flights$month == 1 & flights$day == 1, ] # old style

library(dplyr)
filter(flights, month == 1, day == 1) #dplyr style
flights %>% filter(month == 1, day == 1) # dplyr with piping.
```

More filtering.

```
filter(flights, month == 1 | month == 2) # First OR second month.
slice(flights, 1:10) # selects first ten rows.

arrange(flights, year, month, day) # sort
arrange(flights, desc(arr_delay)) # sort descending

select(flights, year, month, day) # select columns year, month, and day
select(flights, year:day) # select column range
select(flights, -(year:day)) # drop columns
rename(flights, c(tail_num = "tailnum")) # rename column
```

```
# add a new computed colume
mutate(flights,
  gain = arr_delay - dep_delay,
  speed = distance / air_time * 60)

# you can refer to columns you just created! (gain)
mutate(flights,
  gain = arr_delay - dep_delay,
  gain_per_hour = gain / (air_time / 60)
)

# keep only new variables, not all data frame.
transmute(flights,
  gain = arr_delay - dep_delay,
  gain_per_hour = gain / (air_time / 60)
)

# simple statistics
summarise(flights,
  delay = mean(dep_delay, na.rm = TRUE)
  )
# random subsample
sample_n(flights, 10)
sample_frac(flights, 0.01)
```

We now perform operations on subgroups. we group observations along the plane's tail number (`tailnum`), and compute the count, average distance traveled, and average delay. We group with `group_by`, and compute subgroup statistics with `summarise`.

```
by_tailnum <- group_by(flights, tailnum)

delay <- summarise(by_tailnum,
  count = length(),
  avg.dist = mean(distance, na.rm = TRUE),
  avg.delay = mean(arr_delay, na.rm = TRUE))

delay
```

We can group along several variables, with a hierarchy. We then collapse the hierarchy one by one.

```
daily <- group_by(flights, year, month, day)
per_day   <- summarise(daily, flights = n())
per_month <- summarise(per_day, flights = sum(flights))
per_year  <- summarise(per_month, flights = sum(flights))
```

Things to note:

- Every call to `summarise` collapses one level in the hierarchy of grouping. The output of `group_by` recalls the hierarchy of aggregation, and collapses along this hierarchy.

We can use **dplyr** for two table operations, i.e., *joins*. For this, we join the flight data, with the airplane data in `airplanes`.

```
library(dplyr)
airlines
```

```
## # A tibble: 16 x 2
##    carrier name
##    <chr>   <chr>
##  1 9E      Endeavor Air Inc.
```

```
##  2 AA      American Airlines Inc.
##  3 AS      Alaska Airlines Inc.
##  4 B6      JetBlue Airways
##  5 DL      Delta Air Lines Inc.
##  6 EV      ExpressJet Airlines Inc.
##  7 F9      Frontier Airlines Inc.
##  8 FL      AirTran Airways Corporation
##  9 HA      Hawaiian Airlines Inc.
## 10 MQ      Envoy Air
## 11 OO      SkyWest Airlines Inc.
## 12 UA      United Air Lines Inc.
## 13 US      US Airways Inc.
## 14 VX      Virgin America
## 15 WN      Southwest Airlines Co.
## 16 YV      Mesa Airlines Inc.
```

```
# select the subset of interesting flight data.
flights2 <- flights %>% select(year:day, hour, origin, dest, tailnum, carrier)

# join on left table with automatic matching.
flights2 %>% left_join(airlines)
```

```
## Joining, by = "carrier"
```

```
## # A tibble: 336,776 x 9
##     year month   day  hour origin dest  tailnum carrier name
##    <int> <int> <int> <dbl> <chr>  <chr> <chr>   <chr>   <chr>
##  1  2013     1     1     5 EWR    IAH   N14228  UA      United Air Lines I~
##  2  2013     1     1     5 LGA    IAH   N24211  UA      United Air Lines I~
##  3  2013     1     1     5 JFK    MIA   N619AA  AA      American Airlines ~
##  4  2013     1     1     5 JFK    BQN   N804JB  B6      JetBlue Airways
##  5  2013     1     1     6 LGA    ATL   N668DN  DL      Delta Air Lines In~
##  6  2013     1     1     5 EWR    ORD   N39463  UA      United Air Lines I~
##  7  2013     1     1     6 EWR    FLL   N516JB  B6      JetBlue Airways
##  8  2013     1     1     6 LGA    IAD   N829AS  EV      ExpressJet Airline~
##  9  2013     1     1     6 JFK    MCO   N593JB  B6      JetBlue Airways
## 10  2013     1     1     6 LGA    ORD   N3ALAA  AA      American Airlines ~
## # ... with 336,766 more rows
```

```
flights2 %>% left_join(weather)
```

```
## Joining, by = c("year", "month", "day", "hour", "origin")
```

```
## # A tibble: 336,776 x 18
##     year month   day  hour origin dest  tailnum carrier  temp  dewp humid
##    <dbl> <dbl> <int> <dbl> <chr>  <chr> <chr>   <chr>   <dbl> <dbl> <dbl>
##  1  2013     1     1     5 EWR    IAH   N14228  UA       39.0  28.0  64.4
##  2  2013     1     1     5 LGA    IAH   N24211  UA       39.9  25.0  54.8
##  3  2013     1     1     5 JFK    MIA   N619AA  AA       39.0  27.0  61.6
##  4  2013     1     1     5 JFK    BQN   N804JB  B6       39.0  27.0  61.6
##  5  2013     1     1     6 LGA    ATL   N668DN  DL       39.9  25.0  54.8
##  6  2013     1     1     5 EWR    ORD   N39463  UA       39.0  28.0  64.4
##  7  2013     1     1     6 EWR    FLL   N516JB  B6       37.9  28.0  67.2
##  8  2013     1     1     6 LGA    IAD   N829AS  EV       39.9  25.0  54.8
##  9  2013     1     1     6 JFK    MCO   N593JB  B6       37.9  27.0  64.3
## 10  2013     1     1     6 LGA    ORD   N3ALAA  AA       39.9  25.0  54.8
## # ... with 336,766 more rows, and 7 more variables: wind_dir <dbl>,
## #   wind_speed <dbl>, wind_gust <dbl>, precip <dbl>, pressure <dbl>,
## #   visib <dbl>, time_hour <dttm>
```

```
# join with named matching
flights2 %>% left_join(planes, by = "tailnum")
```

```
## # A tibble: 336,776 x 16
##    year.x month   day  hour origin dest  tailnum carrier year.y type
##     <int> <int> <int> <dbl> <chr>  <chr> <chr>   <chr>    <int> <chr>
##  1   2013     1     1     5 EWR    IAH   N14228  UA        1999 Fixe~
##  2   2013     1     1     5 LGA    IAH   N24211  UA        1998 Fixe~
##  3   2013     1     1     5 JFK    MIA   N619AA  AA        1990 Fixe~
##  4   2013     1     1     5 JFK    BQN   N804JB  B6        2012 Fixe~
##  5   2013     1     1     6 LGA    ATL   N668DN  DL        1991 Fixe~
##  6   2013     1     1     5 EWR    ORD   N39463  UA        2012 Fixe~
##  7   2013     1     1     6 EWR    FLL   N516JB  B6        2000 Fixe~
##  8   2013     1     1     6 LGA    IAD   N829AS  EV        1998 Fixe~
##  9   2013     1     1     6 JFK    MCO   N593JB  B6        2004 Fixe~
## 10   2013     1     1     6 LGA    ORD   N3ALAA  AA          NA <NA>
## # ... with 336,766 more rows, and 6 more variables: manufacturer <chr>,
## #   model <chr>, engines <int>, seats <int>, speed <int>, engine <chr>
```

```
# join with explicit column matching
flights2 %>% left_join(airports, by= c("dest" = "faa"))
```

```
## # A tibble: 336,776 x 15
##     year month   day  hour origin dest  tailnum carrier name    lat   lon
##    <int> <int> <int> <dbl> <chr>  <chr> <chr>   <chr>   <chr> <dbl> <dbl>
##  1  2013     1     1     5 EWR    IAH   N14228  UA      Geor~  30.0 -95.3
##  2  2013     1     1     5 LGA    IAH   N24211  UA      Geor~  30.0 -95.3
##  3  2013     1     1     5 JFK    MIA   N619AA  AA      Miam~  25.8 -80.3
##  4  2013     1     1     5 JFK    BQN   N804JB  B6      <NA>   NA    NA
##  5  2013     1     1     6 LGA    ATL   N668DN  DL      Hart~  33.6 -84.4
##  6  2013     1     1     5 EWR    ORD   N39463  UA      Chic~  42.0 -87.9
##  7  2013     1     1     6 EWR    FLL   N516JB  B6      Fort~  26.1 -80.2
##  8  2013     1     1     6 LGA    IAD   N829AS  EV      Wash~  38.9 -77.5
##  9  2013     1     1     6 JFK    MCO   N593JB  B6      Orla~  28.4 -81.3
## 10  2013     1     1     6 LGA    ORD   N3ALAA  AA      Chic~  42.0 -87.9
## # ... with 336,766 more rows, and 4 more variables: alt <int>, tz <dbl>,
## #   dst <chr>, tzone <chr>
```

Types of join with SQL equivalent.

```
# Create simple data
(df1 <- data_frame(x = c(1, 2), y = 2:1))
```

```
## # A tibble: 2 x 2
##       x     y
##   <dbl> <int>
## 1     1     2
## 2     2     1
```

```
(df2 <- data_frame(x = c(1, 3), a = 10, b = "a"))
```

```
## # A tibble: 2 x 3
##       x     a b
##   <dbl> <dbl> <chr>
## 1     1    10 a
## 2     3    10 a
```

```
# Return only matched rows
df1 %>% inner_join(df2) # SELECT * FROM x JOIN y ON x.a = y.a
```

```
## Joining, by = "x"
```

```
## # A tibble: 1 x 4
##       x     y     a b
##   <dbl> <int> <dbl> <chr>
## 1     1     2    10 a
```

```
# Return all rows in df1.
df1 %>% left_join(df2) # SELECT * FROM x LEFT JOIN y ON x.a = y.a
```

```
## Joining, by = "x"
```

```
## # A tibble: 2 x 4
##       x     y     a b
##   <dbl> <int> <dbl> <chr>
## 1     1     2    10 a
## 2     2     1    NA <NA>
```

```
# Return all rows in df2.
df1 %>% right_join(df2) # SELECT * FROM x RIGHT JOIN y ON x.a = y.a
```

```
## Joining, by = "x"
```

```
## # A tibble: 2 x 4
##       x     y     a b
##   <dbl> <int> <dbl> <chr>
## 1     1     2    10 a
## 2     3    NA    10 a
```

```
# Return all rows.
df1 %>% full_join(df2) # SELECT * FROM x FULL JOIN y ON x.a = y.a
```

```
## Joining, by = "x"
```

```
## # A tibble: 3 x 4
##       x     y     a b
##   <dbl> <int> <dbl> <chr>
## 1     1     2    10 a
## 2     2     1    NA <NA>
## 3     3    NA    10 a
```

```
# Like left_join, but returning only columns in df1
df1 %>% semi_join(df2, by = "x")  # SELECT * FROM x WHERE EXISTS (SELECT 1 FROM y WHERE x.a = y.a)
```

```
## # A tibble: 1 x 2
##       x     y
##   <dbl> <int>
## 1     1     2
```

23.3 tidyr

23.4 reshape2

23.5 stringr

23.6 anytime

23.7 Biblipgraphic Notes

23.8 Practice Yourself

Chapter 24

Causal Inferense

Recall this fun advertisement

How come everyone in the past did not know what every kid knows these days: that cigarettes are bad for you. The reason is the difficulty in causal inference. Scientists knew about the correlations between smoking and disease, but no one could prove one caused the other. These could have been nothing more than correlations, with some external cause.

Cigarettes were declared dangerous without any direct causal evidence. It was in the USA's surgeon general report of 1964 that it was decided that despite of the impossibility of showing a direct causal relation, the circumstantial evidence is just too strong, and declared cigarettes as dangerous.

24.1 Causal Inference From Designed Experiments

24.1.1 Design of Experiments

https://cran.r-project.org/web/views/ExperimentalDesign.html

TODO

24.1.2 Randomized Inference

https://dimewiki.worldbank.org/wiki/Randomization_Inference

TODO

24.2 Causal Inference from Observational Data

24.2.1 Principal Stratification

Frumento et al. (2012)

https://en.wikipedia.org/wiki/Principal_stratification

TODO

24.2.2 Instrumental Variables

TODO

24.2.3 Propensity Scores

TODO

24.2.4 Direct Lieklihood

TODO

24.2.5 Regression Discontinuity

24.3 Bibliographic Notes

On the tail behind "smoking causes cancer" see NIH's Reports of the Surgeon General.

24.4 Practice Yourself

Bibliography

Allard, D. (2013). J.-p. chiles, p. delfiner: Geostatistics: Modeling spatial uncertainty.

Analytics, R. and Weston, S. (2015). *foreach: Provides Foreach Looping Construct for R.* R package version 1.4.3.

Anderson-Cook, C. M. (2004). An introduction to multivariate statistical analysis. *Journal of the American Statistical Association*, 99(467):907–909.

Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Arnold, T., Kane, M., and Urbanek, S. (2015). iotools: High-performance i/o tools for r. *arXiv preprint arXiv:1510.00041*.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, pages 311–329.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Bonat, W. H. (2018). Multiple response variables regression models in r: the mcglm package. *Journal of Statistical Software*, 84(1):1–30.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017). *shiny: Web Application Framework for R.* R package version 1.0.0.

Chapple, S. R., Troup, E., Forster, T., and Sloan, T. (2016). *Mastering Parallel Programming with R.* Packt Publishing Ltd.

Christakos, G. (2000). *Modern spatiotemporal geostatistics*, volume 6. Oxford University Press.

Conway, D. and White, J. (2012). *Machine learning for hackers.* " O'Reilly Media, Inc.".

Cotton, R. (2017). *Testing R Code.* Chapman and Hall/CRC.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data.* John Wiley and Sons.

Davis, T. A. (2006). *Direct methods for sparse linear systems.* SIAM.

Diggle, P. J., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Dowle, M. and Srinivasan, A. (2017). *data.table: Extension of 'data.frame'.* R package version 1.10.4.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics*, 3(1):1–21.

Everitt, B. and Hothorn, T. (2011). *An introduction to applied multivariate analysis with R.* Springer Science & Business Media.

Fithian, W. (2015). *Topics in Adaptive Inference*. PhD thesis, STANFORD UNIVERSITY.

Foster, D. P. and Stine, R. A. (2004). Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99(466):303–313.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, 107(498):450–466.

Gentle, J. E. (2012). *Numerical linear algebra for applications in statistics*. Springer Science & Business Media.

Gilbert, J. R., Moler, C., and Schreiber, R. (1992). Sparse matrices in matlab: Design and implementation. *SIAM Journal on Matrix Analysis and Applications*, 13(1):333–356.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Graham, R. (1988). Isometric embeddings of graphs. *Selected Topics in Graph Theory*, 3:133–150.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Izenman, A. J. (2008). Modern multivariate statistical techniques. *Regression, classification and manifold learning*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 6. Springer.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909.

Kalisch, M. and Bühlmann, P. (2014). Causal structure learning and inference: a selective review. *Quality Technology & Quantitative Management*, 11(1):3–21.

Kane, M. J., Emerson, J., Weston, S., et al. (2013). Scalable strategies for computing with massive data. *Journal of Statistical Software*, 55(14):1–19.

Kempthorne, O. (1975). Fixed and mixed models in the analysis of variance. *Biometrics*, pages 473–486.

Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In *Compstat*, pages 575–580. Springer.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge University Press.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3):285–292.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, pages 2791–2817.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pinero, J. and Bates, D. (2000). Mixed-effects models in s and s-plus (statistics and computing).

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabinowicz, A. and Rosset, S. (2018). Assessing prediction error at interpolation and extrapolation points. *arXiv preprint arXiv:1802.00996*.

Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, pages 15–32.

Rosenblatt, J. (2013). A practitioner's guide to multiple testing error rates. *arXiv preprint arXiv:1304.4920*.

Rosenblatt, J., Gilron, R., and Mukamel, R. (2016). Better-than-chance classification for signal detection. *arXiv preprint arXiv:1608.08873*.

Rosenblatt, J. D. and Benjamini, Y. (2014). Selective correlations; not voodoo. *NeuroImage*, 103:401–410.

Rosset, S. and Tibshirani, R. J. (2018). From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, (just-accepted).

Sammut, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5.

Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L., and Mansmann, U. (2009). State of the art in parallel computing with r. *Journal of Statistical Software*, 47(1).

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance components*, volume 391. John Wiley & Sons.

Shah, V. and Gilbert, J. R. (2004). Sparse matrices in matlab* p: Design and implementation. In *International Conference on High-Performance Computing*, pages 144–155. Springer.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Venables, W. N. and Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Venables, W. N., Smith, D. M., Team, R. D. C., et al. (2004). An introduction to r.

Wang, C., Chen, M.-H., Schifano, E., Wu, J., and Yan, J. (2015). Statistical methods and computing for big data. *arXiv preprint arXiv:1502.07989*.

Weihs, C., Mersmann, O., and Ligges, U. (2013). *Foundations of Statistical Algorithms: With References to R Packages*. CRC Press.

Weiss, R. E. (2005). *Modeling longitudinal data*. Springer Science & Business Media.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2011). testthat: Get started with testing. *The R Journal*, 3(1):5–10.

Wickham, H. (2014). *Advanced R*. CRC Press.

Wickham, H. and Francois, R. (2016). *dplyr: A Grammar of Data Manipulation*. R package version 0.5.0.

Wickham, H., Hester, J., and Francois, R. (2016). *readr: Read Tabular Data*. R package version 1.0.0.

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Wilkinson, G. and Rogers, C. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, pages 392–399.

Wilkinson, L. (2006). *The grammar of graphics*. Springer Science & Business Media.

Xie, Y. (2015). *Dynamic Documents with R and knitr*, volume 29. CRC Press.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. CRC Press.